MODELLING CAUSAL MECHANISMS IN ORGANISMAL AGING

by

Glen Pridham

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
August 2024

© Copyright by Glen Pridham, 2024

Contents

List of Tables

List of Figures

Abstract

Geroscience seeks to clarify and explain the connection between chronological age and declining organism health. The field is rich in complex phenomena and increasingly rich in large, multivariate datasets as well, making it fertile ground for quantitative modelling. This thesis outlines the steps needed to develop such models, from contending with aging study data to validating a causal model of effective dynamics. Ultimately, the goal of this thesis is make interpretable predictions for the causes and effects of organism aging, with special attention paid to humans. This goal is met by developing and applying the Stochastic Finite (SF) difference model which uses an interaction network to predict multivariate, future values based on current values. The process by which aging study data facilitates this type of model building involves multiple pitfalls and key assumptions which are addressed in introductory chapters on missing data and unsupervised learning of aging metrics. The SF model is then validated on two mouse datasets and two human datasets, unveiling characteristic dynamical behaviour of aging systems. In particular, we observe mallostasis: the steady-state drift of biomarker values towards worse health. The approach is then applied to specific age metrics, so called "biological ages", to probe more directly the sequence of events which occurs during natural aging. This provides insight into the strengths and weaknesses of existing qualitative theories and provides a path towards more quantitative theories of aging.

List of Abbreviations and Symbols Used

In quasi-alphabetical order:

ADL Activity of Daily Living: essential and routine tasks of daily life [46]. For example, being able to dress oneself. Typically reported on a graded scale ranging from no difficulty to some difficulty to unable to do.

Adverse outcome Any unfavourable, definite change in health. For example, onset of: disability, disease, and death.

Allostasis "Homeostasis through change" [118]: homeostasis but with a set point that can change. This permits adaptive response to environmental demands, but may eventually come at the cost of allostatic load [89].

Allostatic load The cost associated with long term wear-and-tear of the adaptive stress response [89].

AUC Area Under the receiver operating characteristic Curve: a measure of binary prediction performance ranging from 0 (perfectly-incorrect predictor) to 0.5 (guess) to 1 (perfectly-correct predictor). Equivalent to the probability of correctly ranking the afflicted individual ahead of the unaffected in any pair [69].

Biological age A metric sensitive to the effects of aging. I will further assume that it is a measure of overall health. Often reported in units of years, representing an individual's effective age.

Biomarker of aging A metric sensitive to the effects of aging which also satisfies specific criteria [90]. The American Federation for Aging Research suggest four criteria: (i) it must predict exactly where a person is in their lifespan trajectory, and it must predict better than chronological age, (ii) it must monitor the aging process, rather than the effects of disease, (iii) testing must not harm the individual, and (iv) it must works in humans and in laboratory animals, such as mice (for testing). A recent high-profile paper from the Biomarkers of Aging

Consortium proposes to instead define a biomarker of aging as a metric that is sensitive to biological age [128].

C57BL/6 C57, black-6 genetically identical mouse. A commonly-used lab mouse.

CART Classification and regression tree; a machine learning model.

Causal A causal prediction satisfies the Wiener-Granger causality condition. Specifically, $Y \to X$ (Y causes X) if and only if prediction of future values of X is significantly improved by inclusion of past information from Y, as opposed to simply using past values of X [23].

C-index Concordance index: a measure of survival prediction performance, cousin of the AUC. Ranges from 0 (perfectly-incorrect predictor) to 0.5 (guess) to 1 (perfectly-correct predictor) [72]. Equivalent to the probability that the model will correctly predict which individual in any pair will die first.

Eigen-decomposition $\boldsymbol{A} = \sum_{i=1}^{n} \lambda_i P_{\cdot i} \otimes P_{i\cdot}^{-1}$ where $\boldsymbol{A}$ is the decomposed matrix, $\boldsymbol{P}$ is the matrix who's columns are the n eigenvectors and λ_i is the ith eigenvector ($\otimes$ is the outer product).

Eigenvalue The scale, λ, of a linear transformation, $\boldsymbol{A}$, as defined by $\boldsymbol{A}\vec{x} = \lambda\vec{x}$.

Eigenvector The vector, x, of a linear transformation, $\boldsymbol{A}$, as defined by $\boldsymbol{A}\vec{x} = \lambda\vec{x}$.

ELSA English Longitudinal Study of Aging. A large-scale British study of aging including 1000s of variables across domains such as demographic, lifestyle, functional state and lab tests.

FI Frailty Index: the average number of health deficits an individual has out of a set of 30+ health variables subject to selection criteria. The criteria are: (i) must be related to health, (ii) prevalence must increase with age, (iii) prevalence can't saturate at young ages, (iv) must cover a range of biological systems, and (v) if used serially it should include the same specific variables [174]. Usually based on self-reported health variables such as ADLs, IADLs, symptoms, functional limitations, and chronic diseases. For example, "do you have difficulty dressing?"

would be encoded as: 1 if yes (deficit) or 0 if no (healthy); 29 related variables would then be included and the average value would be the individual's FI.

FI CLINIC Frailty index constructed entirely from functional or self-reported health deficit information. This is normal and represents the original FI formulation [174].

FI LAB Frailty index constructed entirely from blood test (lab) values. Blood tests are binarized based on diagnostic thresholds for abnormality, with abnormal values being scored as 1 (deficit) and 0 (health) otherwise.

Gerontology The study of aging, typically from a social science perspective.

Geroscience The study of aging from the perspective of its effects on health [94]. The central geroscience hypothesis is that ameliorating the effects of aging in an individual will prevent or slow the development of most chronic diseases.

GNM Generic Network Model of aging population health and survival [184]. A phenomenon-driven model based on the accumulation and propagation of damage within a binary network of abstract health attributes.

Health deficit An abnormal state of ill-health. Can be graded (ordinal) or binary. For example, angina (chest pain), diabetes, and any difficulty in ADLs or IADLs are all health deficits. In this thesis all deficits will be encoded as 0 if they are perfectly healthy and 1 if they are maximally unhealthy.

Het3 Heterogenous-3 genetically non-identical mouse.

Homeostasis The tendency for biological systems to spontaneously recover from perturbations to an equilibrium state. This means that biomarkers may be perturbed over short timescales, but will tend to return to a steady-state value. For example, white blood cell count may go up during an infection then return to a normal value after the infection is repelled. Self-regulating hormone systems such as the hypothalamic–pituitary–adrenal axis are another example of homeostasis [6].

IADL Instrumental Activity of Daily Living: complex tasks needed to live independently. For example, being able to prepare one's own meals [46]. Typically reported on a graded scale ranging from no difficulty to some difficulty to unable to do.

Imputation Insertion of an invented (instantiated) value to replace a missing value.

Latent variable An unobserved or unobservable variable which affects observed values.

Mallostasis The correlation between steady-state drift rate and survival risk. As individuals age their ("natural") health variables drift towards worse health in the steady-state [147].

MAR Missing At Random: missingness depends on observed values. This means that it is possible to infer the effects of a missing value using only observed values.

MCAR Missing Completely At Random: missingness is independent of observed or unobserved values. Do not incorrectly assume that this means the missing value can be safely ignored, since higher-order patterns can still cause bias Chapter 4.

MNAR Missing Not At Random: missingness depends on unobserved values. Formal inference including the missing value will require more information than is available from the dataset alone.

Measure See metric.

Metric A rule for measurement e.g. an aging metric quantifies the effects of age. Provides a number.

MI Multiple Imputation. Estimates a missing value together with the uncertainty it introduces in the analysis. The procedure is to: (i) generate a collection of new datasets with randomly imputed missing values, (ii) analyse each new dataset, and then (iii) pool the results at the end using Rubin's rules [130].

MSE Mean-Squared Error. Measure of prediction accuracy. MSE $\equiv N^{-1} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$ for known values y_i, predicted values $\hat{y}_i$, and N individuals.

Natural variable Individual scores associated with the eigenvector transformation of a dynamical interaction network. The matrix of eigenvectors, $\boldsymbol{P}$, is used to transform a dataset of individuals $\times$ variables into the basis individuals $\times$ natural variables [147].

NHANES National Health and Nutrition Examination Survey. A large-scale American cross-sectional study that includes 1000s of variables across domains such as diet, lifestyle, functional state and lab tests.

Parseval's theorem Vector length is conserved by isometric transformations, such as orthogonal matrices [26]. This means that the mean plus variance of a dataset must be conserved by an isometric transformation [147].

PC Principal Component. One of the new variables once a dataset has been transformed using principal component analysis.

PCA Principal Component Analysis. The canonical dimensionality reduction algorithm. Equivalently: the independent directions of maximum variance (ranked), the eigen-decomposition of the covariance matrix, and the hyperplane closest to the dataset in terms of mean-squared error (considering all possible linear transformation). Rotates (transforms) a dataset of individuals $\times$ variables into the basis individuals $\times$ principal components.

R Statistical computing language.

Resilience The ability of a system to recover from a perturbation back to its previous state.

Robustness When referring to a living organism: the ability to resist being perturbed by a stressor. When referring to a result: how dependent a result is on the specific dataset under analysis (less dependent means more robust).

RF Random Forest: a non-parametric machine learning model for regression and classification.

RI Rule-based Imputation. A deterministic algorithm for imputing values. Used for imputing gated variables. Gated variables are variables which are not asked during an interview because they are gated by a previous response (e.g. "how many cigarettes do you smoke each day?" is gated by "do you smoke?"; a rule-based imputation would be to insert 0 everywhere the first question is missing and the second question is "no").

RMSE Root Mean-Squared Error. The square-root of the mean-squared error.

SF Model Stochastic Finite difference Model: a dynamical network introduced in Chapter 6. This is the main model of this thesis.

SLAM Study of Longitudinal Aging in Mice. A large-scale American study of aging in mice cared for under lab conditions [136].

x Typically used to indicate an observed auxiliary variable ("covariate").

y Typically used to indicate an observed predictor or essential variable.

z Typically used to indicate a latent/unobserved variable.

Acknowledgements

I am grateful to have benefited from supportive family, friends, and collaborators. The Dalhousie Physics & Atmospheric Science department has also been encouraging and supportive, and is full of wonderful individuals. It has been a long gap between my MSc and now, and I am thankful to have eventually found this opportunity.

A lack of mentorship can be detrimental and I am fortunate to have received an abundance of excellent mentorship at Dalhousie. Ken Rockwood has provided a window into the medical perspective of aging biology, together with generous portions of wisdom and opportunity. I have benefited from wonderful teaching opportunities from Laurent Kreplak, together with his wise council. Finally, my supervisor Andrew Rutenberg has provided more than I could or would ask for. I am fortunate to have benefited from his tutelage, although I feel that understates the scope of his mentorship. He is both wise and kind.

Thank you also to those who played important roles in my academic and research trajectory, including in particular Jason Donev, Robert Pywell, Stephanie Wilson, Alex Medellin, and Yunyan Zhang.

Chapter 1

Introduction

In humans and most other organisms, chronological age is associated with an overall decline in health. This includes an exponential increases in risk of death [99], morbidity [217] and disability [68], Figure 1.1. The effects are nearly universal, affecting almost all known organisms and all biological scales within aging organisms. For example, telomere attrition, DNA damage, epigenetic alterations, mitochondrial dysfunction, altered cellular communication, chronic disease, and ultimately functional decline, disability, and mortality are all strongly affected by the aging process [115, 94]. This makes aging a non-linear, (nearly) universal, multiscale, collective phenomenon [31] of immense practical importance. This thesis seeks to identify and explain essential aging phenomenon using quantitative models. The ultimate goal is to incorporate causal predictions within these models.

There are three major barriers to quantitative model building, which are addressed in this thesis. The first is a lack of good aging metrics which would provide a suitable outcome and foundation for quantitative modelling. This thesis proposes to instead root quantitative models in phenomena and then leverage these models to construct aging metrics (or refine rough metrics). This reveals the second issue, which is that few quantitative models are simultaneously rooted in phenomena and portable to generate aging metrics — they are generally exclusively data or phenomenon-driven and hence new, hybrid models are needed. The underlying reason for this dichotomy appears to be the third major issue, which is that aging data introduce special data handling problems, the foremost of which are survival/censorship and missing data.

Before discussing metrics of aging, it must first be emphasized that *the aging process is distinct from the passage of time*. Being facile, in health terms a 10 year-old dog is "older" than a 10 year-old human. Being exact, the number of health deficits at a specific age varies dramatically across individual humans, as do health trajectories (e.g. survival risk) [125, 30]. However there is no consensus definition of

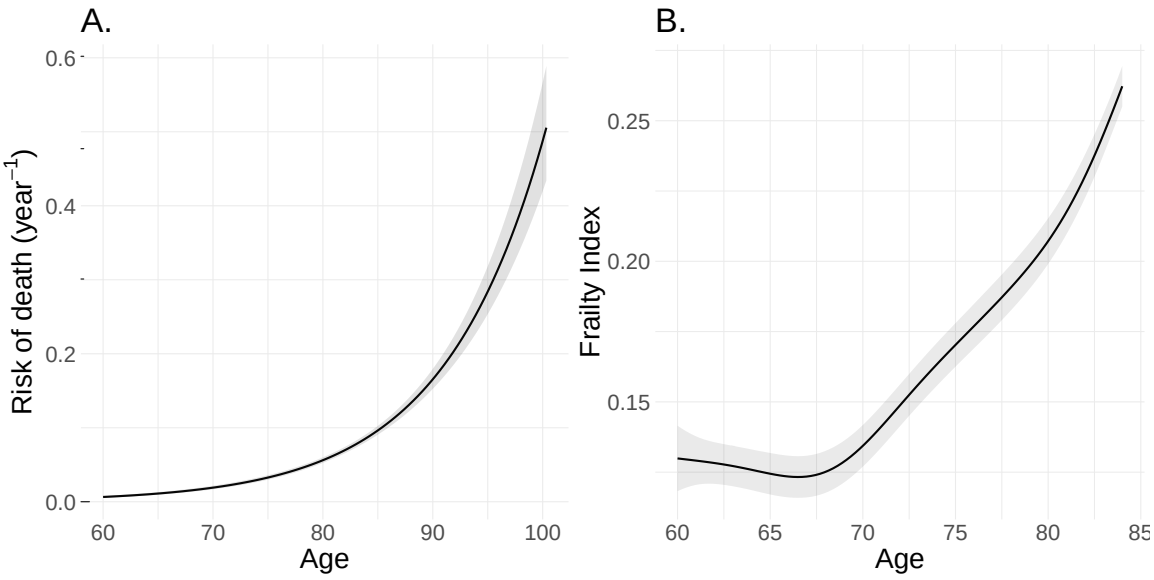

Figure 1.1: Aging includes superlinear increases in risk of death (**A.**) and average number of health deficits (frailty index; **B.**). The health deficits include functional limitations, disability, signs and symptoms, and abnormal lab values (Tables B.1 and B.2). This makes lifespan and declining health characteristic phenomena of aging at sufficiently advanced ages. National Health and Nutrition Examination Survey (NHANES) 2001-02; **A.** is a log-linear fit [53]; **B.** is a spline fit with default parameters [209]; data processing is reported in Chapter 5.

aging [32], and there is no consensus way to quantify the effects of aging. Instead, aging is defined by its phenomena, the foremost of which is an overall decline in health and physical functioning over time. It is this decline which is ultimately of interest [95, 94] and we therefore can use an individual's overall health state as a definition of their *biological age*.

Biological age is usually reported in units of time, representing an individual's health-equivalent, effective age. The distinction between estimator and true value is often neglected, and an algorithm which estimates biological age is also referred to as a (specific) biological age. For historical reasons, a "biomarker of aging" is a biological age, but a biological age is not necessarily a "biomarker of aging". The American Federation for Aging Research requires a "biomarker of aging" to satisfy four criteria: (i) it must predict exactly where a person is in their lifespan trajectory, and it must predict better than chronological age, (ii) it must monitor the aging process, rather than the effects of disease, (iii) testing must not harm the individual, and (iv) it must works in humans and in laboratory animals, such as mice (for testing). While criteria (iii) and (iv) have been met by several candidates, (i) is complicated by evidence indicating that aging is a multidimensional process (e.g. organs may age differentially within a single body), and (ii) may be outright impossible since chronic disease is strongly associated with the aging process [90].[1] For this reason I will refer to a quantitative measure of aging as a "measure", "metric" or "biological age" rather than a "biomarker".

Biological ages address a need for health state variables able to quantify the effects of aging, as well as possible anti-aging interventions [28, 218]. The pragmatic motivation for geroscience is to extend the period of healthy living without disability or affliction, and compress the period of age-related decline [95, 94]. A variety of medical and environmental interventions have been developed that do precisely this in lab animals [143], such as caloric restriction [64] or alpha-ketoglutarate supplementation [7] in mice. Human studies are however constrained by a lack of appropriate outcome. The conventional outcome is lifespan, but there are ethical and practical limitations

[1] Upon revision it has come to my attention that a recent movement alternatively defines a biomarker of aging as a quantitative measure that is sensitive to an individual's level of age-dependent biological changes [128]. While a less stringent definition than the American Federation for Aging Research's is appropriate, the new one seems conspicuously non-specific.

to exposing a human cohort to an anti-aging treatment for the decades required to gather sufficient lifespan statistics. A non-invasive metric sensitive to biological age could provide a faster, potentially more specific outcome for anti-aging studies. Alternatively, from a purely academic standpoint, such a metric is also needed to build quantitative theories of aging, since they require a relevant quantity to model and communicate. At the heart of both understanding aging and anti-aging interventions are cause and effect, leading to overlapping data and theories between the two. Accordingly, I will generally not make a distinction between these objectives. Irrespective of objective there is widespread contemporary interest in prospective biological ages, ranging from epigenetic "clocks" which estimate health from DNA methylation scores at the epigenetic scale [163], to the frailty index (FI) which measures health at the clinically-relevant scale [174]. While the American Federation for Aging Research criteria for a "biomarker of aging" may be unrealistic, it is still true that the existing biological ages have major drawbacks. Epigenetic clocks are noisy, and while they do correctly capture transient changes in health due to random stressors [144], they're also sensitive to impertinent information such as the time of day [102]. Alternatively, functional biological ages such as the FI can work well, but only after a substantial decline in functional state, limiting their informativeness (e.g. for the FI see [122]). It is also true that evidence suggests that health state is multivariate, though not necessarily high-dimensional [141, 52, 66], indicating that there is more than one "true" biological age [110, 87]. Together these conditions drive a high demand for new and improved biological ages.

The FI is one of the few biological ages which is being actively used as a *de facto* biomarker of aging, both as a study outcome and a clinical measure [42] (e.g. [64, 7]). The FI is a number from 0 to 1 that is defined as the average number of health "deficits" an individual has, including signs, symptoms, chronic diseases and functional limitations across multiple domains, with 1 representing all deficit (for a minimum 30 variables) [174]. The FI is a measure of overall health. It can be considered a generalized, continuous measure of frailty: a phenomenon generally described as a loss of physiological "reserves and resilience" [60] meaning that a frail individual is more likely to suffer from a plethora of adverse outcomes [42]. The FI can alternatively be grouped together with biological ages. The FI differs from

typical biological ages in terms of its scale since the FI is in units of health deficiencies whereas biological ages are typically in units of time; this can be ameliorated through a dynamical model [125] (to handle young ages see supplemental of [148]). The FI considers chronic disease to be a sign of aging and therefore cannot satisfy the criterion that a biomarker of aging should differentiate the effects of aging from chronic disease. The FI is safe to measure and easy to calculate across studies and has strong predictive ability for adverse outcomes [101]. What's more, the FI has been ported from humans to lab organisms [80], and is now a popular health measure for aging mice (e.g. see [173, 64, 7]). The FI phenomena have been systematically explored for over a decade [123], providing a treasure trove of phenomena to use when fitting models. This makes the FI popular both experimentally and theoretically (e.g. [108, 194, 184]). The primary limitation of the FI is that it is poor at discriminating health in the young [218, 122] since young people typically have an FI very close to 0 ($\lesssim$ 40 years old for humans). A prospective solution is to build an FI-like metric by binarizing sub-clinical measures of health [160], such as the FI LAB which uses lab blood tests [81]. However, while the FI LAB has been shown to improve survival prediction [20], it has yet to be shown to be more sensitive to health in young individuals (e.g. Figure 1b of [19] suggests that the FI LAB is actually *less* sensitive to changes in young individuals than the conventional FI). The FI plays the role of health state variable in many studies and models, generally complementing survival risk. The FI has not, however, been generally accepted as a "biomarker of aging" [90], ostensibly because it does not discriminate between the effects of chronic disease and those of age, and lacks sensitivity to the sub-clinical drivers of aging.

Without a consensus biomarker of aging [90], data-driven models instead make use of whatever relevant data are available. Survival risk, disability, chronic disease burden and functional limitations (e.g. gait speed and grip strength) are all reasonable proxies for health, particularly in older individuals. In addition, almost every health biomarker measured shows a decline with age [176]. A model that includes both survival prediction and longitudinal biomarker trajectories can then have a well-defined loss function for fitting (e.g. [213, 52]). Alternatively, many models 'fit' instead directly to the phenomenon: performing model selection based on population-level statistics and qualitative behaviour (e.g. [62, 201, 184]). Regardless of the approach, a

good model of aging should capture the salient features of aging, particularly a spontaneous decrease in health and increase in risk of adverse outcomes over time. While metrics of aging are poorly defined, there are lots of metrics available and together with a suite of distinct aging phenomena we can judge model quality and validity. Quantitative models can be used to explore mechanisms of aging in a falsifiable way. In addition, an auxiliary goal is to leverage such models to produce and refine salient metrics of aging (biological ages).

There is a paucity of quantitative models within geroscience, despite an abundance of data. Existing models can be broadly grouped as either phenomenon or data-driven. Phenomenon-driven models seek to recapitulate characteristic population-level aging phenomenon, eschewing specific predictions at the individual-level. A key contribution came in 2001 when Gavrilov and Gavrilova proposed the reliability theory of aging, which derives an exponentially increasing mortality rate (Gompertz' law) with late-life plateau as a consequence of damaging an abstract, redundant system [62]. This formed the foundation for abstract binary network models of damaged/undamaged nodes for human health which recapitulates mortality curves at the population level [201], then more recently both human mortality and the FI by the Generic Network Model (GNM) [184]. While phenomenon-driven models provide important insights into what the important features of aging are and how/why they might occur, their abstract nature makes it impractical to map them into specific biological targets e.g. the nodes of the GNM do not represent specific, observable health variables. Yashin *et al.* proposed an alternative vision using data-driven modelling based on generalized ideas of homeostasis [213]. This idea was eventually applied to multivariate data using deep learning to automatically generate a network of interactions between observable health variables [52]. Unfortunately, this approach is cumbersome and data-hungry, making it inapplicable to small datasets and unlikely to be adopted by (most) aging researchers whom prefer simple models. What's more, while data-driven models can generate precise predictions and mappings into observable variables, their flexibility and agnosticism limits their insights into the aging process i.e. they lack "interpretability". Missing is a hybrid approach which can capture the salient phenomena needed to characterize the aging process but can still be used to analyze specific, observable variables — ideally helping convert individualized

data into biological age estimates.

This thesis proposes that data-driven techniques can be modified to incorporate the missing elements provided by phenomenon-driven techniques. The foremost shortcoming of data-driven models is their connection to causality, which is often tenuous or even explicitly agnostic. Incorporating causality presents the first of two major hurdles, the second being data handling — which is a non-trivial issue for aging data, particularly missing data in aging studies.

Quantitative, causal models are needed to make sense of interventional data [31]. An added benefit is that they can help discriminate between different theories of aging, which have not only proliferated [97] but have become complex to the point where they cannot be explained without employing multi-causal networks [98, 115, 94]. For interventions, causal models are needed to make sense of pleiotropic[2] interventional study data [218, 134]. For example, the prospective anti-aging drug metformin has shown promise, but it has also been shown to reduce visual acuity in mice [134]. Similarly, cellular reprogramming can rejuvenate aged mice but can also cause tumour formation [218]. Ostensibly these treatments are perturbing an underlying biological network resulting in a spectrum of unexpected consequences. Similarly, natural aging is characterized by a spectrum of widespread and diverse symptoms due to dysfunction of the underlying biological network. It is only with quantitative modelling that we can make sense of the biological networks which regulate organism health and across which the effects of aging manifest. The specific effects of either interventions or biological aging processes requires these networks to incorporate causal, directed associations.

I take the Wiener-Granger causality condition as a working definition for causality. For my purposes this means that a vector auto-regressive process[3] is used to show that $Y \to X$ (Y causes X) if and only if prediction of future values of X are improved by inclusion of past values of Y (in addition to past values of X) [23]. The causal connections between variables can then be represented as a directed network of "causal" links. The correctness of such models is to be argued and experimentally verified. By definition, a causal model predicts the behaviour of a system in response

[2]Pleiotropic: many ways.

[3]A vector auto-regressive process is of the form $\vec{x}(t+1) = f(\vec{x}(t), \vec{x}(t-1), ...)$ for some function f.

to an external intervention [45], and hence a grain of salt must accompany our results, which are based on observational data [39]. The conventional gold standard for evidence is the randomized control trial. In Fisher's conceptualization of causality, randomization and repeated measurement removes all confounding effects leaving only causal connections [192]. This conceptualization is far from perfect, however, in particular it has been criticized for relying too heavily on perfect randomization of (repeated) trials without any artifact entering via study design or execution [192]. Wiener-Granger causality based on dynamical modelling provides a complementary approach that emphasizes interpretability. These models are not necessarily correct, but rather they are a reasonable estimate given the available data — in particular, they are an improvement over correlation matrices, which are typically used within geroscience and neglect to take advantage of any evidence of directed associations present in the data.

Aging study data carries with it a number of subtle difficulties which forms a barrier to entry for scientists interested in quantitative modelling. For example, censorship and survival adds a layer of complexity, and most statistical models cannot be reliably used without taking into account these effects (e.g. the fact that an individual survived long enough to reach the end of the study may be informative). In general, missing data are an inevitable presence in any observational study, since individuals may offer only limited time for an interview, or may be unable to complete some tests owing to a state of poor health. This latter point means that missing data in aging studies are often biased [71], such that the reason they were not measured is due in part to the true, unobserved, value. Imputation is the standard statistical approach to avoiding bias due to missing data [112, 180], wherein values are artificially inserted (imputed) into missing entries. Unfortunately, many aging researchers completely ignore the issue of missing data or disclose insufficient detail to reproduce their imputation methods. Needed is a clear example of how imputation biases study conclusions and how to avoid it.

The scope of this thesis is to capture and explain distinct and important aging phenomena. Foremost are describing increasing risk of adverse outcome and functional limitations. Biochemical mechanisms are outside of my scope. All mechanisms are based on effective dynamics and rooted in identifiable phenomena. While this limits

their utility in identifying e.g. pharmacological targets, it increases generalizability and makes use of the available data, which are overwhelmingly large-scale studies of health biomarkers, anthropometric measures and functional limitations.

This thesis reproduces a series of self-contained manuscripts that collectively address the problems endemic to aging data, notably missing data and a lack of metrics, then introduce and apply the causal Stochastic Finite (SF) difference model for biomarker dynamics. The SF model provides both a lightweight tool for estimating networks and an analysis pipeline for producing salient metrics of that network. When applied to aging data it recapitulates the key aging phenomenon of declining health, and predicts an eventual decrease in complexity (dimensionality) of the aging process at advances ages. The remainder of the thesis seeks to unify these results and contextualize them into the broader research field.

Chapter 2

Outline

This manuscript-based thesis is based off of a series of papers designed to simultaneously build skills and convert them into publications, culminating in an interpretable model of effective dynamics which can be used to make causal predictions regarding aging (the SF model). Figure 2.1 illustrates. The pursuit of a causal model of organismal aging proceeds in three steps: (i) identify and resolve problems endemic to aging data, (ii) identify prospective algorithms to quantify the effects of aging on the data, and then (iii) construct a causal model to describe aging trajectories of mice and humans. Each step has an overlying manuscript associated with it and an underlying foundational skill adding additional depth. These manuscripts are reproduced as chapters in the present thesis. Where unclear, I have annotated these manuscripts with footnotes. I conclude with a summary of the key results from an upcoming publication which applies the causal model directly to a collection of biological ages.

I start by tackling missing data in Chapter 4. Missing data are a ubiquitous problem in aging study data. I review the types of missingness using simulated and real data based on National Health and Nutrition Examination Survey (NHANES) study data. I demonstrate that the FI (frailty index), an example measure of interest, is inevitably biased by missing data owing to subtle underlying effects. I show that choice of missing data handling procedure can make the difference between eliminating this bias or making it worse, and offer guidance. Missing data handling is centered around imputation: the instantiation and insertion of values to replace missing data. I compare a set of popular imputation techniques. I demonstrate that rule-based imputation of gated variables is essential for reducing bias. Gated variables are study variables which are not asked because their value can be inferred by other (gating) variables, for example "how many cigarettes do you smoke in a day" is gated by "do you smoke" since a person who does not smoke should therefore smoke 0 cigarettes

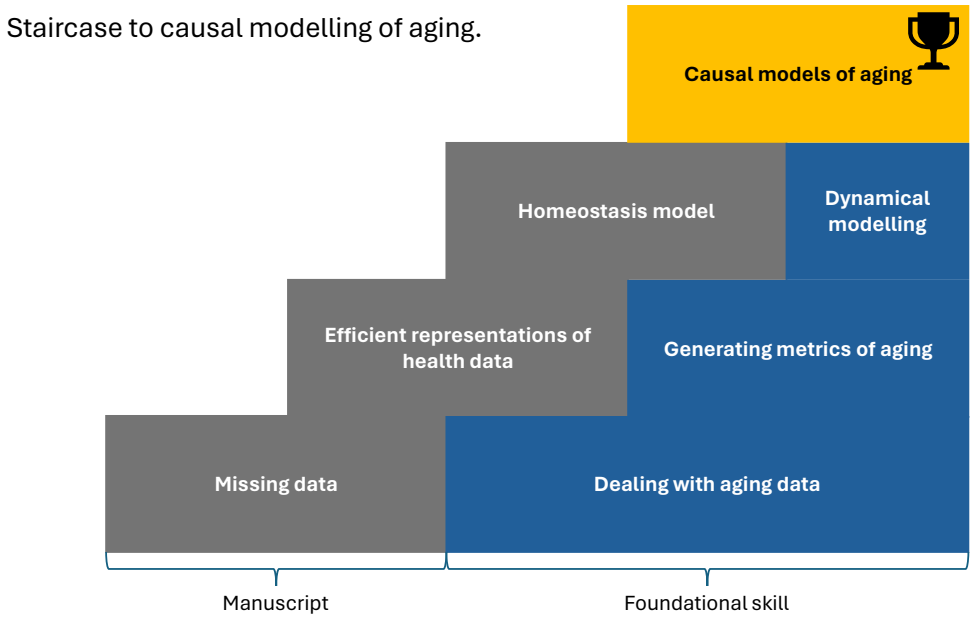

Figure 2.1: Staircase of papers culminating in a causal model of aging. Each publication represents only a portion of a learned fundamental skill.

each day[1]. For missing data which are not gated, choice of imputation algorithm can affect the bias. I provide an exposition for the source of this bias and prospective remedies which can greatly reduce its impact.

In Chapter 5, I investigate the use of unsupervised statistical algorithms to learn biological ages directly from data. These algorithms, collectively known as dimensionality reduction algorithms [86], form efficient representations of the input data, each based on a different statistical criterion (objective function). The FI is used as a target outcome to understand and recapitulate. The FI is compared to three dimensionality reduction algorithms for its ability to: compress input information, and predict adverse outcomes including survival, disability, conditions and chronic diseases. All three algorithms automatically generate a primary metric which is nearly identical to the FI in correlation and performance. I use a simple model to explain how the eigen-decomposition of the covariance matrix, i.e. principal component analysis (PCA), finds the FI as the dominant component. This suggests that pattern

[1]A related, but different, type of missingness is "conditionally relevant" missingness wherein a question isn't asked because it's not relevant, e.g. single people are not asked about their partner's substance use, which can be encoded using dummy variables [44]. Here we refer to variables that are relevant but not asked because the missing value can be safely assumed based on the value of the gating variable — although we will return to this assumption in the Chapter 4 discussion.

recognition algorithms such as the eigen-decomposition may be able to capture the dominant sources of information in aging data. In doing so, they form efficient representations of aging. These are precisely biological ages, and hence we have a path forward to leveraging models to develop biological ages.

In Chapter 6, I develop a causal, dynamical model of aging based off of an operationalization of homeostasis applied to generic health biomarkers such as blood tests and functional limitations. I name this model the Stochastic Finite (SF) difference model, reflecting its relationship to the Stochastic Process Model, which is a continuous version [213]. The SF model is of effective dynamics between health biomarkers. The dynamical behaviour is used to estimate an interaction network which is then eigen-decomposed to produce a new basis, formed out of prospective biological ages which we call "natural variables". The effects of aging can then be interrogated in terms of these natural variables, which naturally compress and simplify the dynamical behaviour of the system of biomarkers. This permits us to make sense of steady-state behaviour as a linear decline towards worse health, a phenomenon we call "mallostasis". The model is lightweight and can be used to estimate a network from any set of continuous, longitudinally-measured variables.

In Chapter 7, I summarize the key results from an application of the SF model and analysis pipeline from Chapter 6. In Chapter 7, a collection of biological ages is used as input to derive an interaction network and subsequent set of natural variables. The interaction network makes definite predictions about how natural aging propagates through the different biological scales in humans. What's more, the natural variables serve as highly efficient meta–biological ages which capture and describe the dominant contribution to decline in health over time. I propose that this tool is a natural step in building and testing theories for aging.

Finally, I summarized with a perspective discussion in Chapter 8 and conclude with Chapter 9

Chapter 3

Dataset Summary

I made use of publicly available datasets. These are summarized in Table 3.1.

Table 3.1: Dataset Summary

Dataset	Ch.[1]	Population	Outcome	N[2]	p[3]	Timepoints (rate)[4]
NHANES[5] 2003-06	4	Human	Survival	9307	68	cross-sectional
NHANES[5] 2001-02	5	Human	Many (47)	1872	55	cross-sectional
ELSA[6]	6	Human	Survival	9330	25	4 (4 years)
Paquid	6	Human	Dementia	500	4	9 (3.2 years)
SLAM[7] C57/BL6	6	Mice	Survival	608	6	20 (6.2 weeks[8])
SLAM[7] Het3	6	Mice	Survival	611	6	27 (4.2 weeks[9])
SATSA[10]	7	Human	FI[11]	845	8	9 (3 years)

[1] Ch.: chapter.
[2] N: number of individuals.
[3] p: number of predictor variables.
[4] Maximum number of timepoints and typical rate (mean or median). The SLAM data were staggered by observation mode and hence the true number of timepoints is less.
[5] NHANES: National Health and Nutrition Examination Survey (biannual).
[6] ELSA: English Longitudinal Study of Ageing.
[7] SLAM: Study of Longitudinal Aging in Mice.
[8] 6.2 weeks $\approx$ 4.9 human-equivalent years (scaling by median lifespan).
[9] 4.2 weeks $\approx$ 3.6 human-equivalent years (scaling by median lifespan).
[10] SATSA: Swedish Adoption/Twin Study of Aging.
[11] Frailty index.

Chapter 4

Strategies for handling missing data that improve frailty index estimation and predictive power: lessons from the NHANES dataset

By Glen Pridham[1], Kenneth Rockwood[2], and Andrew Rutenberg[1].
[1]Department of Physics and Atmospheric Science, Dalhousie University, Halifax, B3H 4R2, Nova Scotia, Canada.
[2]Division of Geriatric Medicine, Dalhousie University, Halifax, B3H 2E1, Nova Scotia, Canada.

Pridham, G., Rockwood, K. & Rutenberg, A. Strategies for handling missing data that improve Frailty Index estimation and predictive power: lessons from the NHANES dataset. GeroScience (2022) doi:10.1007/s11357-021-00489-w [145]

Missing data are ubiquitous in aging studies. Combining the National Health and Nutrition Examination Survey (NHANES) 2003/2004 and 2005/2006 cross-sectional aging studies ($N = 9307$), we investigated the effects of both real and simulated missing data on the Frailty Index (FI) and survival analysis, along with several mitigation strategies. We observed distinct block patterns of missing variables in the dataset. These blocks showed significant hazard rate (HR) differences when they were missing versus present, indicating that missingness cannot be simply ignored. Simulations of this patterned missingness produced a bias of 0.0112 ± 0.0008 to the mean FI when missing values were ignored, representing a change in hazard of 1.09 ± 0.01. A similar bias of 0.0106 ± 0.0001 was estimated in the real missingness. Imputation was able to correct the bias using the multivariate imputation by chained equations (MICE) method via the classification and regression tree (CART) prediction model together with rule-based imputation. Using auxiliary variables (CART+Aux) improved the performance of CART. Well-performing imputation models, especially CART+Aux,

were able to increase the FI predictive power and the reliability of the HR estimates. In contrast, the default MICE models, predictive mean matching/logistic regression (PMM/logreg), caused even stronger biases to the FI. Our results demonstrate that calibration of the FI as a mortality predictor depends on how missing data are handled. Ignoring missing values when calculating the FI may be an acceptable strategy for clinical settings where the FI is used as a rough predictor of adverse outcomes. Where the FI is to be compared across studies or populations, judicious imputation — cognizant of the risks carried by poor imputation — should be used to ensure reliability and precision of statistical estimates and conclusions.

Keywords: imputation, missing data, MICE, Frailty Index, survival, CART.

4.1 Introduction

Imputation uses statistical inference to estimate missing entries in recorded data. Imputation fills gaps that may interfere with or otherwise complicate data analysis. Often, analysis software silently excludes missing data, at times using only the complete cases. This approach can greatly reduce the amount of available data, and can bias statistical conclusions [112, 180]. Although imputation is not typically discussed in the Frailty Index (FI) literature, the most common approach of ignoring missing values is equivalent to individual (row)–mean imputation[1].

For individuals admitted to hospital with an acute stroke, Deng *et al* showed that complete-case analysis determined that none of the four individual history variables — including history of stroke — were significant determiners of the time-to-diagnosis proxy, whereas each of five imputation strategies showed that all variables were both significant and major predictors [41]. However, the choice of imputation strategy can be important. As discussed by Sterne *et al*, multiply-imputed data in a cardiovascular risk study found that cholesterol was unrelated to risk when using initially imputed data, but was a risk factor either when using the available data or when using an improved imputation strategy [180].

[1]Suppose we measured N variables for an individual, $\vec{x}$, but M values are missing. The ignore FI is the mean, $f_{ig} = \sum_{i=n}^{N-M} x_n/(N-M)$. Imputing f_{ig} for missing values also gives $f_{impute} = (\sum_{n=1}^{N-M} x_n + \sum_{m=1}^{M} f_{ig})/N = f_{ig}$.

Imputation is a valid statistical technique [198]: ideal (proper) imputation would introduce no bias and would not under-estimate uncertainties [197]. In contrast, poorly implemented imputation can worsen results [5, 205, 41]. Judiciously implemented imputation strategies, while typically not ideal, can often make significant gains compared to excluding or ignoring.

There are three canonical types of missing data (Section 4.2): missing completely at random (MCAR; independently missing), missing at random (MAR; due to covariates that are not missing), and missing not at random (MNAR; due to covariates that are missing, including the missing value itself) [112]. Higher-order missingness patterns may also be present [172].

Missing data in gerontology are distinctive for the high prevalence of MAR and MNAR missingness. Cognitive and functional deficits that can prevent data collection are common amongst older adults [71], even those dwelling in communities. For example, people living with frailty may be more likely to drop out of longitudinal studies, causing MAR and MNAR missingness in later waves to be more common among the frail [117]. Study designs may also neglect to ask young people about potential deficits that are only prevalent in older adults, a form of MAR missingness. Because of the prevalence of MAR and MNAR missingness, it is important to investigate potential biases when imputing gerontological data.

The FI operationalizes frailty [174] and is associated with adverse outcomes [159]. The FI is a number between 0 and 1 that is the average number of deficit health variables an individual has [159]. When calculating the FI, missing data treatment is typically not disclosed [171], and explicit imputation is seldom performed. Instead, the FI for each individual is typically computed by simply ignoring/dropping missing values, effectively replacing them with the average of the available variables. This is an implicit imputation strategy. A set of heuristics have built up around this ignoring strategy, such as inclusion criteria based on missingness: variables with more than 5% of individuals missing values may be excluded [162], as well as individuals with more than 20% of measurements missing [139, 19].

Per-individual and per-variable missingness can vary substantially between studies, as can the underlying missingness mechanism. As a result, heuristics that improve predictive performance of the FI in one study may affect another study differently.

This heterogeneity is an impediment to translating quantitative heuristics between studies, and limits the development of the FI as a precision tool [82]. An attractive potential alternative is to identify good imputation methods that work for a variety of types and magnitudes of missingness. The Rotterdam study shows that explicit imputation models can improve FI predictive power of mortality [171]. We ask, what is the best available imputation model to use when calculating the FI? More generally, how does the choice of missing data strategy affect the FI?

Multivariate imputation by chained equations (MICE) is a popular multiple imputation (MI) method freely-available in `R` [25]. The underlying engine of MICE is fully conditional specification (FCS), i.e. sequential regression or chained equations [130], which iteratively updates each missing variable or model parameter using the conditional distribution given all other variables and parameter estimates (i.e. Gibbs sampling) [198]. Multiple imputation generates a set of fully-sized, completed datasets which allows estimation of both quantitative results of interest and the uncertainty in those results caused by imputed values.

MICE has been shown to out-perform ignoring missing data [198], classical approaches including kNN (k-nearest neighbours) [12], and even deep learning methods [85, 203]. MICE is popular due to its flexibility, and availability in most statistical software (e.g. `python` [12, 138], `R` [25] and `stata` [205]).

Conversely, MICE can produce strong biases, putatively when too many variables/predictors are included [41, 70]. Since the underlying FCS approach is not theoretically grounded [130], all MICE models must be validated empirically. This may explain why the default MICE option in `R` for treating continuous-valued variables is predictive mean matching (PMM), an *ad hoc* model from the 1980s that has significant limitations [197, 1], but has been widely validated [197] (e.g. [12]).

Here we compare three MICE algorithms for gerontological data: Default (PMM for continuous variables and logistic regression for ordinal variables), CART (classification and regression tree) and RF (random forest) [25]. We also include two single-imputation strategies in our comparison: a classical *de facto* strategy, kNN[2]

[2]Single imputation strategies are appealing since they only require an additional pre-processing step wherein values are imputed, then analysis can continue as usual. kNN (k-nearest neighbours) has become a popular approach since it frequently shows good performance (e.g. [85]). Unfortunately, it has also shown bias in some tests (e.g. [41]).

[103], and a modern machine learning approach, `missForest` [179]. kNN is a popular, conventional approach that has been shown to out-perform individual (row)–mean imputation for gene expression data [195]. In contrast, RF approaches are contemporary machine learning models that have been shown to modestly out-perform kNN in numerous datasets [185, 179]. `missForest` is a variant of FCS that includes an automatic stopping strategy to prevent over-fitting and uses an RF prediction model.

The inclusion of *a priori* expert knowledge may enhance imputation, but presents a barrier-to-entry for non-experts. In the present study we tested inclusion of rule-based imputation (RI) for cases of study design–related missingness. Young, ostensibly healthy individuals were not asked questions specific to older and/or frailer individuals. In RI we assumed these missing values were optimally healthy. Only a subset of the missing values were missing due to study design, and therefore RI was always paired with another missing data handling strategy.

We do not consider other imputation models, including joint modelling, which conventionally requires the underlying distribution [25]. Other recent developments in imputation include tensor factorization [199], and deep learning [85, 203, 65, 151, 52].

We analyzed the effects of missing data and imputation for the National Health and Nutrition Examination Survey (NHANES) cross-sectional data [29]. Our objective was to investigate the effects of missing data and imputation on estimating the FI values and subsequent survival prediction. First, we identified and grouped individuals by their patterns of missingness. We then used these observed patterns to artificially simulate missingness in order to test the performance of imputation strategies when the true values were known. We compared the FI-typical ignoring strategy to several versions of MICE and determined which strategy best reproduced the true FI and which gave the best mortality prediction. Using what we learned, we then applied the most promising imputation strategies to the naturally missing data.

4.2 Missingness Mechanisms

There are three canonical missingness mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [112]. These can be defined in terms of [197],

$$Y_{all} = (Y_{obs}, Y_{mis}), \tag{4.1}$$

where Y_{all} is the matrix of all potentially measured values of interest, including all predictors and outcomes. Y_{obs} are observed values and Y_{mis} are missing. The missingness indicator is a matrix, M, with the same dimensions as Y_{all}, where $M_{ij} = 1$ indicates that variable j is missing for individual i.

By definition, values are MCAR if:

$$MCAR: \quad \Pr(M = 1|Y_{all}) = \Pr(M = 1) \tag{4.2}$$

where Pr indicates the matrix of probabilities[3]. For example, if you are interrupted while entering data and skip an arbitrary entry from an arbitrary individual, then that entry is MCAR. We expect that ignoring MCAR data will produce unbiased results [165].

MAR is defined by values for which:

$$MAR: \quad \Pr(M = 1|Y_{all}) = \Pr(M = 1|Y_{obs}) \tag{4.3}$$

For example, in the personal fitness questionnaire (PFQ) of NHANES 03/04 and 05/06 qualifying participants were asked PFQ061A: "how much difficulty {do you/does SP} have managing {your/his/her} money?" These data are only present for participants whom were aged 60 or older, or answered "yes" to PFQ049, PFQ057 or PFQ059, therefore PFQ is MAR, so long as we know these (auxiliary) variables. When the data are MAR we may produce biases if we ignore the missingness, however, with a sufficiently powerful imputation model we can use Y_{obs} and covariates to estimate the missing values.

Finally, MNAR is defined by:

$$MNAR: \quad \Pr(M = 1|Y_{all}) = \Pr(M = 1|Y_{obs}, Y_{mis}) \tag{4.4}$$

For example, suppose an individual is left to fill out a survey on their own, they read VIQ071: "{have you/Has SP} ever had a cataract operation?", but because they have never had problems with cataracts they skip the question entirely. If the data are MNAR then a proper treatment will require knowledge of the missingness mechanism since the dependence on Y_{mis} could cause severe biases. Nevertheless, due

[3] $\Pr(M = 1|Y_{all}) = \Pr(M = 1)$ reads as: the probability of being missing, conditional on knowing all observed values, is equal to the probability of being missing irrespective of the observed values.

to correlations in the data we may be able to achieve satisfactory results using an imputation model that assumes MAR, such as the imputation models we tested in this study.

Missingness patterns in the missingness matrix, M, may also cause problems. Missingness patterns are a higher-order statistic that represent whether variables tend to go missing together. Such patterns can apply to each of MCAR, MAR, and MNAR. For example, because of study design many of the variables in PFQ are often mutually missing. Similarly, individual limitations may prevent data collection of multiple related variables. In this paper we include a prefix 'p' to indicate the use of patterns (e.g. pMCAR) or 'c' to indicated conventional or cellwise missingness (e.g. cMCAR).

4.3 NHANES Data

We used the combined 2003/04 and 2005/06 NHANES (National Health and Nutrition Examination Survey) cross-sectional study with public-use, linked mortality files from the National Death Index [19], with a total of $N = 9307$ individuals. Inclusion criteria were: over age 20 ($N = 9816$), available survival data ($N = 9310$), and survival at least one year post study date ($N = 9307$). We followed two analysis pipelines: first we investigated real missingness by analyzing the entire, "Full", dataset ($N = 9307$), and then we isolated the $N = 1923$ complete-case, "Complete", dataset (individuals who had all 68 Frailty Index variables reported). The Complete dataset was used to test imputation strategies by simulated missingness together with ground truth (GT) values.

We calculated the combined lab plus self-reported (SR) FI using the methodology of Blodgett *et al* [19]. We included 32 lab variables and 36 SR health variables to calculate the FIs (Appendix Tables A.23 and A.24, respectively). SR health variables were linearly scaled to the range [0,1], while the lab variables were defined as 0 if they were within sex-specific healthy ranges or 1 if they were outside of those ranges (Appendix Table A.23). Lab variables were converted to binary scale *after* imputation to maximize the information available during imputation. SR variables were converted before imputation for coding convenience, but maintained their ordinal type. We used 100 additional variables to test the utility of auxiliary variables for

Table 4.1: NHANES Dataset Summary

	Full	Complete[1]
N	9307	1923
Males	4465 (48.0%)	944 (49.1%)
Females	4842 (52.0%)	979 (50.9%)
Age [median (IQR)]	48 (33-66)	68 (62-76)***
Age 60+	3232 (34.7%)	1635 (85.0%)***
Age under 60	6075 (65.3%)	288 (15.0%)***
Frailty Index[2][mean (sd)]	0.144 (0.078)	0.176 (0.073)***
Deaths	1016 (10.9%)	379 (19.7%)***
Death Age [median (IQR)][3]	81.5 (80.7-82.7)	81.2 (78.3-83.9)***
Missingness[4]	14.5%	0%***
Aux[5]Missingness[4]	12.8%	5.7%***

[1] Comparisons are between individuals in the Complete subset versus the remaining individuals.
[2] Using Ignore.
[3] Log-rank test.
[4] Cellwise missingness rate.
[5] Aux: auxiliary variables.

improving imputation performance (detailed in Appendix A).

Demographic information is summarized in Table 4.1. Individuals in the complete-case dataset were older ($p < 2.2 \cdot 10^{-16}$), frailer ($p < 2.2 \cdot 10^{-16}$), died more often ($p < 2.2 \cdot 10^{-16}$), and had a worse survival curve ($p = 7.2 \cdot 10^{-4}$), relative to the Full dataset.

4.4 Methods

4.4.1 Real Missingness

We directly analyzed missingness of the 68 FI variables in the Full dataset, which we refer to as 'real' missingness. We used the `md.pattern` function in `R` [25] to estimate missingness patterns in the Full dataset.

4.4.2 Simulated Missingness

The process of generating synthetic data with missing values is called "amputation" [172]. Amputation should respect the missingness mechanism (MCAR, MAR, or MNAR) and any salient patterns. MICE incorporates a standardized amputation

approach using missingness patterns [172], which we modified to handle larger quantities of data (see Appendix A). These patterns ensure that amputated data preserve missingness idiosyncrasies. For example, a pair of variables observed with 10% mutual-missingness are amputated together 10% of the time.

To simulate missingness, we took the Complete dataset and amputated values using the missingness patterns of the Full dataset. This generated a new dataset of the same size but with empty cells representing missing data. In contrast to real missingness, we retained the Complete dataset, providing us with a GT against which we compared our imputed values. Figure 4.1 illustrates missingness mechanisms and simulated missingness of mean arterial pressure when no higher-order missingness patterns are present.

Amputation was performed using four missingness mechanisms: cellwise MCAR and MNAR (cMCAR and cMNAR, respectively), and patterned MCAR and MAR (pMCAR and pMAR, respectively). The patterns restricted our maximum simulated missingness to the same level as the real missingness; we chose rates of 5%, 10% and 15% (max). We used the same rates for cellwise missingness, but we were also able to simulate 25%, 50% and 75% missingness for both cMNAR and cMCAR. Selection data were normalized to a $[0, 1]$ deficit scale prior to amputation to prevent problems with the two-sided deficit rule for the lab variables (see Appendix Table A.23). cMCAR randomly, and arbitrarily, selected data points to drop without any patterning. pMCAR and pMAR used the NHANES patterns determined from the Full dataset [172]. We confirmed the patterns were correctly reproduced in the simulated missingness — compare Appendix Figures A.1 versus A.2. We used default settings for both pMCAR and pMAR, with a probabilistic linear decile exclusion rule [172]. cMNAR is a novel cellwise approach wherein we applied cuts directly to the pooled quantiles using the linear decile exclusion rule. Given that the amputation process is stochastic, we generated 10 datasets for each combination of missingness mechanism, patterns, and rate.

4.4.3 Imputation Modelling

We performed imputation using the MICE package (version 3.10.0) [25] in `R` version 4.0.0 [152]. MICE uses FCS to iteratively impute missing data using a prediction

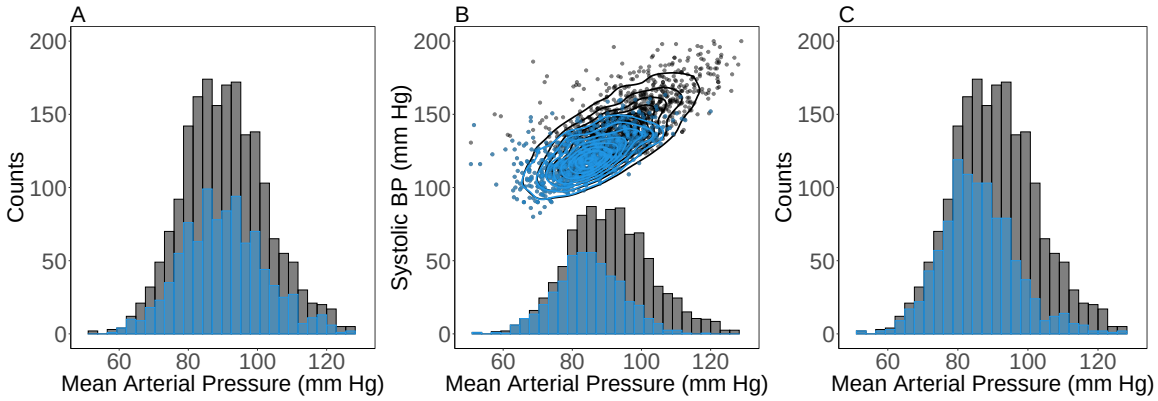

Figure 4.1: Illustration of missingness mechanisms using complete-case NHANES blood pressure (BP) data. Black bars and points reflect the true distribution, blue bars and points are simulated distributions of observed values after applying different missingness mechanisms. A) In missing completely at random (MCAR) the shape of the distribution is preserved but the total amount of data is reduced. B) In missing at random (MAR) data are preferentially excluded according to other related variables. In this case, individuals with large values of systolic BP were preferentially set to missing (points), causing a small bias in the mean arterial pressure distribution (bars). C) In missing not at random (MNAR) the value of missing variables affects the probability they are missing. In this example, we preferentially excluded high mean arterial pressure values.

model. We compared a representative sample of prediction models within MICE: logistic regression (logreg), predictive mean matching (PMM), classification and regression trees (CART), and random forest (RF). Logistic regression is the default for binary, ordinal and categorical data, whereas PMM is the default for continuous variables. CART is the special case of a RF with 1 tree — both accept mixed data types. We imputed the default number of times, m = 5, and combined results using Rubin's rules [205], except when estimating predictive power (we used the average) and visualizing the FI distributions (we used all values).

Rubin's rules describe how to properly aggregate multiple imputations to estimate both the expected effect (the average), and the uncertainty due to missing values, using an analysis of variance (ANOVA)–style decomposition of the between and within–imputation variance. The recommended number of imputations is approximately equal to the percentage of missing data [205], but a smaller number has conventionally been regarded as sufficient [197]. As a sanity check, we have also included a CART m = 15 imputation for each of our $\leq$ 15% simulated missingness

Table 4.2: Imputation Model Summary

Name	Model(s)	MI[1]	Note
RI[2]	–	No	Imputed gated as 0.[3]
Ignore	Row-mean[4]	No	Typical approach
Ignore (Weighted)[5]	Row-mean[4]	No	Linear weights
Ignore20	Row-mean[4]	No	20% missingness cut
RF	RF	No	100 trees
kNN	kNN	Yes	10 trees
MICE Default	PMM/logreg[6]	Yes	–
MICE CART	CART	Yes	1 tree
MICE CART+Aux	CART	Yes	100 auxiliary variables
MICE RF	RF	Yes	10 trees

1 MI: Multiple imputations
2 RI: rule-based imputation.
3 Gated variables were PFQ, RXD and VIQ blocks (Appendix Table A.5).
4 Mean value of the available deficit data for each individual.
5 Results in Appendix A.
6 PMM for continuous (lab) variables, logreg for ordinal/binary (self-reported) variables.

tests.

We also tested two single imputation (non-MICE) algorithms: kNN [103] and RF [179]. Our imputation models are summarized in Table 4.2.

A priori we know that the PFQ061 variables we used — the PFQ variable block (Appendix Table A.5) — and the RXD variable block are all gated variables, meaning data are missing purposely as part of the study design. Individuals under age 60 whom answered 'no' to PFQ049, PFQ057 and PFQ059 were not asked the PFQ block questions. The RXD block was not asked for individuals whom answered 'no' to RXDUSE. In addition, the VIQ variable block was not asked for individuals under age 50 [29]. We considered RI (rule-based imputation) wherein all of the aforementioned types of missing values were assumed to be optimally healthy (0 deficit). We applied RI to the real missingness, supplemented by a variable secondary imputation strategy for the residual missingness. RI was not applied to the simulated missingness because it was based on the Complete dataset which has no missing values and therefore the conditions for RI are not satisfied by any individuals.

We also considered inclusion of 100 auxiliary variables to enhance results. Preliminary results indicated that CART was the best-performing, hence we tested auxiliary variables with CART+Aux.

The FI is typically calculated using available-case analysis, which uses all available data from included individuals [112]. We considered three versions of available-case analysis. In the first, typical, approach missing values were simply ignored when calculating the FI. Second, we considered Ignore20, which excluded individuals with over 20% missingness from analysis and ignored missing values for included individuals [19]. Finally, in Appendix A, we considered weighting individuals in any analysis by the fraction of reported variables each individual has; statistics were only calculated when weighted models were readily available — excluding the area under the receiver operator characteristic curve (AUC) and the hazard rate/ratio (HR).

4.4.4 Statistical Analysis

Our focus was on how imputation strategies affected the FI — including the mean, distribution and downstream measures calculated from it, such as the HR and AUC. Simulated missingness was compared to the GT (ground truth). For real missingness the GT was unknown and we had to infer imputation quality by comparing results to the simulated missingness and assessing survival predictive power.

Survival prediction was based on 4-year-survival using the AUC [155]. 4-year-survival was selected because almost all individuals (excluding 2 in the Full dataset: 1 in the Complete dataset) had survival followup for at least 4 years. Preliminary results showed identical trends using 1, 2 or 4 year survival; final results were confirmed by comparing AUC to the C-index (Appendix A).

We calculated the age/sex adjusted Cox proportional hazards model as was previously done after imputing the Rotterdam study [171]. Analysis of deviance was used to assess predictive power [190]. The FI was scaled by 100 such that the HR was the increase in hazard per 0.01 increase in FI, consistent with most FI survival studies [101]. Differences in survival were tested for using the log-rank test.

To summarize the measures of survival predictive power, we used the AUC, the HR, analysis of deviance and the C-index (Appendix A). The AUC [69] and the C-index [24, 72] are close-relatives, both are descendants of the Wilcoxon non-parametric

statistic. The AUC estimates the probability that a metric will correctly rank the members of the affected group ahead of the members of the unaffected group [69] e.g. the probability that individuals whom will die during the next 4 years currently have higher FIs than non-terminal individuals. The C-index estimates the probability that, for every possible pair of individuals, a metric will correctly rank which individual will be affected first, e.g. die first [72]. Analysis of deviance is a generalization of the residual sum of squares [86] and attributes dispersion (deviance) explained by each variable. The HR is a regression parameter [127] and depends on the quality/validity of the fit and the scale of the data; it is an estimate of the relative change in hazard due to a per unit increment in the predictor variable.

Multiply imputed FIs were aggregated by the mean for each individual when analyzing survival predictive power to allow fair comparison to single imputation strategies, since the multiple imputations artificially increase variability in the FI, and therefore would likely reduce predictive power

FI distributions were compared using the Kolmogorov-Smirnov (KS)–test. Binary group comparisons of continuous variables were made using Mann-Whitney test, avoiding the complication of pre-testing [156]. Categorical vs categorical comparisons used Pearson's χ^2-test. Survival curves were estimated using the Kaplan-Meier estimator with respect to age. AUCs were compared using the Delong-test [40]. Note that the Delong-test includes an additional $1/N$ term in the test statistic which allows significant p-values even when the standard errors overlap [40]. Generic tests for significance used the z-test. Statistical significance is indicated with $* : p < 0.05$, $** : p < 0.01$, or $*** : p < 0.001$. All confidence intervals are 95%. Error bars are standard errors, error is reported in parenthesis from last digit e.g. $0.0034(12) = 0.0034 \pm 0.0012$.

4.5 Results

4.5.1 Missingness Patterns

As illustrated by Figure 4.2, we observed substantial missingness. In the Full dataset we observed an overall missingness of 14.5% (91585 entries), the mean missingness per individual was 9.8 entries, with a median of 12 entries (17.6%) and an inter-quartile

range (IQR) of 1 to 15 entries (1.5-22.1%). Individuals aged 60+ had significantly less missing data than individuals under 60 ($p < 2.2 \cdot 10^{-16}$) and died more often ($p < 2.2 \cdot 10^{-16}$). For individuals at least 60 years old, the mean missingness was 2.5 entries, with a median of 0 entries and an IQR of 0 to 1 entries (0-1.5%), with a death rate during followup of 26.7% versus 2.5% for people under 60. Considering the Full population, while 3606 (38.7%) of individuals were missing more than 20% of their entries only 203/3606 (5.6%) were at least 60 years old. This means that 3403/6075 (56.0%) of individuals under 60 did not pass the Ignore20 cut versus 203/3232 (6.3%) of individuals aged 60+, raising the prospect of age-related biases with Ignore20.

Missingness was not independent across variables, with distinct blocks of missingness forming in the mutually-missing histogram, Figure 4.2, particularly for younger individuals (under 60). Following the NHANES naming convention, these blocks were: personal fitness questionnaire (PFQ), number of prescription drugs taken (RXD), vision questionnaire (VIQ), blood pressure measurement (BPX), lab measurements (LB) and miscellaneous (Misc). As shown in Figure 4.2 the most commonly missing variables overall were the PFQ block of data, with an average cellwise missingness of 53.6% (80.7% for individuals under 60); at least one was missing 61.3% of the time (83.5% for individuals under 60). (See Appendix Table A.6 for block variable demographics.)

As shown in Figure 4.2B, the missingness of older individuals (age 60+) was markedly different. We observed lower overall missingness, higher variance of cellwise missingness within blocks, and no visible block missingness for PFQ or VIQ. These are study-design effects: PFQ was not routinely collected for individuals under age 60, while VIQ was not routinely collected for individuals under age 50 [29].

4.5.2 Missingness-Survival Effects

Kaplan-Meier survival curves showed that the variable blocks had heterogeneous effects on survival, Figure 4.3. With some blocks of variables showing significantly better survival for unmeasured individuals while others showed significantly worse survival. The red curves represent individuals with any entry missing in that block whereas the black curves had all variables observed. The overall missingness (Figure 4.3A) instead compared the individuals with above average missingness (red) vs

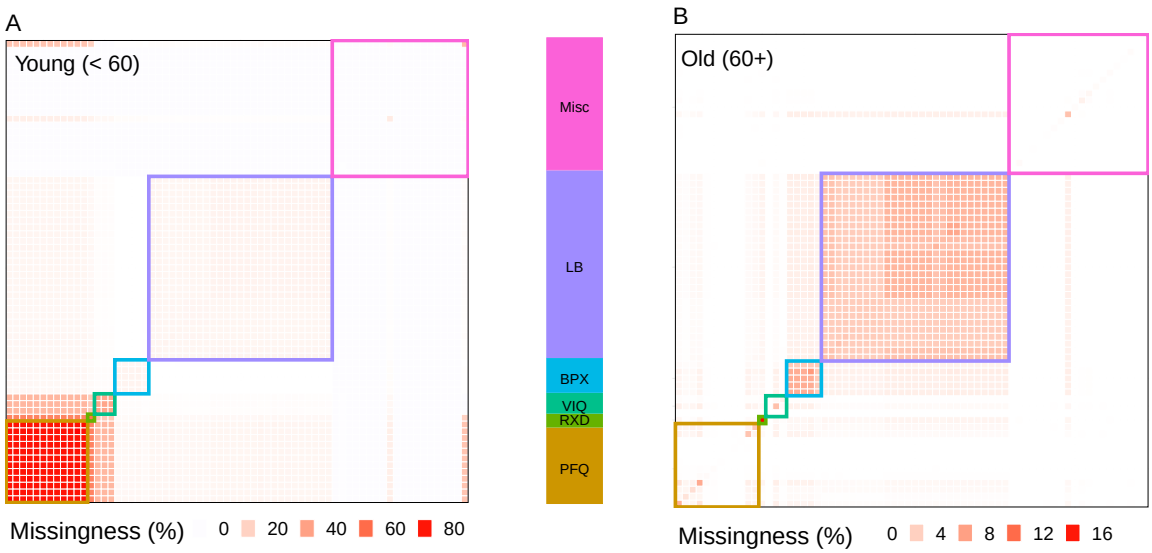

Figure 4.2: Mutual missingness histogram. Missingness fraction of NHANES variables for individuals: A) under age 60 and B) age 60+. These 2D-histograms give the mutual missingness fraction for (row, column) pairs of variables with the diagonal corresponding to each variable's overall missingness. We see a distinct block structure indicating groups of variables that are (almost) always missing together, for example the BPX (blood pressure) 5-variable group appears as a 5x5 block. The variables in each block are provided in Appendix Table A.5. Observe that in B) the LB and BPX blocks dominate whereas the PFQ block is less often missing and contains unpatterned missingness (strong diagonal terms), in contrast to A). Note the scale difference; older individuals had much less missing data. See Appendix Figure A.1 for the pooled young and old, and Figure A.4 for the per-variable labeled result.

below average missingness (black).

Missing LB block meant poorer survival, as did VIQ — for older individuals, and BPX. Conversely, RXD indicated superior survival. The missing PFQ block had crossing survival curves, and was an excellent proxy for the full missingness, showing nearly identical trends for the survival curves. The overall missingness had a complicated effect on survival where missingness was advantageous at young ages but crossed to disadvantageous at older ages.

We also investigated hazard using a blockwise-missingness Cox model with important covariates (sex and age), see the insets of Figure 4.3. The HRs with respect to missingness qualitatively agree with the survival curves: PFQ missingness indicated good survival for the young and poor survival for the old, RXD missingness always

indicated poorer survival but was less severe for the old, and VIQ indicated no change in survival for the young and poor survival for the old. BPX and LB missingness indicated worse survival, with missingness of LB for the young being significantly worse then the old. The overall missingness HR for the young, Figure 4.3A, was less significant than the PFQ, demonstrating that although the PFQ is a good proxy there is a reduction in the strength of the survival effect. In summary, the Cox models confirmed that the HRs were typically significantly different from unity, and differed between the young (< 60) versus old individuals (≥ 60).

4.5.3 Missingness Biases the FI

As shown in Figure 4.4, the blocks did not contribute equally to the FI — in particular the distributions of FI contributions from the blocks are distinct. This suggests that missing an entire block of variables, such as we observed with patterned missingness, will lead to biases in the FI if we simply ignore the missing values (effectively imputing the grey dashed line in Figure 4.4).

This bias could be exacerbated by the Ignore20 exclusion rule. The block sizes were: 12 (PFQ), 1 (RXD), 3 (VIQ), 5 (BPX), and 27 (LB). For 68 variables, the Ignore20 exclusion rule cuts at $N = 13.6$, thus any individual missing the complete LB block would be excluded from analysis.

We can estimate potential bias by using simulated missingness. As shown in Figure 4.5, we note significant and increasing biases of the FI (orange squares, with the implicit ignore imputation strategy) as compared to the ground truth (black dashed line) — for both pMCAR and cMNAR simulated missingness.

For the patterned missingness observed in the NHANES data, we developed a quantitative model of how pMCAR missingness biases the FI. The model details are presented in Appendix 4.8. We see in Figure 4.5A that the approximate model solution (blue line) as well as the more complex exact model solution (red points) agree with the observed FI bias with pMCAR.

4.5.4 Testing Imputation with Simulated Missingness

Using simulated missingness, we explored how common imputation strategies affected the FI. Overall, we found that CART performed the best — and that using auxiliary

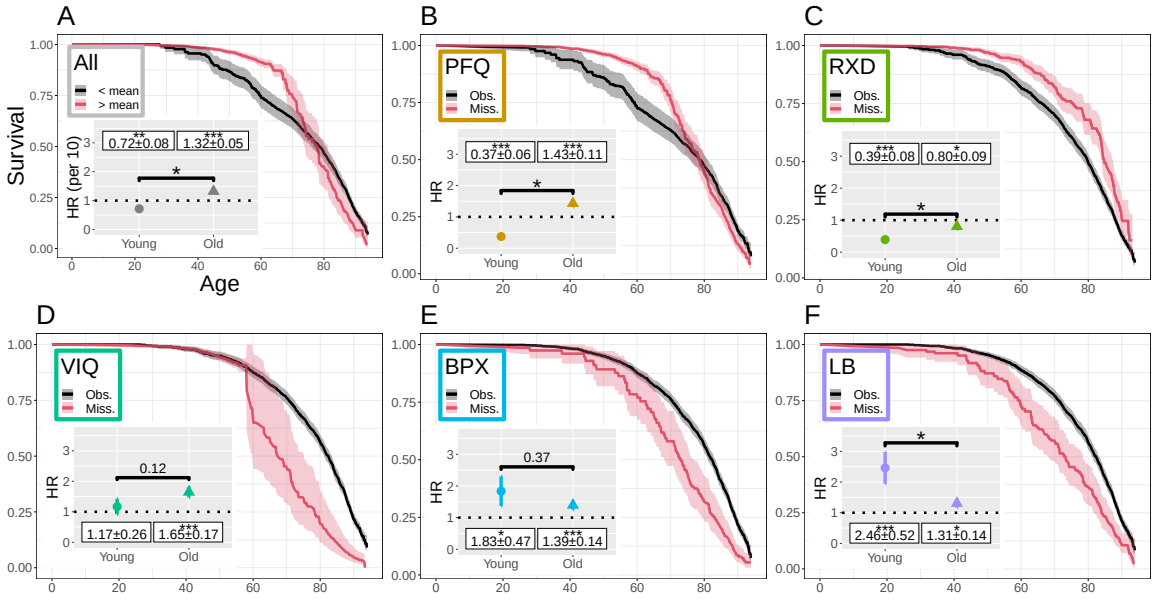

Figure 4.3: Survival and missingness. Survival curves conditioned on missingness show that the block patterns of missingness are strongly related to survival. A) all variables, B) personal fitness (PFQ), C) prescription drugs (RXD), D) vision (VIQ), E) blood pressure (BPX), and F) lab variables (LB). In A) the black line indicates the Kaplain-Meier survival curve for the subpopulation of individuals missing less than the mean (9.8 variables), the red line indicates individuals missing more than the mean. In B)-F), black lines indicate subpopulations without any of the variables in the block missing, red lines have at least one variable in the block missing. Shaded regions indicate 95% confidence intervals. Insets: hazard ratios (HRs) for Cox survival model for individuals stratified by young (< 60) or old (≥ 60), conditioned on age and sex. In A) the Cox model is HR per 10 deficits. In B)-F) each block Cox model was further conditioned on all other blocks (PFQ, RXD, VIQ, BPX and LB). Note the similarity of B) PFQ and A) all, reflecting that PFQ is a large block of variables and is the most commonly missing block. See Appendix Figure A.5 for age cut moved to 50, and Figure A.6 for additional variables.

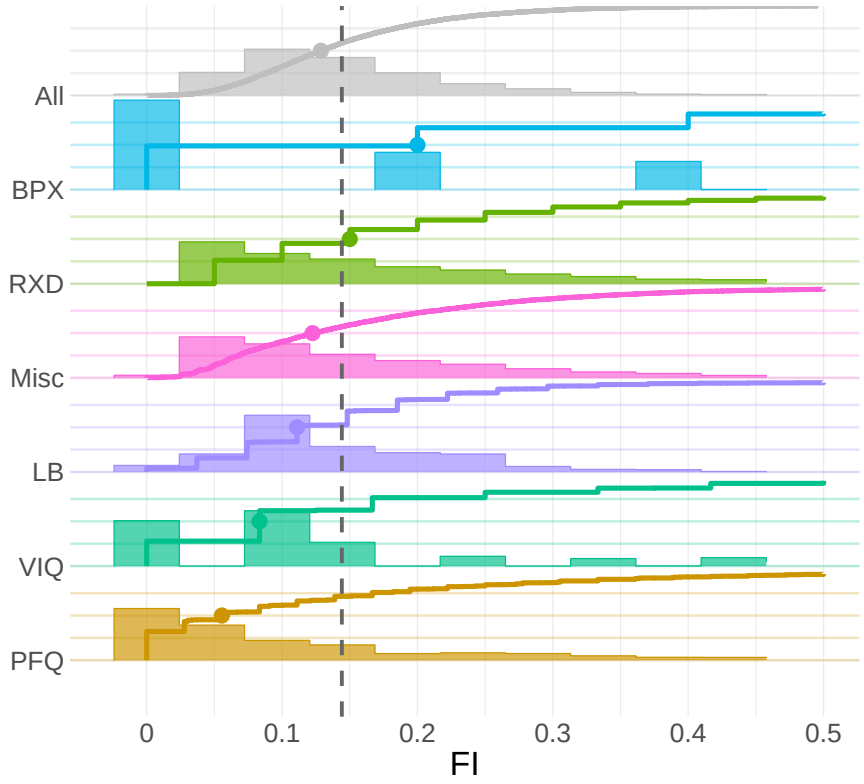

Figure 4.4: The distribution of block-specific FIs for different variable blocks (labels and fill colours correspond to Figures 4.2 and 4.3). Plotted values are the mean block FI across the population: bars indicate the histogram, lines indicate the cumulative distribution and filled circles indicate the median. y-axis grid lines indicate quartiles. The overall population mean FI, which is implicitly imputed by Ignore, is indicated by the dashed vertical grey line. Observe that the distributions vary considerably between blocks and the distributions are strongly skewed so that Ignore (dashed line) is typically well above the median. Plot is truncated at FI $= 0.5$ for visualization.

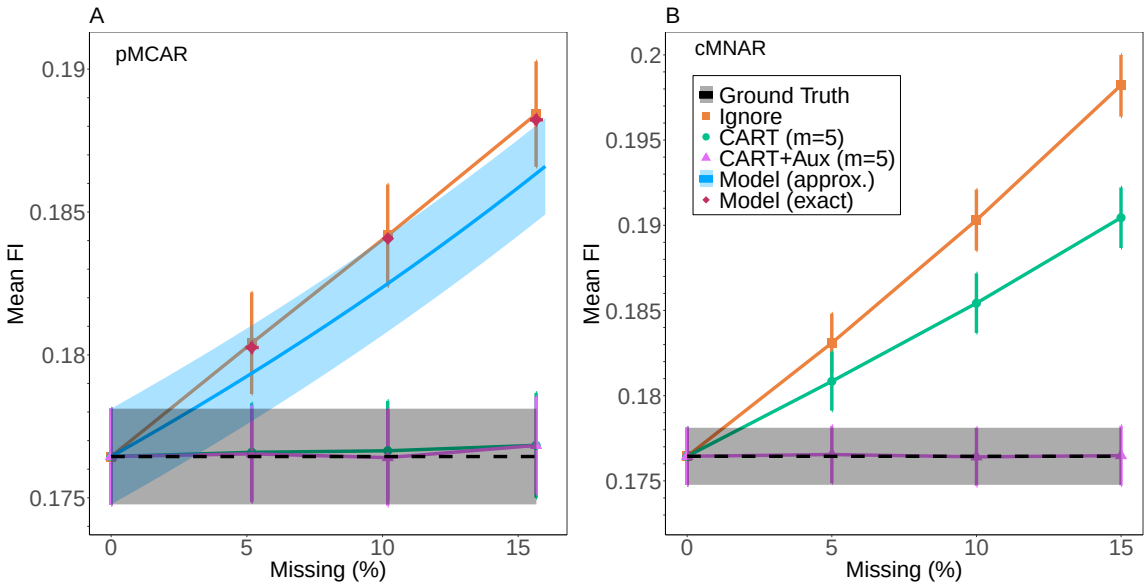

Figure 4.5: Missingness biases the FI. Using different percentages of simulated missingness of type A) pMCAR or B) cMNAR, we show the mean FI for different imputation strategies, as indicated by the legend. The typical, Ignore method (orange squares) shows the largest bias compared to the ground truth (black dashed), and for pMCAR the bias is captured by our approximate (blue line) and exact model (red diamonds), Eqs. 4.7 and 4.9, respectively. The bias is approximately linear in missingness. Our preferred imputation strategy, CART (green circles) eliminates the bias for pMCAR and reduces it for cMNAR. With the addition of auxiliary variables (pink triangles) CART eliminates the bias for both pMCAR and cMNAR. Error bars and intervals are standard errors. Complete plots for all types of simulated missingness and imputation are provided in Appendix Figure A.7.

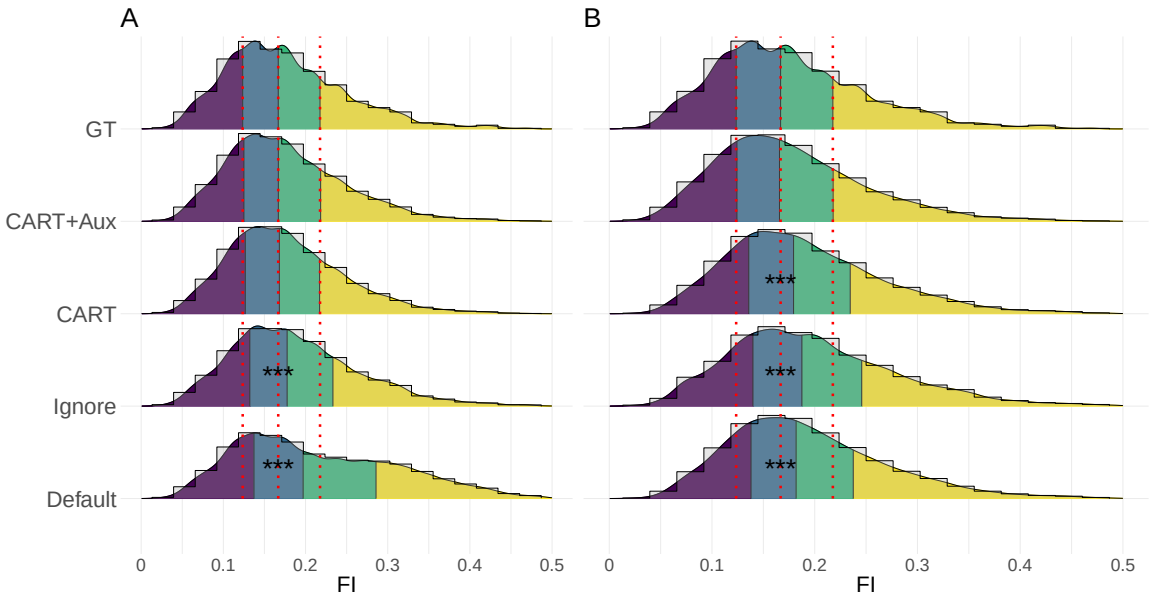

Figure 4.6: FI distributions by imputation type for simulated 15% missingness. A) pMCAR, B) cMNAR. Colours: quartiles. Vertical lines: GT quartiles. Stars: KS-test significance (vs GT). Default was the least similar to the GT for pMCAR whereas Ignore was the least similar for cMNAR. See Appendix Figure A.12 for FI distributions of additional imputation methods. All values from the m = 5 multiple imputations are included for Default, CART and CART+Aux without aggregation.

variables further improved CART performance with no apparent downside. Underperforming imputation strategies, including Ignore, led to significant biases to both the mean and standard deviation (SD) of the FI distributions.

Figure 4.6 shows the distributions of FIs for representative imputation methods at 15% missingness. Imputation of pMCAR caused an increased skew of the FI distributions for both Ignore and Default, but no significant changes when CART or CART+Aux were used. The changes due to the Default (PMM/logreg) imputation were very significant. cMNAR showed a similar pattern, although CART also skewed significantly, and Default skewed less than Ignore. The FI distributions for other imputation strategies are shown in Appendix Figure A.12.

We generally found that the bias in the estimated mean FI was linear for smaller values of missingness ($\leq 15\%$). This is illustrated in Figure 4.5 for CART and Ignore; for other imputation methods see Appendix Figure A.7. Accordingly, we estimated the bias per unit missingness, i.e. the bias rate, using a linear zero-intercept regression model. We also calculated the HR and AUC for each imputed FI at 15% missingness.

Table 4.3: Imputed FI Statistics — Cellwise Simulated Missingness

Imputation	Type	Mean[1]	Bias Rate[2,3]	SD[1]	SD Bias Rate[2,3]	HR[1]	AUC[1,4]
GT	–	0.176	0.000(0)	0.073	0.000(0)	1.075(7)	0.733(36)
Ignore	cMCAR	0.176	**0.000(1)**	0.076	0.014(1)***	1.070(7)	0.728(37)
Ignore20	cMCAR	0.176	**0.000(1)**	0.075	0.013(1)***	1.071(9)	0.729(41)
Default (m=5)	cMCAR	0.193	0.109(1)***	0.078	0.030(1)***	1.071(7)	0.734(37)
MICE RF (m=5)	cMCAR	0.188	0.073(1)***	0.076	0.017(2)***	1.073(7)	0.734(37)
RF	cMCAR	0.177	0.004(0)***	0.074	0.006(0)***	1.074(7)	0.732(37)
kNN	cMCAR	0.179	0.012(1)***	0.072	−0.009(0)***	1.076(7)	0.730(37)
CART (m=5)	cMCAR	0.177	0.002(1)**	0.073	**0.002(1)**	1.075(8)	0.732(37)
CART (m=15)	cMCAR	0.177	0.002(0)***	0.074	0.004(1)***	1.077(7)	0.733(37)
CART+Aux (m=5)	cMCAR	0.177	**0.000(1)**	0.074	0.004(1)*	1.076(8)	0.735(37)
Ignore	cMNAR	0.198	0.142(1)***	0.080	0.046(0)***	1.069(7)	0.732(37)
Ignore20	cMNAR	0.202	0.137(0)***	0.081	0.050(1)***	1.069(7)	0.735(39)
Default (m=5)	cMNAR	0.193	0.109(1)***	0.078	0.029(1)***	1.071(7)	0.733(37)
MICE RF (m=5)	cMNAR	0.188	0.074(1)***	0.076	0.017(1)***	1.073(7)	0.734(37)
RF	cMNAR	0.177	0.004(0)***	0.074	0.006(0)***	1.073(7)	0.733(37)
kNN	cMNAR	0.179	0.012(1)***	0.072	−0.009(0)***	1.076(7)	0.730(37)
CART (m=5)	cMNAR	0.190	0.092(1)***	0.077	0.025(1)***	1.072(7)	0.735(37)
CART (m=15)	cMNAR	0.190	0.092(1)***	0.077	0.023(1)***	1.074(7)	0.735(37)
CART+Aux (m=5)	cMNAR	0.176	**0.000(1)**	0.074	**0.002(1)**	1.075(7)	0.731(37)

[1] At 15% missingness.
[2] The bias rate is the theoretical bias at 100% missingness.
[3] p-value for t-test vs 0.
[4] p-value for vs Ignore; Ignore vs GT.
See Appendix Table A.3 for additional results. See Appendix Figure A.8 for forest plot of HRs.
Bold: noteworthy result.

The results are summarized in Tables 4.3, 4.4, 4.5, and 4.6. Blockwise summaries and the C-index are provided in Appendix Tables A.7 to A.14 (bias) and Tables A.15 to A.22 (predictive power).

As shown in Table 4.3, for the simplest missingness type, cMCAR, all of the imputation strategies except for Ignore and CART+Aux had significant bias rates. Default MICE (PMM) and Mice RF had large biases: > 0.01 for 15% missingness. For cMNAR, all of the bias rates were significant except CART+Aux, although both kNN and RF were small (compared to the SD).

When missingness patterns from NHANES were used to generate either pMCAR or pMAR, they also caused a severe bias in the estimated Ignore FI and an even worse bias in the MICE default, as shown in Table 4.4. The bias rate was significant for all imputation methods including Ignore, but was relatively small for kNN, CART and CART+Aux. CART+Aux achieved a bias of only 2.7% of the SD at the theoretical limit of 100% missingness.

Table 4.4: Imputed FI Statistics — Patterned Simulated Missingness

Imputation	Type	Mean[1]	Bias Rate[2,3]	SD[1]	SD Bias Rate[2,3]	HR[1]	AUC[1,4]
GT	–	0.176	0.000(0)	0.073	0.000(0)	1.075(7)	0.733(36)
Ignore	pMCAR	0.188	0.076(1)***	0.078	0.029(1)***	1.064(7)	0.729(37)
Ignore20	pMCAR	0.181	0.031(1)***	0.075	0.008(2)***	1.073(12)	0.733(50)
Default (m=5)	pMCAR	0.216	0.238(5)***	0.133	0.388(24)***	**1.041(7)**	0.697(40)***
MICE RF (m=5)	pMCAR	0.168	−0.055(1)***	0.068	−0.032(1)***	1.078(8)	0.732(37)
RF	pMCAR	0.161	−0.101(1)***	0.068	−0.035(1)***	1.076(8)	0.726(38)
kNN	pMCAR	0.176	0.014(5)*	0.072	**−0.002(2)**	1.071(8)	0.722(38)
CART (m=5)	pMCAR	0.177	0.002(1)*	0.073	−0.006(2)**	1.075(8)	0.733(37)
CART (m=15)	pMCAR	0.177	0.003(1)**	0.072	−0.009(1)***	1.080(8)	0.733(37)
CART+Aux (m=5)	pMCAR	0.177	0.002(0)***	0.073	**−0.002(1)**	1.076(8)	0.733(37)
Ignore	pMAR	0.187	0.067(1)***	0.075	0.004(1)**	1.070(8)	0.732(37)
Ignore20	pMAR	0.191	0.029(1)***	0.077	0.020(1)***	1.078(10)	0.742(47)
Default (m=5)	pMAR	0.216	0.244(5)***	0.121	0.279(22)***	**1.046(7)**	0.697(41)***
MICE RF (m=5)	pMAR	0.169	−0.044(1)***	0.071	−0.013(2)***	1.074(8)	0.732(37)
RF	pMAR	0.162	−0.092(1)***	0.072	**−0.004(2)**	1.070(7)	0.728(38)
kNN	pMAR	0.179	0.020(4)***	0.074	**0.002(1)**	1.071(7)	0.721(38)
CART (m=5)	pMAR	0.178	0.013(1)***	0.073	−0.004(2)**	1.075(8)	0.733(37)
CART (m=15)	pMAR	0.178	0.013(1)***	0.073	**−0.002(2)**	1.078(8)	0.735(37)
CART+Aux (m=5)	pMAR	0.177	0.005(1)***	0.073	−0.002(1)*	1.075(7)	0.734(37)

[1] At 15% missingness.
[2] The bias rate is the theoretical bias at 100% missingness.
[3] p-value for t-test vs 0.
[4] p-values for vs Ignore; Ignore vs GT.
See Appendix Table A.4 for additional results. See Appendix Figure A.8 for forest plot of HRs.
Bold: noteworthy result.

The SD of the FI was also significantly biased for most of the imputation strategies — including Ignore. CART had small bias rates, though still statistically significant, while kNN performed better than CART for cMNAR, pMCAR and pMAR, but worse for cMCAR. Overall, CART+Aux performed the best, having a consistently small bias rate.

Coverage is the probability that the true value of the mean FI was within the error interval of the imputed mean FI. CART+Aux had 100% coverage for missingness $\leq$ 15%, whereas kNN and the other imputation methods did not (see Appendix Table A.2). Excluding cMNAR, CART also had 100% coverage.

Increasing the number of imputations using CART from 5 to 15 made a trivial difference, yielding nearly identical results, see Tables 4.3 and 4.4. The bias rate of the mean did not change — nor did the coverage (Appendix Table A.2), while the changes to the bias rate of the SD appeared to be random and small.

In Table 4.5 we extended cMCAR to higher rates of missingness. We again observed that the ignore methods are unbiased estimators of the mean, as is CART+Aux. In contrast, kNN showed a large and significant bias rate. Furthermore the SD estimates were biased for all imputation methods. The smallest SD bias rate was observed for Ignore20 and CART+Aux — although Ignore20 excluded all of the data for missingness $\geq$ 50% and therefore could not be calculated. Interestingly, we saw significant reductions in HR and AUC at 50% and 75% missingness for the Ignore methods. Note the increasing HR for kNN likely masked the apparent drop in predictive power observed in the AUC. When 75% of data were missing, the mean FI decreased by 56% for kNN, the HR fit coefficient, $\beta = \log(\text{HR})$, had to increase by 56% to compensate for the shrinking scale, resulting in an expected HR of 1.12 — larger than the observed HR of 1.089 ± 0.021. CART+Aux significantly outperformed Ignore for 50% and 75% missingness (AUC).

Finally, we investigated cMNAR with higher missingness in Table 4.6. We observed that all of the imputation strategies produced large biases in the mean FI, including CART+Aux, illustrating the difficulty of imputing cMNAR.

We found that *when a relatively small fraction of data was missing*, the HR of the FI did not substantially vary across most imputation methods — notably excluding Default, as shown in Tables 4.3, 4.4, 4.5 and 4.6, and Figure A.8. As such, biases

Table 4.5: Imputed FI Statistics for High Simulated cMCAR Missingness

Imputation	Type	Mean[1]	Bias Rate[2,3]	SD[1]	SD Bias Rate[2,3]	HR[1]	AUC[1,4]
GT	0%	0.176	0.000(0)	0.073	0.000(0)	1.075(7)	0.733(36)
Ignore	25%	0.177	**0.000(1)**	0.077	0.035(2)***	1.068(7)	0.723(38)
Ignore20	25%	0.177	**0.000(1)**	0.077	0.013(4)*	1.079(27)	0.741(101)
kNN	25%	0.153	−0.086(1)***	0.061	−0.052(0)***	1.084(9)	0.716(38)
CART+Aux (m=5)	25%	0.177	0.001(0)*	0.073	0.013(4)***	1.076(8)	0.733(37)
Ignore	50%	0.177	**0.000(1)**	0.085	0.035(2)***	**1.055(7)**	0.699(40)**
Ignore20[5]	50%	–	–	–	–	–	–
kNN	50%	0.132	−0.086(1)***	0.048	−0.052(0)***	1.094(14)	0.685(41)
CART+Aux (m=5)	50%	0.176	0.001(0)*	0.079	0.013(4)***	1.074(9)	**0.729(37)*****
Ignore	75%	0.176	**0.000(1)**	0.106	0.035(2)***	**1.035(5)**	0.673(41)***
Ignore20[5]	75%	–	–	–	–	–	–
kNN	75%	0.113	−0.086(1)***	0.034	−0.052(0)***	1.089(21)	0.637(43)*
CART+Aux (m=5)	75%	0.177	0.001(0)*	0.085	0.013(4)***	1.076(11)	**0.732(38)*****

[1] At 15% missingness.
[2] The bias rate is the theoretical bias at 100% missingness.
[3] p-value for t-test vs 0.
[4] p-values for vs Ignore; Ignore vs GT.
[5] Insufficient data due to Ignore20 cut-off rule.
See Appendix Figure A.9 for forest plot of HRs.
Bold: noteworthy result.

in the FI affect the absolute but not relative risk assessed — comparing absolute FI between studies could cause discrepancies, but comparing relative FI within a study appears valid for most imputation strategies. Reinforcing this, the AUC was similar for most imputation strategies.

4.5.5 Imputation of Real Missingness

Given the success of CART when imputing against simulated missingness, we focused on testing this strategy with the observed (real) missingness. Ignore served as the *de facto* standard, and we included Default (PMM/logreg) and kNN for perspective. We also assessed RI as a prospective initial imputation step, which was paired with a subsequent model (Ignore, kNN, etc).

We observed a drop in FI with respect to Ignore for CART, CART+AUX and all of the RI-initialized methods, Table 4.7. In contrast, the FI for Ignore20 and kNN was greater than Ignore. We had no GT with which to directly observe whether the FI was biased for any particular imputation method. Using our quantitative model and assuming MCAR we estimated that the Ignore method should have a bias in the mean FI of 0.0028 using Eq. 4.7 (approximate) or 0.0029 using Eq. 4.9 (exact), both agree

Table 4.6: Imputed FI Statistics for High Simulated cMNAR Missingness

Imputation	Type	Mean[1]	Bias Rate[2,3]	SD[1]	SD Bias Rate[2,3]	HR[1]	AUC[1,4]
GT	0%	0.176	0.000(0)	0.073	0.000(0)	1.075(7)	0.733(36)
Ignore	25%	0.217	0.309(14)***	0.086	0.078(2)***	1.062(6)	0.728(37)
Ignore20	25%	0.252	0.127(1)***	0.095	0.086(2)***	1.068(17)	0.762(72)
kNN	25%	0.183	0.141(12)***	0.071	−0.013(0)***	1.076(7)	0.726(37)
CART+Aux (m=5)	25%	0.200	0.205(11)***	0.079	0.045(5)***	1.071(7)	0.736(37)
Ignore	50%	0.287	0.309(14)***	0.106	0.078(2)***	**1.048(5)**	0.715(37)*
Ignore20[5]	50%	–	–	–	–	–	–
kNN	50%	0.209	0.141(12)***	0.067	−0.013(0)***	1.078(8)	0.713(38)
CART+Aux (m=5)	50%	0.244	0.205(11)***	0.088	0.045(5)***	1.067(7)	**0.734(37)****
Ignore	75%	0.449	0.309(14)***	0.138	0.078(2)***	**1.032(5)**	0.693(39)**
Ignore20[5]	75%	–	–	–	–	–	–
kNN	75%	0.317	0.141(12)***	0.064	−0.013(0)***	1.063(10)	0.668(44)
CART+Aux (m=5)	75%	0.363	0.205(11)***	0.115	0.045(5)***	1.062(8)	**0.730(37)****

[1] At 15% missingness.
[2] The bias rate is the theoretical bias at 100% missingness.
[3] p-value for t-test vs 0.
[4] p-values for vs Ignore; Ignore vs GT.
[5] Insufficient data due to Ignore20 cut-off rule.
See Appendix Figure A.10 for forest plot of HRs.
Bold: noteworthy result.

well with the difference between Ignore and CART or CART+Aux. Notably, this estimate is far smaller than the difference between Ignore and RI-initialized methods, which were all > 0.01.

Based on the observed missingness patterns, however, we suspected that the data were primarily MAR, and hence we also estimated the bias after RI, which should have removed the majority of MAR missingness. The bias in the mean FI for Ignore+RI was -0.00059 (approximate) or -0.00060 (exact), which agrees excellently with the differences between Ignore+RI and CART+RI (-0.006 ± 0.002), and Ignore+RI and CART+Aux+RI (-0.005 ± 0.002).

In summary, CART (with or without Aux) appeared to consistently refine Ignore or Ignore+RI, removing the residual pMCAR-related bias. Our best estimate for the bias in the Ignore mean FI was 0.0106 ± 0.0001, which we calculated by adding the estimated bias in Ignore+RI to the difference between Ignore and Ignore+RI. This effectively assumed MAR missingness was corrected by Ignore+RI and the residual missingness was MCAR and hence could be correctly calculated using our missingness models, Eq. 4.7 and Eq. 4.9. The estimate agrees well with the difference between

Table 4.7: Imputed FI Statistics for Real Missingness

Imputation	Mean	'Bias'[1]	SD	SD 'Bias'[1]	HR[2]	AUC[3]
Ignore	0.1442	0.0000(0)	0.0782	0.0000	1.077(4)	0.832(17)
Ignore20[4]	0.1611	−0.0170(13)***	0.0801	−0.0019	1.078(4)	0.792(21)**
kNN	0.1601	−0.0160(4)***	0.0710	0.0071	1.073(4)	0.773(21)***
Default (m=5)	0.1466	−0.0024(5)***	0.0877	−0.0095	1.077(4)	0.829(18)
CART (m=5)	0.1412	0.0029(4)***	0.0816	−0.0034	1.079(4)	**0.839(17)****
CART+Aux (m=5)	0.1410	0.0031(3)***	0.0784	−0.0003	1.079(4)	**0.841(17)****
Ignore + RI	0.1330	0.0112(1)***	0.0803	−0.0021	1.077(3)	**0.851(16)****
Ignore20 + RI[5]	0.1327	0.0108(1)***	0.0774	0.0008	1.079(4)	0.848(17)
kNN + RI	0.1302	0.0140(2)***	0.0771	0.0011	1.076(4)	**0.841(16)****
Default+RI (m=5)	0.1338	0.0104(2)***	0.0790	−0.0009	1.077(4)	**0.850(16)****
CART+RI (m=5)	0.1336	0.0106(2)***	0.0789	−0.0007	1.079(4)	**0.851(16)****
CART+Aux+RI (m=5)	0.1334	0.0107(2)***	0.0786	−0.0005	1.079(4)	**0.852(16)****

[1] This is the bias proxy: Ignore − Value.
[2] HR per 0.01 increment in FI, conditioned on age and sex.
[3] p-value for vs ignore.
[4] N = 5701 individuals.
[5] N = 8728 individuals.
See Appendix Tables A.28, A.29 and A.30 for additional results.
Bold: noteworthy result.

the FI using Ignore versus either CART+RI or CART+Aux+RI.

In Table 4.10 we report the blockwise FIs for individuals under age 60, without RI. This was used to assess imputation quality. We observed that the blockwise FIs differed between imputation strategies. The qualitative survival effect of missingness ("Survival Frailty" column) was always the same direction as the CART and CART+Aux imputation strategies relative to Ignore, indicating good qualitative performance. For example, BPX missingness has a HR>1 and CART imputations have higher BPX block FI averages than Ignore. RI agreed with the qualitative Survival Frailty for PFQ and RXD, but not VIQ. By design, RI imputed 0 for PFQ, VIQ and RXD, but only PFQ and RXD had HR ≤ 1. Note that it is possible that the correct (latent) values to impute were slightly larger than 0.

In Table 4.11 we report the blockwise FIs for individuals age 60+, without RI. In contrast to Table 4.10, Ignore performed much better for the older individuals, with the FI in the same direction as the survival frailty in 2/5 blocks and for the overall FI, compared to CART and CART+Aux. Importantly, the Ignore strategy got the correct direction of the overall effect.

The FI distributions are in Figure 4.7. In Figure 4.7A we observed that, excluding

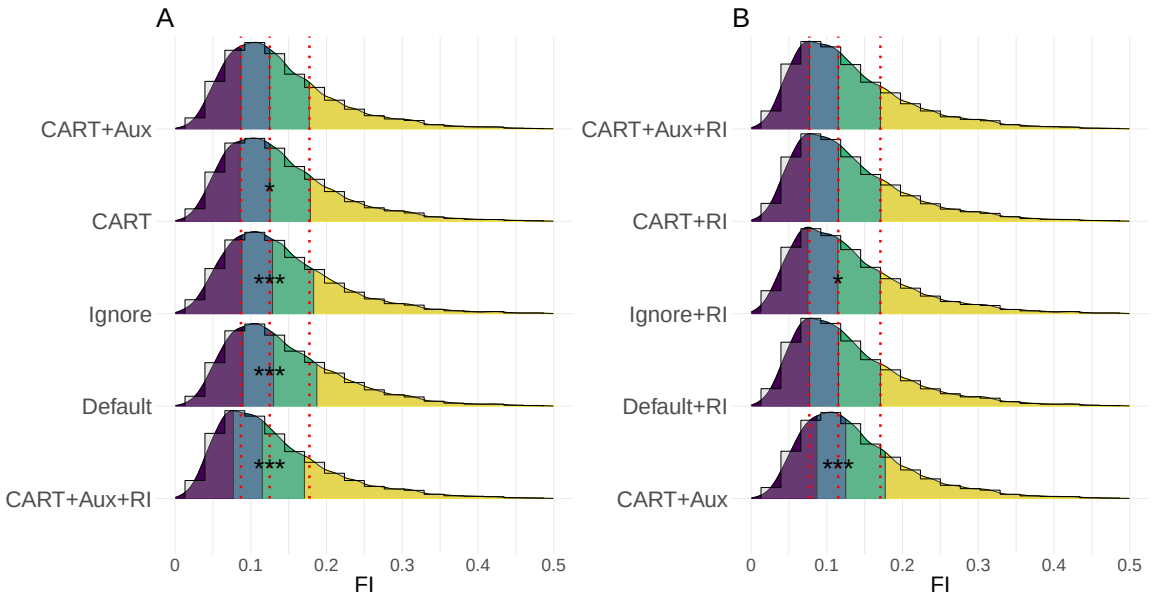

Figure 4.7: FI distributions by imputation type for Full dataset (real missingness). A) without rule-based imputation (RI), B) with RI. Observe that RI shifts the FI distribution to lower values (bottom row is duplicated from the other column for comparison). Colours: quartiles. Vertical lines are quantiles of: CART+Aux (A) or CART+Aux+RI (B). Stars: KS-test significance vs CART+Aux (A) or CART+Aux+RI (B). All values from the m = 5 multiple imputations are included for Default, CART and CART+Aux (including + RI) without aggregation.

RI, the MICE default was the least similar to the surrogate GT (CART+Aux), as was the case with pMCAR simulation — though with less skew than in Figure 4.6. The CART FI distribution was significantly different than CART+Aux, although the difference is not discernible by eye. Taken together, this suggests that the true missingness was somewhere between pMCAR and cMNAR, such as a combination of the two. This is at least partially consistent with our *a priori* expectations that PFQ, VIQ and RXD were pMAR, which was the foundation of our RI strategy.

There was a large shift visible between Figure 4.7A and 4.7B due to RI, as can be seen in the last row. In Figure 4.7B we observed only small differences between the distributions after RI was performed, with only Ignore+RI being significantly different from CART+Aux+RI. It appears that the values imputed by RI were particularly difficult for Ignore and Default to handle, in the latter case we infer that, consistent with Figure 4.6, patterned missingness — which RI imputes — seems to be especially difficult for Default to handle (see also Appendix Figure A.8).

The prediction accuracy for the real missingness is given in Table 4.7. We observed that, relative to Ignore, there was a significant increase in AUC for both CART ($p = 1.7 \cdot 10^{-6}$) and CART+Aux ($p = 5.6 \cdot 10^{-11}$) methods. The largest changes were significant decreases in AUC for the Ignore20 method ($p = 0.0046$, unpaired) and kNN ($p < 2.2 \cdot 10^{-16}$). All of the RI-enhanced imputation strategies out-performed the Ignore method by AUC, except Ignore20+RI. The best AUC belonged to CART+Aux+RI, with an estimated bias of 0.0107 ± 0.0002 versus Ignore — in agreement with our calculated bias, and an HR of 1.079 ± 0.004, implying that the FI hazard would differ by 1.085 ± 0.005 between the two imputation strategies. The HRs are plotted in Figure A.11.

Investigating the effects of missingness via the Cox model, we confirmed that missingness is a significant predictor of mortality — with or without considering age and sex, and had a strong interaction effect at age 60, Tables 4.8 and 4.9. The interaction term causes the direction of the hazard to change from protective (age < 60) to dangerous (age ≥ 60). We also considered changes due to the FI, and considered several imputation strategies (MIs were aggregated as mean). We observed similar results with and without RI.

We observed a large drop in the predictive power of missingness when conditioned on the Ignore FI but not any other FI (Table 4.8), implying that the Ignore FI captured the missingness survival effect. For the other imputation strategies, the FI reduced the predictive power of missingness conditioned on being young. We saw no significant differences in predictive power of the FI between the different imputation methods. The deviance may be less sensitive to differences in predictive power than the AUC, because the deviance carries the underlying assumptions of the Cox model. Note that there was a clear FI position dependence in the predictive power of sex, probably due to sex differences in the FI (e.g. [83]), which appears to have bolstered the predictive power of the FI in Table 4.9, and which complicates direct comparison of the FI deviance between Table 4.8 and Table 4.9.

4.6 Discussion

Deng *et al* [41] and Sterne *et al* [180] showed that either ignoring missing data or carelessly imputing values can adversely affect results. We investigated missingness

Table 4.8: Cox Hazard Analysis of Deviance — FI First

	Model[1]	Miss	Miss|Young[2]	Deviance Age	Sex	FI
{1}.	Miss	15(9)***	–	–	–	–
{2}.	{1}+Miss|Young	15(9)***	26(13)***	–	–	–
{3}.	RIDAGEYR+{2}	15(9)***	24(13)***	7(6)**	–	–
{4}.	RIDAGEYR+RIAGENDR+{2}	19(10)***	27(13)***	7(5)**	42(13)***	–
{5}.	$FI_{(Ignore)}$+{4}	**3(5)**	7(7)**	0(1)	84(19)***	406(43)***
{6}.	$FI_{(Default)}$+{4}	18(10)***	12(9)***	0(1)	80(18)***	399(43)***
{7}.	$FI_{(CART)}$+{4}	18(10)***	12(8)***	0(1)	80(19)***	405(43)***
{8}.	$FI_{(CART+Aux)}$+{4}	17(9)***	12(8)***	0(1)	78(18)***	402(44)***
{9}.	$FI_{(Ignore+RI)}$+{4}	5(6)*	**3(5)**	0(1)	84(19)***	416(42)***
{10}.	$FI_{(Default+RI)}$+{4}	16(9)***	9(7)**	0(1)	79(18)***	404(42)***
{11}.	$FI_{(CART+RI)}$+{4}	17(9)***	8(7)**	0(1)	80(19)***	409(44)***
{12}.	$FI_{(CART+Aux+RI)}$+{4}	19(10)***	9(7)**	0(1)	78(18)***	409(45)***

[1] Deviance was calculated sequentially.
[2] X|Y denotes an interaction between X and Y ("X given Y").
The null model had deviance (−2·log-likelihood) 12480. Young (< age 60) was dropped (never significant).
p-value is z-test versus 0.
Errors by bootstrapping $N = 1000$.
Bold: noteworthy result.

Table 4.9: Cox Hazard Analysis of Deviance — FI Last

	Model[1]	Miss	Miss|Young[2]	Deviance Age	Sex	FI
{1}.	Miss	15(9)***	–	–	–	–
{2}.	{1}+Miss|Young	15(8)***	26(14)***	–	–	–
{3}.	RIDAGEYR+{2}	15(9)***	24(13)***	7(5)**	–	–
{4}.	RIDAGEYR+RIAGENDR+{2}	19(10)***	27(13)***	7(5)**	42(13)***	–
{5}.	{4}+$FI_{(Ignore)}$	19(10)***	27(13)***	7(6)**	42(13)***	405(44)***
{6}.	{4}+$FI_{(Default)}$	19(10)***	27(13)***	7(6)**	42(13)***	413(43)***
{7}.	{4}+$FI_{(CART)}$	19(10)***	27(13)***	7(5)**	42(14)***	419(42)***
{8}.	{4}+$FI_{(CART+Aux)}$	19(10)***	27(13)***	7(5)**	42(14)***	413(44)***
{9}.	{4}+$FI_{(Ignore+RI)}$	19(10)***	27(13)***	7(5)**	42(14)***	412(44)***
{10}.	{4}+$FI_{(Default+RI)}$	19(10)***	27(13)***	7(5)**	42(14)***	412(43)***
{11}.	{4}+$FI_{(CART+RI)}$	19(10)***	27(13)***	7(5)**	42(13)***	419(42)***
{12}.	{4}+$FI_{(CART+Aux+RI)}$	19(10)***	27(13)***	7(5)**	42(13)***	420(43)***

[1] Deviance was calculated sequentially.
[2] X|Y denotes an interaction between X and Y ("X given Y").
The null model had deviance (−2·log-likelihood) 12480. Young (< age 60) was dropped (never significant).
p-value is z-test versus 0.
Errors by bootstrapping $N = 1000$.
Bold: noteworthy result.

with NHANES data to understand if and how the FI changes, and how well the commonly available imputation models perform. We considered both standard Ignore and Ignore20 approaches with the FI, together with a number of explicit imputation strategies including multiple imputation.

The powerful and commonly used imputation strategy, MICE via FCS, is not formally self-consistent. FCS builds predictive distributions for each variable conditioned on the other variables, typically using a modified prediction model. This approach does not represent a general factorization of the true joint distribution [198], and hence a stationary distribution may not exist. As a result, FCS may impute unrealistic values, which can become increasingly unrealistic as more variables are included. Complicating this issue is the underlying prediction model(s) needed by FCS which require separate validation for consistency across datasets. These concerns have been mostly ignored due to its satisfactory empirical performance [198, 130]. By including consistency checks on imputed FI distributions and by quantifying their predictive power we assessed the validity of several common MICE and other imputation models in our study.

Simulated missingness: We observed poor performance for both Default (PMM/logreg) and MICE RF, which both produced biased FI estimates for the simplest simulated missingness, cMCAR, even with $\leq 15\%$ missingness. PMM has previously been shown to produce biased estimates when imputing MCAR data [70], reportedly because of high missingness and too many variables, which were tested up to 64% and 82, respectively. We observed a significant bias even with 15% missingness and 68 variables. MICE RF has also been shown to struggle with large numbers of variables (≥ 200) [41]. Our results indicate that 68 variables may still be too many for either Default or MICE RF.

Increasing cMCAR to 25%, 50%, and ultimately 75% simulated missingness, we also observed a breakdown of both Ignore and kNN. kNN produced a large, significant bias in estimating the mean FI and a drop in the AUC. Ignore showed unbiased estimates of the mean FI but showed a drop in the AUC and HR, with the HR reaching 1.055 for 50% missingness — the same approximate missingness as the PFQ block in the Full dataset, versus the GT value of 1.075 (Appendix Figure A.8). This change is likely due to a noisier FI, as indicated by the significant increase in the SD.

Fewer values available to compute the FI should increase the SD by the Central Limit Theorem. Changes to the SD are important since they affect hypothesis testing, for example the t-test statistic is directly proportional to the inverse of the SD: if the SD is too large our p-values will also be too large (and vice versa). In contrast to Ignore and kNN, imputing with CART+Aux was robust even up to 75% missingness, showing no change in AUC or HR, a trivial change in estimated mean FI and the smallest change in the SD.

There was a significant bias in the Complete-case FI estimates using the Ignore method with NHANES missingness patterns (pMCAR). This bias was absent when the patterns were not used (cMCAR), implying the patterns were the cause. pMAR produced similar results. For 15% missingness, the pMCAR bias was small but visible in the FI distribution, Figure 4.6. The bias was approximately 0.012, but represents a change in HR of 1.09. This suggests that the real missingness data may also produce biased FI estimates and risk assessment when using the Ignore method.

To confirm and better understand why the bias was present in pMCAR data, we modelled it as a consequence of two observations: (1) variables had different frequencies of missingness and (2) variable blocks had different distributions of deficit values (see Appendix 4.8). For example, the PFQ block had the highest probability of missingness (Figure 4.2) and the lowest median deficit/FI value (Figure 4.4). Our calculation agreed perfectly with the observed bias. This confirmed that the pMCAR bias is due to a combination of differences between variables in their likelihood of being both deficit and missing; with a small additional bias due to mutual missingness patterns.

Real missingness: CART and CART+Aux imputed simulated missingness the best, and we have inferred that they also likely performed well with real missingness — and better than either Ignore or Default. The distributions of imputed FIs were very similar to simulated FIs (compare Figure 4.6 vs Figure 4.7) and showed a similar ordering of increasingly skewed FIs from CART+Aux to Default. Further, changes by variable block for younger individuals — representing 65% of our study population, matched changes expected based on survival, where an increased HR due to missingness — and therefore higher frailty [42], correlated perfectly with higher imputed FI values for CART and CART+Aux versus simply ignoring (see Tables 4.10 and

4.11). There was also a small, significant increase in the AUC of the FI for predicting 4-year-survival using CART or CART+Aux versus Ignore, implying these imputed FIs were better measures of frailty than the Ignore FI.

Notably, neither CART nor CART+Aux was able to fully compensate for missing expert knowledge regarding study design, as inferred from RI. In RI we assumed gated variables (PFQ, VIQ and RXD) were all optimally healthy, and in Table 4.7, saw a substantial increase in AUC: confirming RI. Validation of RI can be seen in the survival effects of PFQ and RXD for young people, which strongly imply the missing gated variables were healthy, Table 4.10. VIQ did not follow this trend, however, and therefore may have been better treated using a different imputation model such as CART. After RI was performed, we did observe that CART and CART+Aux appeared to correctly fine-tune the FI such that the residual bias, calculated using Eq. 4.7 and Eq. 4.9 was perfectly cancelled. Based on our results, there appears to be no downside to imputing using CART. The upsides include more accurate FI estimation and improved mortality prediction, especially when auxiliary variables are utilized. Imputing with CART is not a panacea: it did not obviate the need for RI, but it did improve upon it.

Investigating the underlying missingness mechanism, we observe that the real missingness is of mixed type. For example, for younger individuals PFQ was pMAR, since study design skipped those values when specific covariates were not deficits [29]. For older individuals PFQ was cMAR or cMNAR given the lack of patterning and strong relationship with survival (Figures 4.2 and 4.3, respectively). The FI distributions, Figure 4.7A, showed increasing skewness in the same order as the simulated pMCAR — from CART+Aux (least) to Default (most). But the Default distribution was less skewed than pMCAR, and there was a significant change in the distribution of CART vs CART+Aux. Given the similarities of pMCAR and pMAR in our simulations, the real missingness is a combination of patterened pMAR or pMNAR, and cellwise missingness cMAR or cMNAR.

Missingness and survival: What is the expected change in HR per 0.01 increase in FI [101]? This question cannot be answered precisely without good imputation practices since, as we have seen, both the FI and the HR depend on how missing data are handled [171]. High levels of missingness, even in the simplest case: cMCAR,

Table 4.10: FI of Real Missingness Imputation by Blocks, Under Age 60

Block	Ignore	Default	CART	CART+Aux	Survival Frailty[1]
All[2]	0.1218(8)	0.1261(10)	0.1179(8)	**0.1176(8)**	Low
PFQ	0.1151(8)	0.1416(46)	0.0876(32)	**0.0865(23)**	Low
RXD	0.1078(9)	0.0941(18)	0.0962(25)	**0.0909(14)**	Low
VIQ	0.1166(13)	0.0966(34)	0.0947(53)	0.0801(33)	No effect
BPX	0.1377(43)	**0.2454(222)**	0.2367(181)	0.2291(148)	High
LB	0.1232(33)	0.1458(43)	0.1449(46)	**0.1478(40)**	High

[3]

[1] Frailty inferred from Cox model and Kaplan-Meier curves.
[2] All individuals under age 60.
[3] Bold: noteworthy result.

Table 4.11: FI of Real Missingness Imputation by Blocks, Age 60+

Block	Ignore	Default	CART	CART+Aux	Survival Frailty[1]
All[2]	**0.1862(15)**	0.1851(15)	0.1850(15)	0.1852(15)	High
PFQ	**0.2343(39)**	0.2072(84)	0.2042(83)	0.2060(81)	High
RXD	0.1335(28)	**0.1123(53)**	0.1201(61)	0.1302(73)	Low
VIQ	0.2740(73)	0.3580(181)	0.3557(193)	**0.3732(214)**	High
BPX	0.2266(62)	**0.3290(240)**	0.3244(254)	0.3260(245)	High
LB	**0.2097(62)**	0.1605(69)	0.1610(70)	0.1615(67)	High

[3]

[1] Frailty inferred from Cox model and Kaplan-Meier curves.
[2] All individuals age 60+.
[3] Bold: noteworthy result.

can cause significant changes to the estimated HR. We also observed that patterned missingness can bias the FI on the scale of 0.01 in both our simulations and, ostensibly, in the real data. Our simulated patterns were handled well by CART, whereas the real patterns seemed to be better handled using RI then fine-tuning with CART; perhaps due to the increased heterogeneity of the Full population. In general, correct estimation of the HR and optimal reduction of FI bias require a robust imputation strategy such as CART.

We observed large differences in survival based on the missingness of variable blocks, Figure 4.3. For example, individuals under 60 with the personal fitness (PFQ) block missing lived significantly longer than those with the variable reported — with a maximum difference of 17.6 years between the survival curves. In contrast, individuals missing the lab (LB) block tended to die younger than those with the variables reported. We observed heterogeneity between the variable blocks, with some blocks showing longer, shorter or equal survivals when absent, and often showing different survival effects for old versus young individuals.

Very high levels of missingness occur naturally. For example the PFQ block was missing at a rate of over 50% in the Full dataset, and over 80% for individuals under age 60. In the simulated cMNAR, 50% missingness led to a bias in the FI of 0.1110 ± 0.0030 and an HR estimate of 1.048 using Ignore versus 1.074 using CART+Aux (Appendix Figure A.9). Even the relatively benign cMCAR missingness caused the HR to drop to 1.055 at 50% missingness when using Ignore. We observed a decrease in the HR (per 0.01 increase in FI) estimate using Ignore, dropping continuously from the GT value of 1.075 to 1.032 at 75% cMNAR. This suggests that with a high missingness the Ignore method can cause large biases in the HR. No such bias was observed using CART+Aux.

Although the FI was systematically biased when ignoring simulated missing data, there was no significant change in either AUC or HR for $\leq 15\%$ simulated missingness with either Ignore or CART. Increasing the missingness to a very high, 75% cMNAR, only reduced the AUC with Ignore by 1 error bar. Real missingness (at 14.5%) also showed a small effect on survival, although there was a significant increase in AUC when using CART versus Ignore, especially with inclusion of RI and auxiliary variables. The insensitivity of the AUC may be because it describes the predictive

power of all possible FI risk thresholds, and therefore is not sensitive to a systematic bias, furthermore, the scale of the bias may have been too small to significantly change the mortality-risk dichotomization: it was typically much less than the between-individuals variability as measured by the SD. Instead, bias in the FI affects estimates of relative risk.

A previous meta-analysis of adjusted FI-HRs across multiple studies yielded an estimated HR of 1.04 (CI: 1.03-1.04) [101], while our age and sex adjusted FI-HR was 1.08 (Table 4.7). This unexpectedly high HR has previously been attributed to the use of both lab and clinical variables in constructing the FI [19], and is consistent with earlier work [20, 81]. We can speculate about the role of missingness in constructing the FI. The opposing survival effects of missing LB versus PFQ may have helped balance the adverse effects of Ignore. In general, selection bias due to missingness could either enhance or deteriorate the FI. This may help explain the heuristic rules-of-thumb to limit missing data of variables to $< 5\%$ and of individuals to $< 20\%$. The latter could improve prediction of the Ignore FI by preferentially excluding young people, who tend to have bad imputations, Table 4.10. We observed in Table 4.8 that the Ignore FI usurped the predictive power of missingness, but this ability may depend on the variables used to construct the FI. The Ignore method pushed FI values higher for people with missing data, because values likely to be missing, e.g. PFQ, were almost always less than the individual-mean, Figure 4.4. If the individual missing data was older than 60, they were at higher risk of death, Figure 4.3, and therefore the Ignore, and especially Ignore20, methods would have incorporated this missingness-related-risk into the FI. This effect depends on the specific set of variables selected for the FI, and so may limit the utility of quantitative FI comparisons within and between studies.

Imputation strategies with the FI: We observed patterned missingness in the Full dataset with a wide range of missingness from 0% to over 50%. Variables often went missing together as nearly perfectly correlated blocks. We also observed unstructured missingness, particularly for older individuals.

Although the missing gated variables were best handled with RI, they also demonstrate clearly the utility of auxiliary variables. For example, the PFQ block was not reported for individuals under age 60 whom reported "no" to auxiliary variables

PFQ049, PFQ057 and PFQ059. In this case these auxiliary variables are able to convert MNAR (for which there are no general imputation models) to MAR (which many imputation models address). Even with MNAR data, auxiliary variables may still be able to improve imputation by correlating with the latent cause(s) of missingness. With simulated missingness CART+Aux gave excellent performance for low levels of cMNAR missingness. Nevertheless, improvement from auxiliary variables was smaller with real missingness. This may be because simulated missingness was not applied to the auxiliary variables, leading to much lower auxiliary variable missingness in the Complete dataset, Table 4.1. Our simulations should be considered a best-case scenario for auxiliary variable performance.

We expected that RF models would perform well since they are powerful imputation models capable of handling mixed data with non-linearities and interactions between variables [185]. In the present study we compared 1 tree (CART) versus 10 trees (MICE RF) versus 100 trees (`missForest`). We found that using only one tree (CART) consistently performed the best, implying more trees caused over-fitting. Generally speaking, it is expected that more trees should reduce over-fitting [119], though the opposite has been reported for imputation [177]. Similarly, too many predictors can also lead to a biased MICE RF imputation [41]. Often, RFs are built by picking a random subset of input predictor variables for each node i.e. "input selection", whereas CART does not [43]. Input selection could greatly reduce fit quality if there are too many poor predictor variables i.e. spurious covariates [59], since they dilute the pool of available features. This leads to poorly predicting trees and subsequently a poorly predicting forest. Input selection would then reduce accuracy, which could explain the superior performance of CART. A less likely potential source of over-fitting is tree depth [59].

Why do tree-based imputation methods perform better? Imputation strategies typically impute values by randomly drawing or combining 'nearby' observed values. For Ignore, nearby means other values of the same individual, while for other methods nearby is determined by a minimum distance. For PMM the distance is based on linear regression [1], whereas the distance in tree-based methods (CART and RF) is determined by iteratively partitioning the data. As a result, tree-based methods can automatically account for non-linearities, such as interactions between variables.

Previous studies have demonstrated that tree-based methods perform well when interactions are present [43, 177, 79]. Non-linearities are expected in our data due to known interactions, such as the sex-frailty paradox [83], as well as the arbitrary scales used for questionnaire data and the presence of non-normally distributed lab data. These may explain the relatively poor behaviour of the default MICE method versus tree-based methods. Default struggled notably with patterned simulated missingness (pMCAR and pMAR), perhaps because finding a suitable donor (PMM) or set of predictors (logreg) was especially difficult due to the large blocks of mutually missing covariates.

CART was systematically biased high with cMNAR, along with most other strategies; although, CART was less biased than Ignore. kNN and `missForest` were both unbiased for 15% cMNAR, although `missForest` was consistently biased low and hence probably coincidentally biased in the correct direction for cMNAR. kNN performed relatively well with cMNAR but still struggled with $\geq 25\%$ missingness. Our inability to successfully impute for cMNAR reflects the difficulty of the underlying problem, which in general requires knowledge of the biasing mechanism [130]. This may present an opportunity for imputation models designed specifically for aging data (e.g. [52]).

General thoughts. In our study, the 20% exclusion rule preferentially excluded young individuals (under age 60), removing 56% of young individuals versus 6% of older individuals. This radically altered our study population. Since young people had the least realistic blockwise imputation values using the Ignore method (Table 4.10) and Ignore generally imputed higher than the true missing values (Figure 4.4), this suggests that the 20% rule might improve prediction by simply removing individuals for whom Ignore doesn't work well. In our study the 20% exclusion rule also excluded all individuals missing the lab block, which preferentially removed individuals with poor survival prognosis from the analysis (Figure 4.3). The effect of 20% exclusion depends on the specific set of variables used to calculate the FI. If we had used 10 lab variables instead of 27, then the 20% cut would be ≥ 10.2, and only the PFQ block would be excluded, radically changing the survival effect of excluded individuals (Figure 4.3). In the present study, survival prediction dropped significantly when the Ignore20 rule was used versus Ignore, Table 4.7. Given the superior performance of

CART imputation, we see no reason to rely on heuristic rules such as the 20% rule — which biases the study population and could lead to unexpected effects.

Our primary source of error was differences between the Complete-case and Full datasets. We consistently observed that survival, frailty and missingness are interacting variables, and hence the Complete-case data had unavoidable differences in the overall FI, AUC, and mortality rates. Nevertheless, the qualitative results were similar between the simulated and real missingness. We consistently saw that the FI calculated using Default MICE or by ignoring missingness gave higher values than CART and CART+Aux. The latter two matched the GT distribution in the simulated missingness data, were consistent with our bias calculations for real and simulated missingness, and improved predictive power in the real missingness data.

We averaged together multiple imputations when estimating predictive power to estimate the maximum achievable predictive power versus single imputation strategies, but this neglected propagation of error due to imputation hence our confidence intervals were likely too small for the AUC and HR of the real missingness. The simulated missingness used Monte Carlo estimates for the error and therefore should be reliable. Recent results have implied that m = 5 imputations may be far too few for accurate estimation of statistical dispersion [22], however, when we used the recommended m = 15 imputations [205] on the simulated data we saw only a small change in the estimated standard deviation, implying for our low levels of missingness m = 5 was sufficient.

In the future we would like to investigate missingness structures in other common aging studies. It will also be interesting to investigate the 5% missingness-by-variable cut-off that is commonly used in the literature [162]. Further investigation into MICE may prove worthwhile, such as the convergence properties (stability) of FCS and the effect of number of iterations. Tuning of MICE hyperparameters, notably of RF including depth, input selection and number of trees could enhance results, but would require a diverse set of gerontological studies to do reliably.

There is room for improvement from CART+Aux, which had poor performance for high levels of cMNAR, and struggled with the imputation of real missingness both for older individuals and for gated variables better handled using RI. This performance might be improved upon with deep learning models (e.g. [65, 52]), although

scepticism is warranted regarding generalizability across datasets, as lightweight imputation models — including MICE via CART — have been shown to out-perform deep learning in third-party comparison studies [85, 203]. Quantitative, stochastic modelling of aging naturally lends itself both to the development of new imputation strategies and to the ability to generate realistic datasets to validate imputation strategies. This synergy presents an opportunity for quantitative researchers to address a serious pragmatic issue endemic to aging studies: missing data.

4.7 Summary and Conclusions

We considered several types of simulated missingness together with naturally missing data. Imputation of real missingness shared strong similarities with imputing the simulated missingness. Our results indicate that most imputation strategies, including Ignore and the MICE default, are weak against at least one type of missingness. Fortunately, MICE using CART appeared to be robust, and consistently improved estimation and predictive power over simply ignoring missing data.

We observed distinct missingness patterns that bias the standard Ignore (available-case) FI methodology, even when missing completely at random (pMCAR). Imputation with MICE using CART can remove this bias. We advise caution with other MICE models, especially with the default method (PMM/logreg) which made the bias even larger for our simulated missingness. The MICE RF model performed poorly and was unreliable — with performance dependent on the missingness mechanism, as were the popular single-imputation strategies of kNN and `missForest`. kNN did perform well for $\leq 15\%$ missingness, but failed even the simplest test case — cMCAR, for $\geq 25\%$ missingness, and had poor predictive power with the real missingness.

These same patterns of missing variable blocks have a significant effect in survival, with the missingness of some variables being predictive of poor survival, whereas others indicated better survival. These effects are evidence that missingness should not be ignored. The FI tended to cancel out survival effects when using the typical strategy of ignoring missing values, which may suggest an important cancellation in the choice of FI variables. For example, the self-reported and lab variables in this study tended to have opposing survival effects with missingness. What's more, we observed that the heuristic 20% cut-off rule for individuals missing entries can

partially compensate for the limitations of ignoring missingness in certain types of simulated missingness, but can also greatly bias the study population.

The FI prediction of mortality appeared to be robust to missingness, showing only a minor reduction in AUC even when 75% of the data were made missing, however, we observed large changes in the HR estimate for missingness $\geq 25\%$ when missing values were simply ignored. Good HR estimation requires imputation. With inclusion of auxiliary variables, the CART+Aux imputation showed remarkable consistency in both AUC and HR estimation in the simulated missingness, even at 75% missingness. We also observed CART+Aux improved survival prediction (AUC) for the real missingness over ignoring the missing data.

Our observed improvement in survival prediction appears to be consistent with previous work using the Rotterdam study [171], although that study did not provide a direct measure of predictive power such as the AUC or C-index. That study also did not fully report their imputation model — only that they used MICE — but they found a similar bias in the median FI of same scale, 0.01, and in the same direction as the Default MICE imputation in our study. In our study this was the same scale and opposite direction of CART and RI-based imputations, emphasizing the potential for differences between cohorts and the need for full disclosure of imputation models.

Our study indicates a hierarchy of increasingly complex missing data handling for increasingly precise estimation of the FI and subsequent HR. The simplest approach is to use the typical Ignore strategy. The Ignore-FI appears to be a simple, composite health measure of vulnerability to adverse outcomes, suitable for clinical situations. Unfortunately, ignoring missing data makes the FI prone to bias and hence inhibits quantitative FI comparisons across populations and studies. A large improvement in FI precision and predictive power follows if we apply RI. A smaller improvement to FI precision follows if CART is then used to impute remaining values. And finally, inclusion of auxiliary variables with CART can safeguard against low-levels of MNAR without serious risk of over-fitting. In situations where fewer rules are available for RI, imputing with CART using auxiliary variables becomes increasingly important.

Missing data handling can have a significant effect on the precision of the quantitative FI, HR estimate, and its mortality predictive power. A standardized approach for handling missingness is needed to achieve the increasingly high levels of precision

desired in contemporary FI studies, and to facilitate comparisons between studies and translation across populations. Researchers should fully disclose their missing data handling methodology, including imputation model and number of imputations. Basic sanity checks on imputed values are advisable. It is still an open question what effect missingness has across studies and across sets of variables used for the FI. In the present NHANES-based study, imputation using the commonly available CART MICE consistently gave superior FI precision, HR estimation and mortality predictive power over simply ignoring missing values.

Acknowledgments

The authors wish to thank Dr. Joanna Blodgett (University College London) and Dr. Judith Godin (Dalhousie) for their helpful input.

4.8 Appendix: How Missingness Patterns Bias the FI

The complete data matrix B has true elements b_{ij}, where the rows $i \in \{1, 2, ..., N\}$ are over N individuals and the columns $j \in \{1, 2, ..., N_b\}$ are over N_b variables. The missingness matrix $M_{ij} = 1$ if a given entry is missing, and 0 if it was observed. The overall missingness fraction is $\pi = N_m/(NN_b)$ where $N_m = \sum_{ij} M_{ij}$ is the number of missing values in the dataset.

We define $\bar{f}$ as the true average FI (Frailty Index) over the population, so $\bar{f} = \sum_{ij} b_{ij}/(NN_b)$. We define $\bar{f}_{obs}$ as the average observed FI, so $\bar{f}_{obs} = \sum_{ij}(1-M_{ij})b_{ij}/(NN_b - N_m)$. We define $\bar{f}_{miss}$ as the average FI of the missing values, so $\bar{f}_{miss} = \sum_{ij} M_{ij}b_{ij}/N_m$. We then have

$$\bar{f} = (1 - \pi)\bar{f}_{obs} + \pi\bar{f}_{miss}. \tag{4.5}$$

The true population average, $\bar{f}$, only coincides with the observed estimate, $\bar{f}_{obs}$, when $\bar{f}_{miss} = \bar{f}_{obs}$, otherwise there will be a bias (for $\pi > 0$).

To estimate the bias we assume that the distribution of missing data across individuals is P_i, across variables P_j, and across both $P_{i,j}$. We would have $P_{i,j} = \langle M_{ij} \rangle$, where the angle brackets indicates an average over many missingness matrices. If we wanted the distribution of non-missing data P^c_i, P^c_j, and $P^c_{i,j}$ it would be just be $P^c = 1 - P$. Note that $\sum_{ij} P_{i,j} = 1$.

The bias, $\bar{f} - \bar{f}_{obs}$ can be calculated using Bayes' theorem as:

$$\bar{f} - \bar{f}_{obs} = \pi \left(\sum_{ij} b_{ij} P_i P_{j|i} - \sum_{ij} b_{ij} P_i^c P_{j|i}^c \right). \tag{4.6}$$

We assume no individual-specific selection, i.e. $P_i = 1/N$. We can approximate the bias by assuming independence, $P_{j|i} \approx P_j$, then we have:

$$\bar{f} - \bar{f}_{obs} \approx \pi \sum_{j=1}^{N_b} \left(\frac{1}{N} \sum_i b_{ij} \right) (\pi_j - \pi_j^c) \tag{4.7}$$

which is plotted as "Model (approx.)" in Figure 4.5. Note that $1/N \sum_i b_{ij}$ requires knowledge of the grouth truth, unless the data are MCAR. Where $\hat{\pi}_j = P_j = \sum_i M_{ij} / \sum_{ij} M_{ij}$ and $\hat{\pi}_j^c = P_j^c$ is:

$$\hat{\pi}_j^c = \frac{\sum_i (1 - M_{ij})}{\sum_{ij} (1 - M_{ij})} = \frac{1 - \pi N_b \hat{\pi}_j}{N_b - \pi N_b} \tag{4.8}$$

We have M_{ij} directly from the data matrix. Note that if $P_{i,j} = const.$ (cMCAR) then $\bar{f}_{obs} = \bar{f}_{miss} = \bar{f}$, in which case the Ignore method would be unbiased. The independence approximation $P_{j|i} \approx P_j$ is, in light of the strong missingness patterns (Section 4.5.1), unlikely to be exact. We can instead estimate $P_{j|i}$ by assuming independent, identically distributed individuals (pMCAR), as:

$$P_{j|i} = \frac{1}{N} \sum_i \frac{(1 - M_{ij})}{\sum_j (1 - M_{ij})} \tag{4.9}$$

which, after substitution into Eq. 4.6, is plotted as "Model (exact)" in Figure 4.5. The key difference is that $\sum_j (1 - M_{ij})$ varies greatly for patterned missingness. The approximate model posits that the difference between FI contributions between variables and blocks causes a bias, whereas the exact model additionally posits that the specific patterns also contribute to the bias.

Chapter 5

Efficient representations of binarized health deficit data: the frailty index and beyond

By Glen Pridham[1], Kenneth Rockwood[2], and Andrew Rutenberg[1].
[1]Department of Physics and Atmospheric Science, Dalhousie University, Halifax, B3H 4R2, Nova Scotia, Canada.
[2]Division of Geriatric Medicine, Dalhousie University, Halifax, B3H 2E1, Nova Scotia, Canada.

Pridham, G., Rockwood, K. & Rutenberg, A. Efficient representations of binarized health deficit data: the frailty index and beyond. GeroScience (2023) doi:10.1007/s11357-022-00723-z [146]

We investigated efficient representations of binarized health deficit data using the 2001-2002 National Health and Nutrition Examination Survey (NHANES). We compared the abilities of features to compress health deficit data and to predict adverse outcomes. We used principal component analysis (PCA) and several other dimensionality reduction techniques, together with several varieties of the frailty index (FI). We observed that the FI approximates the first — primary — component obtained by PCA and other compression techniques. Most adverse outcomes were well predicted using only the FI. While the FI is therefore a useful technique for compressing binary deficits into a single variable, additional dimensions were needed for high-fidelity compression of health deficit data. Moreover, some outcomes — including inflammation and metabolic dysfunction — showed high-dimensional behavior. We generally found that clinical data were easier to compress than lab data. Our results help to explain the success of the FI as a simple dimensionality reduction technique for binary health data. We demonstrate how PCA extends the FI, providing additional health information, and allows us to explore system dimensionality and complexity.

PCA is a promising tool for determining and exploring collective health features from collections of binarized biomarkers.

Keywords: frailty index, principal component analysis, logistic principal component analysis, dimensionality reduction, biological age, aging.

Table 5.1: Nomenclature

FI	Frailty index
FP	Frailty phenotype
(I)ADL	(Instrumental) activities of daily living
PC(A)	Principal component (analysis)
LPC(A)	Logistic PC (analysis)[1]
LSV(D)	Logistic singular value (decomposition)[2]
GLM	Generalized linear model

[1] Cousin of PCA.
[2] Cousin of LPCA.

5.1 Introduction

Biological dysfunction arising from damage is central to aging [114]. Representing dysfunction requires robust summary measures of aging data, which can then help us to operationalize theories of causal mechanisms [114, 94, 168]. Is there a systematic way to generate summary measures from observed health deficits? How well do they predict a battery of adverse outcomes?

The frailty index (FI) is a simple, robust measure, that is strongly predictive of general adverse outcomes [101, 42]. Dichotomizing data as healthy (0) or deficit (1) probes dysfunction directly. The FI is defined as the average number of dysfunctional (deficit) health variables an individual has [174]. Conventionally, the FI is constructed from self-reported questionnaire — 'clinical' — data, such as (instrumental) activities of daily living (I)ADLs and physical limitations. Recently, the FI has been extended to include 'lab' biomarker data [81, 18].

Aging is widely considered to be multidimensional [97, 114, 56, 31, 168]. The FI is just one of many univariate summary health measures. In particular, many

"biological ages" have been proposed[1]. These measures overlap only moderately, implying that a complete description of "biological age" would require several of them [87, 110, 37]. Machine learning studies also suggest multiple dimensions of health information, though survival information appears to compress into just one or two dimensions [52]. Furthermore, interventional study reviews often report improvement along one dimension at the expense of worsening along other dimensions: for example, mice treated with metformin show improved treadmill performance but reduced visual acuity [134]. For which outcomes is a univariate health measure sufficient? Do integrative hallmarks of aging [114] or bow tie systems[2] [38], which mediate interactions between multiple systems, require multidimensional health measures?

The rapidly increasing dimensionality of 'omics aging data [219] makes these questions pressing. For example, Jansen *et al.* [87] studied over 20,000 gene expressions from fewer than 3,000 individuals. Data with more variables than individuals carry the "curse of dimensionality" which can lead to overfitting and loss of interpretability with standard algorithms [86]. Condensing high-dimensional data into a few salient features simplifies statistical modelling [86, 2]. To achieve this, we need scalable and robust dimensionality reduction techniques.

While the FI is a simple and reproducible dimensionality reduction technique [158] that compresses 30+ binary health variables [174] into a single, graded measure [82], it has not been systematically extended to higher-dimensional health features. *Ad hoc* multivariate extensions such as domain-specific FIs [20, 17] or multiple biological ages [110] neglect the possibilities that these measures may have gaps or redundancies in the information they contain.

The canonical dimensionality reduction technique in machine learning and statistics is principal component analysis (PCA), which is robust, fast, and systematically extensible [86]. PCA linearly combines (rotates) existing health variables into a complete set of new 'latent' health variables — principal components (PCs) — ordered from most to least variance. By construction, the PCs are mutually independent and hence do not suffer the problem of redundant information faced by multiple *ad hoc*

[1]For example, Horvath's biological age predicts chronological age based on individual epigenetic methylation status across a battery of sites. Individuals that are predicted to be older than their true age are inferred to be in worse health. Chapter 7 describes several common biological ages.

[2]Bow tie systems are characterized by many inputs and many outputs. Information fans in and out of the central node, giving it the graphical appearance of a bow tie (the attire), hence the name.

approaches.

PCA has been used to improve epigenetic clock reliability [78], and to analyse raw biomarker data [33, 9] and dysfunction biomarkers [131]. PCA is robust to covariates, including sex, race and study population [33]. When used correctly, PCA summarizes the salient information in a dataset. For example, Entwistle *et al.* [49] applied PCA to NHANES III dietary data and identified the first 4 PCs as being idealized dietary patterns. Nevertheless, studies using PCA to generate new health measures (PCs) from deficit data are rare. Few, if any, have leveraged the extensive literature on health deficit data that surrounds the FI. Furthermore, none of the aforementioned studies have systematically explored multiple dimensionality reduction algorithms nor the effect of modifying the number of PCs on adverse outcome prediction. What are the generic features of dimensionality reduction of health deficit data? How can this help us to understand and build upon the success of the FI?

We should also explore what is the correct number of PCs to use for health deficit data. Arbitrarily restricting which PCs to use has led to serious criticisms of its reproducibility for low-rank projections (e.g. only using the first two PCs) [48]. Others have noted that mortality information can be found in low-variance PCs, which are often neglected [78].

While the FI and PCA are both linear transformations, the FI imposes equal weightings of each variable whereas PCA does not. Accordingly, the FI and PCA need not be related. Nevertheless, previous work has shown that the first PC of biomarker dysfunction data from hemodialysis patients approximately reproduces the two key phenomenon of the FI: approximately equal weightings across input variables, and good prediction of adverse outcomes, including the frailty phenotype (FP) [131]. Furthermore, the FI has been shown to efficiently compress clinical, deficit questionnaire data, with little unexplained residual variance [207]. However, research on biomarker lab deficit data implies the presence of additional dimensions [66]. How many dimensions are relevant in lab biomarker data, and do they overlap with clinical dimensions? Any such information that is shared between lab and clinical domains will affect joint dimensionality reduction.

As with the FI, our primary interest is in damage arising from dysfunction, so we binarize data as either normal (0) or dysfunctional/deficit (1). We expect that

compression of health deficit data will find efficient representations of both dysfunction *and* adverse outcomes because health deficits are themselves adverse outcomes, e.g. ADLs [104]. This improves interpretability — dysfunction is what we care about — and saves us from issues endemic to continuous variables, such as scaling, healthy variability, and non-normal behaviour. Recent advances in PCA specific to binary data provide additional techniques that we also explore: "logistic" PCA [105] and "logistic SVD" [105, 167] (SVD: singular value decomposition).

All of these PCA algorithms are lightweight with minimal assumptions. They compress data into efficient rank-ordered representations, where the first dimension contains the most information and the last contains the least. In contrast, latent variable models such as grade of membership [116, 178, 50], while more directly interpretable, have sub-optimal compression efficiency and do not rank-order their latent space. Efficient, rank-ordered representations will effectively coarse grain the data, allowing us to answer our questions about dimensionality and information flow. Here, we restrict our attention to PCA and its variants.

The goal of this study is to systematically explore the use of PCA in compression and prediction of multidimensional health deficit data, and to compare PCA with the FI. We also examine PCA alternatives. Compression can tell us the maximum number of dimensions required to efficiently represent input data, but can't *a priori* distinguish between useful information and noise. We compare compression of binary deficit data and prediction of adverse outcomes using both outcome associations and a generalized linear model (GLM). We include a battery of adverse outcomes to test predictive power. Finally, we take a deeper look at PCA, fully exploring its utility, its robustness, the patterns it extracts from the data i.e. PCs, and its systematic mode of action. We demonstrate that PCA provides a multidimensional perspective of health not available to univariate health measures.

5.2 Methods

Figure 5.1 outlines the study pipeline. We split the data into three parallel analyses: compression, associations with input/outcome variables, and prediction using generalized linear models (GLMs). We compared compression using the frailty index (FI), principal component analysis (PCA), logistic PCA (LPCA) [105] and logistic singular

value decomposition (LSVD) [105, 167].

5.2.1 Data and preprocessing

We used data from the 2001-2002 NHANES with linked public mortality records [29]. We included individuals over age 60 ($N = 1872$) to focus on older individuals and to avoid problems with gated variables [145]. We used lab and clinical health deficit data from multiple domains to predict multiscale, multidomain outcomes. In total, we included 26 clinical predictors, 29 lab predictors, 47 outcomes, and 7 demographical variables.

The complete list of predictors, outcomes and covariates is provided in Appendix B. We binarized all predictors using standard rules [19] (Table B.1 and B.2). This simplified analysis and forced our dimensionality reduction algorithms to find efficient representations of dysfunction — as desired. Outcomes included health biomarkers, disability, morbidity and mortality. All continuous outcomes were standardized to zero mean and unit variance. We included 7 demographic covariates: age (top-coded at 85), and (binarized) sex, race, family income, education level, smoker status and partner status.

The FI was computed as the average of binarized predictor variables [174]. The FP was included as an alternative frailty measure, defined as 3+ out of 5: low BMI, bottom 20% for gait speed (sex-adjusted), and self-reported: weakness, exhaustion, and low physical activity [208].

Imputation was performed using multivariate imputation by chained equations (MICE) version 3.10.0 [196]. We used the classification and regression tree (CART) method, which performs well with similar NHANES data [145]. We imputed *all* data, including predictors, outcomes, covariates, survival information and auxiliary variables. Imputing outcomes had no significant effect on prediction accuracies: except for gait, which had a higher R^2 by ~ 0.05 (Appendix B). We imputed 15 times, reflecting the $\sim 15\%$ missingness[3] [205]. We propagated the uncertainty in these imputations into our final results using Rubin's rules [205]. We symmetrized and scaled standard errors (assuming normality), applied Rubin's rules, then rescaled to 95%

[3]The rule originates from the observation that the ratio of variance using infinitely-many imputations versus m is $1 + f/m$ where f is the fraction of missingness. According to our rule we use $m = 100f$ giving a 1% ratio between finite versus infinite imputations, small enough to be ignored.

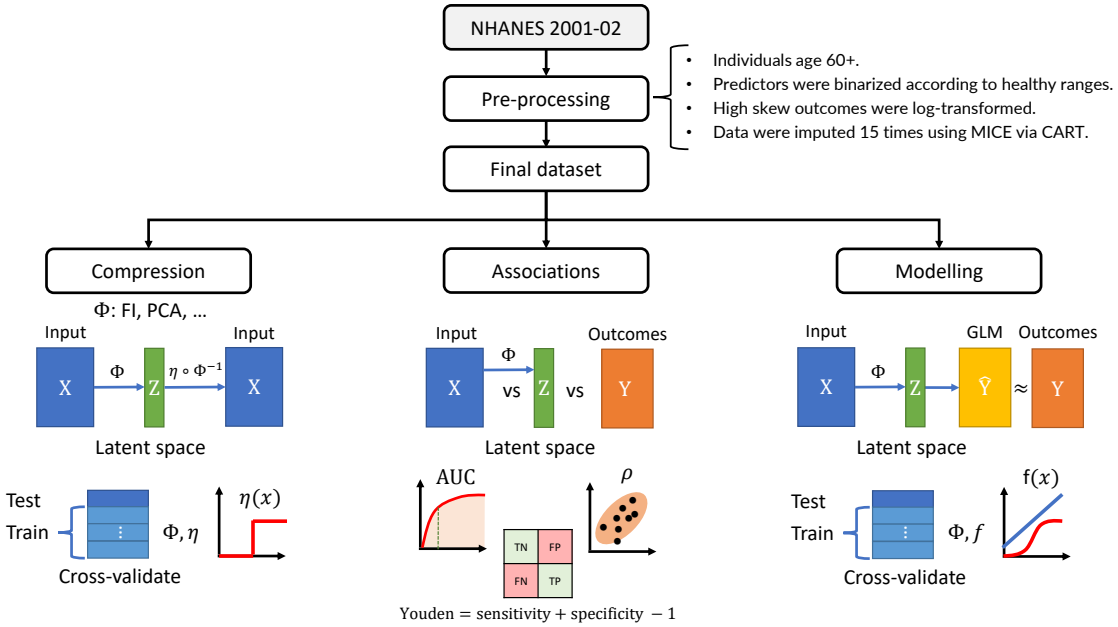

Figure 5.1: Study pipeline. We performed three parallel analyses: compression, feature associations, and outcome modelling. Data were preprocessed, resulting in an input matrix of health deficit data, X, and an outcome matrix of adverse outcomes, Y (rows: individuals, columns: variables). The input was transformed by a dimensionality reduction algorithm, represented by Φ, which was either: the FI (frailty index), PCA (principal component analysis), LPCA (logistic PCA) or LSVD (logistic singular value decomposition). Each algorithm, Φ, generated a matrix of latent features with tunable dimension, Z (dimension: number of columns/features; the FI was not tunable). We tuned the size of this latent feature space, Z, to infer compression efficiency and the maximum dimensions of Z before features became redundant (binarizing with optimal threshold, η). The latent features were then associated with input and outcomes to infer their information content and the flow of information from input to output. The dimension of Z was then again tuned to predict the adverse outcomes. $\hat{Y}$ represents the outcome estimates by the generalized linear model (GLM), which were compared to ground truth, Y, to determine the minimum dimension of Z needed to achieve optimal prediction performance for each outcome. This procedure allowed us to characterize the flow of information through each dimensionality reduction algorithm.

confidence intervals (CIs).

In Appendix B we provide consistency checks on the imputed values and characterized the missing data. Individuals with missing data were older, with median (IQR) age 71 (65-78) vs 76 (67-83) (Wilcox $p = 2 \cdot 10^{-16}$), and had significantly worse survival, hazard ratio[4]: 1.6(1) (p: $7 \cdot 10^{-13}$, log-rank test). This means that the missing data were not missing completely at random and that failure to impute could lead to biased results [180]. We performed our initial analysis using complete case predictor and demographic data (no missingness for each individual), and available case outcome data (individuals were included for any outcome they had reported). Complete case analysis yielded similar results to our full, imputed analysis (Appendix B).

5.2.2 Performance Metrics

Most of the binary outcomes and predictors were rare, with many occurring in less than 10% of study participants (Table B.4). Such unbalanced data poses a problem when measuring binary performance [96]. An uninformative diagnostic test that returns negative regardless of disease status would have 90% accuracy in diagnosing a disease with 10% prevalence. Its Youden index [215], however, would be 0:

$$\text{Youden index} \equiv \text{sensitivity} + \text{specificity} - 1. \tag{5.1}$$

A perfectly informative test would have a Youden index of 1. The Youden index is strongly correlated with the AUC, which estimates the probability that a metric will correctly rank the positive individuals as higher than negative individuals [145]. Assuming case and control are both normality distributed with the same variance, the AUC and Youden index are redundant, for example Youden indexes of 0.2, 0.4, 0.6 and 0.8 correspond to AUCs of 0.64, 0.77, 0.88 and 0.97, respectively [214]; this model fit our data very well (Figure B.15).

When comparing continuous-continuous variable pairs we used Spearman's ρ, a non-parametric measure of correlation [152]; we took the absolute value and estimated the confidence interval using quantiles from bootstrapping (with 2000 resamples). For models predicting continuous outcomes we used R^2, the coefficient of determination, which measures the explained variance as a proportion of the total variance with 1

[4]Note that in this chapter errors are reported in parentheses e.g. $1.234(56) \equiv 1.234 \pm 0.056$.

being perfect. The mean-squared error (MSE) is the average of the squared model residual [86]. We standardized continuous outcomes to zero mean and unit variance, hence useful models have MSE < 1 (assuming unit variance, $R^2 = 1 - \text{MSE}$). Time-to-event outcomes, i.e. survival, were scored using the concordance-index (C-index), a cousin of the AUC [145]. Generalized linear model (GLM) predictive power used R^2, MSE, AUC, Youden and/or the C-index. Outcome associations used Spearman's ρ (specifically $|\rho|$), AUC, Youden or the C-index (specifically $|C - 0.5|$). Feature importance was first inferred from stepwise regression, then validated using selection frequency (Appendix B).

5.2.3 Input Compression

We applied the FI, PCA, LPCA and LSVD to the predictor variables (binarized lab and clinical data) — Figure 5.2 illustrates how PCA compressed the data by decomposing the 2D joint deficit histogram. We treated the binary scale as an absolute scale for dysfunction, akin to the FI, so we did not center variables by their respective means. Lab and clinical data were compressed together and separately. Compression performance was measured by reconstruction accuracy. Data were compressed into a latent space using one of the four algorithms, then mapped back to the inputs using the inverse transform [105] (excluding the FI, which is not invertible). An ROC curve was then trained to map from the reconstruction (PCA, LPCA and LSVD) or latent space (FI) to the inputs, providing an optimal cutting point to reconstruct the original inputs; this step calibrates the reconstruction. Test inputs were then compared to their reconstructions using the Youden index. Note that the Youden index is a (relatively) neutral measure that does not favour PCA, which minimizes the MSE, nor LPCA/LSVD, which minimize the Bernoulli deviance. We progressively increased the size of the latent space to be able to infer the minimum number of dimensions required for high-fidelity reconstruction. This yielded compression plots of increasing fidelity with increasing latent space dimension.

5.2.4 Generalized Linear Models (GLMs)

The primary motivation for using a regression model is to capture conditional effects, including demographical variables and the combined performance of multiple features.

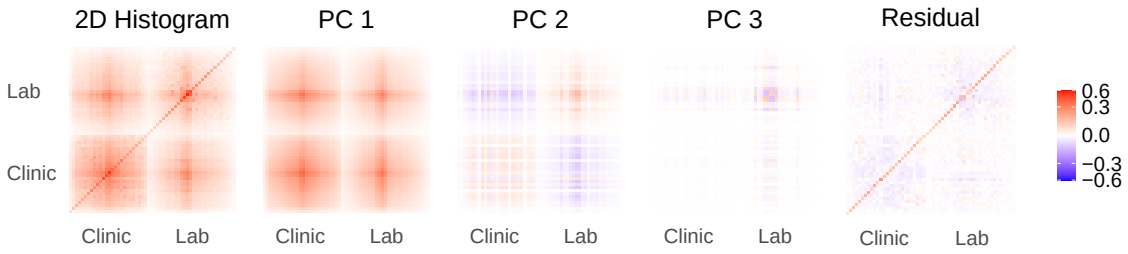

Figure 5.2: Principal component analysis (PCA) of binary data is equivalent to eigen-decomposing the 2D joint deficit histogram. The first column is the complete histogram, the remaining columns sum to the first column (Eq. 5.7). The first PC is clearly dominant and is dense, meaning it is nearly equal weights for each variable (akin to the FI). The eigen-decomposition naturally finds blocks of correlated variables. When it runs out of blocks it looks for strong diagonal terms. This causes PCA to naturally block out like-variables, e.g. lab vs clinical in PC2, similar to an expert choosing to create an FI out of variables from the same domain. Values have been transformed for visualization using $\text{sign}(x)|x|^{\gamma}$, $\gamma = 2/3$, see Figure B.16 for the figure without scaling.

We used generalized linear models (GLMs) [152]. GLMs include linear, logistic and Cox proportional hazard regression [86], allowing us to model each outcome variable with a homologous linear model.

We performed stepwise regression to analyse the effect of iteratively adding variables on the predictive performance, starting with the model that used only demographical information. Our motivation was to determine the optimal number of latent features to include in our models, which are naturally ordered by the dimensionality reduction algorithm, PC1 through to PC55. Stepwise models produced incremental prediction plots for comparison to the compression plots.

We inferred feature importance by building complete models that potentially included all predictors. Feature selection was performed using an L1-penalized GLM (LASSO), with the penalty selected using 10-fold cross-validation to pick the minimum mean-squared error (continuous outcomes) or deviance (binary outcomes) [58]. An L1-penalty penalizes regression coefficients that differ from 0, encouraging the model to retain only the most important features. Selection frequency was used as a measure of feature importance.

GLMs used to predict binary outcomes are known to underestimate the frequency of rare events, even for datasets with 1000s of individuals (such as ours) [96]. In

Appendix B, we studied the use of observation weights to improve the Youden index. We found that the optimal weight of the ith individual was,

$$w_i = \begin{cases} \frac{\text{Frequency of majority class}}{\text{Frequency of minority class}}, & \text{if } i \text{ is in minority.} \\ 1, & \text{if } i \text{ is in the majority.} \end{cases}$$

This choice of weights is equivalent to the "weighted exogenous sampling" method [96], where we have weighted as if the population underlying the sample is perfectly balanced.

All computations were performed using R version 4.0.1 [152]. Error bars are standard errors unless specified otherwise. Errors are reported in parenthesis e.g. $12(3) \equiv 12 \pm 3$. Confidence intervals are 95%. We report out-of-sample performance metrics using 10-fold cross-validation for all parametric models, including compression and prediction. Out-of-sample means that the compression or prediction algorithm is completely ignorant of the testing data[5]. This procedure estimates the expected performance on new, unseen data, from the same population, independent of the training set [11].

5.3 Results

5.3.1 Input Compression

We decomposed and then reconstructed input variables using various dimensionality reduction techniques. In Figure 5.3 we show the out-of-sample Youden index for the FI, PCA, LPCA and LSVD, as indicated. For all, the first dimension dominates and represents $\sim$ 30% of the gain in predictive power over guessing (a guess has Youden index = 0).

LSVD was the most efficient compression technique, having perfect reconstruction after approximately 30 latent dimensions. However, this performance comes at a large cost in terms of number of parameters [105], and with respect to computational resources. Our benchmarks in Appendix B.3.6 indicate that PCA is about 10x faster than LPCA which is itself 10x faster than LSVD.

[5]When data are resampled, such as here, the dataset is split into in and out-of-sample sub-datasets. The in-sample is used to fit the model. The out-of-sample is used to test the model. This means that the model never sees the out-of-sample data prior to testing.

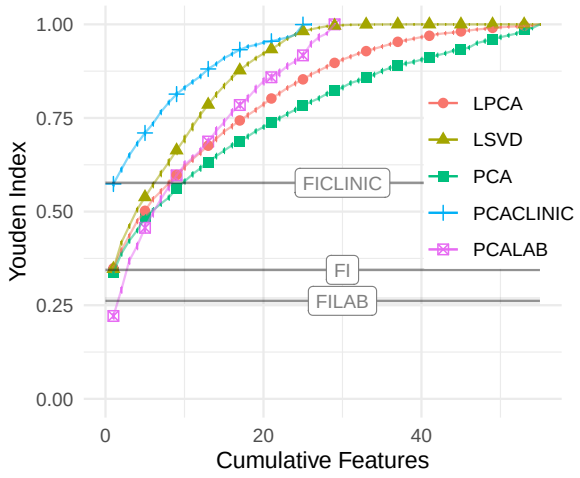

Figure 5.3: Cumulative compression. Tuning the size of the latent dimension bottleneck we inferred the maximum number of dimensions required to efficiently represent the input data. The reader should look for two things: (1) the number of components (dimensions) needed to achieve a relatively high score, and (2) the slope of the curve — when it flattens we can expect the features are noise, variable-specific or otherwise less important. Logistic SVD compresses the input most efficiently, saturating at around 30 features. Note the dramatic difference between lab and clinical compression both for PCA and the FI; the first PC of clinical data scores as well as 9 lab PCs.

The information in the input variables includes both important, latent, information reflecting an individual's health-state and variable-specific information which could be considered noise (i.e. not useful for predicting relevant outcomes). Generally, we see that the first dimension performs similarly for all methods. Additional dimensions are needed for accurate compression. The number of dimensions needed ranges from 30 (LSVD) to all 55 (PCA). This implies the dataset can be fully represented by a manifold of new features with dimensionality at most 30. We also see that clinical data compresses more efficiently than lab data, implying significant correlations between clinical variables. All four dimensionality reduction techniques estimated a very similar first dimension, as indicated by their strong mutual correlations, shown in Figure 5.4. The correlation between the FI and PC1/LPC1/LSV1 is almost perfect, $\rho > 0.95$, with nearly identical age and sex dependencies (Figure B.19). Centering had a negligible affect on results, only reducing the correlation to $\rho > 0.9$. This implies that a very strong signal is present in the data and that it is very close to the FI, particularly the FI CLINIC.

In Appendix 5.6.2 we show how the equivalence between the FI and PC1 can arise from the structure of the joint histogram and provide conditions under which the FI/PC1 is the dominant dimension.

5.3.2 Feature Associations

While compression efficiency identifies the number of dimensions needed to recover the input data, it does not tell us how the information is split across features, nor how that information relates to adverse outcomes.

We explore the flow of information by investigating the associations between each observed variable and each latent feature, e.g. the FI, each PC, etc. We used the metrics describe in Section 5.2.2, which range from 0 (no association) to 1 (perfect association). A score of 1 means that if we know the value of the feature then we can perfectly predict the value of the associated variable. We automatically picked the Youden index greater than 0, reflecting the arbitrary direction of the association.

In Figure 5.5 we present the Youden index for predicting the input variables i.e. compression ability. We can infer the information content of each feature through the

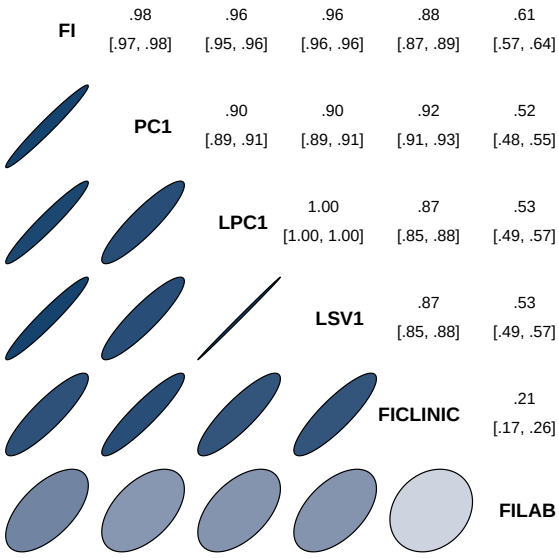

Figure 5.4: Spearman correlation of primary features across algorithms. Diagonal indicates the variable associated with each row and column. Above diagonal are the correlation coefficients between the row and column variables with 95% confidence intervals. Below diagonal are Gaussian contours with the corresponding correlation coefficient [129]. The first latent dimension for either PC, LPC or LSVD correlated strongly with the FI and each other, and correlated more strongly with the FI CLINIC than FI LAB. This implies a strong mutual signal very close to the FI, especially the FI CLINIC. Upper triangle is correlation coefficient with 95% confidence interval.

scores — a higher score implies more information related to a particular input variable. Similarly, in Figure 5.6 we score the association strength of each feature with each outcome. In both figures the inner colour indicates the lower limit of the 95% confidence interval (CI): lighter values are less significant (white is non-significant). Consistent with the compression observations, we see nearly identical patterns between the FI and the first latent dimension: PC1, LPC1 and LSV1; note also the similarity to the FI CLINIC. We have included PCs up to 10 as input variables. We observe higher PCA dimensions tend to be weaker, but also more specific predictors.

5.3.3 Generalized Linear Models (GLMs)

The feature associations give an idea of what information is in each latent variable but they don't consider the contributions that multiple latent variables can make towards prediction. Our GLMs do this, and so allows us to see how many latent dimensions are needed to predict outcomes well.

The cumulative predictive power conditional on all available information up to the N^{th} PC/LPC/LSV is given in Figures 5.7 and 5.8. We have included demographic information as the 0^{th} feature; and are again estimating out-of-sample performance. We see that the discrete outcomes (Figure 5.7), require few dimensions to achieve near-maximum performance. Conversely, continuous outcomes (Figure 5.8) require many dimensions. Overfitting appeared to be present in the highest dimensions, as demonstrated by a drop in performance as the cumulative number of features becomes larger, Figure 5.7. Overfitting was much worse in the complete case data, ostensibly due to outcome rarity (Figure B.41). When choosing the number of PCs to use, the optimal balance between overfitting discrete outcomes and under-fitting continuous outcomes seemed to be at approximately 20 latent dimensions for both PCA and LPCA. It is interesting that LSVD, which was best at compression, required more dimensions, approximately 40, to predict continuous outcomes well. This suggests that LSVD could be susceptible to overfitting when case data are scarce.

We observed strong similarities between PCA and LPCA both in compression, Figure 5.3, and prediction, Figures 5.7 and 5.8. In Appendix 5.6.4 we demonstrate that, under reasonable asssumptions, PCA is the single-iteration approximation of LPCA, which explains the similarities.

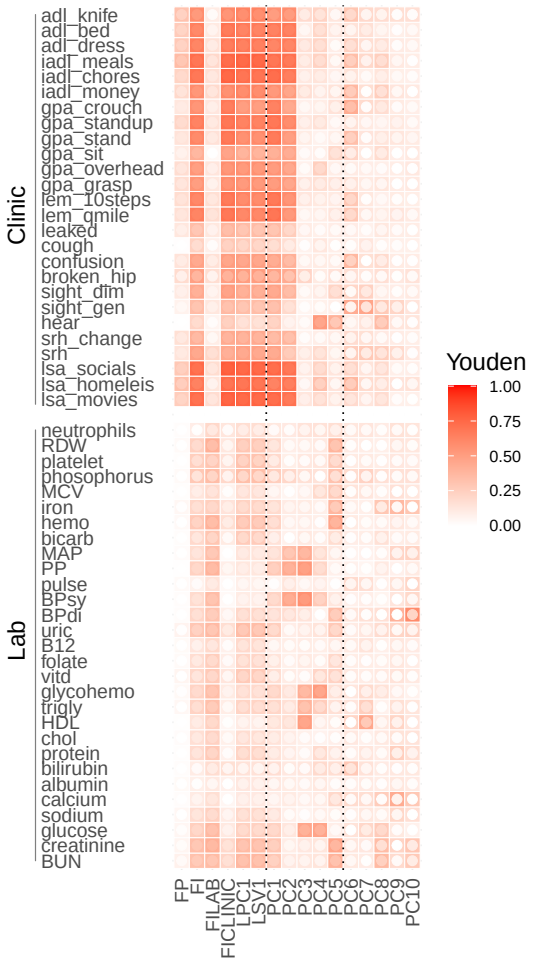

Figure 5.5: Feature associations with individual input variables, i.e. what goes into each feature. Youden index (fill colour) quantifies strength of associations between features (x-axis) and health deficits (y-axis); 0: no association, 1: perfect. Note the similarity of the FI, FI CLINIC, LPC1, LSV1 and PC1. Inner circle fill colour is the lower limit of 95% CI (white is non-significant). Higher PCs show no/low significance.

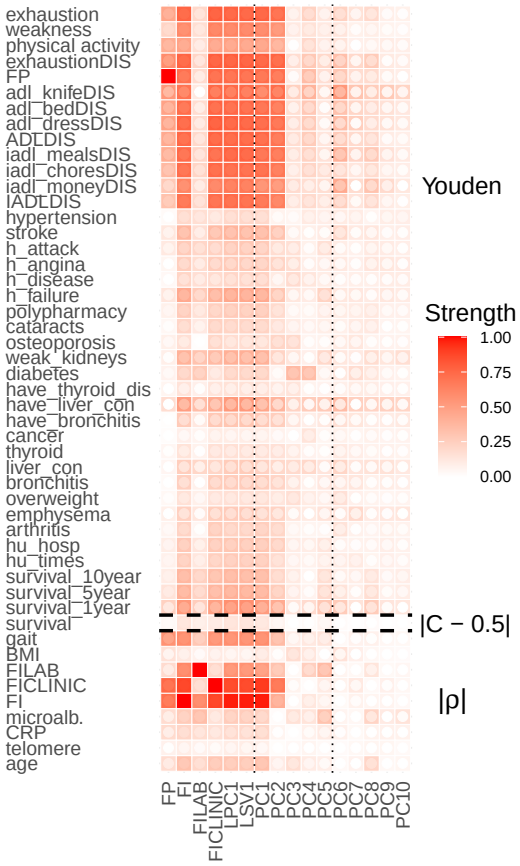

Figure 5.6: Feature associations with individual outcomes, i.e. what we get out of each feature. Association strength (fill colour) between features (x-axis) and adverse outcomes (y-axis); 0: no association, 1: perfect. Note the similarity of the FI, FI CLINIC, LPC1, LSV1 and PC1. Inner circle fill colour is lower limit of 95% CI (white is non-significant). Higher PCs show no/low significance. Text on right denotes accuracy metric used.

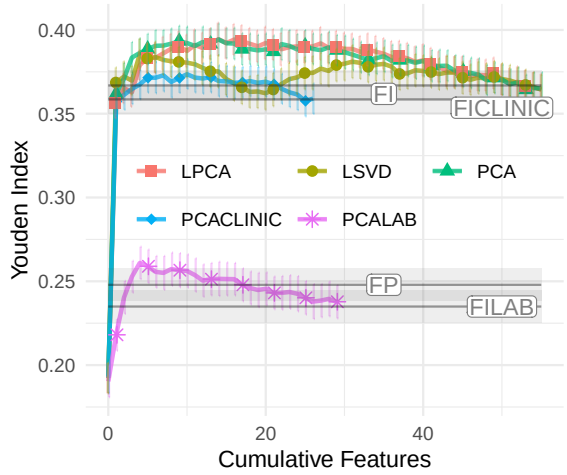

Figure 5.7: Cumulative prediction plot for discrete outcomes (GLM). 0th dimension is demographic information. Increasing the number of features initially improves prediction but eventually it gets worse due to overfitting. LSVD performs notably worse than PCA and LPCA. Youden index: higher is better.

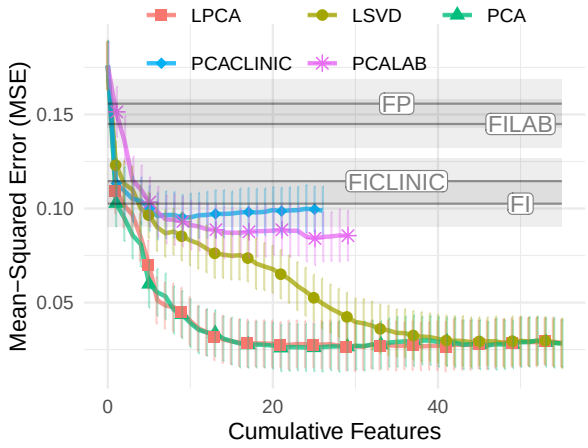

Figure 5.8: Cumulative prediction plot for continuous outcomes (GLM). 0th dimension is demographic information. Increasing the number of features improves prediction monotonically. LSVD performs notably worse than PCA and LPCA. MSE is on standarized scale, therefore $R^2 = 1 - MSE$. MSE: lower is better.

For specific outcomes, the performance of the GLM using PCA is shown in Figure 5.9, grouped by type. Consistent with Figure 5.7, medical conditions, disability and survival (all binary) tend to have low-dimensional representations and do not benefit from more than a few PCs: typically PC1 is sufficient. Note the difference between the FI CLINIC and FI LAB, with the former being perfectly reconstructed by 2 PCs, whereas the later required many more.

In Figure 5.10 we highlight selected outcomes which showed high-dimensional behaviour. These were variables that we visually observed in Figure 5.9 to have positive slopes up to several PCs (excluding the FI LAB because it shares input variables with PCA). We include FP as a reference system that is theoretically high-dimensional [57]. Several of the high-dimensional outcomes are related to biological systems that integrate information from many subsystems: inflammation and metabolism, as well as age itself. Note that microalbuminuria is connected to many different systems as a biomarker of microvasculature damage [208].

In Appendix B we repeated the stepwise GLM using either LPCA or LSVD. We observed only minor differences between LPCA and PCA (Figure B.24). LSVD showed much larger differences than PCA, in particular it achieved lower overall accuracies (Figure B.25). In all cases our qualitative results remain unchanged. We also considered non-linear behaviour by including quadratic and interaction terms between the PCs but found no improvement and a tendency to overfit (Figure B.22), suggesting that the linear model is optimal for the available data.

5.3.4 Robustness Analysis

PCA defines a particular linear transformation (rotation) between the original variables and a new 'latent' space. An important question is reproducibility of this latent space: can it be robustly estimated from the data?

We bootstrapped the sample to estimate the robustness of the linear transformation that rotates the data into PCs. The resulting rotation matrix up to the 10th PC is displayed in Figure 5.11. The exact values are in Table B.7. Note the overall sign of each PC is arbitrary [86]. We observed that the first 4 PCs were reliably estimated, 5 and 6 were marginally robust; the remaining PCs were too noisy to be consistently estimated. The loss of robustness could be due to PC features swapping

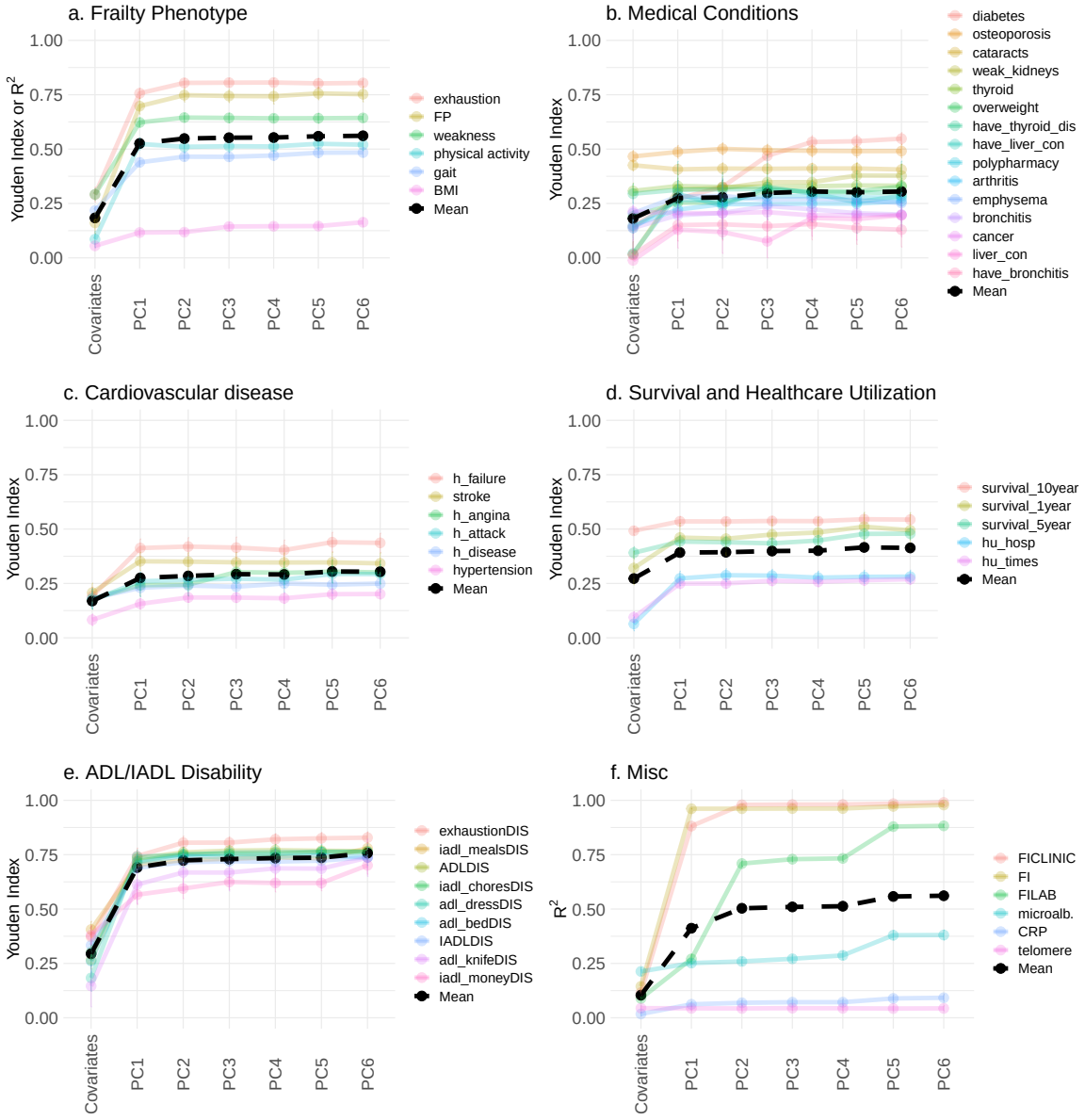

Figure 5.9: Improvement in predictive power as more PCs are included, grouped by outcome type (GLM). Coloured lines indicate specific outcomes, black line indicates the mean for each group. For most outcomes the performance stops improving after a few PCs, hence why we've truncated at PC6. The exceptions are explored in Figure 5.10. Note: legend is sorted from best (top) to worse (bottom) performance of the PC6 model. See Figure B.21 for the complete plots without truncation.

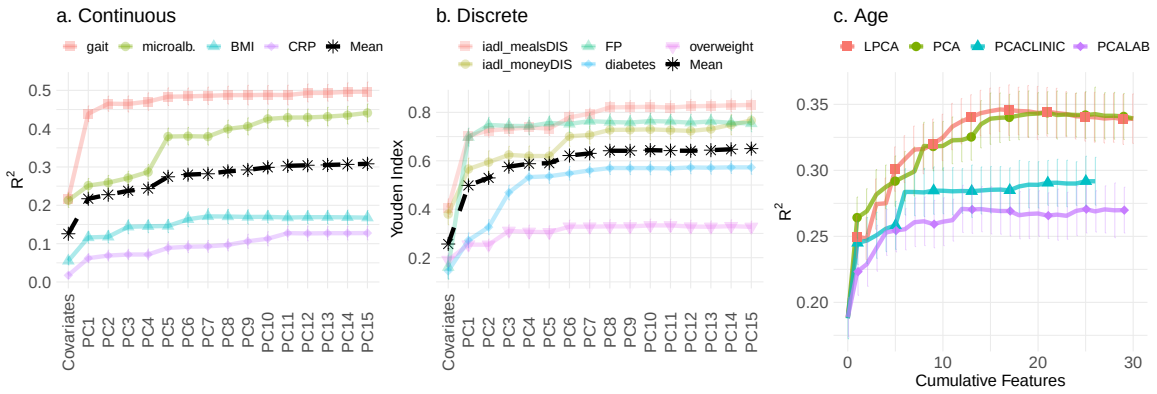

Figure 5.10: Improvement in predictive power as more PCs are included, high-dimensional outcomes (GLM). Outcomes were hand-picked variables based on requiring many PCs to achieve maximum performance. The FP was included for comparison. We tend to see continual improvement for the discrete and continuous outcomes, excluding the FP (up to ∼10). Age appeared to be the highest dimensional.

order due to small changes in their associated eigenvalues (see Figure 5.12), which could be addressed by a matching algorithm. On the other hand, the first 4-6 PCs appear to be robust and generalizable across the sample population.

PC1 is very close to the full FI (lab + clinical), as shown in Section 5.3.1, and we observe in Figure 5.11 that PC1 has nearly uniform weights for each variable, explaining the underlying similarity. Both the FI and PC1 are (nearly) unweighted averages of deficit variables. PC2 suggests that the next most important term to the full FI is a contrast term splitting lab and clinical inputs into their respective domains. PC3 has a similar structure of contrasting domains of blood pressure and metabolism. In Appendix B, we confirmed the robustness of the first 3 PCs to choice of variables by randomly selecting variable subsets of size 30, the remaining PCs did not appear to be robust (Figure B.27).

The corresponding second moments — i.e. eigenvalues — of the PCs are given in Figure 5.12. A bilinear structure was apparent in the log-log plot. In time series analysis, others have attributed this PC structure to fractal dimension [61], indicating a potential connection to complexity [111]. Values curved below the second line after approximately 20 PCs for the complete data, around 15 for the clinical data, and around 12 for the lab data. These values correspond to the end of the optimal-model

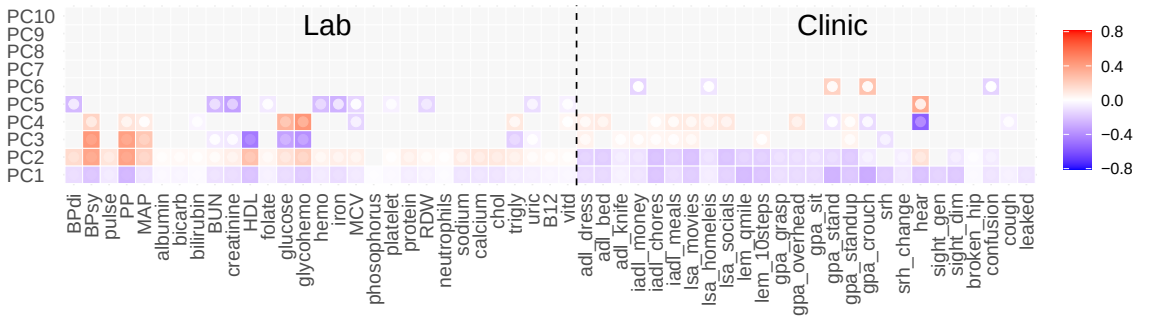

Figure 5.11: PCA robustness. Robustness of the PCA rotation was assessed by randomly sampling which individuals to include (i.e. bootstrapping, $N = 2000$). Left side are lab variables, right are clinical. Inner circle fill colour is 95% CI limit closest to 0. Grayed out tiles were non-significant. The first three PCs were quantitatively robust. We see the robustness drops with increasing PC number. The global sign for each PC were mutually aligned across replicates using the Pearson correlation between individual feature scores. In Figure B.27 we assessed robustness by randomly sub-sampling input variables and again observed that PCs 1-3 were robust.

regions[6] in Figures 5.7 and 5.8 (represented as bands in Figure 5.12); the curved region may therefore provide a useful heuristic for identifying less relevant PCs that exacerbate overfitting.

Note that the PC rotation does not have to be robust for it to be useful, for example PCA can be trained on one sample and used on another. Differences between the samples would then change the eigenvalues (via Eq 5.9) and PC ranks. Practically, this means performing feature selection after PCA, either by inspecting the eigenspectrum (e.g. Figure 5.12) or by using an automated algorithm such as LASSO [58].

5.3.5 Age Stratification

We investigated the effect of cohort age on our results. The joint 2D histogram tended to saturate (increase in magnitude) with age, although the qualitative structure of the histogram was stable (Appendix B). This implies that the PCA features — which are derived from decomposing the 2D histogram — do not change much with age. The increasing saturation does increase the relative contribution of the first PC with age,

[6]Models within this region were within 1 error bar of the best-performing model.

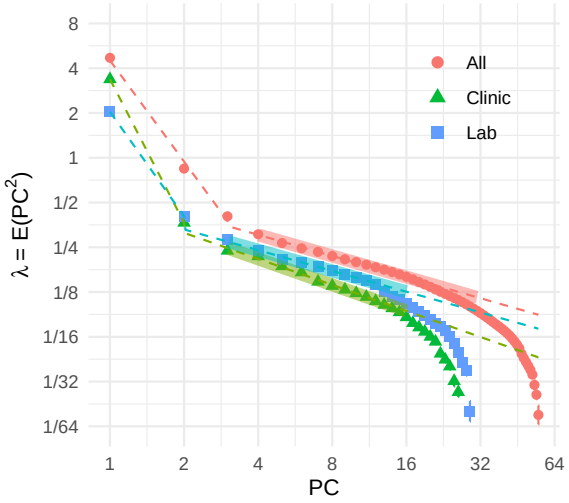

Figure 5.12: PCA second moments (eigenvalues) with bootstrapped standard errors (N=2000). Log-log scales. Note the bilinear structure. Banded region is optimal performance region (±1 error bar from best using Figures 5.7 and 5.8). In all three variable sets, eigenvalues curved away from second line just before overfitting started.

however. The first eigenvalue increased with increasing age quartile from 0.352(3) for ages 60-65 to 0.330(3) for ages 65-72 to 0.405(3) for ages 72-81 to 0.473(3) for ages 81+, as seen in Figure B.30.

To investigate a potential age effect further, we split the population at the median age (72) then redid the analysis using a young cohort (age < 72) and an old cohort (age 72+). (Note that we excluded demographical variables in this comparison because the baseline model, i.e. covariates as the only predictors, may not be equally powerful for both cohorts, confounding direct comparison.) Compression was similar for both cohorts (Figure B.31). Prediction using the GLM, however, was notably different (Figure B.32). For discrete outcomes, the cohorts scored similarly, the young cohort had a maximum Youden index of 0.333(15) compared to the older cohort which scored 0.326(13). For continuous outcomes, the young cohort performed much worse with a minimum MSE of 0.134(20) compared to the older cohort with 0.055(19).

We summarize the variable-specific compression and prediction using GLMs in Appendix B. The results were qualitatively similar, indicating robustness with respect to age. The GLM Youden indexes for compressing each predictor showed a stronger focus on predicting creatinine and BUN in the older cohort than in the younger

cohort. The younger cohort tended to prioritize other predictor variables, e.g. glucose, HDL and iron (Figure B.33). The GLM scores for predicting outcomes also showed microalbuminuria, and to a lesser extend gait, were better predicted in the older cohort (Figure B.34). Most of the differences we observed between the young and old cohorts were strongest in the higher PCs, this reflects the lack of robustness of the higher PCs that we observed in Figure 5.11.

5.4 Discussion

5.4.1 The first latent dimension "is" the frailty index

We performed dimensionality reduction on binarized health data encoded as normal (0) or dysfunctional (1). The first dimension of each algorithm, PCA, LPCA and LSVD, indicated a strong signal with general predictive power both for deficit compression and adverse outcome prediction. This first latent dimension correlated almost perfectly with the FI (Spearman $\rho > 0.95$), reproduced the same gender and age trajectories as the FI (Figure B.19), and had very similar associations (Figures 5.5 and 5.6).

What underlying phenomenon is this first latent dimension capturing? The FI is a measure of frailty [82], making it the primary suspect. Indeed, frailty is strongly associated with adverse outcomes [42] and the first latent dimension strongly predicted almost all outcomes. Specifically, the first latent variable predicted the five key frailty outcomes: exhaustion, weakness, physical inactivity, gait and weight (BMI). Frailty has been strongly associated with inflammation, total and HDL cholesterol, hyperglycemia and insulin resistance [60]. Consistent with this, we observed a (weak) relation between the first latent variable and HDL/total cholesterol, a (weak) relationship with the inflammation biomarker CRP, and a (moderate) relationship with glucose and glycohemoglobin. We saw stronger relationships with clinical measures: IADL/ADL disability, exhaustion and gait, all three of which are important signs/symptoms of frailty [60]. IADL/ADL disability is known to be strongly related to frailty and specifically the FI [189].

What is behind the approximate equality of FI and PC1? We see in Figure 5.2

that the 2D histogram for PC1 is approximately a large block of uniformly correlated variables. In Appendix 5.6.2 we show how an exact block structure leads to PC1 $\approx$ FI and that the FI becomes an increasingly good approximation for the information in the entire 2D histogram as the number of variables increase — saturating past approximately 30 variables. Indeed, others have reported moderate-to-strong correlations between all variables and equal associations/weightings to the first PC [131][7].

We hypothesize that the selection criteria for the FI [174] ensure that the joint histogram has this universal correlation structure between many variables: all deficit variables must (1) be related to health status, (2) increase in prevalence with age, (3) cannot saturate in youth, and (4) should contain "at least 30-40" variables [174]. These conditions are likely to lead to moderate-to-strong correlation between deficits due to their mutual age dependence through an individual's biological age (overall health state) [100]. These correlations then lead to PC1 $\simeq$ FI.

To summarize, the FI is an excellent summary measure for a large collection of moderate-to-highly correlated health deficit variables. That is, the FI acts as a "state variable" which summarizes the health state of an individual [158]. Under such conditions the FI approximately equals PC1 and can describe the collection of health deficit variables with little residual information, as has been empirically observed [207]. In turn, PC1 approximates the more appropriate loss function provided by LPCA (Appendix 5.6.4). The ease with which the FI, PCA, and other methods detect a very similar primary signal suggests that *any* good dimensionality reduction algorithm would identify it as the dominant signal in health deficit data. This signal predicts important outcomes and can be easily estimated via the FI or PC1.

5.4.2 PCs represent scales of dysfunction

PCs should be interpreted as building blocks consisting of coarse grained scales that can be added together to efficiently represent common patterns of dysfunction — adverse outcomes and health deficits. While others have discussed the biological

[7]Nakazato *et al.* (2020) [131] showed that the variability in a battery of blood tests has a correlation structure similar to what we observed, specifically all of the blood test variabilities were moderately or strongly correlated to each other. Their first principal component was roughly equally correlated with each input variable. As we show in the supplemental, this is a natural consequence of the correlation structure.

significance of individual PCs, for example as dietary patterns [49] or up/down inflammation regulators [9], it is unlikely that the PCs in the present study represent specific diseases or adverse conditions (excluding PC1). This is because the PCs must be statistically independent, while the first PC already represents generic dysfunction akin to the FI [101, 42]. Hence, any biological pattern of dysfunction using PCA should include a non-trivial contribution from PC1. Therefore PCs past PC1 are unlikely to represent specific pathways of dysfunction.

Instead, we should look for the minimum number of PCs to combine to construct a known pattern of dysfunction. For example, we can cross-reference Figure 5.11 against upticks in Figure 5.9 or 5.10. Low PC1 plus low PC2 gives global clinical dysfunction with agnostic lab, which is tantamount to the FI CLINIC. This explains why PC1 plus PC2 reconstructed the FI CLINIC. Low PC1 plus high PC2 gives quasi-global lab dysfunction with agnostic clinic, with strong cardiovascular dysfunction, such as would be seen in metabolic syndrome [3]. Adding low PC3 would give metabolic dysfunction alone, this explains why inclusion of PCs 1-3 gives a sudden improvement in BMI, obesity and diabetes prediction. If we then add high PC4, we could get dysfunctional glucose metabolism alone, which explains the uptick in diabetes prediction with inclusion of PC4 [35, 208]. PCA provides an efficient coarse graining procedure such that many common patterns of dysfunction are efficiently represented as sums of PCs.

How does PCA achieve this? PCA identifies domains of variables likely to be mutually deficient, i.e. strongly correlated. In this manner, PCA coarse grains by concatenating domains in a PC (e.g. PC 1 contained all domains), approximating them as a block, and then in the next PC it can contrast those domains with opposing signs to account for the stronger within-domain correlations than between-domains (e.g. PC 2 splitting lab and clinical). In this way the PCs encode domain-specific information, similar to the way experts have manually created domain-specific FIs [17]. Understanding health using multiple domain-specific FIs may be helpful for interpretability but could also make the analysis vulnerable to issues related to collinearity, such as unreliable regression coefficients [86]. In contrast, PCA is high-throughput, and PCs are uncorrelated, making PCA a better foundation for quantitative approaches — including preprocessing [78] before mapping into domains.

An alternative route to improving interpretability is through formal latent variable modelling. For example, grade of membership simultaneously infers health "profiles", along with individual scores for each profile, which are similar to PC scores [116, 50, 178]. The primary advantage that we see in PCA is that it compresses information into the lowest PCs by systematically estimating the direction of highest variance, followed by second highest, etc. This yields a set of optimal representations [86], and makes it particularly easy to quantify the information lost by picking a smaller representation. For example, Figure 5.3 shows the efficiency of each representation from 1 dimension up to the number of input dimensions. PCA also has several practical advantages over formal latent variable models: it is simple, fast, convex, easily tuned, reversible, and standard in statistical software packages, such as R [152]. Because of this, PCA can be easily integrated into an existing analysis pipeline as a preprocessing step.

PCA appears to generalize the action of the FI. The FI treats all health deficits as indistinguishable, such that you can pick *any* 30+ and expect to get the same summary health measure (subject to selection criteria) [174]. PC1 $\simeq$ FI adopts indistinguishability of deficits from the FI. PC2 is able to 'see' (discriminate) the difference between lab and clinical deficits, but can't distinguish individual lab variables from lab nor clinical from clinical. For example, we expect PC1 and PC2 will change little if a new admixture of lab and clinical variables are used — while higher PCs will change more. PC3 is able to 'see' the difference between metabolic, heart-related vs other deficits, and so forth for higher PCs. Within each PC the exact variables used should be unimportant, as long as they come from the same domains.

5.4.3 Domains in lab vs clinical data

We see that dimensionality reduction algorithms treat clinical and lab data domains differently, and are sensitive to domain boundaries. Strongly mutually-dependent variables form block-like domains in the joint histogram, which can be efficiently represented by a single latent dimension, making them preferred targets of PCA (and LPCA).

Clinical variables were strongly associated with a single latent variable whereas lab variables spanned more dimensions. For example, comparing the FI CLINIC to FI

LAB in Figure 5.9: the FI CLINIC was almost completely described by 1 PC whereas the FI LAB required at least 5. Inspecting the 2D histogram, Figure 5.2, we can see that the clinical data have stronger inter-dependencies than the lab. Previous research has shown that clinical variables are sufficiently compressed by a single dimension [207], whereas lab variables need at least two [66]. We did see an indication of high dimensional clinical data in the pooled continuous outcome prediction of clinical PCA, which improved up to 5-6 PCs, probably due to improvements in CRP, BMI, gait and/or age (which were high-dimensional).

Clinical deficits tend to accumulate over time and are efficiently described by the FI CLINIC. In contrast, the lab data are more complex, reflecting the diversity of biological systems the lab data represent, for example: metabolic (e.g. cholesterol and glucose [92]), immune (e.g. neutrophils [109] and CRP), renal (e.g. creatinine and BUN [133]) and cardiovascular (e.g. blood pressure [133]). Ostensibly, there are too many directions for dysfunction to proceed in to be completely captured by a single summary measure such as the FI LAB alone. For example, an individual may be prone to metabolic dysfunction, as indicated by dysfunctional glucose and glycohemoglobin, whereas another may have a weak heart, as indicated by dysfunction blood pressure, or weak kidneys. Why should these individuals accumulate (and propagate) dysfunction, or damage, in the same way? Our results indicated that they don't; multiple dimensions of PCs are needed to represent the diverse phenotypes of dysfunction captured by lab data. In contrast, the clinical data appears considerably more homogeneous.

Clinical data therefore seem to contain more generic (albeit crucial) information than lab data, with only a few dominant PCs in the former but more PCs needed in the latter. This may reflect the improved resolution of biological dysfunction in lab data. For example, lab data can resolve heart disease from hypertension vs from chemotherapy toxicity, but the clinical consequences of heart disease are the same either way. Inclusion of molecular data, such as metabolomics, proteomics or genomics, in future studies would clarify whether this trend towards more PCs continues as biological resolution is further increased.

The underlying univariate structure of the clinical data means that when we calculate the FI with equal weightings we are favouring the strongly-correlated clinical

deficit data over the weakly-correlated lab data. PCA targets large, dense blocks of highly-correlated variables. In the present study, the clinical data formed a dense block of the same size as the lab data and were therefore preferred targets. Conversely, we expect that a very large block of weakly-correlated variables would be a preferable target over the relatively small number of clinical variables. Thus, if we had included an exceptionally large domain, e.g. 'omics data with thousands of features, then it could dominate any much smaller domain. How would we know if there is a problem? We can look for blocks in the 2D joint histogram: a large block indicates strong mutual dependence, which will drag most algorithms towards it. A two-stage, hierarchical dimensionality reduction procedure, homologous to the "bifactor" model of [66], would mediate such an effect, and would be a good starting point for 'omics data. One could do PCA on each domain, take the most important PCs from each domain, and then perform PCA on all of the top PCs.

5.4.4 The dimensionality of integrative systems

"High-dimensional" outcomes required many PCs to fully predict. If an outcome relies on integrating information from many domains then we should see an incremental improvement as we move from PC1 to higher PCs. For example, prediction of age continually improved until about PC20. This is indicative of a high-dimensional, integrative process that accumulates dysfunction over several domains/scales and we therefore surmise, many different pathways of dysfunction. Stated equivalently, these are systems that function in many different ways.

CRP and chronological age showed the highest dimensionality, ostensibly integrating information from many domains. CRP is an inflammation biomarker indicative of altered cellular communication, the latter has been called an "integrative hallmark of aging", meaning that it indicates a phenotypic accumulation of damage [114]. These outcomes may be indicators of accumulated damage across domains and, ostensibly, scales. The approximately 20 PCs needed by age represents information integrated over all domains, probably leaving only noise in the remaining PCs (Figure 5.12). In regressing against age we are generating a biological age model [100]. Such a model is effectively condensing 20 dimensions of age-related decline into a single measure. This explains why there exists many partially-overlapping biological ages [87, 110]:

each biological age represents a different one-dimensional projection from a high-dimensional latent space. All of these ages contain overlapping contributions from the first latent dimension due to its strong explanatory power.

In contrast, medical conditions, ADL/IADL disability, survival and FP all seem to have one dominant dimension: PC1. For these outcomes, the first PC predicted almost as well as including all 55. This means that the only dimension we know is useful for predicting these outcomes is the dimension representing generic health deficits. This implies that our knowledge of these outcomes lies on a line: things go wrong in just one direction.

Money difficulty and difficulty preparing meals, both IADL, were notable exceptions that depended on higher PCs, notably PCs 6-8. These were the most cognitive-intensive clinical outcomes, which suggests that cognitive decline has its own domains of dysfunction captured by later PCs, and is consistent with what others have observed via factor analysis [191]. This highlights the critical difference between outcomes which appear to integrate information across multiple scales/domains, such as chronological age, versus those that depend on specific domains beyond a low rank PC representation, such as difficulty preparing meals. The former should show continual improvement as the number of PCs increases whereas the latter should show sudden improvement when a specific PC is included (e.g. compare the curves for predicting age versus difficulty preparing meals, iadl_mealDIS in Figure 5.10, the former improves with each additional PC).

5.4.5 Practical considerations

The FI's ability to effectively compress the salient information within a set of binary health deficits appears to be due to a dominant underlying signal that is readily identified by various dimensionality reduction techniques. PCA is the most common, robust, simplest and fastest. LPCA is a more complex algorithm that can enhance compression without loss of predictive power. LSVD is too focused on compression to yield good predictive features; it is also much slower. However, any of these techniques can be used to extend the dimensionality of the FI.

A critical aspect of our central hypothesis — that efficient representations of health deficits are efficient representations of adverse outcomes — is that biomarkers must

be converted to a standardized dysfunction scale. Sample-specific scales, such as the standard deviation, run the risk of propagating sample population idiosyncrasies or healthy variation. In contrast, deficit thresholds have been expertly tuned. Applying PCA directly to continuous biomarker data without converting to a standardized dysfunction scale may result in features that primarily capture healthy variation and/or have no clear connection to adverse outcomes.

We have focused on dimensionality reduction using compression algorithms, which do not depend on any specific outcomes. Dimensionality reduction could also be used with specific outcomes [2], or could simply be used with some of the adverse outcomes as input variables — for example medical conditions like diabetes and heart disease [19].

As we observed with LSVD, while compression seeks an efficient representation of the input it may not also be efficient for prediction. We hypothesized that efficient representations of health deficits would also be efficient representations of adverse outcomes. It is thus a surprise that we observed LSVD compressing so well, given its relatively poor predictive performance. Since LSVD has many more parameters than either PCA or LPCA, this could be a manifestation of overfitting to the input data, i.e. finding population-specific features rather than health-specific features.

Both PCA and LPCA are designed to handle cross-sectional data, although we expect they will also be useful for longitudinal data. Both are based on reversible linear transformations which preserve information, and hence they can be applied to new populations or measurement waves without loss of information. If the PCs/LPCs are expected to remain constant over time then we can simply pick a convenient wave to compute the transformation, probably the first, then apply the transformation to all other waves. This would be a viable approach in the present study population, since we observed that the PC transformation did not depend on age (Appendix Section B.3.5). If the transformation changes between waves, then we would suggest to first combine the waves to learn a shared transformation, then apply it separately to each wave.

What additional utility does PCA provide over the FI? Each of the multivariate dimensionality reduction algorithms was at least as good at predicting any given outcome as the FI was. It is also clear from stepwise regression that truncated PCA

can help to avoid overfitting; we can surmise that it would be particularly useful for avoiding the "curse of dimensionality" (when the number of predictors meets or exceeds the number of individuals). The only downside to a multivariate approach is the increased computational complexity, which is minor for PCA.

How do we know the right number of features (PCs) to use? Conventionally, one looks for an 'elbow' in the eigenspectrum, indicating a sudden drop in PC information content, or one picks the optimal number that maximizes prediction accuracy for a desired outcome [86]. PCs with small eigenvalues are, by our choice of normal/dysfunctional scale, minor corrections to the 2D deficit histogram which we hypothesize are unimportant for predicting adverse outcomes. In the present study, we observed that the eigenspectrum had a distinct bilinear structure in the log-log plot, and inferred that when eigenvalues drop below the second line, it is an indication of a drop in PC importance. Others have noted small eigenvalue PCs were significant predictors of mortality [78]; we hypothesize that these PC eigenvalues would lay on the second line (but not below). Our proposed method of finding the 'elbow' can be used to automatically identify the number of PCs to use without needing to refer to any particular outcome. Other fields have shown that there is an unmet need for unbiased PC selection criteria such as ours [48].

We observed little change in PCA performance with age. One minor change was in the relative importance of certain biomarkers/outcomes, as indicated by the PC order in which they appeared. The primary age-effect was a monotonic increase in overall deficit frequency with age, a phenomenon that is well known from the FI [124]. Although the 2D histogram thus became increasingly saturated with age, normalizing by the mean FI showed that the histogram structure was not changing, only the global deficit frequency. This suggests that the canonical pathways of dysfunction do not change with age, rather they saturate.

We note a few sources of error. Data imbalances can negatively affect performance measures and model fits, hence we focused on the Youden index and used a weighting scheme for the GLM. Most of the clinical variables used as inputs were related to physical activity which may skew our interpretation, although other researchers have used a similar set [19]. Our clinical variables were self-reported rather than observer assessed, which may affect performance [187] although it is unclear how

[17]. The main weakness of this study is our use of a single population. We used cross-validation and performed a robustness analysis to estimate precisely the effects for the sample population, but there may be study or population–specific effects in our results. However, the complete case data provided a subpopulation of younger, healthier individuals, and yielded similar conclusions (Appendix B).

We have also neglected to include social vulnerability deficits, which contain additional predictive power over the FI alone [202], although we did include partner status, education and income as covariates. It is curious to consider how PCA would handle grouping domains of information such as social vulnerability and observer vs self-reported health deficits — we expect it would modify the PCs to find these new domain boundaries.

5.4.6 Future Directions

Our work has been motivated by the need for estimating summary measures of health from new domains and with multiple dimensions. PCA may provide a useful extension of the FI to higher dimensions. Other approaches are also worth exploring, such as non-negative matrix decomposition [107], variational autoencoders [119] or kernel PCA [170]. Nested approaches may be useful for dealing with domains with many variables e.g. 'omics data. Our ability to robustly estimate latent dimensions also provides an opportunity for more interpretable latent variable modelling, for example structural equation modelling or factor analysis [66]. The stability of the PC rotation with age may also make it a useful preprocessing step for longitudinal analysis, such as for dynamical modelling. PCs can reduce dimensionality and likely have simplified interactions.

PCA appears to have additional utility. If the 'elbow' we observed in the PC spectrum is caused by a transition from signal to noise, it may be useful for denoising, which has been identified as an important issue with epigenetic clocks [78].

5.5 Conclusion

We compared several dimensionality reduction algorithms for their ability to compress health deficit data and predict adverse outcomes. The FI, PCA, LPCA and LSVD all identified the same dominant signal. This demonstrates and explains the FI's uncanny ability to predict adverse outcomes. We found that the additional dimensions estimated by PCA were helpful for better capturing health outcomes, particularly integrative systems such as inflammation, metabolism, and chronological age. Such systems were sensitive to many dysfunction pathways, domains or scales. PCA is a simple tool that can help researchers to identify and efficiently represent multidimensional biological systems in aging research.

5.6 Appendix: Binary PCA

We seek an efficient representation for binary health data: normal (0) or deficit (1). Equivalently, we are seeking: (1) a basis set, (2) a set of features, or (3) a set of composite health measures. Representation efficiency is commonly measured by *compression*. Compression is the ability to take a set of p input variables, reduce ("compress") them into $k < p$ latent variables, and be able to reconstruct the original p input variables from these k latent variables. Compression fidelity is measured by a loss function that compares the original input variables to the reconstructed inputs.

Each dimensionality reduction technique is the optimal solution to a particular choice of loss function. If we require independent (orthogonal) features and minimize the mean-squared error then the solution is given by PCA (see below). The mean-squared error is not ideal for binary data, however, since we only need to know whether a variable is larger or smaller than 0.5. More appropriate choices for binary data lead to logistic PCA (LPCA) [105] and logistic SVD (LSVD) [105, 167]. The performance gains of these methods over linear PCA have been modest [167, 105], and may not justify the increased algorithmic complexity. When we refer to PCA we mean conventional, linear PCA.

In finding the directions of maximum variance, PCA decomposes the covariance matrix into its eigenvectors. For binary healthy/deficit data the covariance matrix takes on a special meaning. The (uncentered) covariance matrix is equivalent to the

2D pairwise joint deficit frequencies, with the diagonal corresponding to the individual variable's marginal probability (i.e. deficit frequency). Binary PCA effectively compresses all of the marginal and pairwise deficit probabilities into a set of features of decreasing importance. We refer to each feature as a *latent dimension*.

5.6.1 Problem Formalization

We seek an orthogonal basis set of features to efficiently represent the data. Orthogonality ensures that features are independent (uncorrelated) and that each individual has a unique representation in terms of the basis [26].

Let $\vec{\phi}_i$ be the ith basis feature, we want to minimize the reconstruction error between the $N \times p$ data matrix, X, in its original state and after bottlenecking through the (latent) feature-space of size $k \leq p$. The representation of the ith individual and jth deficit from our data in the new space is given by:

$$\hat{X}_{ij} \equiv \sum_{n=1}^{k} Z_{in}\phi_{nj}, \tag{5.2}$$

where Z_{in} represents an individual's feature score and $\hat{X}_{ij}$ is our best estimate of the reconstructed input data, X_{ij}.

Using the orthogonality of the $\vec{\phi}_i$ we estimate the feature scores using the inner product,

$$Z_{in} = \sum_{j=1}^{p} \phi_{nj} X_{ij} \tag{5.3}$$

thus,

$$\begin{aligned} \hat{X}_{ij} &= \sum_{j=1}^{p}\sum_{n=1}^{k} \phi_{nj} X_{ij} \phi_{nj} \\ &= \sum_{j=1}^{p} X_{ij} \sum_{n=1}^{k} \phi_{nj}\phi_{nj} \\ &= \sum_{j=1}^{p} X_{ij} \sum_{n=1}^{k} U_{jn} U^{t}_{nj} \\ \implies \hat{X} &= XUU^{t} \end{aligned} \tag{5.4}$$

where U is the $p \times k$ matrix formed by having ith column equal to the ith basis, $\vec{\phi}_i$, and U^t is the transpose of U.

For simplicity, convexity and robustness, we assume the mean-squared error function, hence we have:

$$\min_{\{\vec{\phi}_i\}} \sum_{i=1}^{N} \sum_{j=1}^{p} \left(X_{ij} - [UU^t \vec{X}_{i\cdot}]_j\right)^2 \text{ with } \sum_{j=1}^{p} U_{ji} U_{jk} = \delta_{ik}. \tag{5.5}$$

This is the Pearson formalism of PCA (where the mean has not been subtracted) [105]. $Z \equiv XU^t$ is the PC score matrix and U is the *rotation matrix*. This formalism can be solved sequentially for each $\vec{\phi}_i$ and is equivalent to picking the rotation of the data such that the first direction, Z_{i1}, has the maximum second moment (eigenvalue), the second direction has the second largest, and so forth [86].

The solution to Eq. 5.5 is found by eigen-decomposition of $X^t X$ [86]. Each of the columns of U, $\vec{\phi}_i$ satisfy

$$\frac{1}{N} X^t X \vec{\phi}_i = \lambda_i \vec{\phi}_i \qquad \text{where } \vec{\phi}_i \equiv U_{\cdot i} \tag{5.6}$$

where λ_i is the ith eigenvalue and $X^t X/N$ is the 2D histogram of joint frequencies of the binary input variables, with the diagonal equal to the 1D frequencies. This implies (using $X^t X \approx X^t \hat{X}$, Eq. 5.4 and Eq. 5.6),

$$\frac{1}{N} X^t X \approx \sum_{i=1}^{k} \lambda_i \left(\vec{\phi}_i \otimes \vec{\phi}_i\right) \tag{5.7}$$

with equality when $k = p$. $\otimes$ denotes the outer/tensor product and the terms are sorted by decreasing strength. Intuitively, we are forming the 2D histogram, $X^t X/N$, then decomposing it into a set of rank 1 matrices — i.e. square blocks — sorted by relative contribution; Figure 5.2 illustrates the process for our dataset.

The principal components (PCs), $P \equiv Z$, are defined as the initial data transformed ("rotated") into the latent space,

$$P_{ij} \equiv \sum_{k=1}^{p} X_{ik} U_{kj} \tag{5.8}$$

using the eigen-decomposition, Eq. 5.6, we can show that the norm of each PC is

determined by its eigenvalue (substituting U for ϕ),

$$\begin{aligned} & \frac{1}{N}\sum_k\sum_j X^t_{nk}X_{kj}U_{ji} = \lambda_i U_{ni} \\ \implies & \frac{1}{N}\sum_n\sum_k\sum_j X^t_{nk}X_{kj}U_{ji}U_{nm} = \lambda_i \sum_n U_{ni}U_{nm} \\ \implies & \frac{1}{N}\sum_k P_{km}P_{ki} = \lambda_i \delta_{im} \end{aligned} \tag{5.9}$$

hence the second moment of each PC determines its eigenvalue, λ, and therefore its order and relative importance. The sum of the second moments is conserved because U is an isometry [26].

5.6.2 Block Histogram

There is a special 2D joint histogram pattern for which the first PC is equal to the FI for both logistic [105] and linear PCA (scaled by an irrelevant constant). When a uniform diagonal is on top of a dense, uniform, off-diagonal, the FI is the dominant eigenvector and is therefore the first PC.

More precisely, suppose the 2D joint frequency histogram, X^tX/N, is given by:

$$\frac{1}{N}X^tX = \begin{bmatrix} a & b & \dots & b \\ b & a & & \vdots \\ \vdots & & \ddots & b \\ b & \dots & b & a \end{bmatrix} \tag{5.10}$$

that is, the diagonal is constant, a, and the off-diagonals are also constant, b. This is a circulant matrix [4]. Note that $a \geq 0$, $b \geq 0$ and $a \geq b$, because they're deficit frequencies ($(X^tX)_{ij} = N\langle x_i x_j\rangle$ for binary variables x_i and x_j, clearly $\langle x_i^2\rangle \geq \langle x_i x_j\rangle$ so $a \geq b$ because $a = \langle x_i^2\rangle$ and $b = \langle x_i x_j\rangle$, where $\langle x_i\rangle$ is the mean of x_i). The eigenvalues of this circulant matrix are [4]:

$$\lambda_k = a - b + b\sum_{j=0}^{p-1}\left(\exp\left(\frac{2\pi}{p}ki\right)\right)^j \tag{5.11}$$

where $k \in [1, p]$ is an integer and p is the number of columns in X (i.e. the number of

variables); $i \equiv \sqrt{-1}$. If $k \neq p$ the sum is a geometric series which converges to [204],

$$\lambda_k = (a-b) + b\left(\frac{1-\exp\left(\frac{2\pi}{p}pki\right)}{1-\exp\left(\frac{2\pi}{p}ki\right)}\right)$$
$$\lambda_k = a - b \qquad\qquad k \neq p \tag{5.12}$$

If $k = p$ we instead have,

$$\lambda_p = a - b + b\sum_{j=0}^{p-1}\left(\exp\left(2\pi i\right)\right)^j$$
$$\lambda_p = a + (p-1)b \tag{5.13}$$

because $a, b \geq 0$ and $a \geq b$ we have that λ_p *must be* the first (largest) eigenvalue (assuming $b > 0$, otherwise it will be a tie).

The associated eigenvectors are given by [4],

$$U_{kl} = \frac{1}{\sqrt{p}} e^{-\frac{2\pi}{p}ikl} \tag{5.14}$$

where k and l are integers. From Eq. 5.13 we know the first eigenvector is,

$$U_{pl} = \frac{1}{\sqrt{p}}. \tag{5.15}$$

Using Eq. 5.8 we can calculate the first principal component,

$$\begin{aligned} P_{\cdot 1} &= \sum_j \frac{1}{\sqrt{p}} X_{ij} \\ &= \sqrt{p}\frac{1}{p}\sum_j X_{ij} \\ &= \sqrt{p} \cdot \text{fraily index}, \end{aligned} \tag{5.16}$$

which is a constant times the FI. Hence if the joint histogram has the form of Eq. 5.10 the FI will coincide with the first PC. In the next section we show the conditions under which the first PC is sufficient.

5.6.3 How well can we approximate the histogram?

The 2D histogram contains all pairwise frequencies (off-diagonals) and individual frequencies, making it an important summary of the information we know about the

deficit statistics. How well does the first eigenvalue/eigenvector pair approximate the complete 2D histogram, given it has the special structure of Eq. 5.10?

From Eq. 5.7, we know that the eigenvalues/eigenvectors approximate the 2D histogram as:

$$\frac{1}{N}X^tX \approx \sum_{i=1}^{k} \lambda_i \left(\vec{\phi}_i \otimes \vec{\phi}_i\right) \tag{5.17}$$

with equality when k is equal to the number of variables, p (equal to the number of columns of X). Since the model is linear, we can summarize the mean-squared error using the coefficient of determination, R^2, and expect $R^2 = 0$ for a useless reconstruction and $R^2 = 1$ for a perfect reconstruction. Specifically,

$$R^2 = 1 - \frac{\sum_i \sum_j \left((X^tX)_{ij}/N - \sum_{l=1}^{k} \lambda_l \left(\vec{\phi}_l \otimes \vec{\phi}_l\right)_{ij}\right)^2}{\sum_i \sum_j \left((X^tX)_{ij}/N\right)^2}. \tag{5.18}$$

Using this, we compute the accuracy of the first eigenvalue/eigenvector pair in approximating the full 2D histogram,

$$\begin{aligned} R^2 &= 1 - \frac{\sum_i \sum_j \left((X^tX)_{ij}/N - \sum_{l=1}^{1} \lambda_l \left(\vec{\phi}_l \otimes \vec{\phi}_l\right)_{ij}\right)^2}{\sum_i \sum_j \left((X^tX)_{ij}/N\right)^2} \\ &= 1 - \frac{\sum_i \sum_j \left((X^tX)_{ij}/N - (a + (p-1)b)(1/p)\right)^2}{\sum_i \sum_j \left((X^tX)_{ij}/N\right)^2}, \end{aligned} \tag{5.19}$$

and substitute in the special form for X^tX/N,

$$\begin{aligned} R^2 &= 1 - \frac{\sum_i \sum_{j\neq i} \left(b - (a + (p-1)b)(1/p)\right)^2}{\sum_i a^2 + \sum_i \sum_{j\neq i} b^2} \\ &\quad + \frac{\sum_i \left(a - (a + (p-1)b)(1/p)\right)^2}{\sum_i a^2 + \sum_i \sum_{j\neq i} b^2} \\ &= 1 - \frac{p-1}{p}(1 - b/a)^2 \frac{1}{p(b^2/a^2) + (1 - b^2/a^2)} \end{aligned} \tag{5.20}$$

where in the last line we emphasize there are only two tunable parameters: b/a is a measure of correlation strength and p is the number of variables. Both $0 \leq a \leq 1$ and $0 \leq b \leq a$ are constrained because X is composed of binary variables.

There are two limits of interest. First, for $b > 0$ if we take $b \to a$,

$$\begin{aligned} \lim_{b\to a} R^2 &= 1 - \frac{p-1}{p^2}\frac{(a-b)^2}{b^2} \\ &= 1 \end{aligned} \tag{5.21}$$

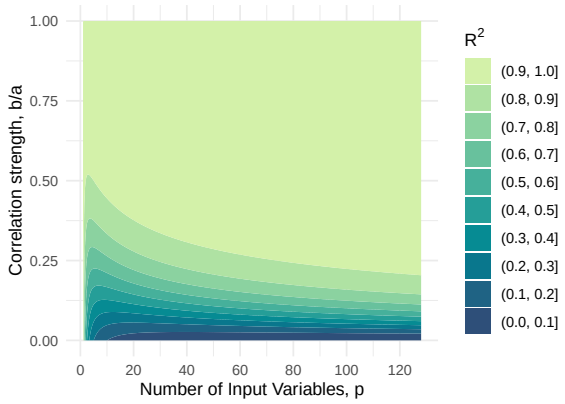

Figure 5.13: Special joint histogram approximation, Eq. 5.20. Fill is the R^2 fit quality for PC1 approximating the full histogram, given the histogram has the special structure given in Eq. 5.10. p is the number of features. a is the deficit frequency. b is the joint deficit frequency.

this corresponds to a 2D histogram of perfectly dependent variables (which would be a rank 1 matrix). The other limit is taking a large number of variables with $b > 0$,

$$\begin{aligned} \lim_{p\to\infty} R^2 &= 1 - \frac{1}{p}\frac{(a-b)^2}{b^2} \\ &= 1 \end{aligned} \tag{5.22}$$

which corresponds to an infinitely large 2D histogram. In both cases $R^2 = 1$ and the first eigenvector — equal to the FI — is sufficient to perfectly estimate the 2D histogram and hence sufficient to completely describe the first and second order statistics. It is interesting to note the compatibility of the two limits which imply that getting a large, but finite, p and having b close to, but not equal to, a is likely to give $R^2 \approx 1$.

In Figure 5.13 we plot Eq. 5.20 for several values of the two free parameters, b/a and p. Nearly perfect R^2 is achieved for fairly modest values of b/a when p is sufficiently large. Interestingly, there is an apparent diminishing return for increasing p with an elbow at $p \approx 25$, this is comparable to the 30+ deficits rule for the FI [174]. The 2D joint histogram in this study had a median diagonal value of $a = 0.20$ and median off-diagonal value of $b = 0.04$, giving $b/a = 0.22$ ($p = 55$; Figure 5.2). We'd then expect an ideal case to have $R^2 = 0.84$, the fit for our data yielded $R^2 = 0.50$ — large, but smaller than the ideal case.

This idealized, "toy", model explains the approximate equivalence of the FI and PC1. What's more, it allows us to estimate how dominant the FI/PC1 is. In the

limit of a large number of variables and/or $b \approx a$ we find that the FI/PC1 becomes a better approximation for the information in the 2D histogram. This is consistent with the observation that the FI is best used to describe a large number of correlated variables.

5.6.4 PCA approximates logistic PCA

Logistic PCA [105] minimizes the Bernoulli deviance, in analogy to the Gaussian formulation of linear (normal) PCA. The optimization problem is not convex but Landgraf and Lee [105] derive an iterative majorization-minimization scheme for solving the problem. We follow their approach and show that the first iteration of their loss function reduces to the same loss function as linear PCA. As a result, the estimated transformation, U, will be the same for either PCA or logistic PCA after the first iteration.

There are four steps to our adaptation of their approach:

1. Initialize $U^{(0)}$ to be an orthogonal matrix. Pick $k = p$. Then $U^{(0)}(U^{(0)})^t = I$.

2. Initialize the mean, $\mu = \text{logit}(\epsilon)$ where $\epsilon \to 0^+$ is a small, positive number. This is akin to not subtracting the mean when we perform PCA.

3. Fix $m \equiv -\mu$. This is the main assumption. m should be a large, positive number [105]. Our definition of μ ensures that m is a large positive number.

4. Iterate the majorization-minimization algorithm [105] exactly once.

The initial $\theta_{ij}^{(1)} = \tilde{\theta}_{ij}$, due to the orthogonality of the initial $U^{(0)}$. Note that $\tilde{\theta}_{ij} \equiv$

$m(2X_{ij}-1)$ [105]. The loss function, Eq. (9) of [105], is then

$$\begin{aligned}
&\min_U \sum_i \sum_j \left([UU^t(\tilde{\theta}_{ij}-\vec{\mu})]_j - (\tilde{\theta}_{ij}-\mu) - 4(X_{ij}-\sigma(\tilde{\theta}_{ij}))\right)^2 \\
&= \min_U \sum_i \sum_j \Bigg([UU^t(2m\vec{X}_{i\cdot} - m\vec{1} - \vec{\mu})]_j \\
&\quad - (2mX_{ij} - m - \mu) - 4(X_{ij}-\sigma(\tilde{\theta}_{ij}))\Bigg)^2 \\
&= \min_U \sum_i \sum_j \Bigg(2m([UU^t\vec{X}_{i\cdot}]_j - X_{ij}) \\
&\quad - ([UU^t(m\vec{1}+\vec{\mu})]_j - m - \mu) - 4(X_{ij}-\sigma(\tilde{\theta}_{ij}))\Bigg)^2 \\
&= 2m \min_U \sum_i \sum_j \left([UU^t\vec{X}_{i\cdot}]_j - X_{ij}\right)^2, \qquad (5.23)
\end{aligned}$$

where we use $m \equiv -\mu$ and $\mu \to -\infty$ in the last line, with $m \to \infty$ ensuring $\sigma(\tilde{\theta}_{ij}) \to X_{ij}$ (σ is the inverse logit). The factor of $2m$ does not affect the position of the minimum and hence Eq. 5.23 finds the same optimal U as the PCA loss function, Eq. 5.5 (recall that U is constructed out of the set of $\vec{\phi}_i$).

Chapter 6

Network dynamical stability analysis reveals key mallostatic natural variables that erode homeostasis and drive age-related decline of health

By Glen Pridham[1] and Andrew Rutenberg[1].
[1]Department of Physics and Atmospheric Science, Dalhousie University, Halifax, B3H 4R2, Nova Scotia, Canada.

Pridham, G. & Rutenberg, A. D. Network dynamical stability analysis reveals key 'mallostatic' natural variables that erode homeostasis and drive age-related decline of health. Sci. Rep. 13, 1–12 (2023) doi:10.1038/s41598-023-49129-7 [147]

Using longitudinal study data, we dynamically model how aging affects homeostasis in both mice and humans. We operationalize homeostasis as a multivariate mean-reverting stochastic process. We hypothesize that biomarkers have stable equilibrium values, but that deviations from equilibrium of each biomarker affects other biomarkers through an interaction network — this precludes univariate analysis. We therefore looked for age-related changes to homeostasis using dynamic network stability analysis, which transforms observed biomarker data into independent "natural" variables[1] and determines their associated recovery rates. Most natural variables remained near equilibrium and were essentially constant in time. A small number of natural variables were unable to equilibrate due to a gradual drift with age in their homeostatic equilibrium, i.e. allostasis. This drift caused them to accumulate over the lifespan course and makes them natural aging variables. Their rate of accumulation was correlated with risk of adverse outcomes: death or dementia onset. We

[1]The natural variables are closely related to principal components, and in this study the two typically coincided, as shown in the supplemental. The natural variables are defined by the eigenvectors of the interaction network whereas the principal components are defined by the eigenvectors of the covariance matrix; the two coincide in the steady-state for a symmetric interaction network.

call this tendency for aging organisms to drift towards an equilibrium position of ever-worsening health "mallostasis". We demonstrate that the effects of mallostasis on observed biomarkers are spread out through the interaction network. This could provide a redundancy mechanism to preserve functioning until multi-system dysfunction emerges at advanced ages.

Keywords: dynamical, aging, homeostasis, allostasis, survival, dementia, mallostasis, allostatic load.

6.1 Introduction

Homeostasis is the self-regulating process that maintains internal stability [15]. Yet as individuals age, it is characteristic for biomarkers to drift away from healthy levels; something about homeostasis is therefore "lost" during the aging process [168]. For example, loss of protein homeostasis is believed to cause the hallmark accumulation of unfolded, misfolded and aggregate proteins with age [114]. Accumulation is observed at multiple biological scales, including oxidative damage [28], epigenetic age [110], senescent cells [93], and regulatory T-cells [154] at the cellular scale, and extending up to the whole organism scale where clinical deficits [123], including chronic diseases [51], accumulate with age. Sehl and Yates performed univariate analysis of 445 health biomarkers and found that almost all of them accumulate negatively with age — typically showing linear decline [176]. Such accumulation of biomarker values in a particular direction appears to be a generic feature of aging. When biomarkers reach abnormal values, they are associated with dysfunction and poor health, independently of age [19, 34]. A general mechanism of how accumulation and poor health emerge from homeostasis has, nevertheless, been missing.

Prior work suggests that accumulation may be a consequence of a drifting equilibrium position. Allostasis, literally "homeostasis through change" [118], describes a version of homeostasis in which the equilibrium position is mutable, adapting as necessary to environmental demands [89]. Over time, "wear-and-tear" of this adaptive stress-response causes a subclinical accumulation of dysfunction known as "allostatic load" [89]. We hypothesize that these allostatic changes may be asymmetric, causing a coherent, population-level drift in equilibrium biomarker values with age, and

ultimately leading to accumulating biomarker values in particular directions.

Directly estimating an individual's allostatic load remains an open challenge [89], owing to the confounding effects of the underlying interaction networks [34]. Instead, most algorithms infer allostatic load by outlier detection [89, 34] or other symmetric indicators, agnostic to any preferred biomarker accumulation direction [113, 213]. These approaches have not been reconciled with generic, age-associated biomarker accumulation, which proceeds in preferred directions [176]. It therefore remains unclear how allostatic load leads to worsening health. Other theories posit that outlying biomarker values indicate damage, which promotes further damage e.g. as quantified by the number of health deficits ("frailty index", FI)[122, 19, 181]. Needed is direct evidence of allostasis and how it is associated with worsening health.

Instability is another mechanism for accumulation. While linear accumulation is the norm [176], some biomarkers accumulate exponentially with age e.g. senescent cells [93] and the FI [123]. Exponential growth indicates an instability [8]. Nevertheless, a weak instability can appear linear until advanced ages. As a result, it remains unclear whether age-related accumulation proceeds due to a shifting equilibrium, a weak instability, or some hybrid of the two.

Operationalizing and quantifying homeostatic changes is challenging [89] because homeostasis is a property of the whole system, not individual constituent parts [15, 34]. In the language of complexity science [36], homeostasis is an emergent property of a network of interacting variables. Each variable measures a part of the system, but changes to one part can be balanced by other parts. For example, heart rate declines with age but can be compensated for by increased stroke volume [176] in order to maintain arterial blood pressure [15]. The essential aspects of homeostasis are: (i) a multivariate interacting dynamical system, (ii) an equilibrium state, which may vary with age (allostasis), (iii) the system spontaneously returns to the equilibrium state (dynamical stability), and (iv) stresses (and interventions) provide random shocks to the system. Altogether, homeostasis can be operationalized as a multivariate, mean-reverting stochastic process [213].

Dynamical stability analysis uses eigen-decomposition to probe the stability of arbitrary systems [106, 84]. The system is first linearized around an equilibrium

position [106]. Orthogonal eigenvectors are then identified that decouple the interactions between variables. Eigenvectors are composite health measures that serve as *natural variables* since they do not interact or compensate for each other, and so can be analyzed individually. Each such natural variable has an associated eigenvalue that determines its stability via a characteristic recovery rate or timescale ($-$eigenvalue = rate = timescale^{-1}). A system is stable if and only if all recovery rates are positive [106]. Conversely, dynamical instability arises only if at least one recovery rate is negative.

We confront homeostasis with minimal assumptions. We seek generic changes to biomarker equilibrium and stability within aging organisms. We investigate multiple longitudinal datasets with multiple organisms (mice and humans) and multiple outcomes (dementia and death). In contrast to earlier work by Sehl and Yates [176], our model is multivariate and generic so that we can model homeostasis without constraining its dynamical behaviour. We find that allostatic drift is consistent with the observed data. Importantly, we find that a small set of natural variables drive mortality and can be used to characterize an individual's health state. We do not observe any dynamical instabilities.

6.2 Model

To analyse stability for deterministic [106] or stochastic [84] dynamics, we use a linear approximation near a stable point,

$$\begin{aligned}
\vec{y}_{in+1} &= \vec{y}_{in} + \boldsymbol{W}\Delta t_{in+1}(\vec{y}_{in} - \vec{\mu}_{in}) + \vec{\epsilon}_{in+1}, \\
\vec{\epsilon}_{in+1} &\sim \mathcal{N}(0, \boldsymbol{\Sigma}|\Delta t|_{in+1}) \\
\vec{\mu}_{in} &\equiv \vec{\mu}_0 + \boldsymbol{\Lambda}\vec{x}_{in} + \vec{\mu}_{age}t_{in}
\end{aligned} \tag{6.1}$$

where i indexes the individual, n indexes the timepoint, t_n is the age, $\vec{y}$ is a vector of observed biomarkers, and $\vec{x}$ is a vector of covariates that includes sex. $\boldsymbol{W}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Lambda}$ are constant matrices, independent of i and n. If we take the average over individuals (indicated by angled brackets) then we can obtain rates of average change

$$\frac{\langle \Delta y_{ijn+1} \rangle}{\langle \Delta t_{in+1} \rangle} = W_{jj}\langle y_{ijn} - \mu_{ijn} \rangle + \sum_{k \neq j} W_{jk}\langle y_{ikn} - \mu_{ikn} \rangle. \tag{6.2}$$

We see that changes to the mean of a particular biomarker, y_j, are due either to recovery of y_j towards the equilibrium position, μ_j, or because of interactions with a compensating variable, $y_{k\neq j}$ through off-diagonal elements of $\boldsymbol{W}$. This provides both a mechanism for biological redundancy — if the organism can actively influence some of the y_k then it can use them to steer others — and a mechanism for mutual dysfunction — since $\boldsymbol{W}$ couples dysregulation of $y_{k\neq j}$ to that of y_j. In Appendix Section C.8 we show that equation (6.1) approximates general nonlinear dynamics [106]. We also specifically show that it approximates the stochastic process model [213], a framework for aging biomarker dynamics. Note that the model permits unequally-spaced sampling of individuals through Δt_{in+1}, which is the time between measurements of individual i at time t_n and t_{n+1}.

The stability of the model depends on the eigenvalues of $\boldsymbol{W}$. We can decouple variable means with the eigenvector transformation matrix $\boldsymbol{P}$. We obtain

$$z_{ijn+1} = z_{ijn} + \lambda_j \Delta t_{in+1}(z_{ijn} - \tilde{\mu}_{ijn}) + \tilde{\epsilon}_{ij}, \tag{6.3}$$

where $\vec{z}_n \equiv \boldsymbol{P}^{-1}\vec{y}_n$, $\lambda_j \equiv P_{j\cdot}^{-1}\boldsymbol{W}P_{\cdot j}$, $\tilde{\vec{\mu}}_n \equiv \boldsymbol{P}^{-1}\mu_n$ and $\tilde{\vec{\epsilon}} \equiv \boldsymbol{P}^{-1}\vec{\epsilon}$. We refer to $\vec{z}$ as *natural variables*. The natural variables build correlations only through the noise term, $\tilde{\epsilon}$ — in addition to any correlated initial conditions. The system is mutually-diagonal if $\tilde{\epsilon}$ is uncorrelated. While our dynamics are discrete, it is also helpful to consider continuous dynamics corresponding to the limit $\Delta t \to 0$; see Figure 6.1 and Box 1, Chapter 6 (also Appendix Section C.8 for more details).

The parameters $\boldsymbol{W}$, $\vec{\mu}_0$, $\vec{\mu}_{age}$ and $\boldsymbol{\Lambda}$ are estimated from the data ($\boldsymbol{\Sigma}$ can also be). The stochastic term, $\vec{\epsilon}$, is assumed to be normally distributed and independent across timepoints. See Appendix Section C.6 for details. (For the remainder of the paper, we simplify notation by dropping the tilde and suppressing the individual i and timepoint n indices.) Optimal parameter values are selected by maximizing the likelihood. For uncorrelated noise this reduces to weighted linear regression.

If a mutually-diagonal system reaches steady-state — having run long enough to forget initial conditions — then the natural variables, $\vec{z}$, are the principal components, ranked by stability (Appendix equation (C.49)). We used principal component analysis (PCA) as a preprocessing step. If $\boldsymbol{P}$ is orthogonal (which it is from PCA) then Parseval's theorem states that $\langle \sum_j y_{jn}^2 \rangle = \langle \sum_j z_{jn}^2 \rangle = \sum_j (\mathrm{Var}(z_j) + \langle z_j \rangle^2)$: this means that a single z_k with large mean and variance can dominate that of the $\vec{y}$.

Box 1. Ordinary differential equation (ODE) behaviour

Consider a 1-dimensional space, z. If we take the limit $\Delta t \to 0$ then equation (6.3) is a modified Ornstein-Uhlenbeck process. The mean and variance are solutions to ordinary differential equations. The mean is described by

$$\begin{aligned}\frac{d}{dt}\langle z\rangle &= \lambda(\langle z\rangle - \mu(t)) \\ &= \lambda\langle z\rangle - \lambda\mu_0 - \lambda\mu_{age}t \end{aligned} \tag{6.4}$$

where $\mu_{age}t$ is the time-dependent part of μ_n and μ_0 is the remaining part. The general solution of equation (6.4) is

$$\begin{aligned}\langle z\rangle(t) = \big(\langle z_0\rangle - \frac{\mu_{age}}{\lambda} - \mu_0\big)e^{\lambda t} \\ + \frac{\mu_{age}}{\lambda} + \mu_0 + \mu_{age}t \end{aligned} \tag{6.5}$$

where $\langle z_0\rangle$ is the initially observed mean at $t = 0$. The exponential factors dampen or aggregate the mean depending on the sign of λ. If $\lambda < 0$ the system is stable and once $|\lambda|t \gg 1$ a dynamical steady-state (ss) is reached,

$$\langle z\rangle_{ss}(t) = \frac{\mu_{age}}{\lambda} + \mu_0 + \mu_{age}t = \mu(t) - \frac{\mu_{age}}{|\lambda|}. \tag{6.6}$$

The steady-state is equivalent to the system forgetting its initial conditions. This steady-state behaviour can explain the drift observed by Sehl and Yates [176] (Appendix Section C.8.6). In the steady-state, the mean drifts at a constant rate,

$$\frac{d}{dt}\langle z\rangle_{ss}(t) = \mu_{age}, \tag{6.7}$$

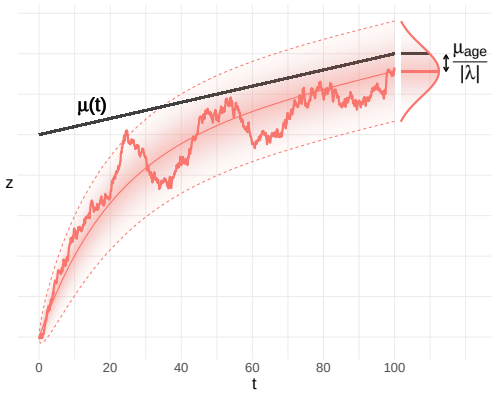

Figure 6.1: Simulation example of a stable system, with $\lambda < 0$. Initial conditions can differ from $\mu(t)$. A stable system is attracted to $\mu(t)$ (black line), but will be offset by $-\mu_{age}/|\lambda|$ in the steady-state. ODE solutions are super imposed for mean and variance (dotted lines are 95% interval). Fill density is proportional to probability density. Observing an ensemble at any time will yield Gaussian statistics.

but there is a constant lag of $\langle z-\mu\rangle_{ss} = \mu_{age}/\lambda$; Figure 6.1 illustrates. Only when $\mu_{age} = 0$ (no drift) is $\mu(t)$ the steady-state position. Outside of steady-state, the mean is displaced by

$$\langle z-\mu\rangle(t) = \langle z-\mu\rangle(t_0)e^{\lambda(t-t_0)} + \frac{\mu_{age}}{\lambda}(1-e^{\lambda(t-t_0)}) = \text{Memory} + \text{Drift} \tag{6.8}$$

for reference time t_0. The first term encodes the system's initial conditions, whereas the last term encodes long-time drifting behaviour. Systems near steady-state exhibit Memory $\ll$ Drift.

If $\lambda = 0$ the system is marginally stable and preserves its initial conditions. If $\lambda > 0$ the system is unstable and the initial conditions grow exponentially over time. In either case the steady-state is never reached. In contrast to the mean, for $\lambda < 0$ the variance eventually equilibrates, reaching a constant value. The variance is described by

$$\frac{d}{dt}\text{Var}(z) = 2\lambda\text{Var}(z) + \sigma^2 \tag{6.9}$$

where σ^2 is the noise strength. The general solution is given by

$$\text{Var}(z)(t) = \text{Var}(z_0)e^{2\lambda t} - \frac{\sigma^2}{2\lambda}(1-e^{2\lambda t}) \xrightarrow[\lambda<0,\ t\to\infty]{\text{steady state}} \frac{\sigma^2}{2|\lambda|}. \tag{6.10}$$

Approaching instability, with $\lambda \to 0$, the system accumulates noise.

6.3 Results

We analysed four datasets originating from three studies: two human and two mouse. Our analysis focused on the key properties of homeostasis: stability and equilibrium position. We used model selection to compare our model to the null hypothesis and to pick an optimal model form (Appendix Section C.5).

We observed that both an interaction network, $\boldsymbol{W}$, and an equilibrium term, $\vec{\mu}$, were needed to optimally predict future biomarker values. We saw no evidence of nonlinear terms in the dynamics. We found that fitting equation (6.3) using principal components (PCs) yielded equivalent performance to the model with full flexibility, equation (6.1), but was already in diagonal form. Hence for each dataset we analysed a set of decoupled, one-dimensional equations in z_j (with j sorted by stability, so that

$j = 1$ is the least stable in each study).

For covariates, we generally found non-significant improvements in prediction — though we kept them to improve interpretability (to reduce confounding). The exception was the age covariate, μ_{age}, which significantly improved the fit of the SLAM Het3 mice (SLAM C57/BL6 were almost significant). The presence of μ_{age} indicates allostasis in the form of a time-dependent homeostasis.

The interaction networks between variables can be represented by the respective weight matrices, e.g. Figure 6.2A. For ELSA we see expected relationships, e.g. total/HDL/LDL cholesterol, and non-dominant/dominant grip strength. ELSA also shows a block of like-variables including the FI-ADL, FI-IADL, self-reported health (srh), and gait speed, which could relate to frailty [146]. See Appendix Figure C.9 for networks from the other datasets.

The interactions between observed biomarkers prevents us from assessing stability directly. However, we can eigen-decompose the networks to yield an equivalent non-interacting network of natural variables. Each natural variable has a characteristic recovery rate, Figure 6.2B. All natural variables were stable, with $\lambda < 0$. Faster recovery rates indicate higher stability (resilience) [84]. It takes 3 timescales for the system mean to recover 95% of the way to equilibrium. For each mortality dataset the recovery timescale of its slowest natural variable was comparable to the organism's lifespan, ≈ 40 human-equivalent years; only the mental acuity dataset (Paquid) was faster (≈ 20 years). In all datasets the rates for the natural variables extended to higher and lower values than the diagonal elements of the observed biomarkers (compare solid to dotted lines) — indicating that network interactions play an important role in recovery dynamics.

We summarize homeostasis in Figure 6.3A, using the population means. If variables are in homeostasis then the mean should be close to μ_n, where the scale is determined by the native dispersion. Each dataset had most natural variables near 0 with a few outliers, such as z_1 for all datasets. (In contrast, the majority of observed biomarkers had large deviations from equilibrium — see Appendix Figure C.10). We characterized the natural variable dynamics using equation (6.8) in Figure 6.3B. Excluding ELSA, most data points were in a steady-state, as indicated by their small memory term (relative to drift). The steady-state mean includes a drift caused by

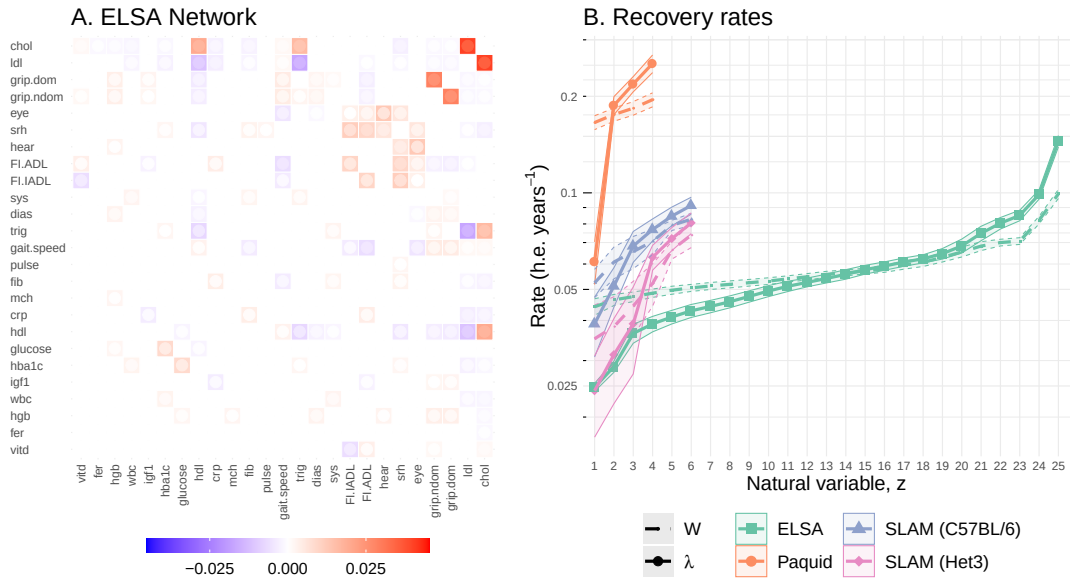

Figure 6.2: **A.** ELSA interaction network. Tile colour indicates interaction strength (saturation) and direction (colour) of the interaction from the y-axis variable to the x-axis variable. Inner dot colour indicates the limit of the 95% confidence interval (CI) closest to zero (more visible point indicates lower significance). Non-significant interactions have been whited-out. Diagonal has been suppressed for visualization (see dotted lines in B). The matrix is real and symmetric because the data were diagonalized by an orthogonal matrix (PCA). Variables are sorted by diagonal strength in both A. and B. (increasing rate). **B.** Recovery rates in human-equivalent (h.e.) years i.e. negative eigenvalues ($-\lambda$). The smallest recovery rates determine system stability [106]. A recovery rate of 0.025 implies $1 - e^{-1} = 63\%$ recovery after $-\lambda^{-1} = 40$ years (95% recovery after 120 years). The survival data all have similar minimum rates near 0.025, whereas the dementia data was faster (Paquid). The dotted lines are network diagonals ($-W_{jj}$); the solid lines are rates ($-\lambda_j$).

μ_{age}. Across variables, the deviations from equilibrium observed in Figure 6.3A, $\langle z_n - \mu_n \rangle$, were very strongly correlated with μ_{age}, with correlation coefficients: -0.988 ($p = 2 \cdot 10^{-4}$, SLAM BL/6), -0.947 ($p = 10^{-3}$, SLAM Het3), -0.989 ($p = 0.01$, Paquid), and -0.302 ($p = 0.14$, ELSA). This is consistent with equation (6.6), and supports our use of an allostatic model with equilibrium drifts given by μ_{age}. The smaller correlations observed with the ELSA dataset are consistent with the strong memory effect seen in Figure 6.3B — violating the steady-state assumption of equation (6.6). ELSA may have failed to reach steady-state due to the limited followup period, which was the shortest of all datasets by a factor of 2, or could indicate the confounding effects of medical interventions, which are not relevant for the other

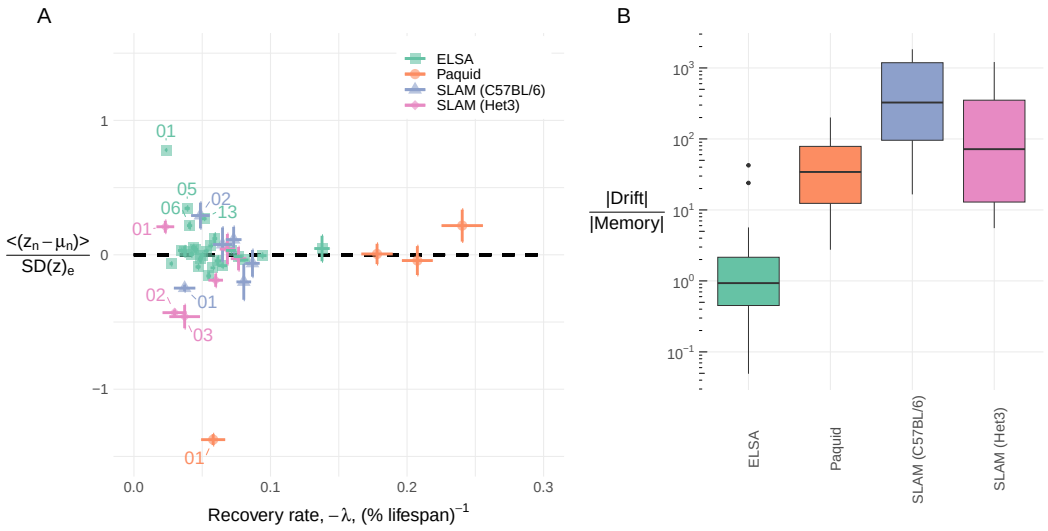

Figure 6.3: **A.** Position relative to equilibrium vs recovery rate. Most natural variables were homeostatic (near equilibrium at 0). Some (labeled) variables were observed to be far from equilibrium; variables are labelled by rank e.g. $01 \equiv z_{01}$ has the fastest recovery (furthest left). **B.** Characterization of natural variable deviations from equilibrium using equation (6.8). Observe that ELSA is the only dataset where memory may dominate the system behaviour (ratio $\lesssim 1 = 10^0$), indicating that the followup period may have been too short to reach a steady-state. In both figures only mouse (SLAM) data points over age 80 weeks were used since biomarkers had a u-shaped curve over the lifespan [135].

datasets.

Most natural variables have small drift and are effectively homeostatic — with only a few strongly drifting allostatic natural variables. The steady-state drift rate of natural variables, μ_{age}, was correlated with the survival risk for each dimension: Figure 6.4A. The correlations were typically strong: -0.958 ($p = 0.002$, SLAM BL/6), -0.713 ($p = 0.1$ SLAM Het3), -0.987 ($p = 0.01$, Paquid), and -0.534 ($p = 0.006$ ELSA); overall: -0.742 ($p = 3 \cdot 10^{-8}$). The correlation was weakest for ELSA, which had not reached steady-state. The Cox proportional hazards coefficients, conditioned on age and sex, showed a similarly strong correlation with μ_{age}, 0.70 ($p = 10^{-7}$, all data) (Appendix Figure C.14). Furthermore, we see that the drift direction, sign(μ_{age}), is the same as the risk direction ($p = 0.0003$, Fisher test). Hence, not only does homeostasis drift with age, the direction of the drift is *towards ill-health*. The primary risk directions were z_1 for ELSA and Paquid and z_2 for SLAM. Interestingly, z_2 of the Het3 mice is nearly identical to z_1 of the C57BL/6 mice in terms of covariates

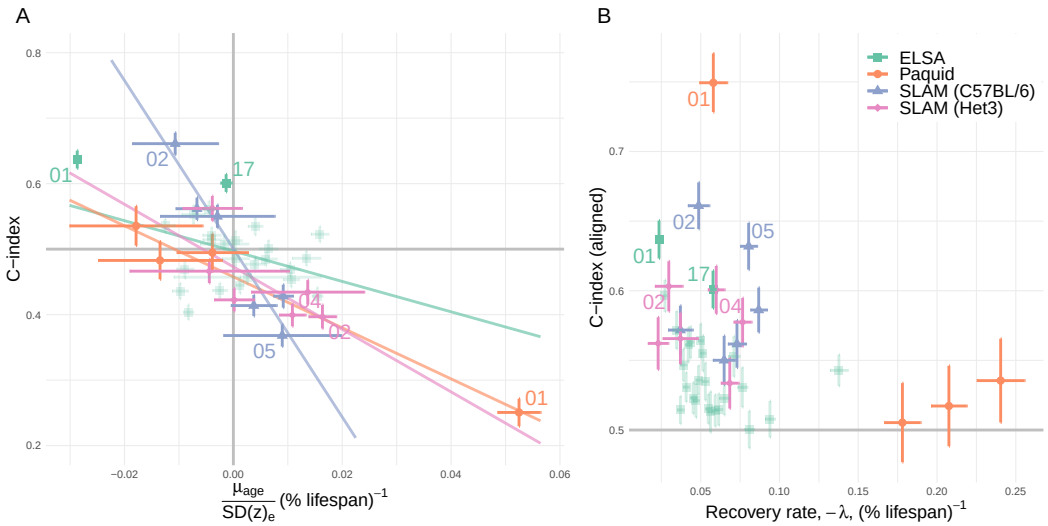

Figure 6.4: Survival effects. **A.** Allostasis drifts towards the risk direction, "mallostasis". The relationship appears to be linear (lines), with strong correlations: -0.96 (SLAM BL/6), -0.71 (SLAM Het3), -0.99 (Paquid), and -0.53 (ELSA). The equilibrium dispersion provides a native scale for each variable. High risk natural variables for each dataset have been labelled by eigenvalue rank (e.g. $z_1 \equiv 01$ has the smallest eigenvalue, $z_2 \equiv 02$ the second smallest, etc). **B.** Recovery rate (-eigenvalue), $-\lambda$, has an ambiguous relationship with survival. Smaller eigenvalues appear to be important survival dimensions (e.g. 01 for ELSA and Paquid), but the overall correlation is weak ($\rho = -0.254$, $p = 0.1$). The C-index measures the relative risk for pairs of individuals based on the value of z_j (C-index of 0.5 indicates no risk; C-index larger than 0.5 means small values are bad).

and survival effect — hence z_1 of the C57BL/6 is also likely a key risk direction (Appendix Figures C.11 and C.18). Regardless, z_1 or z_2 exhibited the strongest survival effect for their each dataset (Figure 6.4A). These variables also both had small eigenvalues (z_1 is rank 1 and z_2 is rank 2). However, this relationship between survival and eigenvalue magnitude does not appear to generalize, see Figure 6.4B.

As an illustration of the utility of the correlation between survival and μ_{age}, we consider a simple summary health measure. The Cox proportional hazards model assumes the hazard scales as $\exp(\vec{\beta}^T \vec{z})$, where the jth coefficient, β_j, is the log–hazard ratio per unit increase of z_j. As mentioned in the previous paragraph, $\vec{\beta} \sim \vec{\mu}_{age}$ were correlated; the relative hazard can therefore be approximated by $\exp(\vec{\mu}_{age}^T \vec{z})$. Indeed, we observed that $b \equiv \vec{\mu}_{age}^T \vec{z}$ is an excellent predictor of survival, see Figure 6.5A and Appendix Figure C.15.

The natural variables with large $|\mu_{age}|$ will eventually experience the largest drift,

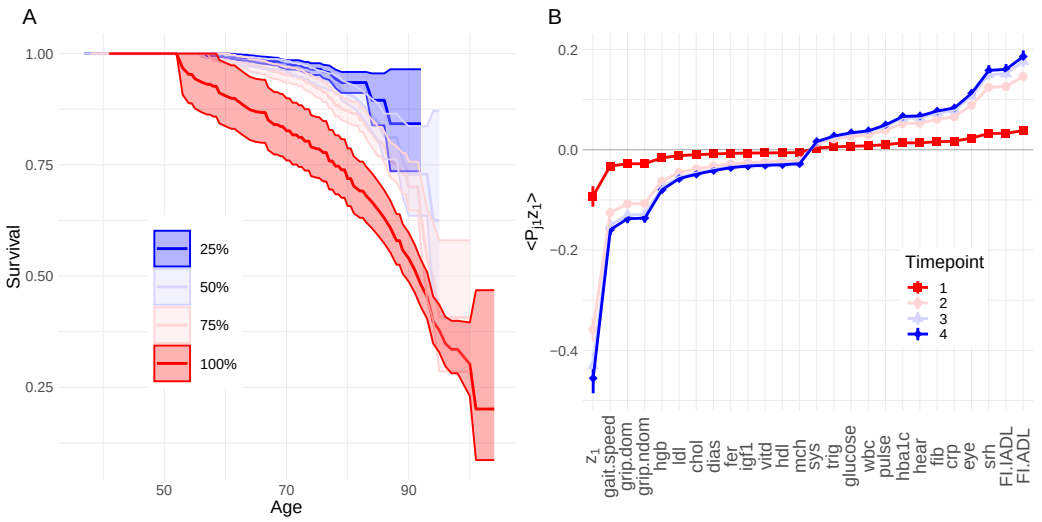

Figure 6.5: **A.** Composite health measure of survival $b \equiv (\vec{\mu}_{age}^{T}\vec{z})$, stratified by quartile (ELSA). Separation is excellent, indicating a strong survival predictor. Fill is 95% confidence interval. See Appendix Figure C.15 for the other datasets. **B.** Natural variables can drive changes in observable biomarkers. The z_1 mean is accumulating in the negative direction. This accumulation is mapped into observable variables with $\langle P_{j1}z_1\rangle$ for indicated timepoints each separated by approximately 4 years. The drift direction is overwhelmingly unhealthy: increased disability measures (srh, eye, hear, FI.ADL and FI.IADL — high is bad), decreased physical ability scores (gait and grip), increased inflammation (crp), increased glucose, etc. The effect of the drift is concentrated in z_1 but dilute across its covariates, which could make the effect of unhealthy z_1 subclinical in the observed biomarkers. All variables are on standardized scale. Similar effects were observed for the other datasets (Appendix Figure C.13).

according to equation (6.7). z_1 in Figure 6.5B is an example of such an accumulating variable. The other variables with large $|\mu_{age}|$ experienced a similar accumulation (Appendix Figure C.13). For an orthogonal transformation such as $\boldsymbol{P}^{-1}$, the sum of the variance and squared mean is conserved (Parseval's theorem). Natural variables with large means and variances will therefore disproportionately affect the means and variances of observed biomarkers. The effect is demonstrated for ELSA in Figure 6.5B. As the dominant natural variable drifts it influences observable biomarkers to drift as well.

Slower recovery rates (eigenvalues) take longer to forget perturbations, causing the associated natural variables to accumulate variance due to noise. Recall that

the slowest recovery rates were on the order of a lifespan (Figure 6.2B). The Pearson correlations between the estimated variance and rate (-eigenvalue) were strong: -0.852 ($p = 0.03$, SLAM BL/6), -0.802 ($p = 0.05$ SLAM Het3), -0.998 ($p = 0.002$, Paquid), and -0.764 ($p = 9 \cdot 10^{-6}$ ELSA) (log-log scale; see Appendix Figure C.16). Hence the variances we observe at old ages will be dominated by the variables with the smallest eigenvalues, λ (e.g. z_1 and z_2). As we have seen before, these variables are often — but not always — strongly associated with adverse effects, depending on the drift rate μ_{age}. This suggests that most of the age-related changes to health were concentrated in a few z_k which drive both biomarker drift (mean) and dispersion (variance). Growing variance along these dimensions may capture individual accumulation of stochastic damage, such as genetic damage or disease.

6.4 Discussion

We fit a homeostasis model of equilibrium and stability to four longitudinal aging datasets (two mouse and two human) using generic health biomarkers. Our model is lightweight, can be estimated using standard statistical algorithms, and is sufficient to capture essential information about the aging process. Health biomarkers have an equilibrium position, $\vec{\mu}$. Their corresponding stochastic term, which has covariance $\boldsymbol{\Sigma}$, represents random stresses that drive individuals away from equilibrium; as well as residual effects such as individual variability and nonlinearities. An interaction network, $\boldsymbol{W}$, pulls individuals towards equilibrium either through recovery (diagonal terms) or by one variable compensating for another (off-diagonal terms). By eigen-decomposing $\boldsymbol{W}$ we transformed the dataset into non-interacting, natural variables — which are linear combinations of the input biomarkers. This increases interpretability and simplifies analysis. The stability of the system is described by the recovery rates of the natural variables — which are the corresponding eigenvalues with flipped signs, $-\lambda$.

We modelled homeostasis as stability around an equilibrium mean value. Stability can only be assessed using the natural variables because the original biomarkers have interactions between them. Homeostasis was violated by some natural variables (Figure 6.3A). Although most natural variables had average values near the homeostatic equilibrium — indicative of homeostasis — several were far away. We determined that

this latter group were out of homeostasis because they were chasing drifting equilibrium positions from behind. This equilibrium drift with age represents allostasis, i.e. a changeable equilibrium position.

Allostatic variables accumulated over the course of the study period as they chased after the drifting equilibrium, systematically increasing or decreasing. This was facilitated by an age-dependent equilibrium position, governed by μ_{age}, and was typically accompanied by a small eigenvalue. The gap between the population and allostatic equilibrium position is governed by μ_{age}/λ, so a small λ enables a large gap such that the entire population drifts coherently towards the moving equilibrium — causing population-level accumulation. This makes the linear drift term, μ_{age}, the primary culprit for causing biomarkers to drift with age. Presumably μ_{age} arises from either (a) the effects of unknown biomarkers/mechanisms not included in the model (i.e. $\mu_{age}t \approx \sum_{k\neq j} W_{jk}\langle y_{ikn} - \mu_{ikn}\rangle$ in equation (6.2) for a set of unknown y_{ikn}), or (b) asymmetric stressors, which cannot be captured by our symmetric stochastic term (e.g. there is no such thing as negative damage so health deficits skew positive [123]).

The transformation to natural variables effectively compressed the drifting (accumulating) mean of many variables into a small number of natural variables. The natural variables can be thought of as the underlying cause of the observed biomarker drift (Figure 6.5B). In this manner, the widely observed age-related decline in biomarkers [176] are governed by a few natural variables — which are not directly observed. The effect is spread out by the transformation, potentially hiding the observed biomarker decline below diagnostic thresholds. This may be a redundancy mechanism: the network permits the biological system to spread out the age-related decline to keep biomarkers in healthy ranges for longer. The trade-off may be that many biomarkers would reach unhealthy ranges concurrently, leading to multisystem dysfunction. For example, the effects of chronic kidney disease are mild and non-specific until the patient nears kidney failure — at which point multisystem failure is imminent, typically leading to death via cardiovascular disease [126]. This tradeoff assumes that diagnostic thresholds represent critical values beyond which deterioration of a biological system accelerates. Univariate dynamical modelling of senescent cell count in mice [93] and *E. coli* membrane integrity [212] supports the existence of such critical values, where repair mechanisms saturate and decline accelerates. From this perspective, an

individual's robustness would depend on their buffer space available to absorb new insults, which could be quantified by the natural variable scores together with the stressor effect strength which should be proportional to the noise σ.

Consistent with this perspective [93], the allostatic drift rate, μ_{age}, strongly correlated with the mortality/dementia risk associated with each natural variable (Figure 6.4A and Appendix Figure C.14). Since μ_{age} is the steady-state drift rate of the mean, the steady-state behaviour is continually worsening health due to the drifting mean. Prior work on operationalizing allostasis has neglected the existence of a preferred risk direction, instead using the absolute distance from allostasis as a mortality factor [213, 34, 113], irrespective of whether biomarkers are high or low. In contrast, our results indicate that numerous natural variables do, in fact, have preferred risk directions. Aging researchers should be aware of this symmetry breaking. This means that the adaptive changes due to allostasis at best mitigate declining health and, at worst, lead to a further decline in health. We refer to this phenomenon as "mallostasis": the tendency of an aging biological system towards an ever-worsening equilibrium position. We have used this phenomenon both to identify important survival variables and to generate a novel composite health measure.

Our key quantitative results coincide with three key qualitative predictions made by allostatic load theory: (i) a shifting equilibrium position for biomarkers indicative of adaptive changes (allostasis, Figure 6.3A), (ii) the shift is associated with adverse outcomes (mallostasis, Figure 6.4A), and (iii) the shift is subclinical due to compensating mechanisms between biomarkers (transformation, Figure 6.5B) [89]. This is compelling evidence that allostasis is a generic aging phenomenon, rather than being specific to neuroendocrinology. Our proposed composite health measure is therefore a novel estimator of allostatic load. In contrast to conventional estimators [89], we were able to estimate allostatic drift directly as μ_{age}. Our results rely on using natural variables, which are canonical coordinates that greatly simplify analysis.

Allostatic load is believed to arise from the long-term costs of short-term protection against stressors [118], making it an example of antagonistic pleiotropy [63]. Alternatively, long-term costs could reflect imperfect repair. Regardless, long-term costs that accumulate in a given direction would lead to the allostatic drift which we have observed and characterized. Furthermore, we observed slow dynamical rates for

the dominant mortality-risk natural variables (Figure 6.2B). Accordingly, the dynamical timescale of these effects are comparable to the organismal lifetime — consistent with long-term costs.

Interestingly, we did not observe any instabilities or nonlinearities. We had expected that "allostatic overload" — the final state of allostatic load theory [89] — would operationalize as an instability. However, instabilities, and their associated exponential growth, are rare among health biomarkers [176]; although they are observed in summary measures of health such as the FI (frailty index) [123]. Other unstable, FI-like variables can be extracted from generic biomarkers using nonlinear techniques such as a deep neural network [8], diagnostic thresholds [19], or quantile-based preprocessing [181, 182]. Since we did not see evidence of instabilities or other nonlinearities in the natural variables, the nonlinear embedding or discretization should be considered as a possible cause for observed FI-like instabilities. It may be that biological systems naturally suppress nonlinear effects of aging — obscuring the effects — or, conversely, that aging is primarily a linear phenomenon that slowly pushes individuals towards nonlinear tolerance thresholds for dysfunction/damage, e.g. saturation of repair [93, 212] and/or emergence of chronic disease. A non-trivial issue is that exponential growth often appears linear, for example the FI in mice and younger humans ($\lesssim$ 85 years old) [157]. Nonlinear effects in biomarker dynamics may require special populations, such as the ill or exceptionally old, to be observed.

The key model variables, z_1 and z_2, dominate the aging process. These natural variables with smaller λ carried the majority of the variance and become the dominant principal components in the steady-state model (Appendix Figure C.16 and equation (C.49), respectively). Applying Parseval's theorem, these variables will control the variance of directly observed biomarkers. Since they also dominate the means via allostatic drift, they will determine the aging phenotype that we observe. Both effects get stronger with age. This means that the empirically observed age-related changes in the mean and variance of biomarkers will be predominantly caused by only a few key natural variables. Hence the nearly-universal linear decline in health biomarkers observed by Sehl and Yates [176] may simply be a few declining natural variables spread across the observed biomarkers (Appendix Section C.8.6). Furthermore, this implies that a single dimensional decline can drive many observed biomarkers, which

is the foundational assumption of "biological age" estimators [100, 90]. Our results provide much needed support for such low dimensional representations of aging — which should become increasingly accurate with advancing age since the means of the key natural variables grow fastest, and their variances grow largest.

The natural variables, z, should be good choices for targeting and monitoring interventions. They are prospective biomarkers with the convenient property that if you can intervene on one it will not affect the others. In contrast, we know from the network of interactions that intervening on any single biomarker is likely to affect many other biomarkers. In the steady-state, the mallostatic drift rate, controlled by μ_{age}, is a proxy for the hazard and therefore identifies the most important targets of intervention. The coefficients of the transformation, $\boldsymbol{P}$, provides both hints at what mechanisms each z_j is capturing as well as a map for which biomarkers will be affected by interventions on z_j. For example, z_1 of ELSA shares many features with frailty: strong age dependence, large effect in gait, weakness (grip strength), disability and self-reported health, and large survival effect [146]. z_1 is thus a prospective biomarker of frailty and can be used both to monitor an individual's frailty and to engineer interventions[2]. The strong signals we see in Figure 6.5B for gait, grip strength and activities of daily living are hints that loss of physical fitness is one mechanism by which frailty proceeds and therefore one mechanism by which we can intervene, consistent with a meta-analysis which has shown that physical activity can reduce frailty in humans [132]. Remarkably, we observed other prospective targets in addition to z_1 of ELSA. Given an organism and set of biomarkers, each z_j with substantial drift should be considered a prospective intervention target, with the faster drifting being the most important.

We note a few limitations of our study. We assumed linear, time-invariant interactions through the network, $\boldsymbol{W}$ — following previous work that suggested that interactions are linear and time-invariant [52] (as are the principal components [146]). The networks we extracted were symmetric and hence acausal[3] due to our use of PCA

[2]The key advantage of z_1 over the frailty index is that z_1 uses continuous input variables, giving it a larger dynamic range. The frailty index is zero until the emergence of signs, symptoms, impairment or morbidity, making it insensitive at young ages ($\lesssim$ 40 years old). z_1 is based on continuous lab values and hence does not necessarily suffer from this limitation.

[3]Upon revision, this is inaccurate. Symmetric networks still satisfy Granger causality, but there is an ambiguity as to whether the undirected links approximate directed ones.

as a preprocessing step, although we did estimate more general networks and found they performed no better. This could be a consequence of the data which were entirely observational, obfuscating causality. Understanding interventions using our results is similarly subject to the caveat that biomarkers behave the same whether they are observed or intervened upon [39]. This seems plausible since observational studies include everyday interventions such as disease, medicine, or lifestyle changes. Finally, our model is at the population-level and hence we cannot resolve homeostatic changes at the individual level.

We see exciting opportunities for future work. Our observation that principal components could be effectively used as independent variables suggests that more complex statistical models could also be applied. For example, individual-level model parameter estimates via mixed-effects modelling would help to determine whether individual health changes are gradual or sudden (or possibly critical [166]). Changes in our model parameters due to age, chronic or acute illness, or medical interventions is particularly interesting, but will require specialized datasets to assess. Fortunately, small datasets are tractable with our linearized model. The generic nature of the model and its ability to find accumulating natural variables could also be applied to other biological or temporal scales. Others have postulated that damage aggregates due to dysfunction in regulatory systems or other intermediate scales [114], which could be tested. Composite health measures, including biological age [90], are also interesting to explore using our approach. Applying our approach using multiple biological ages as biomarkers [110] will naturally extract salient information regarding stability and mallostasis, as well as a smaller set of essential natural variables. New datasets will open up new opportunities for this analysis pipeline. It is interesting to consider leveraging the effect of the natural variables to intervene and observe in clever ways. For example, z_1 appears to be a biomarker of frailty, which affects both mental and physical health [132], hence we could potentially intervene based on a physical mechanism but monitor using cognitive changes.

We have developed and applied a lightweight network model that includes the salient features of homeostasis: equilibrium values and recovery rates. Equilibrium values are allowed to drift, to accommodate allostatic changes. Across datasets and species we consistently observed that the linear decline of biomarkers with age was

governed by a small set of accumulating natural aging variables. This accumulation can be described as mallostasis: homeostatic dysfunction and associated declining health. These variables appear to be important measures of age-related decline, including health and mortality. Their effects are spread out by a network of interactions, driving drift in the observed biomarkers, and potentially diluting and obfuscating the effects of age. We find that generic biomarkers spontaneously move towards an equilibrium position which is itself continuously drifting towards ill-health. Mallostasis is a generic feature of the aging process.

6.5 Methods

6.5.1 Materials

We used 4 longitudinal datasets originating from 3 studies (organism, primary outcome): Paquid (human, dementia) [150], SLAM (mouse, death) [135, 136] and ELSA (human, death) [10]. We directly modelled biomarkers, $\vec{y}$, and included covariates, $\vec{x}$, in the homeostatic term, $\vec{\mu}$, using equation (6.1).

The Paquid dataset is a random subset of 500 humans (212 males and 288 females) from the Paquid prospective cohort study, enriched in dementia prevalence [150]. Age range: 66-95 years-old. Individuals were measured on average every 3.2 years for a maximum of 9 timepoints. We modelled four ordinal variables, including three measures of mental acuity: mini-mental state examination (MMSE), Benton visual retention test (BVRT) and Isaacs set test (IST), along with a self-reported depression score (CESD). We considered for covariates: sex, age and education level (completed primary vs not).

The Study of Longitudinal Aging in Mice (SLAM) includes two datasets, one for each mouse strain. Both include body composition measures and glucose serum at 12 week intervals starting at 7 weeks of age and continuing for the lifespan of each mouse [136]. Body composition and serum measurements were staggered and had to be imputed. Covariates included age and sex. We dropped 538/66138 measurements that were recorded after death, ostensibly these were coding errors. After preprocessing, the first dataset included 608 C57BL/6 mice (303 male and 305 female) measured on average every 6.2 weeks for a maximum of 20 timepoints (every

4.9 human-equivalent years). C57BL/6 mice are genetically similar (inbred) and prone to lymphoma and metabolic dysfunction [121]. The second included 611 Het3 mice (304 male and 307 female) measured on average every 4.2 weeks for a maximum of 27 timepoints (every 3.6 human-equivalent years). Het3 mice are a genetically heterogeneous cross of four inbred mice (including C57BL/6) [121]. We converted to human-equivalent years using the ratio of median survival times of each strain to ELSA. Full details of the study are described elsewhere [136, 135].

The English Longitudinal Study of Ageing (ELSA) is a representative sample of English people aged 50 and over (with some younger) [10]. We used physical functioning questionnaire data and blood tests for 9330 humans (4063 males and 5267 females), reported at 4 timepoints, each separated by approximately 4 years. Our choice of 25 variables includes frailty measures, cardiometabolic biomarkers, and immune biomarkers (Appendix Table C.1). We considered waves 2, 4, 6 and 8, since only these contained the full suite of biomarkers. Covariates included age and sex. We considered only individuals whom were present both in wave 2 and in subsequent waves, thus excluding new recruits. Despite the large number of individuals, ELSA appeared to have the worst quality data due to high individual heterogeneity and low number of timepoints.

6.5.2 Data handling

All missing data were imputed. Dead individuals were also imputed, as it reduced bias due to mortality in simulated data (Appendix Section C.4). We compared several imputation strategies, including carry forward/back, multivariate imputation using chained equations (MICE) [196], and using our model to impute the model mean. Ultimately, we used carry forward/back followed by the model mean, except for ELSA which used each individual's mean biomarker value followed by the population multivariate normal mean then model mean. See Appendix Section C.4 for details.

Estimation of our model, equation (6.1), used Appendix Algorithm C.1, which iteratively ($\times 5$) applied the maximum likelihood estimator:

$$\begin{aligned}\hat{\boldsymbol{\Lambda}} = &\langle|\Delta t_{in+1}|\vec{y}_{in}\vec{x}_{in}^T\rangle_{i,n}\left(\langle|\Delta t_{in+1}|\vec{x}_{in}\vec{x}_{in}^T\rangle_{i,n}\right)^{-1} \\ &- \boldsymbol{W}^{-1}\langle\text{sign}(\Delta t_{in+1})(\vec{y}_{in+1}-\vec{y}_{in})\vec{x}_{in}^T\rangle_{i,n}\left(\langle|\Delta t_{in+1}|\vec{x}_{in}\vec{x}_{in}^T\rangle_{i,n}\right)^{-1}\end{aligned} \tag{6.11}$$

for $\boldsymbol{\Lambda}$ (which includes μ_0 through $x_0 = 1$), and

$$\hat{\boldsymbol{W}} = \langle \text{sign}(\Delta t_{in+1})(\vec{y}_{in+1} - \vec{y}_{in})(\vec{y}_{in} - \vec{\mu}_{in})^T \rangle_{i,n} \big(\langle |\Delta t_{in+1}| (\vec{y}_{in} - \vec{\mu}_{in})(\vec{y}_{in} - \vec{\mu}_{in})^T \rangle_{i,n} \big)^{-1} \tag{6.12}$$

for $\boldsymbol{W}$, where the expectation values are to be taken over times, n, and individuals, i. For the diagonal models we instead used weighted linear regression. Missing values were imputed with the model prediction after each iteration (except ELSA). Estimators are described and validated in Appendix Sections C.6 and C.7, respectively. We used a time-dependent Cox model to assess survival. We assumed stepwise constant covariates via start-stop formatting [127]. All correlations are Pearson. All errorbars are standard errors unless stated otherwise.

6.5.3 Model assessment

We simultaneously estimated both parameter uncertainty and model performance using the standard deviation from 100× repeat bootstrap resampling. We compared model performance using the root-mean squared error (RMSE) and mean absolute error (MAE). Both were estimated using out-of-sample bootstrap [74]. In validation tests we found that a simple 632 estimator i.e., $\text{RMSE}_{632} \equiv 0.632 \cdot \text{RMSE}_{\text{test}} + 0.368 \cdot \text{RMSE}_{\text{train}}$, provided a good estimate for the true values of both performance metrics (Appendix Figure C.7). 0.632 is the expected fraction of unique individuals in each bootstrap [74].

Acknowledgements

ELSA is funded by the National Institute on Aging (R01AG017644), and by UK Government Departments coordinated by the National Institute for Health and Care Research (NIHR). A.R. thanks the Natural Sciences and Engineering Research Council of Canada (NSERC) for operating Grant RGPIN-2019-05888.

Data availability

All data used are publicly available. The SLAM datasets are available from a previous publication [135]. Paquid is available from a software package [150]. ELSA [10]

is available from the UK Data Service https://ukdataservice.ac.uk/. Software for fitting and simulating our model is available at https://github.com/GlenPr/stochastic_finite_model.

Chapter 7

Application of the SF model to multiple biological ages provides a framework for building and testing network theories of aging

We named our model from Chapter 6 the Stochastic Finite (SF) difference model. Using the SF model we can test causal theories of aging using collections of biological ages. In this chapter I present some of the salient results from a publication to appear in mid to late 2024 [148].

A collection of biological ages, one per biological system, could be used to understand interacting systems (e.g. [175]). If we had a biological age for each theoretical factor of aging, we could directly test the proposed interactions made in contemporary network theories of aging (e.g. [115, 94]). In support of this proposal, biological ages are becoming increasingly available, and are improving in their specificity and sensitivity to the aging process [163]. Furthermore, multivariate analyses of multiple biological ages suggest that they contain non-redundant aging information [110, 87].

Unfortunately, biological ages are unsatisfactory biomarkers of aging. For example, a landmark 2023 paper by Gladyshev's group confirmed that several DNA methylation–based biological ages are sensitive to anti-aging intervention as well as adverse stressor events — but the effects were small and not consistently discernible from noise, even under lab conditions [144]. It is therefore prudent to consider two possible futures for biological ages: a future wherein aging can be measured directly using specific biomarkers, or a future wherein it is conclusively shown that aging cannot be directly measured and an additional layer of inference is always necessary between the raw data and its implications for underlying the state of aging. In a proof-of-concept paper, I show that the SF model and analysis pipeline from Chapter 6 can take advantage of either possible futures by generating an interaction network then decomposing it into natural variables — which are meta–biological ages.

We modelled multivariate, longitudinal biological age data from the Swedish

Adoption/Twins Study of Aging (SATSA) [110]. The dataset includes 8 biological ages of varying biological scales together with the FI, which I used as an outcome measure of health (I also considered including the FI in the network during sensitivity analysis). This includes first generation DNA methylation ages: Horvath and Hannum, as well as second generation: GrimAge and PhenoAge (first generation are regression models that predict chronological age, second generation predict survival risk converted to the scale of age [163, 148]). Telomere length is also included — telomere erosion is associated with the aging process, primarily cellular senescence (arrest of cellular division) [164]. At the systems-level, PhysioAge is included as a generic biological age using the Klemera-Doubal method [100] and a set of biomarkers primarily related to cardiometabolic health (which differed for males and females). Also at the systems-level was the first principal component from a battery of cognitive tests, which I call Cognition. Finally, the functional aging index (FAI) and frailty index (FI) were included as functional-level biological ages. The FAI is based on sensory, grip, pulmonary and gait metrics, averaged together, and is known to correlate with the FI [54]. For most biological ages, higher values indicate worse health but for Telomere and Cognition it is the opposite. Table 7.1 summarizes.

My approach mirrors that of Chapter 6, where I first estimate a network and then decompose it into natural variables. The network here is of particular interest and I will compare it to the well-known Hallmark [114, 115] and Pillar [94] theories of aging. Then I will investigate each of the natural variables as a possible underlying driver of the dynamical changes observed in the network. This will identify prospective meta-biological ages.

In Figure 7.1 I present the interaction network derived from the SF model. PhysioAge and GrimAge occupied central locations in the network with many outgoing connections. The two have feedbacks between each other, suggesting that age-related dysfunction enters near PhysioAge and GrimAge, then propagates outwards with reverberations between GrimAge and PhysioAge which persistently push dysfunction into other biological systems. This can be seen in the correlation matrix: chronological age Spearman correlates the strongest with PhysioAge at 0.90, then GrimAge at 0.79, and then Hannum at 0.73, ostensibly the correlation with age drops with increasing proximity from the underlying cause.

Table 7.1: Summary of biological ages used

Biological age	Summary
Frailty index (FI)[1]	average number of health deficits[2]
Functional aging index (FAI)	sensory, pulmonary, grip and gait[3]
Cognition	PC1 from cognitive tests (down is bad)
Physiological Age (PhysioAge)	cardiometabolic Klemera-Doubal[4]
GrimAge	epigenetic mortality risk[5]
PhenoAge	epigenetic mortality risk[6]
Hannum	epigenetic chronological age regression
Horvath	epigenetic chronological age regression
Telomere	telomere length[7](down is bad)

[1] Held aside as outcome measure of health.

[2] Score from 0 (none) to 1 (full): disability, disease, and self-reported health.

[3] Average score from: self-reported hearing/vision, grip strength, lung strength and gait speed.

[4] Biomarkers: male: body mass index, waist-to-height ratio, weight, systolic blood pressure, diastolic blood pressure, hemoglobin, serum glucose (log), cholesterol, and apolipoprotein B. Female: hip circumference, waist circumference systolic blood pressure, serum glucose (log), and triglycerides (log). Transformed by PCA then the Klemera-Doubal method was used to estimate the latent biological age.

[5] Input is CpGs (epigenetic data) trained to emulate: adrenomedullin, beta-2-microglobulim, cystatin C, GDF-15, leptin, PAI-1, and tissue inhibitor metalloproteinases 1, and smoking pack-years. GrimAge is mortality risk scaled to mean/sd of chronological age.

[6] Input is CpGs (epigenetic data) trained to predict time-to-death (Gompertz) including a proportional hazard from: albumin, creatinine, serum glucose, C-reactive protein, lymphocytes (%), mean red cell volume, red cell distribution width, alkaline phosphatase, white blood cell count, and age. PhenoAge is inverted 10-year multivariate mortality risk (solved for age).

[7] Normalized by standard deviation.

The popular Hallmark theory of aging speculates that dysregulated nutrient sensing may play an important role as an antagonist of primary damage [114], which would be consistent with our results since PhysioAge has a strong cardiometabolic component. Their theory speculates that epigenetic alterations, genomic instability and loss of proteiostasis are all primary sources of damage, and hence a feedback from GrimAge — which could capture any of these three — and PhysioAge would

be consistent with accumulating damage (a theoretical driver of aging [184, 194]). The Hallmark theory also considers telomere attrition to be a hallmark of aging and a primary source of damage that drives aging. To the contrary, our results suggest that Telomere was a spectator driven by changes to GrimAge and, surprisingly, FAI (suggesting damage back-propagating from the functional scale, the exact opposite of the Hallmark theory proposal for telomere length). Another popular theory, the Pillar theory is agnostic to specific network connections only that 7 key pillars are interconnected: metabolism, macromolecular damage, epigenetics, inflammation, adaptation to stress, proteostasis, and stem cells. While this acknowledges the central importance of PhysioAge (metabolism) and GrimAge (epigenetics), the Pillar theory is non-specific regarding network topology. For true negatives, both theories are consistent with loss of physiological function as being peripheral to the aging process, as captured by FAI. Together our results support the notion that both the Hallmark and Pillar theories of aging have identified important contributors to the aging process, but their proposed connectivity between those contributors is lacking in accuracy and specificity.

Decomposing the network, I observed that the first one or two natural variables were able to capture the dominant effects of aging, particularly past age 80. This is readily seen in the equilibrium behaviour of the natural variables. For both pragmatic and conceptual reasons, we restricted μ to be a sex-dependent constant, meaning that each natural variable will eventually reach an equilibrium mean ($\mu_{age} = 0$ in Box 1). For the raw biological ages, Figure 7.2, the equilibrium positions were well outside of normal human lifespan ensuring that each biological age will increase continuously over time (in the case of Telomere and Cognition they will (correctly) decrease over time). In contrast, most of the natural variables, Figure 7.3, reached their equilibrium before age 80, with only z_1 and z_2 continuing to increase beyond that age. This indicates that the age-related changes become lower dimensional over time and z_1 becomes increasingly dominant in terms of the mean, and likely also the variance, which grows until the equilibrium position is reached (the variance doesn't have to grow with age, but we expect it to e.g. the FI increases in mean and variance with age [161]). When we compared the z to the FI — a proxy for health — we found that z_1 contained most of the information and the strongest correlation, and that this

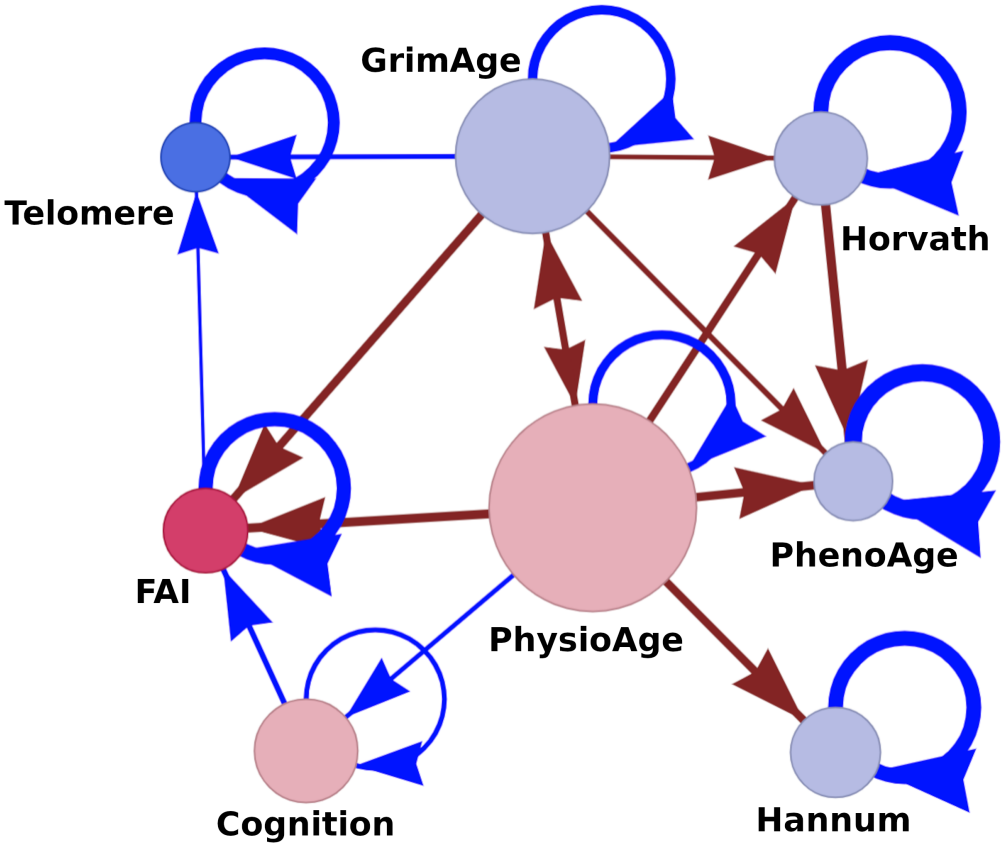

Figure 7.1: Estimated interaction network between biological ages (nodes) using the SF model. Node colour reflects biological scale: blue is genetic (telomere length), blue-gray are epigenetic, pink is "system" level (cardiometabolic: PhysioAge, or cognitive: Cognition), and red is the whole organism's functional ability (e.g. gait speed). For links, red links are positive associations, blue links are negative associations. PhysioAge formed a central node, with GrimAge forming an important secondary node. We inferred that age-related changes originate close to PhysioAge and/or GrimAge then propagated outwards, driving the peripheral biological ages. In this manner one dysfunctional sub-system (metabolism) can propagate dysfunction into other sub-systems, driving them awry. Self-loops control stability; large and negative (blue) indicates strong stability. See [148] for full details. Note that the network is not symmetrical, it was permitted full flexibility during the estimation process (in contrast to Chapter 6). Node size is $n_k \equiv \sqrt{\sum_{j \neq k} W_{jk}^2}$ (outgoing strength).

dropped with increasing z.

These data were originally processed by Li *et al* [110], who performed a correlation analysis on the biological ages and the residual (having independently adjusted each biological age for chronological age). They found strong correlations across biological ages and marginal correlations across residuals. (We know from Chapter 5 that this is because the "true" biological age, which is close to chronological age, causes the mutual correlation.) This would seem to indicate that each biological age contains unique information, and while our network analysis confirms that this is true it is a highly misleading result. Most of the multivariate information appears to be unrelated to the aging process, and by age 80 there are really only 2 degrees-of-freedom of interest, out of a possible 8. Something fundamentally changes about human aging between ages 60-80, as observed in Figure 7.3, but the correlation analysis by Li *et al* cannot see this effect since it flattened out age information rather than stratifying by age. Older individuals become increasingly driven by z_1, just as we saw for humans in Chapters 5 and 6 (PC1).

The relationship between z_1 and network topology indicates that GrimAge and PhysioAge plays a central role in the dominance of z_1. When we visualized z_1 using the outer product of eigenvector 1, we saw that it is based on a block of outgoing connections from GrimAge and PhysioAge, with important contributions from Cognition. We know from Chapter 5 that the eigen-decomposition seeks low-rank blocks, and we can surmise that the feedback between GrimAge and PhysioAge helps attract the algorithm and thus makes it the key feature picked up by eigenvector 1 (a bi-directional feedback will create a block in the network). This comports with our human understanding of the dynamical behaviour of the system. Information enters via $\mathbf{\Sigma}$ and bounces back and forth between GrimAge and PhysioAge and continuously outwards into peripheral nodes, leading to a persistent signal with a long auto-correlation time (hence small eigenvalue i.e. weak stability). This signal is aging, or perhaps more specifically some form of advanced aging. In either case, the signal becomes z_1 and dominates the aging process at advanced ages and is able to drive an inevitable decline in health. The network and natural variable–pictures are therefore self-consistent and we can build an intuition both in terms of experimentally-meaningful biological ages, and the dynamically-meaningful natural variables. The former are likely easier to

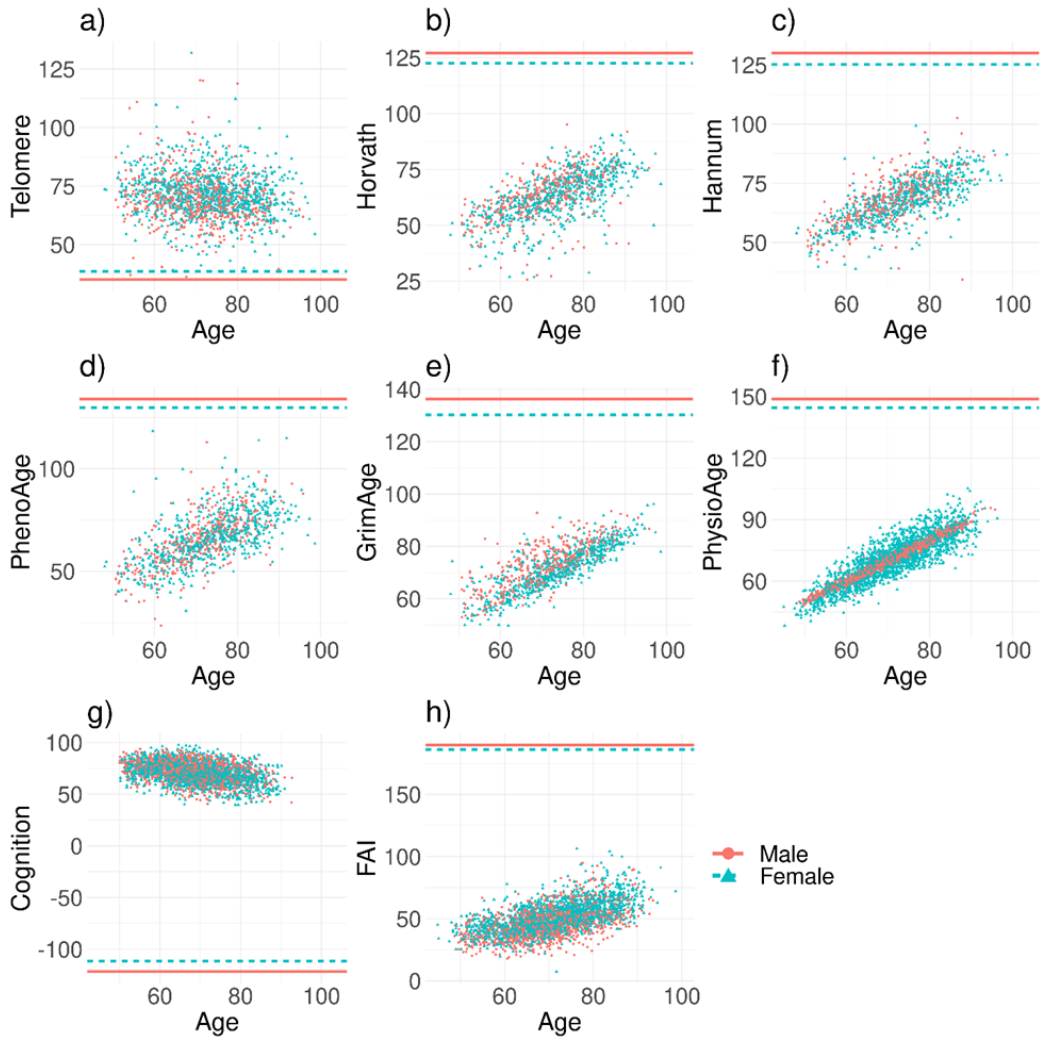

Figure 7.2: Scatterplot of biological ages versus age. Horizontal lines are equilibrium positions (μ). All of the biological ages are visibly correlated with age, except Telomere. Furthermore, owing to the gap between the equilibrium and the data, values increase continuously over time in our dynamical model. Telomere and Cognition were scaled to mean/sd of chronological age. For PhysioAge, males had a different set of variables hence the different with females [110].

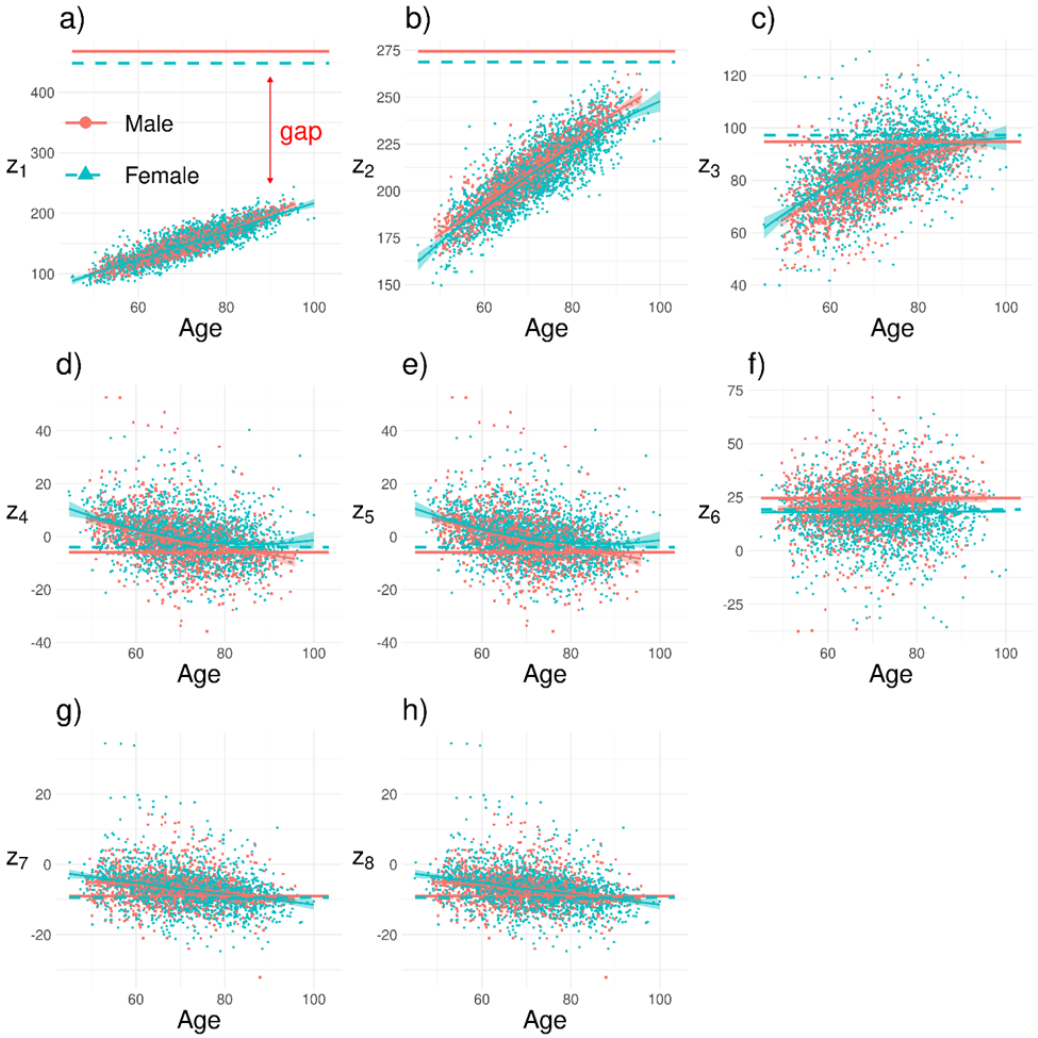

Figure 7.3: Scatterplot of natural variables versus age. Horizontal lines are equilibrium positions (μ). In contrast to the input biological ages (Figure 7.2) only the first three natural variables are visibly moderate-to-strongly correlated with chronological age. z_1 and z_2 were particularly strongly correlated with chronological age. The gap (indicated) ensures that those natural variables will not equilibrate, instead drifting up for the entire human lifespan. Biological ages were transformed using the eigen-decomposition transformation from the network (Figure 7.1).

map into known biological processes, whereas the latter are likely better at capturing the widespread, collective effects of the aging process (e.g. making them strong independent predictors).

How do we know a network such as Figure 7.1 is correct? We can test the model accuracy directly using a measure of accuracy such as the RMSE (root mean-squared error) for trajectories, as we did in Chapter 6 (this yielded $RMSE_{632} = 5.75 \pm 0.09$, $R^2_{train} \approx R^2_{test} = 0.65 \pm 0.01$). Biological ages are used as health state variables, however, and hence we are are concerned primarily when an individual gets worse (goes up) or better (does down) between now and some followup time. Testing this prognostic ability yielded $AUC = 0.764 \pm 0.005$ for predicting worsening. This test seems to be more sensitive to the use of directed links, and we found that we could better predict worsening between time points if we used a network with fully flexible, directed links [148] (asymmetrical).

I consider next a sensitivity analysis of the network to infer if any of the links are a spurious consequence of the data — either the variable suite or individuals within the sample. In Chapter 5 we were faced with this same problem applied to PCA, where we resorted to bootstrapping both by individuals and by input variables to test the robustness of the PCA mapping. The latter test proved to be particularly conservative. A more interesting way to perform sensitivity analysis is to add interesting additional variables and check whether they alter the meaning of the network. In the case of the biological ages study, we considered adding chronological age and the transformed FI, Figure 7.4. The similarity between Figures 7.1 and 7.4 provides us some confidence that these structural elements are real, including the central roles for PhysioAge and GrimAge, and many of the shared links. It also tells us that there is likely missing information, picked up by Age, but it may not be essential (since Age is not a central node). Furthermore it suggests that the FI (and related FAI) have only weak feedbacks into the lower scales. This supports the interpretation that functional decline is an effect of aging, rather than a significant cause. The lack of any major change in network structure reassures us that the initial network, Figure 7.1, is fairly robust given the data.

Networks have become the *de facto* tool for organising and communicating contemporary theories of aging [94, 115]. We propose a method to directly test these theories

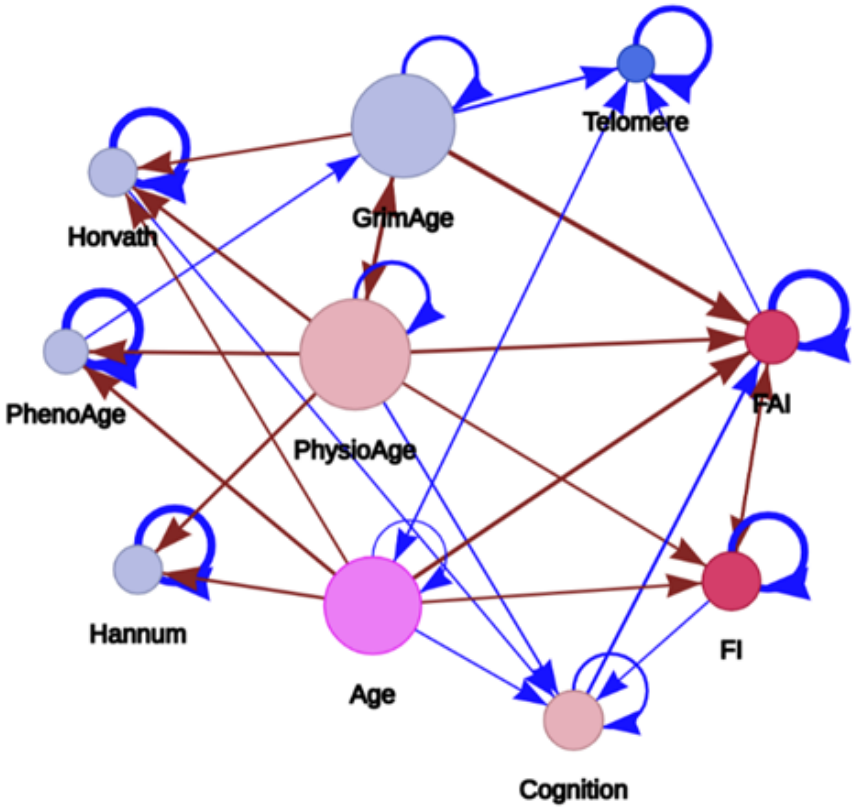

Figure 7.4: Estimated interaction network between biological ages (nodes) using the SF model — expanded to include chronological age and the FI. The network is surprisingly similar to Figure 7.1 despite adding two new nodes. In particular, PhysioAge and GrimAge occupy central positions in both networks (lots of outgoing links, proportional to node size). Note: the FI, f, was transformed by $\log(f + 0.065)$ to improve normality (then scaled to the mean and standard deviation of chronological age). Node size is $n_k \equiv \sqrt{\sum_{j \neq k} W_{jk}^2}$ (outoing strength).

by constructing networks of biological ages directly from aging study data and then comparing predicted interactions and essential behaviour to theory. What's more, as interventional data become available we propose this approach may be useful for understanding the effects of anti-aging interventions, since it provides a system-level (holistic) picture of health. By decomposing the network into its natural variables we can understand widespread effects while simultaneously having simple dynamical behaviour. We envision a future where natural variables, such as z_1, are specifically targeted by anti-aging interventions. More broadly, we may find that these natural variables generalize across species and hence can be used to interrogate how evolution has tuned longevity across species. At its core the approach is a bare-bones, essential model of aging that captures spontaneous aging decline as a consequence of simple dynamical behaviour. This essential simplicity permits a deeper level of understanding of the underlying causal mechanisms of aging.

Chapter 8

Discussion

The mainstream contemporary theories of aging are not accompanied by quantitative models [115, 94]. In Chapter 7 I addressed this paucity with a prospective framework for quantifying and testing contemporary theories using a dynamical network model based on a multivariate biological age representation. Quantitative models have the advantage of making exact predictions, which are better able to falsify theories, and can appropriately weigh competing effects such as those seen in interventions. Inclusion of dynamical behaviour enhances interpretability. For example, in Chapter 6 declining biomarker values with increasing age are interpreted as an inevitable consequence of a drifting homeostatic set-point. What's more, in Chapters 7, 6, and 5 I showed that decomposing a model into its constituents can be an effective strategy for quantifying an individual's biological age, maximizing the utility of whatever data are available. The search for biomarkers of aging and the search for quantitative models of aging can be viewed as common goals since they share significant overlap. Good models of aging naturally condense aging information, and salient biomarkers are easier to model. Unfortunately, data wrangling serves as a major barrier. Lab data are limited and large-scale studies are prone to bias, such as due to missing data (Chapter 4). This is particularly important for humans, which do not have lab data. This thesis delineates the steps needed to build quantitative models: from missing data to decomposition methods, and ultimate modelling.

Missing data are ubiquitous in aging studies and must be dealt with. In its rawest form, this problem appears in quantitative estimators since they need to be told how to handle gaps in the data (some software doesn't even have such options and will only accept complete data). In Chapter 4 I showed a variety of missing data handling strategies for aging study data, outlining the basic problem and prospective *post hoc* solutions. Multiple imputation emerged as an important tool since it is agnostic to the analysis being performed and is able to self-consistently estimate and propagate the

uncertainty in imputed values. In Chapter 5 multiple imputation was used together with standard algorithms, with the final results then pooled using Rubin's rules. In Chapter 6 I used both the iterative expectation-maximization procedure (impute, fit, impute) and derived estimators based on covariance matrices which can be computed using pairwise observations (excluding missing values). In all cases, sanity checks were performed to ensure that the imputed values were reasonable. The most basic check is to plot imputed values together with observed values to verify that they look reasonable in terms of mean and dispersion (typically plotted versus age). In many cases the individuals missing data are in particularly good or poor health, and we'd expect to see a slight shift in the imputed values reflecting that. Missing data handling is an important consideration in any analysis pipeline and one which has no easy answer. The key considerations are: how could missing values potentially bias study conclusions, can missing values be incorporated directly into the model, how plausible do impute values look, and how do results change when missing values are handled differently?

The next serious issue is estimating the signal associated with aging given a collection of health biomarkers which may have varying specificity to the aging process. It appears that automated methods do work but likely rely on *a priori* knowledge, including data curation. In Chapter 5 we looked to the FI for inspiration, since it is probably the most successful biological age to date. As demonstrated by Klemera and Doubal [100], a biological age which independently affects multiple biomarker trajectories will introduce a correlation between those variables, even if they would otherwise be uncorrelated. In their model, biological age plays the role of a hidden variable, inserting coherent information about the individual's health state into the observable variables causing an observable correlation. This means that biomarkers connected via a biological age will form a large block in the correlation matrix, possibly in addition to smaller blocks due to local biological ages of sub-systems, just as we saw in Figure 5.2 for binary health deficits. The eigen-decomposition, in this case PCA, then naturally found this 'global' mutual correlation and subsequent sub-blocks of increasingly-specific biological ages for sub-systems. The eigen-decomposition is trying to find low rank representations by adding together blocks to approximate the input matrix. The strategy in this paper was to rely on strong patterns present in

the data caused by the effects of aging, such that the underlying biological age naturally emerged. Critically, this approach depends on good data curation, which in the case of Chapter 5 was borrowed from the wisdom of preprocessing for the purpose of building an FI (see [188] for an expert summary of the procedure). The variables included have been pre-screened to be relevant to health, to have a strong association with chronological age, and to be not too correlated with each other such that a sub-block of deficits would overwhelm the global block. In the case of the FI LAB, deciding when continuous biomarker values are indicative of dysfunction rather than simply individual variability is another critical step. Given this *a priori* wisdom, the FI is an excellent compression and prediction algorithm for health deficit data. Unsupervised dimensionality reduction algorithms can easily capture this effect and offer a hands-off, automated approach to aging metrics — albeit one dependent on prior data curation.

Data curation is challenging and potentially biased, however, since it depends on the vagaries of the curation and the specific dataset(s) that motivated the curation. Ideally we'd like to relax our dependency on data curation such that we can explore a wider space of possibilities and reduce, or at least change, the risk of bias. In Chapter 6 I observed the mallostasis phenomenon wherein steady-state drift rate correlates with survival risk. This provides an alternative way to rank metrics rather than looking, for example, at the variance as is done with PCA. While in Chapter 5 I brushed aside concerns that small variance PCs may contain important, but rare, age-related health information (at least for health deficit data). In Chapter 6 I found that PC rank was not a great indicator of relevance, as illustrated by $z_4 \approx$ PC4 for the Het3 strain of mice which was found to be an important survival predictor despite its low PC rank (4th out of 6). By using steady-state drift rate we are leveraging an aging phenomenon, in this case that the steady-state behaviour should be drift towards worsening health, to automatically curate data in an unbiased manner. Our pipeline permits us to automatically generate aggregate aging metrics, and then rank them by relative importance.

We consistently observed that only a few, salient aging metrics are responsible for controlling most of the observed decline in health with age. Popular qualitative theories have proposed that aging emerges from a multi-causal network of influences

— which are believed to be needed to explain the widespread, interacting phenomena observed [94, 114, 115]. Flexible, quantitative studies indicate that there are at least two dimensions of aging, but they are less clear past two [141, 52, 66]. In Chapter 5 we found that a few special biological systems required many PCs (principal components) for optimal prediction, whereas most outcomes only depended on PC1 (which was almost perfectly correlated with the FI). This approach didn't account for the PC basis being oblique to the outcome, though, and a dependency on many PCs might simply indicate a non-canonical basis for that outcome of interest e.g. diabetics can be predicted very well based on the presence of dysfunctional glucose and glycohemoglobin. In Chapter 6 we observed that as age advances, Parseval's theorem ensures that a small set of mallostatic variables are able to drive age-related decline (often just one variable). Consistent with this picture, in analyzing the eigen-spectrum of Chapter 7 we found the first natural variable was dominant with a small correction in the second natural variable and smaller in the third and so forth. At advanced ages the mean and the variance of the first few natural variables can be very large due to equilibration and/or mallostasis, whereas most of the higher natural variables quickly reach a steady state. This permits the effects of aging to become increasingly dominated by a single variable at advanced age. Fedichev's group has suggested that a single order parameter similar to the FI — the "dynamical" FI — could be responsible for driving aging at advanced ages in mice [8]. Others have suggested that aging may become 'simpler' over time, for example as indicated by the decreasing coefficient of variation of the FI [161] and other aging metrics [93, 212], or in the apparent loss of "complexity" (e.g. fractal dimension) in dendrites, heart rate and other physiological functions [111]. PC1 in Chapter 5, z_1 or z_2 in Chapter 6 (depending on the dataset), and z_1 from Chapter 7, all demonstrated that they were well-connected variables carrying a great deal of information and driving widespread increases in risk of many adverse outcomes. In summary, it seems that there is a global "decline" variable which is associated with universal dysfunction across all biological systems and which is at least partially, if not wholly, captured by the FI at advanced ages. The FI is the average number of health deficits, and this apparent univariate-simplicity of the aging process may reflect that irrespective of how many causes there are, when health fails it always leads to the same generic loss

of physiological functioning, and gain of chronic diseases.

While this thesis has used agnostic models, several theory-driven models produce consistent or overlapping results. Probably the simplest theory is entropy-driven decline. There are many ways for things to go awry, and this should lead to a net preference for dysfunction to occur. This can be seen in DNA methylation and health deficit data which shows a linear drift term with age, similar to mallostasis, and can be argued to originate from entropy [186]. Biological systems are robust and resilient, and it is also conceivable that dysfunction promotes dysfunction, particularly when it occurs in special protective (robustness) or repair (resilience) mechanisms. This would explain the apparent exponential growth of the FI [123], and is a central assumption in the GNM (generic network model) of aging which recovers population-level survival and FI statistics [184]. It also seems likely that biological systems have finite resources and increasing demand with age, which is the foundation of the saturation-repair model — which correctly predicted the existence of mallostasis [93, 212, 6]. These theories are not entirely fleshed out, however, although there is widespread recognition that resilience and robustness are affected by the aging process [51, 31, 60]. This introduces an interesting question about the nature of dynamical phase(s) related to loss of robustness and resilience. In Chapters 6 and 7, I assumed constant resilience ($\boldsymbol{W}$) and robustness/stress (noise, $\boldsymbol{\Sigma}$) parameters — which implies that dysfunction should accumulate at a constant rate. Alternatively, robustness and/or resilience could change, for example there may be a critical value at which dysfunction accumulates faster than it can be repaired or averted, and decline becomes inevitable as suggested for the "dynamical" FI [8], and the saturating-repair model [93, 212]. These are exciting questions for future research.

This thesis has left a number of open questions that might be answered by future researchers. In Chapter 4 I noted that the *ad hoc* exclusion criteria used for variable selection in creating an FI often includes a cap on maximum missingness at 5%. I did not address this directly and it may be worth continuing the analysis down to the individual variable level to determine if variables with high missingness can be high leverage points in the analysis process i.e. that greatly affect results. Another issue is the robustness of the underlying Gibbs' sampler and its sensitivity to MNAR or sampling-related biases, which are worth exploring.

Moving on to Chapter 5, two tacit hypotheses were that (1) aging should be a big effect and it should be universal and easy to pick up, and (2) there should be biologically-meaningful signals in the data that we can estimate. The first we demonstrated was correct, with the first principal component (PC1) picking up a strong signal related to age-related decline. Conversely the second hypothesis appears to be correct only for PC1, with the remaining PCs serving as an efficient basis that is non-specific to biological processes (i.e. PC1 means something, it's a measure of overall functional state, but PC2 and onwards look non-specific). Major chronic diseases might nevertheless affect the PCs, and it would be interesting to consider the relationship between the two. Permitting different PCs for different individuals would be useful here, as is done in biclustering [153]. My approach of stepwise regression to test PC relevance is computationally intensive, and it should be made clear that there are simpler methods [14]. What's more, the dimensionality reduction can be circumvented entirely when testing for associations with outcomes [216]. These are useful considerations for followup work.

Chapter 6 provides a novel approach to dimensionality reduction, including feature selection. Specifically, the eigen-decomposition appears to push homeostatic dysfunction into the smallest eigenvalues, and in the exceptional cases the relevant variables with large eigenvalues can be identified by a strong drift rate, $|\mu_{age}|$. We have returned to modelling homeostasis with the SF model in a publication currently under review [149]. In the article we make use of high quality data from dialysis patients with standard blood tests measured every 6 weeks. This permit us to resolve a second mechanism for natural variable relevance, small eigenvalues are more important λ (we also recapitulated μ_{age}). What's more the article highlights the utility of using the natural variables over simply a multivariate analysis. In particular, nothing is lost by using the natural variables since the mapping is invertible. This means we can simplify the dynamical behaviour, or even make it specific to syndromes, and then map those components individually or as a whole back into observable signs. This addresses the problem of syndromes which are collections of medical signs without a clear cause [27]. In practical terms this means that each syndrome is associated with multiple disparate abnormal biomarker values, e.g. protein energy wasting syndrome is directly

associated with: low serum albumin, low serum cholesterol, changes to serum creatinine, and changes to a number of indirect biomarkers [55]. When we look directly at how specific health biomarkers change with age, e.g. hemoglobin [200], we can tap into existing literature on interpreting the associated biological significance but this doesn't help to understand syndromes. In contrast, the natural variables appear to be naturally sensitive to syndromes, while preserving their mapping into observable biomarkers. In the future, it would be interesting to see if the natural variables help disambiguate confounding biological effects, for example creatinine drops due to loss of lean mass and rises due to kidney disease [193]. These effects might appear as different natural variables permitting an individual to simultaneously suffer from both, despite normal creatinine levels.

Finally Chapter 7 opens the door for a number of possibilities, the foremost being direct quantitative analysis of networks theories of aging. First I offer some insight into the SF model. The model parameters for the network ($\boldsymbol{W}$) tend to be robust across stratifying variables, including: sex, frailty status and age, whereas the set point positions ($\vec{\mu}$) tend to change across stratifying variables. These latter effects can be captured at least in part by including covariates in $\vec{\mu}$ (using $\boldsymbol{\Lambda}$, Eq. 6.1). This may be because the model fits to differences, $\vec{y}_{n+1} - \vec{y}_n$, thus stripping out much of the individual effects (or perhaps only in the z with small eigenvalues, since there is a $\lambda\mu\Delta t$ term). Individualizing the models is thus better done through the set points, μ, and can be done through including covariates either as linear terms or as generalized additive terms [86]. The network typically doesn't change enough to justify stratifying before fitting, although this can be done as well.

We found that the natural variables z_1 and z_2 were driving the age-related changes in the Chapter 7 dataset, but these data were only a small subset of the available biomarkers of aging. In particular, despite having multiple epigenetic ages these are not sensitive to either cellular senescence or genomic instability [91], both of which could be sources of damage. Future research should considered all of the possible sources of damage and their interactions, for example those outlined by the Hallmarks [114, 115] or Pillar [94] theories. It is exciting to consider increasingly specific molecular biological ages that may point to important underlying drivers of aging. I picture these as being included in the network, but they may instead

modulate the rate of aging as covariates in μ (from Eq. 6.1 we have a velocity term $\boldsymbol{W}\vec{\mu}\Delta t$, hence μ controls the aging rate).

Chapter 9

Conclusion

Understanding geroscience at the systems level is an emerging field [98, 94, 115, 31, 6]. While several "biological ages" have shown they contain condensed aging information and are sensitive to stressors and interventions, there is significant room for improvement. Without reliable metrics of the effects of aging, an additional level of analysis is needed to extract age-specific signals from more generic biomarkers of health. This approach permits us to make progress with whatever data are available. The effective behaviour of those data, irrespective of the underlying biological mechanisms, can be sufficiently robust in its phenomena to make rigorous interpretations and predictions. I have demonstrated that important aspects of aging can be understood through the dynamical behaviour of either generic health biomarkers or specific biological ages. The underlying SF model permits two complementary perspectives: a multidimensional network representation of the interactions between observable biomarkers, together with a set of latent natural variables capturing the effective dynamical behaviour.

The network encodes conditional dependencies as a highly interpretable graphical model. This permit a holistic view that identifies the key observable variables and interactions between them, while faithfully capturing the complexity. This approach should prove useful in building and testing quantitative theories of aging. Networks also serve as the simplest sufficient foundation for *in silico* experiments that incorporate known multivariate interactions.

The natural variables emerge from the underlying interactions between biological systems, as identified by the network. The natural variables are efficient representations that address the need for quantitative measures of aging. They are ostensive, and as such other approaches can identify similar patterns, e.g. labelling them as syndromes. The weakness of the natural variables is that they need not be simple in terms of modifiable or observable biology. As such, natural variables may represent

combinations of fundamental mechanisms. For example, clever combinations of anti-aging interventions may be necessary to target specific natural variables. This would be challenging, but the expected payoff is broad downstream effects with simple dynamical trajectories. The natural variables should be viewed as a quantitative tool that may be useful for efficiently characterizing and communicating the effects and mechanisms involved in aging.

The natural variables appear to be a bottleneck of simple behaviour between the multiple causes and multiple effects of aging — although the former is speculative. I consistently observed that the aging process can be efficiently represented by well-connected natural variables which drive coherent decline across multiple biological systems. This means that aging may not be high dimensional and much of the apparent dimensionality comes from our naïve perspective. When we look at the dynamical behaviour in terms of natural variables we see a perturbative process, that is: one which is increasingly dominated by a single underlying variable, particularly past age 75 in humans. This gives hope for understanding aging through modelling of this relatively small subset of natural variables.

Bibliography

[1] Imputation by predictive mean matching: promise & peril. https://statisticalhorizons.com/predictive-mean-matching. Published: 05-03-2015. Accessed: 04-08-2020.

[2] Kofi P Adragni and R Dennis Cook. Sufficient dimension reduction and prediction in regression. *Philos. Trans. A Math. Phys. Eng. Sci.*, 367(1906):4385–4405, 2009.

[3] K G M M Alberti, P Zimmet, and J Shaw. Metabolic syndrome–a new worldwide definition. a consensus statement from the international diabetes federation. *Diabet. Med.*, 23(5):469–480, May 2006.

[4] Ruben Aldrovandi. *Special Matrices of Mathematical Physics: Stochastic, Circulant, and Bell Matrices*. World Scientific, 2001.

[5] Paul D Allison. Multiple imputation for missing data: a cautionary tale. *Sociological Methods & Research*, 28(3):301–309, 2000.

[6] Uri Alon. *Systems Medicine: Physiological Circuits and the Dynamics of Disease*. CRC Press, 2023.

[7] Azar Asadi Shahmirzadi, Daniel Edgar, Chen-Yu Liao, Yueh-Mei Hsu, Mark Lucanic, Arash Asadi Shahmirzadi, Christopher D Wiley, Garbo Gan, Dong Eun Kim, Herbert G Kasler, Chisaka Kuehnemann, Brian Kaplowitz, Dipa Bhaumik, Rebeccah R Riley, Brian K Kennedy, and Gordon J Lithgow. Alpha-Ketoglutarate, an endogenous metabolite, extends lifespan and compresses morbidity in aging mice. *Cell Metab.*, 32(3):447–456.e6, September 2020.

[8] Konstantin Avchaciov, Marina P Antoch, Ekaterina L Andrianova, Andrei E Tarkhov, Leonid I Menshikov, Olga Burmistrova, Andrei V Gudkov, and Peter O Fedichev. Unsupervised learning of aging principles from longitudinal data. *Nat. Commun.*, 13(1):6529, November 2022.

[9] Karen Bandeen-Roche, Jeremy D Walston, Yi Huang, Richard D Semba, and Luigi Ferrucci. Measuring systemic inflammatory regulation in older adults: evidence and utility. *Rejuvenation Res.*, 12(6):403–410, December 2009.

[10] J Banks, G D Batty, J J F Breedvelt, K Coughlin, R Crawford, M Marmot, J Nazroo, Z Oldfield, N Steel, A Steptoe, and Others. English longitudinal study of ageing: Waves 0–9, 1998–2019 [data collection]. UK data service. SN: 5050., 2021.

[11] Stephen Bates, Trevor Hastie, and Robert Tibshirani. Cross-validation: what does it estimate and how well does it do it? 2021.

[12] Brett K Beaulieu-Jones, Daniel R Lavage, John W Snyder, Jason H Moore, Sarah A Pendergrass, and Christopher R Bauer. Characterizing and managing missing structured data in electronic health records: Data analysis. *JMIR Med Inform*, 6(1):e11, February 2018.

[13] Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate cox proportional hazards models. *Stat. Med.*, 24(11):1713–1723, June 2005.

[14] Dominique Bertrand, El Mostafa Qannari, and Evelyne Vigneau. Latent root regression analysis: an alternative method to PLS. *Chemometr. Intell. Lab. Syst.*, 58(2):227–234, October 2001.

[15] George E Billman. Homeostasis: The underappreciated and far too often ignored central organizing principle of physiology. *Front. Physiol.*, 11:200, March 2020.

[16] Joanna Blodgett, Olga Theou, Susan Kirkland, Pantelis Andreou, and Kenneth Rockwood. Frailty in NHANES: comparing the frailty index and phenotype. *Arch. Gerontol. Geriatr.*, 60(3):464–470, May 2015.

[17] Joanna M Blodgett, Mario U Pérez-Zepeda, Judith Godin, D Scott Kehler, Melissa K Andrew, Susan Kirkland, Kenneth Rockwood, and Olga Theou. Frailty indices based on self-report, blood-based biomarkers and examination-based data in the canadian longitudinal study on aging. *Age Ageing*, 51(5), 2022.

[18] Joanna M Blodgett, Kenneth Rockwood, and Olga Theou. Changes in the severity and lethality of age-related health deficit accumulation in the USA between 1999 and 2018: A population-based cohort study. *The Lancet Healthy Longevity*, 2(2):e96–e104, February 2021.

[19] Joanna M Blodgett, Olga Theou, Susan E Howlett, and Kenneth Rockwood. A frailty index from common clinical and laboratory tests predicts increased risk of death across the life course. *Geroscience*, 39(4):447–455, August 2017.

[20] Joanna M Blodgett, Olga Theou, Susan E Howlett, Frederick C W Wu, and Kenneth Rockwood. A frailty index based on laboratory deficits in community-dwelling men predicted their risk of adverse health outcomes. *Age Ageing*, 45(4):463–468, July 2016.

[21] Joanna M Blodgett, Olga Theou, Arnold Mitnitski, Susan E Howlett, and Kenneth Rockwood. Associations between a laboratory frailty index and adverse health outcomes across age and sex. *Aging Med (Milton)*, 2(1):11–17, 2019.

[22] Todd E Bodner. What improves with increased missing data imputations? *Struct. Equ. Modeling*, 15(4):651–675, October 2008.

[23] Steven L Bressler and Anil K Seth. Wiener-Granger causality: a well established methodology. *Neuroimage*, 58(2):323–329, September 2011.

[24] Byron W Brown, Jr, Myles Hollander, and Ramesh M Korwar. Nonparametric tests of independence for censored data with application to heart transplant studies. Technical report, Florida State University, September 1973.

[25] S van Buuren and Karin Groothuis-Oudshoorn. MICE: multivariate imputation by chained equations in R. *Journal of Statistical Software*, pages 1–68, 2010.

[26] Frederick W Byron and Robert W Fuller. *Mathematics of Classical and Quantum Physics*. Dover, 1992.

[27] Franz Calvo, Bryant T Karras, Richard Phillips, Ann Marie Kimball, and Fred Wolf. Diagnoses, syndromes, and diseases: a knowledge representation problem. *AMIA Annu. Symp. Proc.*, 2003:802, 2003.

[28] Judith Campisi, Pankaj Kapahi, Gordon J Lithgow, Simon Melov, John C Newman, and Eric Verdin. From discoveries in ageing research to therapeutics for healthy ageing. *Nature*, 571(7764):183–192, July 2019.

[29] Centers for Disease Control and Prevention (CDC). National health and nutrition examination survey data. https://www.cdc.gov/nchs/nhanes/index.htm.

[30] Andrew Clegg, Chris Bates, John Young, Ronan Ryan, Linda Nichols, Elizabeth Ann Teale, Mohammed A Mohammed, John Parry, and Tom Marshall. Development and validation of an electronic frailty index using routine primary care electronic health record data. *Age Ageing*, 45(3):353–360, May 2016.

[31] Alan A Cohen, Luigi Ferrucci, Tamàs Fülöp, Dominique Gravel, Nan Hao, Andres Kriete, Morgan E Levine, Lewis A Lipsitz, Marcel G M Olde Rikkert, Andrew Rutenberg, Nicholas Stroustrup, and Ravi Varadhan. A complex systems approach to aging biology. *Nature Aging*, 2(7):580–591, July 2022.

[32] Alan A Cohen, Brian K Kennedy, Ulrich Anglas, Anne M Bronikowski, Joris Deelen, Frédérik Dufour, Gerardo Ferbeyre, Luigi Ferrucci, Claudio Franceschi, Daniela Frasca, Bertrand Friguet, Pierrette Gaudreau, Vadim N Gladyshev, Efstathios S Gonos, Vera Gorbunova, Philipp Gut, Mikhail Ivanchenko, Véronique Legault, Jean-François Lemaître, Thomas Liontis, Guang-Hui Liu, Mingxin Liu, Andrea B Maier, Otávio T Nóbrega, Marcel G M Olde Rikkert, Graham Pawelec, Sylvie Rheault, Alistair M Senior, Andreas Simm, Sonja Soo, Annika Traa, Svetlana Ukraintseva, Quentin Vanhaelen, Jeremy M Van Raamsdonk, Jacek M Witkowski, Anatoliy I Yashin, Robert Ziman, and Tamàs Fülöp. Lack of consensus on an aging biology paradigm? a global survey reveals an agreement to disagree, and the need for an interdisciplinary framework. *Mech. Ageing Dev.*, 191:111316, October 2020.

[33] Alan A Cohen, Emmanuel Milot, Qing Li, Patrick Bergeron, Roxane Poirier, Francis Dusseault-Bélanger, Tamàs Fülöp, Maxime Leroux, Véronique Legault, E Jeffrey Metter, Linda P Fried, and Luigi Ferrucci. Detection of a novel, integrative aging process suggests complex physiological integration. *PLoS One*, 10(3):e0116489, March 2015.

[34] Alan A Cohen, Emmanuel Milot, Jian Yong, Christopher L Seplaki, Tamàs Fülöp, Karen Bandeen-Roche, and Linda P Fried. A novel statistical approach shows evidence for multi-system physiological dysregulation during aging. *Mech. Ageing Dev.*, 134(3-4):110–117, March 2013.

[35] D G Cook, A G Shaper, D S Thelle, and T P Whitehead. Serum uric acid, serum glucose and diabetes: relationships in a population study. *Postgrad. Med. J.*, 62(733):1001–1006, 1986.

[36] Peter A Corning. The re-emergence of "emergence": A venerable concept in search of a theory. *Complexity*, 7(6):18–30, July 2002.

[37] Eileen M Crimmins, Bharat Thyagarajan, Morgan E Levine, David R Weir, and Jessica Faul. Associations of age, sex, Race/Ethnicity, and education with 13 epigenetic clocks in a nationally representative U.S. sample: The health and retirement study. *J. Gerontol. A Biol. Sci. Med. Sci.*, 76(6):1117–1123, May 2021.

[38] Marie Csete and John Doyle. Bow ties, metabolism and disease. *Trends Biotechnol.*, 22(9):446–450, September 2004.

[39] A Philip Dawid. Beware of the DAG! In *Causality: objectives and assessment*, pages 59–86. MLR proceedings, 2010.

[40] E R DeLong, D M DeLong, and D L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, September 1988.

[41] Yi Deng, Changgee Chang, Moges Seyoum Ido, and Qi Long. Multiple imputation for general missing data patterns in the presence of high-dimensional data. *Sci. Rep.*, 6:21689, February 2016.

[42] Elsa Dent, Paul Kowal, and Emiel O Hoogendijk. Frailty measurement in research and clinical practice: a review. *Eur. J. Intern. Med.*, 31:3–10, June 2016.

[43] L L Doove, S Van Buuren, and E Dusseldorp. Recursive partitioning for missing data imputation in the presence of interaction effects. *Comput. Stat. Data Anal.*, 72:92–104, April 2014.

[44] John J Dziak and Kimberly L Henry. Two-part predictors in regression models. *Multivariate Behav. Res.*, 52(5):551–561, September 2017.

[45] Frederick Eberhardt. Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics*, 3(2):81–91, March 2017.

[46] Peter F Edemekong, Deb L Bomgaars, Sukesh Sukumaran, and Shoshana B Levy. *Activities of Daily Living*. StatPearls Publishing, Treasure Island (FL), 2021.

[47] Monika Eichholzer, Aline Barbir, Shehzad Basaria, Adrian S Dobs, Manning Feinleib, Eliseo Guallar, Andy Menke, William G Nelson, Nader Rifai, Elizabeth A Platz, and Sabine Rohrmann. Serum sex steroid hormones and frailty in older american men of the third national health and nutrition examination survey (NHANES III). *Aging Male*, 15(4):208–215, 2012.

[48] Eran Elhaik. Principal component analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. *Sci. Rep.*, 12(1):14683, August 2022.

[49] Marcela R Entwistle, Donald Schweizer, and Ricardo Cisneros. Dietary patterns related to total mortality and cancer mortality in the united states. *Cancer Causes Control*, 32(11):1279–1288, 2021.

[50] Elena A Erosheva, Stephen E Fienberg, and Cyrille Joutard. Describing disability through individual-level mixture models for multivariate binary data. *Ann. Appl. Stat.*, 1(2):346–384, 2007.

[51] Elisa Fabbri, Marco Zoli, Marta Gonzalez-Freire, Marcel E Salive, Stephanie A Studenski, and Luigi Ferrucci. Aging and multimorbidity: New tasks, priorities, and frontiers for integrated gerontological and clinical research. *J. Am. Med. Dir. Assoc.*, 16(8):640–647, August 2015.

[52] Spencer Farrell, Arnold Mitnitski, Kenneth Rockwood, and Andrew D Rutenberg. Interpretable machine learning for high-dimensional trajectories of aging health. *PLoS Comput. Biol.*, 18(1):e1009746, 2022.

[53] Mathieu Fauvernier, Laurent Remontet, Zoé Uhry, Nadine Bossard, and Laurent Roche. survpen: an r package for hazard and excess hazard modelling with multidimensional penalized splines. *Journal of Open Source Software*, 4(40):1434, 2019.

[54] Deborah Finkel, Ola Sternäng, Juulia Jylhävä, Ge Bai, and Nancy L Pedersen. Functional aging index complements frailty in prediction of entry into care and mortality. *J. Gerontol. A Biol. Sci. Med. Sci.*, 74(12):1980–1986, November 2019.

[55] D Fouque, K Kalantar-Zadeh, J Kopple, N Cano, P Chauveau, L Cuppari, H Franch, G Guarnieri, T A Ikizler, G Kaysen, B Lindholm, Z Massy, W Mitch, E Pineda, P Stenvinkel, A Trevinho-Becerra, and C Wanner. A proposed nomenclature and diagnostic criteria for protein–energy wasting in acute and chronic kidney disease. *Kidney Int.*, 73(4):391–398, February 2008.

[56] Adam Freund. Untangling aging using dynamic, organism-level phenotypic networks. *Cell Systems*, 8(3):172–181, Mar 2019.

[57] Linda P Fried, Alan A Cohen, Qian-Li Xue, Jeremy Walston, Karen Bandeen-Roche, and Ravi Varadhan. The physical frailty syndrome as a transition from homeostatic symphony to cacophony. *Nature Aging*, 1(1):36–46, 2021.

[58] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1):1, 2010.

[59] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*, volume 1. Springer, New York, 2001.

[60] T Fulop, A Larbi, J M Witkowski, J McElhaney, M Loeb, A Mitnitski, and G Pawelec. Aging, frailty and age-related diseases. *Biogerontology*, 11(5):547–563, 2010.

[61] J B Gao, Yinhe Cao, and Jae-Min Lee. Principal component analysis of $1/f^{\alpha}$ noise. *Phys. Lett. A*, 314(5):392–400, 2003.

[62] L A Gavrilov and N S Gavrilova. The reliability theory of aging and longevity. *J. Theor. Biol.*, 213(4):527–545, December 2001.

[63] Leonid A Gavrilov and Natalia S Gavrilova. Evolutionary theories of aging and longevity. *ScientificWorldJournal*, 2:339–356, February 2002.

[64] Benedikt Gille, Annika Müller-Eigner, Shari Gottschalk, Erika Wytrwat, Martina Langhammer, and Shahaf Peleg. Titan mice as a model to test interventions that attenuate frailty and increase longevity. *Geroscience*, January 2024.

[65] Lovedeep Gondara and Ke Wang. Mida: Multiple imputation using denoising autoencoders. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 260–272. Springer, 2018.

[66] Alden L Gross, Michelle C Carlson, Nadia M Chu, Mara A McAdams-DeMarco, Dan Mungas, Eleanor M Simonsick, Ravi Varadhan, Qian-Li Xue, Jeremy Walston, and Karen Bandeen-Roche. Derivation of a measure of physiological multisystem dysregulation: Results from WHAS and health ABC. *Mech. Ageing Dev.*, 188:111258, June 2020.

[67] Yian Gu, Jose A Luchsinger, Yaakov Stern, and Nikolaos Scarmeas. Mediterranean diet, inflammatory and metabolic biomarkers, and risk of alzheimer's disease. *J. Alzheimers. Dis.*, 22(2):483–492, 2010.

[68] Manon Guay, Marie-France Dubois, María Corrada, Marie-Pierre Lapointe-Garant, and Claudia Kawas. Exponential increases in the prevalence of disability in the oldest old: a canadian national survey. *Gerontology*, 60(5):395–401, May 2014.

[69] J A Hanley and B J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, April 1982.

[70] Jochen Hardt, Max Herke, and Rainer Leonhart. Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. *BMC Med. Res. Methodol.*, 12:184, December 2012.

[71] Susan E Hardy, Heather Allore, and Stephanie A Studenski. Missing data: a special challenge in aging research. *J. Am. Geriatr. Soc.*, 57(4):722–729, April 2009.

[72] F E Harrell, Jr, R M Califf, D B Pryor, K L Lee, and R A Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, May 1982.

[73] Aisha Harun, Yevgeniy R Semenov, and Yuri Agrawal. Vestibular function and activities of daily living: Analysis of the 1999 to 2004 national health and nutrition examination surveys. *Gerontol Geriatr Med*, 1, 2015.

[74] T Hastie, R Tibshirani, and J Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2nd. Springer, 2017.

[75] Waylon J Hastings, Idan Shalev, and Daniel W Belsky. Comparability of biological aging measures in the national health and nutrition examination study, 1999–2002. *Psychoneuroendocrinology*, 106:171–178, 2019.

[76] Leonhard Held and Daniel Sabanés Bové. *Applied Statistical Inference: Likelihood and Bayes*. Springer, Berlin, Heidelberg, 2014.

[77] Douglas Henderson and Peter Plaschko. *Stochastic Differential Equations In Science And Engineering (with cd-rom)*. World Scientific, August 2006.

[78] Albert T Higgins-Chen, Kyra L Thrush, Yunzhang Wang, Pei-Lun Kuo, Meng Wang, Christopher J Minteer, Ann Zenobia Moore, Stefania Bandinelli, Christiaan H Vinkers, Eric Vermetten, Bart P F Rutten, Elbert Geuze, Cynthia Okhuijsen-Pfeifer, Marte Z van der Horst, Stefanie Schreiter, Stefan Gutwinski, Jurjen J Luykx, Luigi Ferrucci, Eileen M Crimmins, Marco P Boks, Sara Hägg, Tina T Hu-Seliger, and Morgan E Levine. A computational solution for bolstering reliability of epigenetic clocks: Implications for clinical trials and longitudinal tracking. *Nature Aging*, 2:644–661, 2022.

[79] Shangzhi Hong and Henry S Lynn. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC Med. Res. Methodol.*, 20(1):199, July 2020.

[80] Susan E Howlett. Assessment of frailty in animal models. *Interdiscip Top Gerontol Geriatr*, 41:15–25, 2015.

[81] Susan E Howlett, Michael R H Rockwood, Arnold Mitnitski, and Kenneth Rockwood. Standard laboratory tests to identify older adults at increased risk of death. *BMC Med.*, 12:171, October 2014.

[82] Susan E Howlett, Andrew D Rutenberg, and Kenneth Rockwood. The degree of frailty as a translational measure of health in aging. *Nature Aging*, 1(8):651–665, 2021.

[83] Ruth E Hubbard. Sex differences in frailty. *Interdiscip Top Gerontol Geriatr*, 41:41–53, July 2015.

[84] Anthony R Ives. Measuring resilience in stochastic systems. *Ecol. Monogr.*, 65(2):217–233, May 1995.

[85] Sebastian Jäger, Arndt Allhorn, and Felix Bießmann. A benchmark for data imputation methods. *Front Big Data*, 4:693674, July 2021.

[86] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, New York, 2013.

[87] Rick Jansen, Laura Km Han, Josine E Verhoeven, Karolina A Aberg, Edwin Cgj van den Oord, Yuri Milaneschi, and Brenda Wjh Penninx. An integrative study of five biological clocks in somatic and mental health. *Elife*, 10, 2021.

[88] Kazuaki Jindai, Carrie M Nielson, Beth A Vorderstrasse, and Ana R Quiñones. Multimorbidity and functional limitations among adults 65 or older, NHANES 2005-2012. *Prev. Chronic Dis.*, 13(160174):E151, 2016.

[89] Robert-Paul Juster, Bruce S McEwen, and Sonia J Lupien. Allostatic load biomarkers of chronic stress and impact on health and cognition. *Neurosci. Biobehav. Rev.*, 35(1):2–16, September 2010.

[90] Juulia Jylhävä, Nancy L Pedersen, and Sara Hägg. Biological age predictors. *EBioMedicine*, 21:29–36, July 2017.

[91] Sylwia Kabacik, Donna Lowe, Leonie Fransen, Martin Leonard, Siew-Lan Ang, Christopher Whiteman, Sarah Corsi, Howard Cohen, Sarah Felton, Radhika Bali, Steve Horvath, and Ken Raj. The relationship between epigenetic age and the hallmarks of ageing in human cells. *Nature Aging*, pages 1–10, May 2022.

[92] A K Kant, M I Whitley, and B I Graubard. Away from home meals: associations with biomarkers of chronic disease and dietary intake in american adults, NHANES 2005–2010. *Int. J. Obes.*, 39(5):820–827, October 2014.

[93] Omer Karin, Amit Agrawal, Ziv Porat, Valery Krizhanovsky, and Uri Alon. Senescent cell turnover slows with age providing an explanation for the Gompertz law. *Nat. Commun.*, 10(1):5495, December 2019.

[94] Brian K Kennedy, Shelley L Berger, Anne Brunet, Judith Campisi, Ana Maria Cuervo, Elissa S Epel, Claudio Franceschi, Gordon J Lithgow, Richard I Morimoto, Jeffrey E Pessin, Thomas A Rando, Arlan Richardson, Eric E Schadt, Tony Wyss-Coray, and Felipe Sierra. Geroscience: linking aging to chronic disease. *Cell*, 159(4):709–713, 2014.

[95] Donald Kennedy. Longevity, quality, and the one-hoss shay. *Science*, 305(5689):1369, September 2004.

[96] Gary King and Langche Zeng. Logistic regression in rare events data. *Polit. Anal.*, 9(2):137–163, 2001.

[97] Thomas B L Kirkwood. Understanding the odd science of aging. *Cell*, 120(4):437–447, February 2005.

[98] Thomas B L Kirkwood. Systems biology of ageing and longevity. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 366(1561):64–70, January 2011.

[99] Thomas B L Kirkwood. Deciphering death: a commentary on gompertz (1825) 'on the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies'. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 370(1666):20140379, April 2015.

[100] Petr Klemera and Stanislav Doubal. A new approach to the concept and computation of biological age. *Mech. Ageing Dev.*, 127(3):240–248, March 2006.

[101] Gotaro Kojima, Steve Iliffe, and Kate Walters. Frailty index as a predictor of mortality: a systematic review and meta-analysis. *Age Ageing*, 47(2):193–200, March 2018.

[102] Karolis Koncevičius, Akhil Nair, Aušrinė Šveikauskaitė, Agnė Šeštokaitė, Auksė Kazlauskaitė, Audrius Dulskas, and Artūras Petronis. Epigenetic age oscillates during the day. *Aging Cell*, page e14170, April 2024.

[103] Alexander Kowarik and Matthias Templ. Imputation with the R package VIM. *Journal of Statistical Software*, 74(7):1–16, 2016.

[104] H-K Kuo, S G Leveille, Y-H Yu, and W P Milberg. Cognitive function, habitual gait speed, and Late-Life disability in the national health and nutrition examination survey (NHANES) 1999–2002. *Gerontology*, 53(2):102–110, 2007.

[105] Andrew J Landgraf and Yoonkyung Lee. Dimensionality reduction for binary data through the projection of natural parameters. *J. Multivar. Anal.*, 180:104668, 2020.

[106] Glenn Ledder. *Mathematics for the Life Sciences*. Springer New York, 2013.

[107] D D Lee and H S Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[108] Caitlin Lees, Judith Godin, Janet E McElhaney, Shelly A McNeil, Mark Loeb, Todd F Hatchette, Jason LeBlanc, William Bowie, Guy Boivin, Allison McGeer, André Poirier, Jeff Powis, Makeda Semret, Duncan Webster, and Melissa K Andrew. Frailty hinders recovery from influenza and acute respiratory illness in older adults. *J. Infect. Dis.*, 222(3):428–437, July 2020.

[109] Qing Li, Shengrui Wang, Emmanuel Milot, Patrick Bergeron, Luigi Ferrucci, Linda P Fried, and Alan A Cohen. Homeostatic dysregulation proceeds in parallel in multiple physiological systems. *Aging Cell*, 14(6):1103–1112, December 2015.

[110] Xia Li, Alexander Ploner, Yunzhang Wang, Patrik Ke Magnusson, Chandra Reynolds, Deborah Finkel, Nancy L Pedersen, Juulia Jylhävä, and Sara Hägg. Longitudinal trajectories, correlations and mortality associations of nine biological ages across 20-years follow-up. *Elife*, 9, 2020.

[111] Lewis A Lipsitz and Ary L Goldberger. Loss of 'complexity' and aging: potential applications of fractals and chaos theory to senescence. *JAMA*, 267(13):1806–1809, 1992.

[112] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, Hoboken, NJ, 3 edition, 2020.

[113] Mingxin Liu, Véronique Legault, Tamàs Fülöp, Anne-Marie Côté, Dominique Gravel, F Guillaume Blanchet, Diana L Leung, Sylvia Juhong Lee, Yuichi Nakazato, and Alan A Cohen. Prediction of mortality in hemodialysis patients using moving multivariate distance. *Front. Physiol.*, 12:612494, March 2021.

[114] Carlos López-Otín, Maria A Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. The hallmarks of aging. *Cell*, 153(6):1194–1217, 2013.

[115] Carlos López-Otín, Maria A Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. Hallmarks of aging: An expanding universe. *Cell*, 186(2):243–278, January 2023.

[116] K G Manton and M A Woodbury. Grade of membership generalizations and aging research. *Exp. Aging Res.*, 17(4):217–226, 1991.

[117] Kieran A McCaul, Osvaldo P Almeida, Paul E Norman, Bu B Yeap, Graeme J Hankey, Jon Golledge, and Leon Flicker. How many older people are frail? Using multiple imputation to investigate frailty in the population. *J. Am. Med. Dir. Assoc.*, 16(5):439.e1–7, May 2015.

[118] B S McEwen. Allostasis and allostatic load: implications for neuropsychopharmacology. *Neuropsychopharmacology*, 22(2):108–124, February 2000.

[119] Pankaj Mehta, Ching-Hao Wang, Alexandre G R Day, Clint Richardson, Marin Bukov, Charles K Fisher, and David J Schwab. A high-bias, low-variance introduction to machine learning for physicists. *Phys. Rep.*, 810:1–124, 2019.

[120] Olaf Mersmann. *microbenchmark: Accurate Timing Functions*, 2019. R package version 1.4-7.

[121] Sarah J Mitchell, Morten Scheibye-Knudsen, Dan L Longo, and Rafael de Cabo. Animal models of aging research: implications for human aging and age-related diseases. *Annu Rev Anim Biosci*, 3:283–303, 2015.

[122] A B Mitnitski, A D Rutenberg, S Farrell, and K Rockwood. Aging, frailty and complex networks. *Biogerontology*, 18(4):433–446, August 2017.

[123] Arnold Mitnitski and Kenneth Rockwood. Aging as a process of deficit accumulation: its utility and origin. *Interdiscip. Top. Gerontol.*, 40:85–98, 2015.

[124] Arnold Mitnitski and Kenneth Rockwood. The rate of aging: the rate of deficit accumulation does not change over the adult life span. *Biogerontology*, 17(1):199–204, 2016.

[125] Arnold B Mitnitski, Janice E Graham, Alexander J Mogilner, and Kenneth Rockwood. Frailty, fitness and late-life mortality in relation to chronological and biological age. *BMC Geriatr.*, 2:1, February 2002.

[126] Maja Mizdrak, Marko Kumrić, Tina Tičinović Kurir, and Joško Božić. Emerging biomarkers for early detection of chronic kidney disease. *J Pers Med*, 12(4):548, March 2022.

[127] Dirk F Moore. *Applied Survival Analysis Using R*. Springer, 2016.

[128] Mahdi Moqri, Chiara Herzog, Jesse R Poganik, Jamie Justice, Daniel W Belsky, Albert Higgins-Chen, Alexey Moskalev, Georg Fuellen, Alan A Cohen, Ivan Bautmans, Martin Widschwendter, Jingzhong Ding, Alexander Fleming, Joan Mannick, Jing-Dong Jackie Han, Alex Zhavoronkov, Nir Barzilai, Matt Kaeberlein, Steven Cummings, Brian K Kennedy, Luigi Ferrucci, Steve Horvath, Eric Verdin, Andrea B Maier, Michael P Snyder, Vittorio Sebastiano, and Vadim N Gladyshev. Biomarkers of aging for the identification and evaluation of longevity interventions. *Cell*, 186(18):3758–3775, August 2023.

[129] Duncan Murdoch and E. D. Chow. *ellipse: Functions for Drawing Ellipses and Ellipse-Like Confidence Regions*, 2020. R package version 0.4.2.

[130] Jared S Murray. Multiple imputation: A review of practical and theoretical findings. *Stat. Sci.*, 33(2):142–159, May 2018.

[131] Yuichi Nakazato, Tomoko Sugiyama, Rena Ohno, Hirofumi Shimoyama, Diana L Leung, Alan A Cohen, Riichi Kurane, Satoru Hirose, Akihisa Watanabe, and Hiromi Shimoyama. Estimation of homeostatic dysregulation and frailty using biomarker variability: a principal component analysis of hemodialysis patients. *Sci. Rep.*, 10(1):10314, June 2020.

[132] Ahmed M Negm, Courtney C Kennedy, Lehana Thabane, Areti-Angeliki Veroniki, Jonathan D Adachi, Julie Richardson, Ian D Cameron, Aidan Giangregorio, Maria Petropoulou, Saad M Alsaad, Jamaan Alzahrani, Muhammad Maaz, Muhammad M Ahmed, Eileen Kim, Hadi Tehfe, Robert Dima, Kalyani Sabanayagam, Patricia Hewston, Hajar Abu Alrob, and Alexandra Papaioannou. Management of frailty: A systematic review and network meta-analysis of randomized controlled trials. *J. Am. Med. Dir. Assoc.*, 20(10):1190–1198, October 2019.

[133] Chao Nie, Yan Li, Rui Li, Yizhen Yan, Detao Zhang, Tao Li, Zhiming Li, Yuzhe Sun, Hefu Zhen, Jiahong Ding, Ziyun Wan, Jianping Gong, Yanfang Shi, Zhibo Huang, Yiran Wu, Kaiye Cai, Yang Zong, Zhen Wang, Rong Wang, Min Jian, Xin Jin, Jian Wang, Huanming Yang, Jing-Dong J Han, Xiuqing Zhang, Claudio Franceschi, Brian K Kennedy, and Xun Xu. Distinct biological ages of organs and systems identified from a multi-omics study. *Cell Rep.*, 38(10), March 2022.

[134] Dushani L Palliyaguru, Jacqueline M Moats, Clara Di Germanio, Michel Bernier, and Rafael de Cabo. Frailty index as a biomarker of lifespan and healthspan: Focus on pharmacological interventions. *Mech. Ageing Dev.*, 180:42–48, 2019.

[135] Dushani L Palliyaguru, Eric J Shiroma, John K Nam, Eleonora Duregon, Camila Vieira Ligo Teixeira, Nathan L Price, Michel Bernier, Simonetta Camandola, Kelli L Vaughan, Ricki J Colman, Andrew Deighan, Ron Korstanje, Luanne L Peters, Stephanie L Dickinson, Keisuke Ejima, Eleanor M Simonsick, Lenore J Launer, Chee W Chia, Josephine Egan, David B Allison, Gary A Churchill, Rozalyn M Anderson, Luigi Ferrucci, Julie A Mattison, and Rafael de Cabo. Fasting blood glucose as a predictor of mortality: Lost in translation. *Cell Metab.*, 33(11):2189–2200.e3, November 2021.

[136] Dushani L Palliyaguru, Camila Vieira Ligo Teixeira, Eleonora Duregon, Clara di Germanio, Irene Alfaras, Sarah J Mitchell, Ignacio Navas-Enamorado, Eric J Shiroma, Stephanie Studenski, Michel Bernier, Simonetta Camandola, Nathan L Price, Luigi Ferrucci, and Rafael de Cabo. Study of longitudinal aging in mice: Presentation of experimental techniques. *J. Gerontol. A Biol. Sci. Med. Sci.*, 76(4):552–560, March 2021.

[137] Soo Kyung Park, Caroline R Richardson, Robert G Holleman, and Janet L Larson. Frailty in people with COPD, using the national health and nutrition evaluation survey dataset (2003–2006). *Heart Lung*, 42(3):163–170, 2013.

[138] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[139] Fernando G Peña, Olga Theou, Lindsay Wallace, Thomas D Brothers, Thomas M Gill, Evelyne A Gahbauer, Susan Kirkland, Arnold Mitnitski, and Kenneth Rockwood. Comparison of alternate scoring of variables on the performance of the frailty index. *BMC Geriatr.*, 14:25, February 2014.

[140] Petersen, Kaare, Brandt and Pedersen, Michael, Syskind. The matrix cookbook. Online https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf, 2012.

[141] Emma Pierson, Pang Wei Koh, Tatsunori Hashimoto, Daphne Koller, Jure Leskovec, Nicholas Eriksson, and Percy Liang. Inferring multidimensional rates of aging from Cross-Sectional data. *Proc Mach Learn Res*, 89:97–107, 2019.

[142] Maik Pietzner, Anne Kaul, Ann-Kristin Henning, Gabi Kastenmüller, Anna Artati, Markus M Lerch, Jerzy Adamski, Matthias Nauck, and Nele Friedrich. Comprehensive metabolic profiling of chronic low-grade inflammation among generally healthy individuals. *BMC Med.*, 15(1):210, 2017.

[143] Jason N Pitt and Matt Kaeberlein. Why is aging conserved and what can we do about it? *PLoS Biol.*, 13(4):e1002131, April 2015.

[144] Jesse R Poganik, Bohan Zhang, Gurpreet S Baht, Alexander Tyshkovskiy, Amy Deik, Csaba Kerepesi, Sun Hee Yim, Ake T Lu, Amin Haghani, Tong Gong, Anna M Hedman, Ellika Andolf, Göran Pershagen, Catarina Almqvist, Clary B Clish, Steve Horvath, James P White, and Vadim N Gladyshev. Biological age is increased by stress and restored upon recovery. *Cell Metab.*, 35(5):807–820.e5, May 2023.

[145] Glen Pridham, Kenneth Rockwood, and Andrew Rutenberg. Strategies for handling missing data that improve frailty index estimation and predictive power: lessons from the nhanes dataset. *GeroScience*, 44(2):897–923, 2022.

[146] Glen Pridham, Kenneth Rockwood, and Andrew Rutenberg. Efficient representations of binarized health deficit data: the frailty index and beyond. *Geroscience*, 45:1687–1711, January 2023.

[147] Glen Pridham and Andrew D Rutenberg. Network dynamical stability analysis reveals key "mallostatic" natural variables that erode homeostasis and drive age-related decline of health. *Sci. Rep.*, 13(1):1–12, December 2023.

[148] Glen Pridham and Andrew D Rutenberg. Dynamical network stability analysis of multiple biological ages provides a framework for understanding the aging process. *The Journals of Gerontology: Series A*, page glae021, 01 2024.

[149] Glen Pridham, Karthik K Tennankore, Kenneth Rockwood, George Worthen, and Andrew D Rutenberg. Systems-level health of patients living with end-stage kidney disease using standard lab values. May 2024.

[150] Cécile Proust-Lima, Viviane Philipps, and Benoit Liquet. Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. *J. Stat. Softw.*, 78:1–56, June 2017.

[151] Yeping Lina Qiu, Hong Zheng, and Olivier Gevaert. Genomic data imputation with variational auto-encoders. *Gigascience*, 9(8), August 2020.

[152] R Core Team. R: A language and environment for statistical computing, 2020.

[153] Aaditya V Rangan, Caroline C McGrouther, John Kelsoe, Nicholas Schork, Eli Stahl, Qian Zhu, Arjun Krishnan, Vicky Yao, Olga Troyanskaya, Seda Bilaloglu, Preeti Raghavan, Sarah Bergen, Anders Jureus, Mikael Landen, and Bipolar Disorders Working Group of the Psychiatric Genomics Consortium. A loop-counting method for covariate-corrected low-rank biclustering of gene-expression and genome-wide association study data. *PLoS Comput. Biol.*, 14(5):e1006105, May 2018.

[154] Jana Raynor, Celine S Lages, Hesham Shehata, David A Hildeman, and Claire A Chougnet. Homeostasis and function of regulatory T cells in aging. *Curr. Opin. Immunol.*, 24(4):482–487, August 2012.

[155] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77, 2011.

[156] Justine Rochon, Matthias Gondan, and Meinhard Kieser. To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Med. Res. Methodol.*, 12:81, June 2012.

[157] K Rockwood, J M Blodgett, O Theou, M H Sun, H A Feridooni, A Mitnitski, R A Rose, J Godin, E Gregson, and S E Howlett. A frailty index based on deficit accumulation quantifies mortality risk in humans and in mice. *Sci. Rep.*, 7:43068, February 2017.

[158] Kenneth Rockwood and Arnold Mitnitski. Frailty, fitness, and the mathematics of deficit accumulation. *Rev. Clin. Gerontol.*, 17(1):1–12, 2007.

[159] Kenneth Rockwood and Arnold Mitnitski. Frailty in relation to the accumulation of deficits. *J. Gerontol. A Biol. Sci. Med. Sci.*, 62(7):722–727, July 2007.

[160] Kenneth Rockwood, Arnold Mitnitski, and Susan E Howlett. Frailty: Scaling from cellular deficit accumulation? *Interdiscip Top Gerontol Geriatr*, 41:1–14, July 2015.

[161] Kenneth Rockwood, Alexander Mogilner, and Arnold Mitnitski. Changes with age in the distribution of a frailty index. *Mech. Ageing Dev.*, 125(7):517–519, July 2004.

[162] Kenneth Rockwood, Xiaowei Song, and Arnold Mitnitski. Changes in relative fitness and frailty across the adult lifespan: Evidence from the Canadian National Population Health Survey. *CMAJ*, 183(8):E487–94, May 2011.

[163] Jarod Rutledge, Hamilton Oh, and Tony Wyss-Coray. Measuring biological age using omics data. *Nat. Rev. Genet.*, 23(12):715–727, December 2022.

[164] Jason L Sanders and Anne B Newman. Telomere length in epidemiology: a biomarker of aging, age-related disease, both, or neither? *Epidemiol. Rev.*, 35:112–131, January 2013.

[165] Joseph L Schafer and John W Graham. Missing data: our view of the state of the art. *Psychological Methods*, 7(2):147, 2002.

[166] Marten Scheffer, Stephen R Carpenter, Timothy M Lenton, Jordi Bascompte, William Brock, Vasilis Dakos, Johan van de Koppel, Ingrid A van de Leemput, Simon A Levin, Egbert H van Nes, Mercedes Pascual, and John Vandermeer. Anticipating critical transitions. *Science*, 338(6105):344–348, October 2012.

[167] Andrew I Schein, Lawrence K Saul, and Lyle H Ungar. A generalized linear model for principal component analysis of binary data. In Christopher M Bishop and Brendan J Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, volume R4 of *Proceedings of Machine Learning Research*, pages 240–247. PMLR, 2003.

[168] Tomas Schmauck-Medina, Adrian Molière, Sofie Lautrup, Jianying Zhang, Stefan Chlopicki, Helena Borland Madsen, Shuqin Cao, Casper Soendenbroe, Els Mansell, Mark Bitsch Vestergaard, Zhiquan Li, Yosef Shiloh, Patricia L Opresko, Jean-Marc Egly, Thomas Kirkwood, Eric Verdin, Vilhelm A Bohr, Lynne S Cox, Tinna Stevnsner, Lene Juel Rasmussen, and Evandro F Fang. New hallmarks of ageing: a 2022 Copenhagen ageing meeting summary. *Aging*, 14(16):6829–6839, August 2022.

[169] Kerry Schnell, Carlos O Weiss, Todd Lee, Jerry A Krishnan, Bruce Leff, Jennifer L Wolff, and Cynthia Boyd. The prevalence of clinically-relevant comorbid conditions in patients with physician-diagnosed COPD: a cross-sectional study using data from NHANES 1999-2008. *BMC Pulm. Med.*, 12(1):26, 2012.

[170] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10(5):1299–1319, July 1998.

[171] Josje D Schoufour, Nicole S Erler, Loes Jaspers, Jessica C Kiefte-de Jong, Trudy Voortman, Gijsbertus Ziere, Jan Lindemans, Caroline C Klaver, Henning Tiemeier, Bruno Stricker, Arfan M Ikram, Joop S E Laven, Guy G O Brusselle, Fernando Rivadeneira, and Oscar H Franco. Design of a frailty index among community living middle-aged and older people: the Rotterdam study. *Maturitas*, 97:14–20, March 2017.

[172] Rianne Margaretha Schouten, Peter Lugtig, and Gerko Vink. Generating missing values for simulation purposes: a multivariate amputation procedure. *J. Stat. Comput. Simul.*, 88(15):2909–2930, October 2018.

[173] Michael B Schultz, Alice E Kane, Sarah J Mitchell, Michael R MacArthur, Elisa Warner, David S Vogel, James R Mitchell, Susan E Howlett, Michael S Bonkowski, and David A Sinclair. Age and life expectancy clocks based on machine learning analysis of mouse frailty. *Nat. Commun.*, 11(1):4618, September 2020.

[174] Samuel D Searle, Arnold Mitnitski, Evelyne A Gahbauer, Thomas M Gill, and Kenneth Rockwood. A standard procedure for creating a frailty index. *BMC Geriatr.*, 8:24, September 2008.

[175] Raghav Sehgal, Margarita Meer, Aladdin H Shadyab, Ramon Casanova, Joann E Manson, Parveen Bhatti, Eileen M Crimmins, Themistocles L Assimes, Eric A Whitsel, Albert T Higgins-Chen, and Morgan Levine. Systems age: A single blood methylation test to quantify aging heterogeneity across 11 physiological systems. July 2023.

[176] M E Sehl and F E Yates. Kinetics of human aging: I. rates of senescence between ages 30 and 70 years in healthy people. *J. Gerontol. A Biol. Sci. Med. Sci.*, 56(5):B198–208, May 2001.

[177] Anoop D Shah, Jonathan W Bartlett, James Carpenter, Owen Nicholas, and Harry Hemingway. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *Am. J. Epidemiol.*, 179(6):764–774, March 2014.

[178] Eric Stallard. Trajectories of morbidity, disability, and mortality among the u.s. elderly population. *N. Am. Actuar. J.*, 11(3):16–53, July 2007.

[179] Daniel J Stekhoven and Peter Bühlmann. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, October 2011.

[180] Jonathan A C Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338:b2393, June 2009.

[181] Garrett Stubbings, Spencer Farrell, Arnold Mitnitski, Kenneth Rockwood, and Andrew Rutenberg. Informative frailty indices from binarized biomarkers. *Biogerontology*, 21(3):345–355, 2020.

[182] Garrett Stubbings, Kenneth Rockwood, Arnold Mitnitski, and Andrew Rutenberg. A quantile frailty index without dichotomization. *Mechanisms of Ageing and Development*, 199:111570, 2021. MAD-D-21-00175R1.

[183] Bonnielin K Swenor, Moon J Lee, Jing Tian, Varshini Varadaraj, and Karen Bandeen-Roche. Visual impairment and frailty: Examining an understudied relationship. *J. Gerontol. A Biol. Sci. Med. Sci.*, 75(3):596–602, 2019.

[184] Swadhin Taneja, Arnold B Mitnitski, Kenneth Rockwood, and Andrew D Rutenberg. Dynamical network model for age-related health deficits and mortality. *Phys Rev E*, 93(2):022309, February 2016.

[185] Fei Tang and Hemant Ishwaran. Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6):363–377, 2017.

[186] Andrei E Tarkhov, Kirill A Denisov, and Peter O Fedichev. Aging clocks, entropy, and the limits of age-reversal. October 2022.

[187] O Theou, M D L O'Connell, B L King-Kallimanis, A M O'Halloran, K Rockwood, and R A Kenny. Measuring frailty using self-report and test-based health measures. *Age Ageing*, 44(3):471–477, 2015.

[188] Olga Theou, Clove Haviva, Lindsay Wallace, Samuel D Searle, and Kenneth Rockwood. How to construct a frailty index from an existing dataset in 10 steps. *Age Ageing*, 52(12), December 2023.

[189] Olga Theou, Michael R H Rockwood, Arnold Mitnitski, and Kenneth Rockwood. Disability and co-morbidity in relation to frailty: How much do they overlap? *Arch. Gerontol. Geriatr.*, 55(2):e1–e8, September 2012.

[190] Terry M Therneau. *A Package for Survival Analysis in R*, 2020. R package version 3.1-12.

[191] V S Thomas, K Rockwood, and I McDowell. Multidimensionality in instrumental and basic activities of daily living. *J. Clin. Epidemiol.*, 51(4):315–321, April 1998.

[192] R Paul Thompson. Causality, mathematical models and statistical association: dismantling evidence-based medicine. *J. Eval. Clin. Pract.*, 16(2):267–275, April 2010.

[193] Charat Thongprayoon, Wisit Cheungpasitporn, Wonngarm Kittanamongkolchai, Andrew M Harrison, and Kianoush Kashani. Prognostic importance of low admission serum creatinine concentration for mortality in hospitalized patients. *Am. J. Med.*, 130(5):545–554.e1, May 2017.

[194] Rebecca Tobin, Glen Pridham, and Andrew D Rutenberg. Modelling lifespan reduction in an exogenous damage model of generic disease. *Sci. Rep.*, 13(1):16304, September 2023.

[195] O Troyanskaya, M Cantor, G Sherlock, P Brown, T Hastie, R Tibshirani, D Botstein, and R B Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, June 2001.

[196] S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.*, pages 1–68, 2010.

[197] Stef Van Buuren. *Flexible imputation of missing data*. CRC press, Boca Raton, FL, 2 edition, 2018.

[198] Stef Van Buuren, Jaap PL Brand, Catharina GM Groothuis-Oudshoorn, and Donald B Rubin. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12):1049–1064, 2006.

[199] Mahin Vazifehdan, Mohammad Hossein Moattar, and Mehrdad Jalali. A hybrid bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction. *Journal of King Saud University-Computer and Information Sciences*, 31(2):175–184, 2019.

[200] Stephanie Venn-Watson, Eric D Jensen, and Nicholas J Schork. A 25-y longitudinal dolphin cohort supports that long-lived individuals in same environment exhibit variation in aging rates. *Proc. Natl. Acad. Sci. U. S. A.*, 117(34):20950–20958, August 2020.

[201] Dervis C Vural, Greg Morrison, and L Mahadevan. Aging in complex interdependency networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 89(2):022811, February 2014.

[202] Lindsay M K Wallace, Olga Theou, Fernando Pena, Kenneth Rockwood, and Melissa K Andrew. Social vulnerability as a predictor of mortality and disability: cross-country differences in the survey of health, aging, and retirement in europe (SHARE). *Aging Clin. Exp. Res.*, 27(3):365–372, 2015.

[203] Zhenhua Wang, Olanrewaju Akande, Jason Poulos, and Fan Li. Are deep learning models superior for missing data imputation in large surveys? evidence from an empirical comparison, 2021.

[204] Eric W Weisstein. Geometric series. https://mathworld.wolfram.com/GeometricSeries.html.

[205] Ian R White, Patrick Royston, and Angela M Wood. Multiple imputation using chained equations: issues and guidance for practice. *Stat. Med.*, 30(4):377–399, February 2011.

[206] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[207] I S Widagdo, N Pratt, M Russell, and E E Roughead. Construct validity of four frailty measures in an older australian population: A rasch analysis. *J Frailty Aging*, 5(2):78–81, 2016.

[208] Emilee R Wilhelm-Leen, Yoshio N Hall, Manjula K Tamura, and Glenn M Chertow. Frailty and chronic kidney disease: the third national health and nutrition evaluation survey. *Am. J. Med.*, 122(7):664–71.e2, 2009.

[209] Simon N Wood. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Am. Stat. Assoc.*, 99(467):673–686, September 2004.

[210] Xin And Xiaogang, Yan. *Linear Regression Analysis: Theory And Computing*. World Scientific, 2009.

[211] Yachen Yan. *rBayesianOptimization: Bayesian Optimization of Hyperparameters*, 2016. R package version 1.1.0.

[212] Yifan Yang, Omer Karin, Avi Mayo, Xiaohu Song, Peipei Chen, Ana L Santos, Ariel B Lindner, and Uri Alon. Damage dynamics and the role of chance in the timing of e. coli cell death. *Nat. Commun.*, 14(1):2209, April 2023.

[213] Anatoli I Yashin, Konstantin G Arbeev, Igor Akushevich, Aliaksandr Kulminski, Lucy Akushevich, and Svetlana V Ukraintseva. Stochastic model for analysis of longitudinal data on aging and mortality. *Math. Biosci.*, 208(2):538–551, August 2007.

[214] Jingjing Yin and Lili Tian. Joint confidence region estimation for area under ROC curve and youden index. *Stat. Med.*, 33(6):985–1000, 2014.

[215] W J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

[216] Matthew A Zapala and Nicholas J Schork. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proc. Natl. Acad. Sci. U. S. A.*, 103(51):19430–19435, December 2006.

[217] Aleksandr Zenin, Yakov Tsepilov, Sodbo Sharapov, Evgeny Getmantsev, L I Menshikov, Peter O Fedichev, and Yurii Aulchenko. Identification of 12 genetic loci associated with human healthspan. *Commun Biol*, 2:41, January 2019.

[218] Bohan Zhang, Alexandre Trapp, Csaba Kerepesi, and Vadim N Gladyshev. Emerging rejuvenation strategies—reducing the biological age. *Aging Cell*, December 2021.

[219] Jonas Zierer, Cristina Menni, Gabi Kastenmüller, and Tim D Spector. Integration of 'omics' data in aging research: from biomarkers to systems biology. *Aging Cell*, 14(6):933–944, 2015.

Appendix A

Supplemental Information for Strategies for handling missing data that improve Frailty Index estimation and predictive power: lessons from the NHANES dataset

A.1 `ampute`

The MICE package in `R` (version 3.10.0) [25] includes a function for generating missing data: `ampute`. `ampute` preserves missingness patterns and allows the specification of missingness proportion. We observed two primary limitations to the function: (1) the cellwise missingness proportion option does not reliably generate the correct cellwise missingness proportion — forcing us to use the patternwise missingness proportion option instead, and (2) the quantile-based missingness rule does not work properly when patterns have small frequencies.

A.1.1 Issue (1) Cellwise Missingness

Based on Schouten *et al*'s approach [172], given a set of missingness patterns with missingness, M_i (the number of values/cells missing for the ith pattern), and a set of frequencies for each of the i patterns, f_i, for a dataset with R rows (entries), C columns (variables) and a total of N_{miss} cells missing, the cellwise missingness proportion, P, is:

$$P = \frac{N_{miss}}{RC}. \tag{A.1}$$

Let N_i be the (stochastic) number of cells missing due to the ith pattern and let $\langle N_i \rangle$ be the average of N_i, then we have:

$$\begin{aligned} \langle N_i \rangle &= R f_i M_i \pi \\ N_{miss} &= \sum_i N_i \approx \sum_i \langle N_i \rangle, \end{aligned} \tag{A.2}$$

where N_{miss} is the total number of missing values, and π is the casewise/patternwise missingness proportion (`prop`) parameter of `ampute`. The parameters taken by `ampute` are: f_i, M_i (indirectly via patterns), and π. Given we want a particular P we can solve for π as:

$$\pi = \frac{PC}{\sum_i f_i M_i}. \tag{A.3}$$

Using Eq. A.3 we calculated the correct $\pi \equiv$ `prop` parameter, to get the desired cellwise missingness, P. When `prop` is significantly greater than the real missingness used to estimate the patterns, $\pi > 1$ and there is no valid solution. Note that $0 \leq f_i \leq 1$ and $\sum_i f_i = 1$.

A.1.2 Issue (2) Insufficient Data

`ampute` is severely limited when dealing with small pattern frequencies, which is unavoidable when there is a large number of patterns and/or small amount of data. Quantiles can be used to bias the data, however, this method is invalid for the aforementioned case. Quantiles are calculated *after* data have been split into pattern-specific subsets. For example, the 1923 individuals in our Complete dataset would be split into the 751 observed missingness patterns, meaning that quantiles would be calculated using data-subsets of average size 2.6: far too few to accurately estimate quantiles. We adjusted the `ampute` code to calculate quantiles first, using the Complete dataset of 1923 individuals, and then subset the dataset into patterns. This allowed the quantiles to be accurately calculated.

A.2 Extended Results

For simulated missingness, the ground truth values were known and we were able to directly compare the accuracy of the imputed values. It is not clear how important it is for the individual imputations to be close to the true values, as opposed to them being unbiased and accurately reproducing the dispersion of values. This is because we are performing some post-imputation-processing to calculate the FI, and we are often performing multiple imputations – for which the dispersion of values is important. Regardless, for discrete-valued variables, i.e. the 36 self-reported/demographic health variables, we calculated the accuracy of getting the exact value and calculated the confidence interval assuming a binomial distribution. For continuous-valued variables, i.e. the 32 lab variables, we calculated the unadjusted R^2 value with the confidence interval estimated via bootstrapping ($N = 1000$). Multiple imputations were aggregated by majority vote for discrete variables and by mean for continuous variables for fair comparison to the single imputation strategies. Table A.1 reports the imputation accuracy for each type of simulated missingness. The accuracy seems to have little connection with the overall performance of the subsequent FI. In light of our other results, it appears that being close to the true values doesn't guarantee a good imputation, perhaps due to bias or inaccurate dispersion — at least for values subsequently processed into the FI. We conjecture that the FI procedure is insensitive to some random error in imputed values as long as the underlying distributions are captured.

The coverage probability for the true mean FI is given in Table A.2 ($N = 10$ for each). CART with auxiliary variables (Aux) and 5 imputations had 100% coverage in all cases. Ideally, the coverage should be 95%, however, our confidence interval (CI) estimates did not include shared values between the true dataset and the dataset with missing values, and hence our CIs are likely too large and the coverage should be used as a relative, not absolute, measure of imputation accuracy (bigger implies better).

The extended results for bias and predictive power for the simulated missingness is given for cMCAR and cMNAR in Table A.3, and for pMCAR and pMAR in Table A.4. Note that the predictive power is missing for Ignore (weighted) because the packages used did not allow for weights.

In Figure A.5 we present the survival curves with the young/old cut moved from age 60 to age 50. The VIQ block was collected at age 50, hence the VIQ results should be more accurate in this figure than in Figure 4.3, note the significant difference in HRs for young vs old.

In Figure A.6 we present extended survival curves from Figure 4.3, including the Misc variables which were commonly missing: RDQ031 and KIQ046.

The FI bias for all imputation methods is illustrated in Figure A.7. Note the strong linearity of results, supporting out use of linear bias rates.

Forest plots of the HR for simulated 15% missingness are given in Figure A.8.

Forest plots of the HR for simulated higher levels of cMCAR are given in Figure A.9. Ignore shows an underestimate of the HR, while kNN appears to frequently overestimate the HR with huge variation between the $N = 10$ simulated datasets.

Forest plots for simulated higher levels of cMNAR are given in Figure A.10. Ignore greatly underestimated the HR.

Forest plots for the real missingness are given in Figure A.11. kNN appeared to underestimate the HR without rule-based imputation (RI).

Figure A.12 gives the ridge plot of the FIs calculating using all imputation strategies considered.

A.3 FI Blocks

We illustrate the Full dataset missingness patterns, including young and old patients, in Figure A.1. The exact variables in each block are described in Table A.5.

To confirm that the amputation of simulated missingness was successful, we compared the 15% simulated missingness patterns to the real missingness. The simulated patterned missingness are displayed in Figure A.2, and the cellwise unpatterned missingness are displayed in Figure A.3. Compared to the real missingness in Figure A.1 we see excellent agreement for pMCAR and pMAR, indicated successful amputation.

In Figure A.4 we give the missingness patterns with full variable names rather than just blocks, the data are identical to Figure 4.2, only the plot aesthetics differ. All missingness block pattern figures followed this variable order.

The demographics of the missingness blocks are given in Table A.6. Individuals with the PFQ, RXD or VIQ blocks missing were significantly younger and had lower Ignore FIs (all $p < 2.2 \cdot 10^{-16}$) and individuals missing the BPX block were older ($p = 4.2 \cdot 10^{-7}$) and had higher Ignore FIs ($p = 1.2 \cdot 10^{-14}$).

The complete blockwise summary of the FI bias to the mean and SD in the simulated missingness are tabulated below from Table A.7 to Table A.14. The bias rates were fit using 5%, 10% and 15% missingness. The missingness column quantifies the exact rate of missingness in the simulation.

The complete blockwise predictive power for each type of simulated 15% missingness are tabulated below from Table A.15 to Table A.22. The blockwise summary gives some idea which component variables of the FI are most important for prediction. No statistical testing was applied to either the C-index or HR.

The extended FI statistics for real missingness by block are given in Table A.28. Tables A.29 and A.30 include RI (rule-based imputation).

A.4 FI Variables

The FI was calculated using 68 variables: 36 self-reported and 32 lab. The 32 lab variables used are described in Table A.23. The low and high healthy ranges used for binarization are included for males (M) and females (F). The 36 self-reported health variables used to calculate the FI are described in Table A.24.

A.5 Auxiliary Variables

We tested the utility of 100 auxiliary variables: 73 self-reported health variables and 27 lab variables. These variables were not re-coded, representing *ad hoc* inclusions. We selected as many variables as possible using the following selection criteria. These criteria are not, in general, good selection criteria for imputing real missingness. They include criteria for simplicity and ease of application, and criteria to explicitly forbid trivially-easy imputations: to make the simulated missingness more realistic.

1. The variable belongs to either demographical data or one of the NHANES documents already being used to calculate the FI.

2. The variable exists in *both* NHANES 03/04 and 05/06.

3. The variable measures something unique from the other variables.

4. The variable is not a follow-up question to another auxiliary variable. E.g. MCQ0170L: "Do you/Does SP still ... have any kind of liver condition?"

5. The variable is not continuous-valued with missingness encoded as a number greater than the maximum e.g. missing OSQ020c values were encoded as 9999. These values would have to be manually converted to NA, we excluded these variables for convenience.

6. The variable is relevant for the individuals used in our study, e.g. PFQ020: "{Do you/Does SP} have an impairment or health problem that limits {your/his/her} ability to {crawl, walk or play} {walk, run or play} {walk or run}?" was not included because our population was over age 20.

7. The variable has unique values (at least 2) but not an excessive number of categories, e.g. RXDDRUG had 287 unique drugs and was not included. A variable with only 1 value has no utility. A categorical variable with many values is inconvenient to work with.

8. The variable does not relate to a specific variant condition, e.g. type of cancer and type of bone-break were excluded. This was primarily to reduce the number of auxiliary variables for convenience.

Ordinal-type variables were converted to categorical variables. This was necessary because missingness was left numerically encoded. Many of the variables included are unlikely to have any utility in imputation but that's deliberate: we want our imputation strategy to be robust enough that we can mindlessly pick auxiliary variables to improve results. The 27 auxiliary lab variables used are reported in Table A.25. The 73 auxiliary self-reported health variables are reported in Table A.26 and Table A.27.

A.6 Tables

Table A.1: Imputation Accuracy — Simulated 15% Missingness

Imputation	Missingness	Accuracy (95% CI)	R^2 (95% CI)
Default (m=5)	pMCAR	53.5% (53.4-53.7%)	.41 (.28-.63)
MICE RF (m=5)	pMCAR	75.7% (75.5-75.8%)	.40 (.28-.64)
RF	pMCAR	75.9% (75.8-76.1%)	.47 (.32-.72)
kNN	pMCAR	60.0% (59.9-60.2%)	.46 (.31-.71)
CART (m=5)	pMCAR	73.7% (73.5-73.8%)	.40 (.27-.64)
CART+Aux (m=5)	pMCAR	76.0% (75.9-76.2%)	.41 (.28-.66)
Default (m=5)	pMAR	54.7% (54.5-54.8%)	.51 (.35-.75)
MICE RF (m=5)	pMAR	77.9% (77.8-78.1%)	.51 (.35-.75)
RF	pMAR	78.1% (78.0-78.3%)	.61 (.42-.85)
kNN	pMAR	62.8% (62.7-63.0%)	.59 (.40-.83)
CART (m=5)	pMAR	75.6% (75.5-75.8%)	.53 (.36-.75)
CART+Aux (m=5)	pMAR	77.9% (77.8-78.1%)	.47 (.32-.71)
Default (m=5)	cMCAR	98.2% (98.2-98.3%)	.95 (.88-.99)
MICE RF (m=5)	cMCAR	98.7% (98.7-98.8%)	.95 (.88-.99)
RF	cMCAR	98.7% (98.7-98.8%)	.96 (.89-.99)
kNN	cMCAR	98.5% (98.4-98.5%)	.96 (.89-.99)
CART (m=5)	cMCAR	74.8% (74.6-75.0%)	.39 (.27-.54)
CART+Aux (m=5)	cMCAR	75.3% (75.1-75.5%)	.45 (.34-.60)
Default (m=5)	cMNAR	84.2% (84.0-84.4%)	.69 (.60-.76)
MICE RF (m=5)	cMNAR	88.4% (88.3-88.6%)	.72 (.64-.79)
RF	cMNAR	88.6% (88.5-88.7%)	.82 (.74-.87)
kNN	cMNAR	86.3% (86.1-86.4%)	.80 (.73-.85)
CART (m=5)	cMNAR	85.5% (85.4-85.7%)	.47 (.30-.61)
CART+Aux (m=5)	cMNAR	98.4% (98.3-98.4%)	.98 (.97-.99)

= [1]

[1] Score was (unadjusted) R^2 for continuous (lab) variables, accuracy for discrete (SR) variables. m = 5 multiple imputations were used where applicable, aggregated by majority vote (discrete) or mean (continuous) for fair comparison to single imputation methods.

Table A.2: Coverage probability for mean FI by missingness (10 repeats).

Imputation	Missingness	Coverage[1] 5%	10%	15%
Ignore	pMCAR	0.1	0	0
Ignore20	pMCAR	0.1	0	0
Ignore (weighted)	pMCAR	0.8	0	0
Default (m=5)	pMCAR	0	0	0
kNN	pMCAR	0.7	0.4	0.9
RF	pMCAR	0	0	0
MICE RF (m=5)	pMCAR	0.8	0	0
CART (m=5)	pMCAR	1	1	1
CART (m=15)	pMCAR	1	1	1
CART+Aux (m=5)	pMCAR	1	1	1
Ignore	pMAR	1	0	0
Ignore20	pMAR	1	0	0
Ignore (weighted)	pMAR	0	0	0
Default (m=5)	pMAR	0	0	0
kNN	pMAR	0.9	0.5	0.7
RF	pMAR	0	0	0
MICE RF (m=5)	pMAR	1	0	0
CART (m=5)	pMAR	1	1	1
CART (m=15)	pMAR	1	1	0.99[2]
CART+Aux (m=5)	pMAR	1	1	1
Ignore	cMNAR	0	0	0
Ignore20	cMNAR	0	0	0
Ignore (weighted)	cMNAR	0	0	0
Default (m=5)	cMNAR	0	0	0
kNN	cMNAR	1	1	1
RF	cMNAR	1	1	1
MICE RF (m=5)	cMNAR	0.22[2]	0	0
CART (m=5)	cMNAR	0	0	0
CART (m=15)	cMNAR	0	0	0
CART+Aux (m=5)	cMNAR	1	1	1
Ignore	cMNAR	1	1	1
Ignore20	cMCAR	1	1	1
Ignore (weighted)	cMCAR	1	1	1
Default (m=5)	cMCAR	0	0	0
kNN	cMCAR	1	1	1
RF	cMCAR	1	1	1
MICE RF (m=5)	cMCAR	0.14[2]	0	0
CART (m=5)	cMCAR	1	1	1
CART (m=15)	cMCAR	1	1	1
CART+Aux (m=5)	cMCAR	1	1	1

[1] Coverage is the probability that the true value was found within the 95% CI. Higher is better. CIs may have been overestimated, see text.

[2] MI can be fractional e.g. 10 repeats of m = 15 gives 150 possible values.

Table A.3: Imputed FI Summary — Cellwise Simulated Missingness (Supplemental)

Imputation	Type	N	Missingness	Mean FI	Bias	FI SD	SD Bias	C-index	HR	AUC
GT	cMNAR	1923	0.150 ± 0.000	0.176	0.000 ± 0.000	0.073	0.000 ± 0.000	0.654 ± 0.016	1.075 ± 0.007	0.733 ± 0.036
Ignore	cMNAR	1923	0.150 ± 0.000	0.198	$0.022 \pm 0.000^{***}$	0.080	$0.007 \pm 0.000^{***}$	0.653 ± 0.016	1.069 ± 0.007	0.732 ± 0.037
Ignore (weighted)	cMNAR	1923	0.150 ± 0.000	0.199	$0.023 \pm 0.002^{***}$	0.081	$0.007 \pm 0.000^{***}$	–	–	–
Ignore20	cMNAR	1658.6	0.150 ± 0.000	0.202	$0.020 \pm 0.000^{***}$	0.081	$0.008 \pm 0.000^{***}$	0.654 ± 0.017	1.069 ± 0.007	0.735 ± 0.039
kNN	cMNAR	1923	0.150 ± 0.000	0.179	$0.002 \pm 0.000^{***}$	0.072	$-0.001 \pm 0.000^{***}$	0.652 ± 0.016	1.076 ± 0.007	0.730 ± 0.037
RF	cMNAR	1923	0.150 ± 0.000	0.177	$0.001 \pm 0.000^{*}$	0.074	$0.001 \pm 0.000^{**}$	0.653 ± 0.016	1.073 ± 0.007	0.733 ± 0.037
Default (m=1)	cMNAR	1923	0.150 ± 0.000	0.188	$0.011 \pm 0.000^{***}$	0.076	$0.002 \pm 0.000^{***}$	0.653 ± 0.016	1.073 ± 0.007	0.733 ± 0.037
Default (m=5)	cMNAR	1923	0.150 ± 0.000	0.193	$0.017 \pm 0.001^{***}$	0.078	$0.004 \pm 0.001^{***}$	0.654 ± 0.016	1.071 ± 0.007	0.733 ± 0.037
MICE RF (m=1)	cMNAR	1923	0.150 ± 0.000	0.181	$0.005 \pm 0.000^{***}$	0.074	0.000 ± 0.000	0.653 ± 0.016	1.074 ± 0.007	0.733 ± 0.037
MICE RF (m=5)	cMNAR	1923	0.150 ± 0.000	0.188	$0.011 \pm 0.001^{***}$	0.076	$0.002 \pm 0.001^{**}$	0.654 ± 0.016	1.073 ± 0.007	0.734 ± 0.037
CART (m=1)	cMNAR	1923	0.150 ± 0.000	0.185	$0.008 \pm 0.000^{***}$	0.075	$0.001 \pm 0.000^{***}$	0.654 ± 0.016	1.073 ± 0.007	0.734 ± 0.037
CART (m=5)	cMNAR	1923	0.150 ± 0.000	0.190	$0.014 \pm 0.001^{***}$	0.077	$0.004 \pm 0.001^{***}$	0.655 ± 0.016	1.072 ± 0.007	0.735 ± 0.037
CART+Aux (m=1)	cMNAR	1923	0.150 ± 0.000	0.173	$-0.003 \pm 0.001^{***}$	0.072	$-0.002 \pm 0.000^{***}$	0.651 ± 0.016	1.076 ± 0.007	0.730 ± 0.037
CART+Aux (m=5)	cMNAR	1923	0.150 ± 0.000	0.176	0.000 ± 0.001	0.074	0.000 ± 0.001	0.652 ± 0.016	1.075 ± 0.007	0.731 ± 0.037
GT	cMCAR	1923	0.150 ± 0.000	0.176	0.000 ± 0.000	0.073	0.000 ± 0.000	0.654 ± 0.016	1.075 ± 0.007	0.733 ± 0.036
Ignore	cMCAR	1923	0.150 ± 0.000	0.176	0.000 ± 0.000	0.076	$0.002 \pm 0.000^{***}$	0.648 ± 0.016	1.070 ± 0.007	0.728 ± 0.037
Ignore (weighted)	cMCAR	1923	0.150 ± 0.000	0.176	0.000 ± 0.002	0.075	$0.002 \pm 0.000^{***}$	–	–	–
Ignore20	cMCAR	1670.3	0.150 ± 0.000	0.176	0.000 ± 0.000	0.075	$0.002 \pm 0.001^{**}$	0.648 ± 0.020	1.071 ± 0.009	0.729 ± 0.041
kNN	cMCAR	1923	0.150 ± 0.000	0.179	$0.002 \pm 0.000^{***}$	0.072	$-0.001 \pm 0.000^{***}$	0.652 ± 0.016	1.076 ± 0.007	0.730 ± 0.037
RF	cMCAR	1923	0.150 ± 0.000	0.177	$0.001 \pm 0.000^{*}$	0.074	$0.001 \pm 0.000^{**}$	0.653 ± 0.016	1.074 ± 0.007	0.732 ± 0.037
Default (m=1)	cMCAR	1923	0.150 ± 0.000	0.188	$0.011 \pm 0.000^{***}$	0.076	$0.002 \pm 0.000^{***}$	0.654 ± 0.016	1.073 ± 0.007	0.734 ± 0.037
Default (m=5)	cMCAR	1923	0.150 ± 0.000	0.193	$0.017 \pm 0.001^{***}$	0.078	$0.005 \pm 0.001^{***}$	0.654 ± 0.016	1.071 ± 0.007	0.734 ± 0.037
MICE RF (m=1)	cMCAR	1923	0.150 ± 0.000	0.181	$0.004 \pm 0.000^{***}$	0.074	0.000 ± 0.000	0.653 ± 0.016	1.074 ± 0.007	0.733 ± 0.037
MICE RF (m=5)	cMCAR	1923	0.150 ± 0.000	0.188	$0.011 \pm 0.001^{***}$	0.076	$0.002 \pm 0.001^{*}$	0.654 ± 0.016	1.073 ± 0.007	0.734 ± 0.037
CART (m=1)	cMCAR	1923	0.150 ± 0.000	0.172	$-0.004 \pm 0.000^{***}$	0.072	$-0.002 \pm 0.000^{***}$	0.651 ± 0.016	1.076 ± 0.007	0.730 ± 0.037
CART (m=5)	cMCAR	1923	0.150 ± 0.000	0.177	0.000 ± 0.001	0.073	-0.000 ± 0.001	0.652 ± 0.016	1.075 ± 0.008	0.732 ± 0.037
CART+Aux (m=1)	cMCAR	1923	0.150 ± 0.000	0.173	$-0.003 \pm 0.001^{***}$	0.072	$-0.002 \pm 0.000^{***}$	0.653 ± 0.016	1.076 ± 0.007	0.733 ± 0.037
CART+Aux (m=5)	cMCAR	1923	0.150 ± 0.000	0.177	0.000 ± 0.001	0.074	0.000 ± 0.001	0.654 ± 0.016	1.076 ± 0.008	0.735 ± 0.037

Table A.4: Imputed FI Summary — Patterned Simulated Missingness (Supplemental)

Imputation	Type	N	Missingness	Mean FI	Bias	FI SD	SD Bias	C-index	HR	AUC
GT	pMCAR	1923	0.157 ± 0.002	0.176	0.000 ± 0.000	0.073	0.000 ± 0.000	0.654 ± 0.016	1.075 ± 0.007	0.733 ± 0.036
Ignore	pMCAR	1923	0.157 ± 0.002	0.188	$0.012 \pm 0.001^{***}$	0.078	$0.004 \pm 0.001^{***}$	0.648 ± 0.016	1.064 ± 0.007	0.729 ± 0.037
Ignore (weighted)	pMCAR	1923	0.157 ± 0.002	0.186	$0.010 \pm 0.002^{***}$	0.077	$0.003 \pm 0.000^{***}$	–	–	–
Ignore20	pMCAR	1117	0.157 ± 0.002	0.181	$0.005 \pm 0.001^{***}$	0.075	0.001 ± 0.001	0.655 ± 0.025	1.073 ± 0.012	0.733 ± 0.050
kNN	pMCAR	1923	0.157 ± 0.002	0.176	-0.000 ± 0.003	0.072	-0.001 ± 0.002	0.644 ± 0.016	1.071 ± 0.008	0.722 ± 0.038
RF	pMCAR	1923	0.157 ± 0.002	0.161	$-0.016 \pm 0.001^{***}$	0.068	$-0.006 \pm 0.001^{***}$	0.647 ± 0.016	1.076 ± 0.008	0.726 ± 0.038
Default (m=1)	pMCAR	1923	0.157 ± 0.002	0.213	$0.037 \pm 0.002^{***}$	0.087	$0.013 \pm 0.002^{***}$	0.630 ± 0.019	1.053 ± 0.007	$0.695 \pm 0.041^{***}$
Default (m=5)	pMCAR	1923	0.157 ± 0.002	0.216	$0.040 \pm 0.003^{***}$	0.133	$0.060 \pm 0.020^{**}$	0.631 ± 0.019	1.041 ± 0.007	$0.697 \pm 0.040^{***}$
MICE RF (m=1)	pMCAR	1923	0.157 ± 0.002	0.165	$-0.011 \pm 0.001^{***}$	0.067	$-0.006 \pm 0.000^{***}$	0.648 ± 0.016	1.078 ± 0.008	0.728 ± 0.037
MICE RF (m=5)	pMCAR	1923	0.157 ± 0.002	0.168	$-0.008 \pm 0.001^{***}$	0.068	$-0.005 \pm 0.001^{***}$	0.651 ± 0.016	1.078 ± 0.008	0.732 ± 0.037
CART (m=1)	pMCAR	1923	0.157 ± 0.002	0.174	$-0.002 \pm 0.001^{*}$	0.069	$-0.005 \pm 0.000^{***}$	0.649 ± 0.016	1.077 ± 0.008	0.730 ± 0.038
CART (m=5)	pMCAR	1923	0.157 ± 0.002	0.177	0.000 ± 0.001	0.073	-0.001 ± 0.002	0.652 ± 0.016	1.075 ± 0.008	0.733 ± 0.037
CART+Aux (m=1)	pMCAR	1923	0.157 ± 0.002	0.175	$-0.002 \pm 0.000^{***}$	0.071	$-0.002 \pm 0.000^{***}$	0.655 ± 0.016	1.078 ± 0.007	0.732 ± 0.037
CART+Aux (m=5)	pMCAR	1923	0.157 ± 0.002	0.177	0.000 ± 0.001	0.073	-0.000 ± 0.001	0.655 ± 0.016	1.076 ± 0.008	0.733 ± 0.037
GT	pMAR	1923	0.157 ± 0.003	0.176	0.000 ± 0.000	0.073	0.000 ± 0.000	0.654 ± 0.016	1.075 ± 0.007	0.733 ± 0.036
Ignore	pMAR	1923	0.157 ± 0.003	0.187	$0.011 \pm 0.001^{***}$	0.075	0.001 ± 0.001	0.649 ± 0.016	1.070 ± 0.008	0.732 ± 0.037
Ignore (weighted)	pMAR	1923	0.157 ± 0.003	0.188	$0.012 \pm 0.002^{***}$	0.075	$0.001 \pm 0.001^{*}$	–	–	–
Ignore20	pMAR	1117.4	0.157 ± 0.003	0.191	$0.005 \pm 0.001^{***}$	0.077	$0.003 \pm 0.001^{***}$	0.666 ± 0.023	1.078 ± 0.010	0.742 ± 0.047
kNN	pMAR	1923	0.157 ± 0.003	0.179	0.002 ± 0.003	0.074	0.000 ± 0.001	0.645 ± 0.016	1.071 ± 0.007	0.721 ± 0.038
RF	pMAR	1923	0.157 ± 0.003	0.162	$-0.015 \pm 0.001^{***}$	0.072	$-0.002 \pm 0.000^{***}$	0.645 ± 0.017	1.070 ± 0.007	0.728 ± 0.038
Default (m=1)	pMAR	1923	0.157 ± 0.003	0.214	$0.037 \pm 0.004^{***}$	0.076	0.002 ± 0.002	0.633 ± 0.019	1.065 ± 0.009	$0.695 \pm 0.042^{***}$
Default (m=5)	pMAR	1923	0.157 ± 0.003	0.216	$0.040 \pm 0.004^{***}$	0.121	$0.048 \pm 0.018^{**}$	0.634 ± 0.018	1.046 ± 0.007	$0.697 \pm 0.041^{***}$
MICE RF (m=1)	pMAR	1923	0.157 ± 0.003	0.166	$-0.010 \pm 0.001^{***}$	0.070	$-0.003 \pm 0.000^{***}$	0.646 ± 0.017	1.073 ± 0.008	0.728 ± 0.037
MICE RF (m=5)	pMAR	1923	0.157 ± 0.003	0.169	$-0.007 \pm 0.001^{***}$	0.071	$-0.003 \pm 0.001^{***}$	0.649 ± 0.016	1.074 ± 0.008	0.732 ± 0.037
CART (m=1)	pMAR	1923	0.157 ± 0.003	0.176	-0.001 ± 0.001	0.070	$-0.004 \pm 0.000^{***}$	0.650 ± 0.017	1.076 ± 0.008	0.730 ± 0.037
CART (m=5)	pMAR	1923	0.157 ± 0.003	0.178	$0.002 \pm 0.001^{*}$	0.073	-0.001 ± 0.001	0.652 ± 0.016	1.075 ± 0.008	0.733 ± 0.037
CART+Aux (m=1)	pMAR	1923	0.157 ± 0.003	0.175	$-0.001 \pm 0.000^{**}$	0.072	$-0.002 \pm 0.000^{***}$	0.654 ± 0.016	1.076 ± 0.007	0.732 ± 0.037
CART+Aux (m=5)	pMAR	1923	0.157 ± 0.003	0.177	0.001 ± 0.001	0.073	-0.000 ± 0.001	0.655 ± 0.016	1.075 ± 0.007	0.734 ± 0.037

Table A.5: Missingness Block Variables

Block	Variables
PFQ	PFQ061A,PFQ061D,PFQ061E,PFQ061G,PFQ061H,PFQ061I,PFQ061J,PFQ061K,PFQ061L,PFQ061P, PFQ061R,PFQ061T
RXD	RXDCOUNT
VIQ	VIQ071,VIQ051C,VIQ031
BPX	BPXMeanArterialPressure,BPXPulsePressure,BPXSY,BPXDI,BPXPLS
LB	LBXCRP,LBDVIDMS,LBDRBFSI,LBXNEPCT,LBXGH,LBXHGB,LBXMCVSI,LBXPLTSI,LBXRDW,LBXSNASI, LBDSALSI,LBDSBUSI,LBDSCASI,LBDSCRSI,LBDSGLSI,LBXSAPSI,LBDB12SI,LBDHDDSI,LBXSLDSI, LBXSC3SI,LBDSTPSI,LBDSIRSI,LBDSTBSI,LBDSTRSI,LBDSUASI,LBDSPHSI,LBDSCHSI
Misc	AUQ,BPQ020,DIQ010,HUQ010,HUQ020,HUQ050,HUQ071,KIQ022,KIQ046,MCQ160A,MCQ160C, MCQ160D,MCQ160E,MCQ160F,MCQ160M,MCQ220,OSQ010A,OSQ060,PFQ057,RDQ031

Table A.6: Demographics of Missingness Blocks

Block	N[1]	Size	Mean FI (SD)[2]	Mean Age (SD)	Males (%)	Deaths (%)
PFQ	5703	12	0.128 (0.073)***	41.5 (15.9)***	2722 (47.7)***	350 (6.1)*
RXD	4038	1	0.111 (0.054)***	39.8 (15.2)***	2133 (52.8)***	137 (3.4)***
VIQ	2540	3	0.129 (0.078)***	37.9 (14.2)***	1194 (47.0)***	162 (6.4)***
BPX	643	5	0.176 (0.102)***	54.0 (21.3)***	266 (41.4)***	143 (22.2)***
LB	916	27	0.154 (0.101)	49.9 (20.0)	420 (45.9)	135 (14.7)***
All	9307	68	0.144 (0.078)	49.9 (19.0)	4465 (48.0)	1016 (10.9)

[3]

[1] Number of individuals with this block missing.

[2] FI calculated using Ignore.

[3] p-values compare individuals missing the block versus not missing the block.

Table A.7: Imputed FI Bias Summary by Block - 15% pMCAR (1/2)

Imputation	Block	N	Missingness	Mean FI	Bias	Bias Rate	FI SD	SD Bias	SD Bias Rate
GT	All	1923	0.157 ± 0.002	0.176	0.000 ± 0.000	0.000 ± 0.000	0.073	0.000 ± 0.000	0.000 ± 0.000
Ignore	All	1923	0.157 ± 0.002	0.188	$0.012 \pm 0.001^{***}$	$0.076 \pm 0.001^{***}$	0.078	$0.004 \pm 0.001^{***}$	$0.029 \pm 0.001^{***}$
Ignore (weighted)	All	1923	0.157 ± 0.002	0.186	$0.010 \pm 0.002^{***}$	$0.062 \pm 0.001^{***}$	0.077	$0.003 \pm 0.000^{***}$	$0.021 \pm 0.000^{***}$
Ignore20	All	1117	0.157 ± 0.002	0.181	$0.005 \pm 0.001^{***}$	$0.031 \pm 0.001^{***}$	0.075	0.001 ± 0.001	$0.008 \pm 0.002^{***}$
kNN	All	1923	0.157 ± 0.002	0.176	-0.000 ± 0.003	$0.014 \pm 0.005^{*}$	0.072	-0.001 ± 0.002	-0.002 ± 0.002
RF	All	1923	0.157 ± 0.002	0.161	$-0.016 \pm 0.001^{***}$	$-0.101 \pm 0.001^{***}$	0.068	$-0.006 \pm 0.001^{***}$	$-0.035 \pm 0.001^{***}$
Default (m=1)	All	1923	0.157 ± 0.002	0.213	$0.037 \pm 0.002^{***}$	$0.219 \pm 0.005^{***}$	0.087	$0.013 \pm 0.002^{***}$	$0.081 \pm 0.002^{***}$
Default (m=5)	All	1923	0.157 ± 0.002	0.216	$0.040 \pm 0.003^{***}$	$0.238 \pm 0.005^{***}$	0.133	$0.060 \pm 0.020^{**}$	$0.388 \pm 0.024^{***}$
MICE RF (m=1)	All	1923	0.157 ± 0.002	0.165	$-0.011 \pm 0.001^{***}$	$-0.073 \pm 0.001^{***}$	0.067	$-0.006 \pm 0.000^{***}$	$-0.039 \pm 0.001^{***}$
MICE RF (m=5)	All	1923	0.157 ± 0.002	0.168	$-0.008 \pm 0.001^{***}$	$-0.055 \pm 0.001^{***}$	0.068	$-0.005 \pm 0.001^{***}$	$-0.032 \pm 0.001^{***}$
CART (m=1)	All	1923	0.157 ± 0.002	0.174	$-0.002 \pm 0.001^{*}$	$-0.015 \pm 0.001^{***}$	0.069	$-0.005 \pm 0.000^{***}$	$-0.031 \pm 0.001^{***}$
CART (m=5)	All	1923	0.157 ± 0.002	0.177	0.000 ± 0.001	$0.002 \pm 0.001^{*}$	0.073	-0.001 ± 0.002	$-0.006 \pm 0.002^{**}$
CART+Aux (m=1)	All	1923	0.157 ± 0.002	0.175	$-0.002 \pm 0.000^{***}$	$-0.011 \pm 0.001^{***}$	0.071	$-0.002 \pm 0.000^{***}$	$-0.014 \pm 0.000^{***}$
CART+Aux (m=5)	All	1923	0.157 ± 0.002	0.177	0.000 ± 0.001	$0.002 \pm 0.000^{***}$	0.073	-0.000 ± 0.001	-0.002 ± 0.001
GT	BPX	1923	0.064 ± 0.003	0.289	0.000 ± 0.000	0.000 ± 0.000	0.263	0.000 ± 0.000	0.000 ± 0.000
Ignore	BPX	1832.4	0.064 ± 0.003	0.287	-0.002 ± 0.002	$-0.033 \pm 0.004^{***}$	0.266	$0.004 \pm 0.001^{***}$	$0.056 \pm 0.003^{***}$
Ignore (weighted)	BPX	1832.4	0.064 ± 0.003	0.287	-0.002 ± 0.009	$-0.027 \pm 0.007^{***}$	0.267	$0.004 \pm 0.001^{***}$	$0.055 \pm 0.005^{***}$
Ignore20	BPX	1082.4	0.064 ± 0.003	0.286	-0.002 ± 0.002	$-0.031 \pm 0.005^{***}$	0.268	$0.006 \pm 0.003^{*}$	$0.076 \pm 0.008^{***}$
kNN	BPX	1923	0.064 ± 0.003	0.284	-0.004 ± 0.003	$-0.069 \pm 0.006^{***}$	0.259	$-0.003 \pm 0.001^{***}$	$-0.053 \pm 0.002^{***}$
RF	BPX	1923	0.064 ± 0.003	0.272	$-0.016 \pm 0.003^{***}$	$-0.260 \pm 0.005^{***}$	0.263	0.000 ± 0.001	0.003 ± 0.002
Default (m=1)	BPX	1923	0.064 ± 0.003	0.283	$-0.005 \pm 0.002^{*}$	$-0.080 \pm 0.005^{***}$	0.260	$-0.003 \pm 0.001^{***}$	$-0.048 \pm 0.002^{***}$
Default (m=5)	BPX	1923	0.064 ± 0.003	0.290	0.002 ± 0.003	$0.027 \pm 0.005^{***}$	0.272	0.009 ± 0.008	$0.171 \pm 0.028^{***}$
MICE RF (m=1)	BPX	1923	0.064 ± 0.003	0.283	$-0.005 \pm 0.003^{*}$	$-0.086 \pm 0.005^{***}$	0.259	$-0.003 \pm 0.001^{***}$	$-0.052 \pm 0.002^{***}$
MICE RF (m=5)	BPX	1923	0.064 ± 0.003	0.289	0.001 ± 0.003	0.005 ± 0.005	0.270	0.008 ± 0.008	$0.124 \pm 0.019^{***}$
CART (m=1)	BPX	1923	0.064 ± 0.003	0.284	-0.004 ± 0.003	$-0.071 \pm 0.007^{***}$	0.259	$-0.003 \pm 0.001^{***}$	$-0.049 \pm 0.002^{***}$
CART (m=5)	BPX	1923	0.064 ± 0.003	0.290	0.001 ± 0.003	$0.015 \pm 0.004^{**}$	0.273	0.010 ± 0.006	$0.154 \pm 0.018^{***}$
CART+Aux (m=1)	BPX	1923	0.064 ± 0.003	0.283	-0.005 ± 0.003	$-0.088 \pm 0.007^{***}$	0.259	$-0.003 \pm 0.001^{***}$	$-0.049 \pm 0.002^{***}$
CART+Aux (m=5)	BPX	1923	0.064 ± 0.003	0.290	0.002 ± 0.003	$0.017 \pm 0.006^{**}$	0.275	0.012 ± 0.007	$0.176 \pm 0.019^{***}$
GT	LB	1923	0.065 ± 0.003	0.147	0.000 ± 0.000	0.000 ± 0.000	0.087	0.000 ± 0.000	0.000 ± 0.000
Ignore	LB	1821.8	0.065 ± 0.003	0.147	0.000 ± 0.000	0.001 ± 0.001	0.089	$0.002 \pm 0.001^{**}$	$0.022 \pm 0.002^{***}$
Ignore (weighted)	LB	1821.8	0.065 ± 0.003	0.147	-0.000 ± 0.003	0.001 ± 0.002	0.089	0.001 ± 0.001	$0.017 \pm 0.003^{***}$
Ignore20	LB	1117	0.065 ± 0.003	0.146	0.000 ± 0.000	$0.001 \pm 0.000^{**}$	0.086	-0.001 ± 0.001	-0.008 ± 0.004
kNN	LB	1923	0.065 ± 0.003	0.140	$-0.007 \pm 0.002^{***}$	$-0.107 \pm 0.003^{***}$	0.088	0.001 ± 0.001	$0.022 \pm 0.004^{***}$
RF	LB	1923	0.065 ± 0.003	0.139	$-0.008 \pm 0.001^{***}$	$-0.117 \pm 0.002^{***}$	0.090	$0.002 \pm 0.000^{***}$	$0.037 \pm 0.001^{***}$
Default (m=1)	LB	1923	0.065 ± 0.003	0.142	$-0.005 \pm 0.001^{***}$	$-0.077 \pm 0.001^{***}$	0.088	$0.001 \pm 0.000^{**}$	$0.011 \pm 0.001^{***}$
Default (m=5)	LB	1923	0.065 ± 0.003	0.148	0.001 ± 0.001	$0.015 \pm 0.002^{***}$	0.089	$0.002 \pm 0.001^{**}$	$0.031 \pm 0.003^{***}$
MICE RF (m=1)	LB	1923	0.065 ± 0.003	0.141	$-0.006 \pm 0.001^{***}$	$-0.093 \pm 0.001^{***}$	0.088	$0.001 \pm 0.000^{***}$	$0.015 \pm 0.001^{***}$
MICE RF (m=5)	LB	1923	0.065 ± 0.003	0.147	-0.000 ± 0.001	0.001 ± 0.001	0.089	0.002 ± 0.002	$0.027 \pm 0.005^{***}$
CART (m=1)	LB	1923	0.065 ± 0.003	0.142	$-0.005 \pm 0.001^{***}$	$-0.080 \pm 0.002^{***}$	0.088	0.000 ± 0.000	$0.006 \pm 0.002^{***}$
CART (m=5)	LB	1923	0.065 ± 0.003	0.148	0.001 ± 0.001	$0.010 \pm 0.002^{***}$	0.090	0.003 ± 0.003	$0.039 \pm 0.007^{***}$
CART+Aux (m=1)	LB	1923	0.065 ± 0.003	0.143	$-0.004 \pm 0.001^{***}$	$-0.057 \pm 0.001^{***}$	0.088	0.000 ± 0.000	$0.004 \pm 0.001^{***}$
CART+Aux (m=5)	LB	1923	0.065 ± 0.003	0.147	0.000 ± 0.001	$0.004 \pm 0.001^{**}$	0.089	0.002 ± 0.002	$0.027 \pm 0.005^{***}$

Table A.8: Imputed FI Bias Summary by Block - 15% pMCAR (2/2)

Imputation	Block	N	Missingness	Mean FI	Bias	Bias Rate	FI SD	SD Bias	SD Bias Rate
GT	PFQ	1923	0.577 ± 0.010	0.095	0.000 ± 0.000	0.000 ± 0.000	0.125	0.000 ± 0.000	0.000 ± 0.000
Ignore	PFQ	837.1	0.577 ± 0.010	0.095	-0.000 ± 0.000	$-0.000 \pm 0.000^{***}$	0.127	0.002 ± 0.003	0.002 ± 0.001
Ignore (weighted)	PFQ	837.1	0.577 ± 0.010	0.095	0.000 ± 0.005	0.000 ± 0.001	0.128	0.002 ± 0.003	$0.003 \pm 0.001^{*}$
Ignore20	PFQ	777.8	0.577 ± 0.010	0.095	-0.000 ± 0.000	$-0.000 \pm 0.000^{*}$	0.127	0.002 ± 0.004	0.003 ± 0.001
kNN	PFQ	1923	0.577 ± 0.010	0.120	0.025 ± 0.014	$0.067 \pm 0.007^{***}$	0.123	-0.003 ± 0.014	$0.014 \pm 0.006^{*}$
RF	PFQ	1923	0.577 ± 0.010	0.042	$-0.053 \pm 0.003^{***}$	$-0.092 \pm 0.001^{***}$	0.095	$-0.031 \pm 0.002^{***}$	$-0.049 \pm 0.001^{***}$
Default (m=1)	PFQ	1923	0.577 ± 0.010	0.310	$0.215 \pm 0.014^{***}$	$0.349 \pm 0.007^{***}$	0.271	$0.145 \pm 0.008^{***}$	$0.257 \pm 0.004^{***}$
Default (m=5)	PFQ	1923	0.577 ± 0.010	0.310	$0.215 \pm 0.019^{***}$	$0.349 \pm 0.007^{***}$	0.647	$0.522 \pm 0.138^{***}$	$0.949 \pm 0.046^{***}$
MICE RF (m=1)	PFQ	1923	0.577 ± 0.010	0.053	$-0.041 \pm 0.002^{***}$	$-0.072 \pm 0.001^{***}$	0.091	$-0.034 \pm 0.002^{***}$	$-0.056 \pm 0.001^{***}$
MICE RF (m=5)	PFQ	1923	0.577 ± 0.010	0.053	$-0.041 \pm 0.003^{***}$	$-0.072 \pm 0.001^{***}$	0.099	$-0.026 \pm 0.005^{***}$	$-0.043 \pm 0.002^{***}$
CART (m=1)	PFQ	1923	0.577 ± 0.010	0.095	0.001 ± 0.004	0.001 ± 0.001	0.091	$-0.034 \pm 0.002^{***}$	$-0.057 \pm 0.001^{***}$
CART (m=5)	PFQ	1923	0.577 ± 0.010	0.095	0.001 ± 0.005	0.001 ± 0.001	0.140	0.014 ± 0.014	$0.026 \pm 0.004^{***}$
CART+Aux (m=1)	PFQ	1923	0.577 ± 0.010	0.096	0.001 ± 0.002	0.001 ± 0.001	0.108	$-0.018 \pm 0.003^{***}$	$-0.030 \pm 0.001^{***}$
CART+Aux (m=5)	PFQ	1923	0.577 ± 0.010	0.096	0.001 ± 0.002	0.001 ± 0.001	0.125	-0.001 ± 0.009	-0.002 ± 0.003
GT	RXD	1923	0.469 ± 0.006	0.207	0.000 ± 0.000	0.000 ± 0.000	0.140	0.000 ± 0.000	0.000 ± 0.000
Ignore	RXD	1021.6	0.469 ± 0.006	0.209	0.000 ± 0.000	0.000 ± 0.000	0.141	0.000 ± 0.001	0.001 ± 0.001
Ignore (weighted)	RXD	1021.6	0.469 ± 0.006	0.209	0.002 ± 0.006	0.003 ± 0.001	0.141	0.001 ± 0.002	-0.000 ± 0.001
Ignore20	RXD	801.6	0.469 ± 0.006	0.208	0.000 ± 0.000	0.000 ± 0.000	0.140	-0.001 ± 0.003	-0.000 ± 0.001
kNN	RXD	1923	0.469 ± 0.006	0.188	$-0.019 \pm 0.003^{***}$	$-0.041 \pm 0.001^{***}$	0.118	$-0.022 \pm 0.002^{***}$	$-0.045 \pm 0.001^{***}$
RF	RXD	1923	0.469 ± 0.006	0.192	$-0.015 \pm 0.004^{***}$	$-0.034 \pm 0.001^{***}$	0.133	$-0.007 \pm 0.002^{***}$	$-0.016 \pm 0.001^{***}$
Default (m=1)	RXD	1923	0.469 ± 0.006	0.219	$0.012 \pm 0.003^{***}$	$0.024 \pm 0.001^{***}$	0.132	$-0.008 \pm 0.002^{**}$	$-0.018 \pm 0.001^{***}$
Default (m=5)	RXD	1923	0.469 ± 0.006	0.219	$0.012 \pm 0.005^{**}$	$0.024 \pm 0.001^{***}$	0.197	$0.057 \pm 0.025^{*}$	$0.128 \pm 0.010^{***}$
MICE RF (m=1)	RXD	1923	0.469 ± 0.006	0.199	$-0.008 \pm 0.004^{*}$	$-0.019 \pm 0.001^{***}$	0.113	$-0.027 \pm 0.001^{***}$	$-0.056 \pm 0.001^{***}$
MICE RF (m=5)	RXD	1923	0.469 ± 0.006	0.199	-0.008 ± 0.004	$-0.019 \pm 0.001^{***}$	0.157	0.016 ± 0.012	$0.041 \pm 0.005^{***}$
CART (m=1)	RXD	1923	0.469 ± 0.006	0.207	-0.000 ± 0.003	-0.001 ± 0.001	0.119	$-0.021 \pm 0.001^{***}$	$-0.044 \pm 0.001^{***}$
CART (m=5)	RXD	1923	0.469 ± 0.006	0.207	-0.000 ± 0.004	-0.001 ± 0.001	0.172	0.032 ± 0.025	$0.060 \pm 0.009^{***}$
CART+Aux (m=1)	RXD	1923	0.469 ± 0.006	0.207	-0.000 ± 0.004	-0.001 ± 0.001	0.119	$-0.022 \pm 0.002^{***}$	$-0.045 \pm 0.001^{***}$
CART+Aux (m=5)	RXD	1923	0.469 ± 0.006	0.207	-0.000 ± 0.005	-0.001 ± 0.001	0.183	0.042 ± 0.026	$0.074 \pm 0.009^{***}$
GT	VIQ	1923	0.278 ± 0.011	0.193	0.000 ± 0.000	0.000 ± 0.000	0.189	0.000 ± 0.000	0.000 ± 0.000
Ignore	VIQ	1406.1	0.278 ± 0.011	0.194	0.001 ± 0.001	$0.004 \pm 0.000^{***}$	0.191	0.002 ± 0.002	$0.009 \pm 0.001^{***}$
Ignore (weighted)	VIQ	1406.1	0.278 ± 0.011	0.194	0.001 ± 0.006	$0.004 \pm 0.002^{*}$	0.192	0.003 ± 0.002	$0.010 \pm 0.002^{***}$
Ignore20	VIQ	1026.3	0.278 ± 0.011	0.193	0.001 ± 0.001	$0.005 \pm 0.000^{***}$	0.192	0.003 ± 0.003	$0.011 \pm 0.003^{***}$
kNN	VIQ	1923	0.278 ± 0.011	0.172	$-0.020 \pm 0.004^{***}$	$-0.072 \pm 0.002^{***}$	0.173	$-0.016 \pm 0.002^{***}$	$-0.056 \pm 0.002^{***}$
RF	VIQ	1923	0.278 ± 0.011	0.162	$-0.030 \pm 0.003^{***}$	$-0.110 \pm 0.002^{***}$	0.170	$-0.019 \pm 0.002^{***}$	$-0.065 \pm 0.002^{***}$
Default (m=1)	VIQ	1923	0.278 ± 0.011	0.207	$0.014 \pm 0.004^{***}$	$0.051 \pm 0.002^{***}$	0.176	$-0.013 \pm 0.003^{***}$	$-0.045 \pm 0.002^{***}$
Default (m=5)	VIQ	1923	0.278 ± 0.011	0.207	$0.014 \pm 0.005^{**}$	$0.051 \pm 0.002^{***}$	0.226	$0.037 \pm 0.016^{*}$	$0.150 \pm 0.013^{***}$
MICE RF (m=1)	VIQ	1923	0.278 ± 0.011	0.176	$-0.017 \pm 0.003^{***}$	$-0.063 \pm 0.002^{***}$	0.166	$-0.022 \pm 0.002^{***}$	$-0.078 \pm 0.002^{***}$
MICE RF (m=5)	VIQ	1923	0.278 ± 0.011	0.176	$-0.017 \pm 0.004^{***}$	$-0.063 \pm 0.002^{***}$	0.200	0.011 ± 0.017	$0.031 \pm 0.010^{**}$
CART (m=1)	VIQ	1923	0.278 ± 0.011	0.192	-0.000 ± 0.003	-0.000 ± 0.002	0.168	$-0.020 \pm 0.002^{***}$	$-0.072 \pm 0.002^{***}$
CART (m=5)	VIQ	1923	0.278 ± 0.011	0.192	-0.000 ± 0.004	-0.000 ± 0.002	0.214	0.026 ± 0.018	$0.087 \pm 0.011^{***}$
CART+Aux (m=1)	VIQ	1923	0.278 ± 0.011	0.192	-0.000 ± 0.002	0.000 ± 0.001	0.174	$-0.014 \pm 0.002^{***}$	$-0.049 \pm 0.002^{***}$
CART+Aux (m=5)	VIQ	1923	0.278 ± 0.011	0.192	-0.000 ± 0.003	0.000 ± 0.001	0.201	0.013 ± 0.008	$0.051 \pm 0.006^{***}$

Table A.9: Imputed FI Bias Summary by Block - 15% pMAR (1/2)

Imputation	Block	N	Missingness	Mean FI	Bias	Bias Rate	FI SD	SD Bias	SD Bias Rate
GT	All	1923	0.157 ± 0.003	0.176	0.000 ± 0.000	0.000 ± 0.000	0.073	0.000 ± 0.000	0.000 ± 0.000
Ignore	All	1923	0.157 ± 0.003	0.187	$0.011 \pm 0.001^{***}$	$0.067 \pm 0.001^{***}$	0.075	0.001 ± 0.001	$0.004 \pm 0.001^{**}$
Ignore (weighted)	All	1923	0.157 ± 0.003	0.188	$0.012 \pm 0.002^{***}$	$0.076 \pm 0.001^{***}$	0.075	$0.001 \pm 0.001^{*}$	$0.007 \pm 0.001^{***}$
Ignore20	All	1117.4	0.157 ± 0.003	0.191	$0.005 \pm 0.001^{***}$	$0.029 \pm 0.001^{***}$	0.077	$0.003 \pm 0.001^{***}$	$0.020 \pm 0.001^{***}$
kNN	All	1923	0.157 ± 0.003	0.179	0.002 ± 0.003	$0.020 \pm 0.004^{***}$	0.074	0.000 ± 0.001	0.002 ± 0.001
RF	All	1923	0.157 ± 0.003	0.162	$-0.015 \pm 0.001^{***}$	$-0.092 \pm 0.001^{***}$	0.072	$-0.002 \pm 0.000^{***}$	-0.004 ± 0.002
Default (m=1)	All	1923	0.157 ± 0.003	0.214	$0.037 \pm 0.004^{***}$	$0.225 \pm 0.005^{***}$	0.076	0.002 ± 0.002	0.006 ± 0.003
Default (m=5)	All	1923	0.157 ± 0.003	0.216	$0.040 \pm 0.004^{***}$	$0.244 \pm 0.005^{***}$	0.121	$0.048 \pm 0.018^{**}$	$0.279 \pm 0.022^{***}$
MICE RF (m=1)	All	1923	0.157 ± 0.003	0.166	$-0.010 \pm 0.001^{***}$	$-0.063 \pm 0.001^{***}$	0.070	$-0.003 \pm 0.000^{***}$	$-0.014 \pm 0.002^{***}$
MICE RF (m=5)	All	1923	0.157 ± 0.003	0.169	$-0.007 \pm 0.001^{***}$	$-0.044 \pm 0.001^{***}$	0.071	$-0.003 \pm 0.001^{***}$	$-0.013 \pm 0.002^{***}$
CART (m=1)	All	1923	0.157 ± 0.003	0.176	-0.001 ± 0.001	$-0.004 \pm 0.001^{***}$	0.070	$-0.004 \pm 0.000^{***}$	$-0.021 \pm 0.001^{***}$
CART (m=5)	All	1923	0.157 ± 0.003	0.178	$0.002 \pm 0.001^{*}$	$0.013 \pm 0.001^{***}$	0.073	-0.001 ± 0.001	$-0.004 \pm 0.002^{**}$
CART+Aux (m=1)	All	1923	0.157 ± 0.003	0.175	$-0.001 \pm 0.000^{**}$	$-0.008 \pm 0.000^{***}$	0.072	$-0.002 \pm 0.000^{***}$	$-0.010 \pm 0.001^{***}$
CART+Aux (m=5)	All	1923	0.157 ± 0.003	0.177	0.001 ± 0.001	$0.005 \pm 0.001^{***}$	0.073	-0.000 ± 0.001	$-0.002 \pm 0.001^{*}$
GT	BPX	1923	0.063 ± 0.006	0.289	0.000 ± 0.000	0.000 ± 0.000	0.263	0.000 ± 0.000	0.000 ± 0.000
Ignore	BPX	1832.7	0.063 ± 0.006	0.287	-0.001 ± 0.002	$-0.023 \pm 0.004^{***}$	0.267	$0.004 \pm 0.002^{*}$	$0.063 \pm 0.005^{***}$
Ignore (weighted)	BPX	1832.7	0.063 ± 0.006	0.291	0.002 ± 0.009	$0.057 \pm 0.008^{***}$	0.267	0.004 ± 0.002	$0.064 \pm 0.006^{***}$
Ignore20	BPX	1085.6	0.063 ± 0.006	0.300	-0.002 ± 0.003	$-0.023 \pm 0.006^{***}$	0.269	$0.006 \pm 0.003^{*}$	$0.097 \pm 0.007^{***}$
kNN	BPX	1923	0.063 ± 0.006	0.285	-0.004 ± 0.003	$-0.066 \pm 0.006^{***}$	0.259	$-0.003 \pm 0.001^{**}$	$-0.049 \pm 0.003^{***}$
RF	BPX	1923	0.063 ± 0.006	0.272	$-0.017 \pm 0.003^{***}$	$-0.270 \pm 0.004^{***}$	0.263	0.000 ± 0.001	0.005 ± 0.003
Default (m=1)	BPX	1923	0.063 ± 0.006	0.283	$-0.006 \pm 0.003^{*}$	$-0.085 \pm 0.006^{***}$	0.260	$-0.003 \pm 0.001^{**}$	$-0.046 \pm 0.003^{***}$
Default (m=5)	BPX	1923	0.063 ± 0.006	0.290	0.001 ± 0.003	$0.026 \pm 0.005^{***}$	0.269	0.006 ± 0.004	$0.177 \pm 0.039^{***}$
MICE RF (m=1)	BPX	1923	0.063 ± 0.006	0.283	$-0.006 \pm 0.002^{*}$	$-0.088 \pm 0.006^{***}$	0.259	$-0.003 \pm 0.001^{***}$	$-0.050 \pm 0.003^{***}$
MICE RF (m=5)	BPX	1923	0.063 ± 0.006	0.289	0.000 ± 0.003	0.009 ± 0.005	0.272	0.009 ± 0.007	$0.159 \pm 0.017^{***}$
CART (m=1)	BPX	1923	0.063 ± 0.006	0.284	-0.004 ± 0.003	$-0.065 \pm 0.006^{***}$	0.259	$-0.003 \pm 0.001^{***}$	$-0.050 \pm 0.003^{***}$
CART (m=5)	BPX	1923	0.063 ± 0.006	0.290	0.001 ± 0.003	$0.021 \pm 0.005^{***}$	0.270	0.007 ± 0.004	$0.134 \pm 0.015^{***}$
CART+Aux (m=1)	BPX	1923	0.063 ± 0.006	0.283	$-0.005 \pm 0.002^{*}$	$-0.080 \pm 0.006^{***}$	0.259	$-0.003 \pm 0.001^{**}$	$-0.050 \pm 0.003^{***}$
CART+Aux (m=5)	BPX	1923	0.063 ± 0.006	0.289	0.001 ± 0.003	$0.016 \pm 0.005^{**}$	0.276	0.014 ± 0.012	$0.193 \pm 0.028^{***}$
GT	LB	1923	0.065 ± 0.005	0.147	0.000 ± 0.000	0.000 ± 0.000	0.087	0.000 ± 0.000	0.000 ± 0.000
Ignore	LB	1823	0.065 ± 0.005	0.148	0.000 ± 0.000	0.000 ± 0.001	0.089	0.002 ± 0.001	$0.021 \pm 0.002^{***}$
Ignore (weighted)	LB	1823	0.065 ± 0.005	0.150	0.003 ± 0.003	$0.050 \pm 0.003^{***}$	0.088	0.001 ± 0.001	$0.016 \pm 0.002^{***}$
Ignore20	LB	1117.4	0.065 ± 0.005	0.156	0.000 ± 0.000	$0.001 \pm 0.000^{**}$	0.091	$0.004 \pm 0.001^{***}$	$0.063 \pm 0.003^{***}$
kNN	LB	1923	0.065 ± 0.005	0.141	$-0.006 \pm 0.001^{***}$	$-0.096 \pm 0.004^{***}$	0.089	0.001 ± 0.001	$0.023 \pm 0.004^{***}$
RF	LB	1923	0.065 ± 0.005	0.140	$-0.007 \pm 0.001^{***}$	$-0.110 \pm 0.001^{***}$	0.090	$0.003 \pm 0.000^{***}$	$0.044 \pm 0.001^{***}$
Default (m=1)	LB	1923	0.065 ± 0.005	0.142	$-0.005 \pm 0.001^{***}$	$-0.074 \pm 0.001^{***}$	0.088	0.001 ± 0.000	$0.014 \pm 0.002^{***}$
Default (m=5)	LB	1923	0.065 ± 0.005	0.148	0.001 ± 0.001	$0.018 \pm 0.001^{***}$	0.090	0.002 ± 0.002	$0.039 \pm 0.005^{***}$
MICE RF (m=1)	LB	1923	0.065 ± 0.005	0.141	$-0.006 \pm 0.001^{***}$	$-0.086 \pm 0.001^{***}$	0.088	$0.001 \pm 0.000^{**}$	$0.020 \pm 0.002^{***}$
MICE RF (m=5)	LB	1923	0.065 ± 0.005	0.148	0.001 ± 0.001	$0.009 \pm 0.001^{***}$	0.088	0.001 ± 0.002	$0.023 \pm 0.006^{***}$
CART (m=1)	LB	1923	0.065 ± 0.005	0.142	$-0.005 \pm 0.001^{***}$	$-0.074 \pm 0.001^{***}$	0.088	0.001 ± 0.001	$0.011 \pm 0.002^{***}$
CART (m=5)	LB	1923	0.065 ± 0.005	0.148	0.001 ± 0.001	$0.014 \pm 0.001^{***}$	0.089	0.001 ± 0.001	$0.025 \pm 0.004^{***}$
CART+Aux (m=1)	LB	1923	0.065 ± 0.005	0.144	$-0.003 \pm 0.001^{***}$	$-0.053 \pm 0.001^{***}$	0.088	0.000 ± 0.000	$0.007 \pm 0.002^{***}$
CART+Aux (m=5)	LB	1923	0.065 ± 0.005	0.148	0.001 ± 0.001	$0.009 \pm 0.002^{***}$	0.089	0.001 ± 0.001	$0.021 \pm 0.003^{***}$

Table A.10: Imputed FI Bias Summary by Block - 15% pMAR (2/2)

Imputation	Block	N	Missingness	Mean FI	Bias	Bias Rate	FI SD	SD Bias	SD Bias Rate
GT	PFQ	1923	0.583 ± 0.012	0.095	0.000 ± 0.000	0.000 ± 0.000	0.125	0.000 ± 0.000	0.000 ± 0.000
Ignore	PFQ	824.5	0.583 ± 0.012	0.107	-0.000 ± 0.001	$-0.001 \pm 0.000^{***}$	0.134	0.009 ± 0.005	$0.016 \pm 0.001^{***}$
Ignore (weighted)	PFQ	824.5	0.583 ± 0.012	0.109	$0.014 \pm 0.006^{*}$	$0.025 \pm 0.001^{***}$	0.134	0.009 ± 0.005	$0.017 \pm 0.001^{***}$
Ignore20	PFQ	763.5	0.583 ± 0.012	0.109	-0.000 ± 0.001	$-0.001 \pm 0.000^{***}$	0.136	$0.010 \pm 0.005^{*}$	$0.019 \pm 0.001^{***}$
kNN	PFQ	1923	0.583 ± 0.012	0.132	$0.037 \pm 0.018^{*}$	$0.073 \pm 0.006^{***}$	0.133	0.007 ± 0.011	$0.024 \pm 0.004^{***}$
RF	PFQ	1923	0.583 ± 0.012	0.046	$-0.049 \pm 0.003^{***}$	$-0.081 \pm 0.001^{***}$	0.102	$-0.024 \pm 0.004^{***}$	$-0.034 \pm 0.002^{***}$
Default (m=1)	PFQ	1923	0.583 ± 0.012	0.313	$0.218 \pm 0.020^{***}$	$0.357 \pm 0.007^{***}$	0.257	$0.132 \pm 0.011^{***}$	$0.237 \pm 0.005^{***}$
Default (m=5)	PFQ	1923	0.583 ± 0.012	0.313	$0.218 \pm 0.023^{***}$	$0.357 \pm 0.007^{***}$	0.591	$0.465 \pm 0.112^{***}$	$0.817 \pm 0.044^{***}$
MICE RF (m=1)	PFQ	1923	0.583 ± 0.012	0.059	$-0.036 \pm 0.003^{***}$	$-0.060 \pm 0.001^{***}$	0.098	$-0.028 \pm 0.004^{***}$	$-0.042 \pm 0.002^{***}$
MICE RF (m=5)	PFQ	1923	0.583 ± 0.012	0.059	$-0.036 \pm 0.003^{***}$	$-0.060 \pm 0.001^{***}$	0.107	$-0.018 \pm 0.003^{***}$	$-0.027 \pm 0.002^{***}$
CART (m=1)	PFQ	1923	0.583 ± 0.012	0.103	$0.008 \pm 0.003^{*}$	$0.014 \pm 0.001^{***}$	0.096	$-0.029 \pm 0.004^{***}$	$-0.046 \pm 0.001^{***}$
CART (m=5)	PFQ	1923	0.583 ± 0.012	0.103	0.008 ± 0.004	$0.014 \pm 0.001^{***}$	0.145	0.019 ± 0.019	$0.031 \pm 0.005^{***}$
CART+Aux (m=1)	PFQ	1923	0.583 ± 0.012	0.097	0.002 ± 0.002	$0.004 \pm 0.001^{***}$	0.110	$-0.016 \pm 0.004^{***}$	$-0.025 \pm 0.001^{***}$
CART+Aux (m=5)	PFQ	1923	0.583 ± 0.012	0.097	0.002 ± 0.003	$0.004 \pm 0.001^{***}$	0.129	0.004 ± 0.010	0.005 ± 0.003
GT	RXD	1923	0.472 ± 0.011	0.207	0.000 ± 0.000	0.000 ± 0.000	0.140	0.000 ± 0.000	0.000 ± 0.000
Ignore	RXD	1014.7	0.472 ± 0.011	0.218	0.000 ± 0.000	0.000 ± 0.000	0.146	0.006 ± 0.003	$0.013 \pm 0.001^{***}$
Ignore (weighted)	RXD	1014.7	0.472 ± 0.011	0.222	$0.015 \pm 0.006^{*}$	$0.036 \pm 0.002^{***}$	0.146	$0.006 \pm 0.003^{*}$	$0.012 \pm 0.001^{***}$
Ignore20	RXD	800.9	0.472 ± 0.011	0.226	0.000 ± 0.000	0.000 ± 0.000	0.149	$0.009 \pm 0.002^{***}$	$0.020 \pm 0.001^{***}$
kNN	RXD	1923	0.472 ± 0.011	0.189	$-0.018 \pm 0.003^{***}$	$-0.037 \pm 0.001^{***}$	0.122	$-0.018 \pm 0.003^{***}$	$-0.035 \pm 0.001^{***}$
RF	RXD	1923	0.472 ± 0.011	0.190	$-0.017 \pm 0.003^{***}$	$-0.038 \pm 0.001^{***}$	0.133	-0.007 ± 0.004	$-0.013 \pm 0.001^{***}$
Default (m=1)	RXD	1923	0.472 ± 0.011	0.217	$0.010 \pm 0.003^{***}$	$0.020 \pm 0.001^{***}$	0.130	$-0.010 \pm 0.002^{***}$	$-0.023 \pm 0.001^{***}$
Default (m=5)	RXD	1923	0.472 ± 0.011	0.217	$0.010 \pm 0.004^{*}$	$0.020 \pm 0.001^{***}$	0.190	$0.050 \pm 0.025^{*}$	$0.098 \pm 0.008^{***}$
MICE RF (m=1)	RXD	1923	0.472 ± 0.011	0.202	-0.005 ± 0.004	$-0.008 \pm 0.001^{***}$	0.117	$-0.023 \pm 0.003^{***}$	$-0.046 \pm 0.001^{***}$
MICE RF (m=5)	RXD	1923	0.472 ± 0.011	0.202	-0.005 ± 0.005	$-0.008 \pm 0.001^{***}$	0.163	$0.023 \pm 0.011^{*}$	$0.051 \pm 0.005^{***}$
CART (m=1)	RXD	1923	0.472 ± 0.011	0.210	0.003 ± 0.003	$0.009 \pm 0.001^{***}$	0.122	$-0.018 \pm 0.003^{***}$	$-0.037 \pm 0.001^{***}$
CART (m=5)	RXD	1923	0.472 ± 0.011	0.210	0.003 ± 0.004	$0.009 \pm 0.001^{***}$	0.170	0.030 ± 0.016	$0.058 \pm 0.006^{***}$
CART+Aux (m=1)	RXD	1923	0.472 ± 0.011	0.210	0.003 ± 0.003	$0.009 \pm 0.001^{***}$	0.122	$-0.018 \pm 0.003^{***}$	$-0.037 \pm 0.001^{***}$
CART+Aux (m=5)	RXD	1923	0.472 ± 0.011	0.210	0.003 ± 0.004	$0.009 \pm 0.001^{***}$	0.166	$0.026 \pm 0.010^{*}$	$0.064 \pm 0.006^{***}$
GT	VIQ	1923	0.274 ± 0.013	0.193	0.000 ± 0.000	0.000 ± 0.000	0.189	0.000 ± 0.000	0.000 ± 0.000
Ignore	VIQ	1415.1	0.274 ± 0.013	0.196	0.001 ± 0.001	$0.004 \pm 0.000^{***}$	0.193	$0.004 \pm 0.002^{**}$	$0.019 \pm 0.001^{***}$
Ignore (weighted)	VIQ	1415.1	0.274 ± 0.013	0.199	0.006 ± 0.007	$0.032 \pm 0.003^{***}$	0.194	$0.005 \pm 0.002^{*}$	$0.020 \pm 0.002^{***}$
Ignore20	VIQ	1034	0.274 ± 0.013	0.204	0.001 ± 0.001	$0.005 \pm 0.000^{***}$	0.198	$0.009 \pm 0.004^{*}$	$0.034 \pm 0.002^{***}$
kNN	VIQ	1923	0.274 ± 0.013	0.172	$-0.021 \pm 0.004^{***}$	$-0.071 \pm 0.002^{***}$	0.173	$-0.016 \pm 0.001^{***}$	$-0.052 \pm 0.002^{***}$
RF	VIQ	1923	0.274 ± 0.013	0.163	$-0.029 \pm 0.003^{***}$	$-0.102 \pm 0.002^{***}$	0.172	$-0.016 \pm 0.001^{***}$	$-0.055 \pm 0.002^{***}$
Default (m=1)	VIQ	1923	0.274 ± 0.013	0.205	$0.012 \pm 0.004^{**}$	$0.047 \pm 0.002^{***}$	0.176	$-0.013 \pm 0.002^{***}$	$-0.046 \pm 0.001^{***}$
Default (m=5)	VIQ	1923	0.274 ± 0.013	0.205	$0.012 \pm 0.005^{*}$	$0.047 \pm 0.002^{***}$	0.228	$0.039 \pm 0.016^{*}$	$0.141 \pm 0.012^{***}$
MICE RF (m=1)	VIQ	1923	0.274 ± 0.013	0.176	$-0.016 \pm 0.003^{***}$	$-0.055 \pm 0.002^{***}$	0.169	$-0.020 \pm 0.001^{***}$	$-0.068 \pm 0.002^{***}$
MICE RF (m=5)	VIQ	1923	0.274 ± 0.013	0.176	$-0.016 \pm 0.004^{***}$	$-0.055 \pm 0.002^{***}$	0.203	0.014 ± 0.020	$0.046 \pm 0.012^{***}$
CART (m=1)	VIQ	1923	0.274 ± 0.013	0.192	-0.001 ± 0.004	0.002 ± 0.002	0.170	$-0.019 \pm 0.001^{***}$	$-0.065 \pm 0.001^{***}$
CART (m=5)	VIQ	1923	0.274 ± 0.013	0.192	-0.001 ± 0.004	0.002 ± 0.002	0.205	0.016 ± 0.011	$0.077 \pm 0.009^{***}$
CART+Aux (m=1)	VIQ	1923	0.274 ± 0.013	0.192	-0.000 ± 0.003	0.001 ± 0.002	0.176	$-0.013 \pm 0.002^{***}$	$-0.046 \pm 0.001^{***}$
CART+Aux (m=5)	VIQ	1923	0.274 ± 0.013	0.192	-0.000 ± 0.004	0.001 ± 0.002	0.206	0.017 ± 0.012	$0.056 \pm 0.007^{***}$

Table A.11: Imputed FI Bias Summary by Block - 15% cMNAR (1/2)

Imputation	Block	N	Missingness	Mean FI	Bias	Bias Rate	FI SD	SD Bias	SD Bias Rate
GT	All	1923	0.150 ± 0.000	0.176	0.000 ± 0.000	0.000 ± 0.000	0.073	0.000 ± 0.000	0.000 ± 0.000
Ignore	All	1923	0.150 ± 0.000	0.198	$0.022 \pm 0.000^{***}$	$0.142 \pm 0.001^{***}$	0.080	$0.007 \pm 0.000^{***}$	$0.046 \pm 0.000^{***}$
Ignore (weighted)	All	1923	0.150 ± 0.000	0.199	$0.023 \pm 0.002^{***}$	$0.147 \pm 0.001^{***}$	0.081	$0.007 \pm 0.000^{***}$	$0.048 \pm 0.000^{***}$
Ignore20	All	1658.6	0.150 ± 0.000	0.202	$0.020 \pm 0.000^{***}$	$0.137 \pm 0.000^{***}$	0.081	$0.008 \pm 0.000^{***}$	$0.050 \pm 0.001^{***}$
kNN	All	1923	0.150 ± 0.000	0.179	$0.002 \pm 0.000^{***}$	$0.012 \pm 0.001^{***}$	0.072	$-0.001 \pm 0.000^{***}$	$-0.009 \pm 0.000^{***}$
RF	All	1923	0.150 ± 0.000	0.177	$0.001 \pm 0.000^{*}$	$0.004 \pm 0.000^{***}$	0.074	$0.001 \pm 0.000^{**}$	$0.006 \pm 0.000^{***}$
Default (m=1)	All	1923	0.150 ± 0.000	0.188	$0.011 \pm 0.000^{***}$	$0.072 \pm 0.001^{***}$	0.076	$0.002 \pm 0.000^{***}$	$0.015 \pm 0.001^{***}$
Default (m=5)	All	1923	0.150 ± 0.000	0.193	$0.017 \pm 0.001^{***}$	$0.109 \pm 0.001^{***}$	0.078	$0.004 \pm 0.001^{***}$	$0.029 \pm 0.001^{***}$
MICE RF (m=1)	All	1923	0.150 ± 0.000	0.181	$0.005 \pm 0.000^{***}$	$0.030 \pm 0.000^{***}$	0.074	0.000 ± 0.000	$0.003 \pm 0.000^{***}$
MICE RF (m=5)	All	1923	0.150 ± 0.000	0.188	$0.011 \pm 0.001^{***}$	$0.074 \pm 0.001^{***}$	0.076	$0.002 \pm 0.001^{**}$	$0.017 \pm 0.001^{***}$
CART (m=1)	All	1923	0.150 ± 0.000	0.185	$0.008 \pm 0.000^{***}$	$0.055 \pm 0.000^{***}$	0.075	$0.001 \pm 0.000^{***}$	$0.009 \pm 0.001^{***}$
CART (m=5)	All	1923	0.150 ± 0.000	0.190	$0.014 \pm 0.001^{***}$	$0.092 \pm 0.001^{***}$	0.077	$0.004 \pm 0.001^{***}$	$0.025 \pm 0.001^{***}$
CART+Aux (m=1)	All	1923	0.150 ± 0.000	0.173	$-0.003 \pm 0.001^{***}$	$-0.021 \pm 0.001^{***}$	0.072	$-0.002 \pm 0.000^{***}$	$-0.010 \pm 0.001^{***}$
CART+Aux (m=5)	All	1923	0.150 ± 0.000	0.176	0.000 ± 0.001	0.000 ± 0.001	0.074	0.000 ± 0.001	0.002 ± 0.001
GT	BPX	1923	0.140 ± 0.003	0.289	0.000 ± 0.000	0.000 ± 0.000	0.263	0.000 ± 0.000	0.000 ± 0.000
Ignore	BPX	1923	0.140 ± 0.003	0.316	$0.027 \pm 0.002^{***}$	$0.194 \pm 0.002^{***}$	0.289	$0.026 \pm 0.002^{***}$	$0.187 \pm 0.003^{***}$
Ignore (weighted)	BPX	1923	0.140 ± 0.003	0.317	$0.028 \pm 0.009^{**}$	$0.199 \pm 0.002^{***}$	0.289	$0.026 \pm 0.002^{***}$	$0.186 \pm 0.003^{***}$
Ignore20	BPX	1658.6	0.140 ± 0.003	0.321	$0.025 \pm 0.002^{***}$	$0.185 \pm 0.002^{***}$	0.288	$0.025 \pm 0.002^{***}$	$0.181 \pm 0.003^{***}$
kNN	BPX	1923	0.140 ± 0.003	0.299	$0.010 \pm 0.003^{***}$	$0.067 \pm 0.004^{***}$	0.256	$-0.007 \pm 0.001^{***}$	$-0.044 \pm 0.002^{***}$
RF	BPX	1923	0.140 ± 0.003	0.288	-0.000 ± 0.001	-0.002 ± 0.001	0.263	0.000 ± 0.001	$0.003 \pm 0.001^{*}$
Default (m=1)	BPX	1923	0.140 ± 0.003	0.290	0.002 ± 0.001	$0.013 \pm 0.001^{***}$	0.263	0.001 ± 0.001	$0.005 \pm 0.001^{***}$
Default (m=5)	BPX	1923	0.140 ± 0.003	0.295	$0.007 \pm 0.002^{***}$	$0.048 \pm 0.001^{***}$	0.269	0.006 ± 0.003	$0.040 \pm 0.004^{***}$
MICE RF (m=1)	BPX	1923	0.140 ± 0.003	0.292	$0.004 \pm 0.001^{*}$	$0.025 \pm 0.002^{***}$	0.264	0.001 ± 0.001	$0.009 \pm 0.002^{***}$
MICE RF (m=5)	BPX	1923	0.140 ± 0.003	0.303	$0.015 \pm 0.002^{***}$	$0.101 \pm 0.002^{***}$	0.273	$0.010 \pm 0.005^{*}$	$0.072 \pm 0.007^{***}$
CART (m=1)	BPX	1923	0.140 ± 0.003	0.292	$0.003 \pm 0.001^{**}$	$0.022 \pm 0.001^{***}$	0.264	0.002 ± 0.001	$0.011 \pm 0.001^{***}$
CART (m=5)	BPX	1923	0.140 ± 0.003	0.298	$0.009 \pm 0.002^{***}$	$0.066 \pm 0.001^{***}$	0.269	$0.006 \pm 0.002^{**}$	$0.047 \pm 0.003^{***}$
CART+Aux (m=1)	BPX	1923	0.140 ± 0.003	0.284	$-0.005 \pm 0.002^{**}$	$-0.034 \pm 0.002^{***}$	0.261	$-0.002 \pm 0.001^{*}$	$-0.011 \pm 0.001^{***}$
CART+Aux (m=5)	BPX	1923	0.140 ± 0.003	0.288	-0.000 ± 0.002	$-0.003 \pm 0.001^{*}$	0.267	0.004 ± 0.003	$0.030 \pm 0.004^{***}$
GT	LB	1923	0.162 ± 0.001	0.147	0.000 ± 0.000	0.000 ± 0.000	0.087	0.000 ± 0.000	0.000 ± 0.000
Ignore	LB	1923	0.162 ± 0.001	0.168	$0.021 \pm 0.001^{***}$	$0.125 \pm 0.001^{***}$	0.099	$0.012 \pm 0.000^{***}$	$0.070 \pm 0.001^{***}$
Ignore (weighted)	LB	1923	0.162 ± 0.001	0.168	$0.021 \pm 0.003^{***}$	$0.129 \pm 0.001^{***}$	0.099	$0.012 \pm 0.000^{***}$	$0.072 \pm 0.001^{***}$
Ignore20	LB	1658.6	0.162 ± 0.001	0.171	$0.020 \pm 0.001^{***}$	$0.121 \pm 0.001^{***}$	0.100	$0.012 \pm 0.001^{***}$	$0.073 \pm 0.001^{***}$
kNN	LB	1923	0.162 ± 0.001	0.147	0.000 ± 0.001	0.001 ± 0.001	0.086	$-0.001 \pm 0.000^{***}$	$-0.008 \pm 0.001^{***}$
RF	LB	1923	0.162 ± 0.001	0.147	0.000 ± 0.001	0.001 ± 0.001	0.089	$0.002 \pm 0.000^{***}$	$0.011 \pm 0.001^{***}$
Default (m=1)	LB	1923	0.162 ± 0.001	0.152	$0.005 \pm 0.001^{***}$	$0.031 \pm 0.001^{***}$	0.090	$0.002 \pm 0.001^{***}$	$0.014 \pm 0.001^{***}$
Default (m=5)	LB	1923	0.162 ± 0.001	0.166	$0.019 \pm 0.001^{***}$	$0.112 \pm 0.001^{***}$	0.095	$0.007 \pm 0.002^{***}$	$0.045 \pm 0.002^{***}$
MICE RF (m=1)	LB	1923	0.162 ± 0.001	0.149	$0.002 \pm 0.001^{***}$	$0.013 \pm 0.001^{***}$	0.088	$0.001 \pm 0.001^{*}$	$0.007 \pm 0.001^{***}$
MICE RF (m=5)	LB	1923	0.162 ± 0.001	0.164	$0.017 \pm 0.001^{***}$	$0.105 \pm 0.001^{***}$	0.094	$0.007 \pm 0.002^{***}$	$0.045 \pm 0.003^{***}$
CART (m=1)	LB	1923	0.162 ± 0.001	0.151	$0.004 \pm 0.001^{***}$	$0.025 \pm 0.001^{***}$	0.089	$0.002 \pm 0.001^{***}$	$0.011 \pm 0.001^{***}$
CART (m=5)	LB	1923	0.162 ± 0.001	0.164	$0.017 \pm 0.001^{***}$	$0.105 \pm 0.001^{***}$	0.095	$0.008 \pm 0.002^{***}$	$0.049 \pm 0.002^{***}$
CART+Aux (m=1)	LB	1923	0.162 ± 0.001	0.140	$-0.007 \pm 0.001^{***}$	$-0.045 \pm 0.001^{***}$	0.086	-0.001 ± 0.001	$-0.008 \pm 0.001^{***}$
CART+Aux (m=5)	LB	1923	0.162 ± 0.001	0.147	-0.000 ± 0.001	-0.000 ± 0.001	0.092	$0.005 \pm 0.002^{*}$	$0.024 \pm 0.002^{***}$

Table A.12: Imputed FI Bias Summary by Block - 15% cMNAR (2/2)

Imputation	Block	N	Missingness	Mean FI	Bias	Bias Rate	FI SD	SD Bias	SD Bias Rate
GT	PFQ	1923	0.158 ± 0.002	0.095	0.000 ± 0.000	0.000 ± 0.000	0.125	0.000 ± 0.000	0.000 ± 0.000
Ignore	PFQ	1923	0.158 ± 0.002	0.103	$0.008 \pm 0.001^{***}$	$0.052 \pm 0.001^{***}$	0.133	$0.007 \pm 0.001^{***}$	$0.046 \pm 0.001^{***}$
Ignore (weighted)	PFQ	1923	0.158 ± 0.002	0.104	$0.009 \pm 0.004^{*}$	$0.059 \pm 0.001^{***}$	0.134	$0.008 \pm 0.001^{***}$	$0.052 \pm 0.001^{***}$
Ignore20	PFQ	1658.6	0.158 ± 0.002	0.109	$0.008 \pm 0.001^{***}$	$0.051 \pm 0.001^{***}$	0.136	$0.011 \pm 0.001^{***}$	$0.061 \pm 0.002^{***}$
kNN	PFQ	1923	0.158 ± 0.002	0.093	$-0.002 \pm 0.001^{***}$	$-0.012 \pm 0.000^{***}$	0.122	$-0.003 \pm 0.001^{***}$	$-0.019 \pm 0.001^{***}$
RF	PFQ	1923	0.158 ± 0.002	0.094	-0.001 ± 0.001	$-0.005 \pm 0.000^{***}$	0.125	-0.000 ± 0.000	-0.001 ± 0.001
Default (m=1)	PFQ	1923	0.158 ± 0.002	0.105	$0.010 \pm 0.001^{***}$	$0.062 \pm 0.001^{***}$	0.127	$0.002 \pm 0.001^{***}$	$0.014 \pm 0.001^{***}$
Default (m=5)	PFQ	1923	0.158 ± 0.002	0.105	$0.010 \pm 0.001^{***}$	$0.062 \pm 0.001^{***}$	0.130	$0.005 \pm 0.001^{***}$	$0.031 \pm 0.001^{***}$
MICE RF (m=1)	PFQ	1923	0.158 ± 0.002	0.097	$0.002 \pm 0.000^{***}$	$0.014 \pm 0.000^{***}$	0.125	-0.000 ± 0.000	$-0.002 \pm 0.001^{*}$
MICE RF (m=5)	PFQ	1923	0.158 ± 0.002	0.097	$0.002 \pm 0.001^{***}$	$0.014 \pm 0.000^{***}$	0.127	0.001 ± 0.001	$0.009 \pm 0.001^{***}$
CART (m=1)	PFQ	1923	0.158 ± 0.002	0.101	$0.006 \pm 0.000^{***}$	$0.039 \pm 0.001^{***}$	0.126	$0.001 \pm 0.000^{*}$	$0.007 \pm 0.001^{***}$
CART (m=5)	PFQ	1923	0.158 ± 0.002	0.101	$0.006 \pm 0.001^{***}$	$0.039 \pm 0.001^{***}$	0.129	$0.004 \pm 0.002^{*}$	$0.022 \pm 0.002^{***}$
CART+Aux (m=1)	PFQ	1923	0.158 ± 0.002	0.095	-0.000 ± 0.001	-0.001 ± 0.001	0.123	$-0.003 \pm 0.001^{***}$	$-0.019 \pm 0.001^{***}$
CART+Aux (m=5)	PFQ	1923	0.158 ± 0.002	0.095	-0.000 ± 0.001	-0.001 ± 0.001	0.125	-0.001 ± 0.000	$-0.005 \pm 0.001^{***}$
GT	RXD	1923	0.055 ± 0.005	0.207	0.000 ± 0.000	0.000 ± 0.000	0.140	0.000 ± 0.000	0.000 ± 0.000
Ignore	RXD	1816.4	0.055 ± 0.005	0.208	0.000 ± 0.000	0.000 ± 0.000	0.141	0.001 ± 0.001	$0.008 \pm 0.002^{**}$
Ignore (weighted)	RXD	1816.4	0.055 ± 0.005	0.209	0.002 ± 0.005	$0.037 \pm 0.003^{***}$	0.141	0.000 ± 0.001	$0.006 \pm 0.003^{*}$
Ignore20	RXD	1576.6	0.055 ± 0.005	0.213	0.000 ± 0.000	0.000 ± 0.000	0.143	$0.003 \pm 0.001^{**}$	$0.036 \pm 0.005^{***}$
kNN	RXD	1923	0.055 ± 0.005	0.205	-0.002 ± 0.001	$-0.035 \pm 0.003^{***}$	0.138	$-0.002 \pm 0.001^{*}$	$-0.037 \pm 0.002^{***}$
RF	RXD	1923	0.055 ± 0.005	0.206	-0.001 ± 0.001	$-0.018 \pm 0.003^{***}$	0.140	-0.000 ± 0.001	$-0.009 \pm 0.003^{**}$
Default (m=1)	RXD	1923	0.055 ± 0.005	0.208	0.001 ± 0.001	$0.022 \pm 0.002^{***}$	0.139	-0.001 ± 0.001	$-0.020 \pm 0.002^{***}$
Default (m=5)	RXD	1923	0.055 ± 0.005	0.208	0.001 ± 0.001	$0.022 \pm 0.002^{***}$	0.146	0.005 ± 0.003	$0.089 \pm 0.010^{***}$
MICE RF (m=1)	RXD	1923	0.055 ± 0.005	0.207	0.000 ± 0.001	-0.002 ± 0.003	0.138	$-0.002 \pm 0.001^{**}$	$-0.042 \pm 0.002^{***}$
MICE RF (m=5)	RXD	1923	0.055 ± 0.005	0.207	0.000 ± 0.001	-0.002 ± 0.003	0.144	0.004 ± 0.003	$0.071 \pm 0.010^{***}$
CART (m=1)	RXD	1923	0.055 ± 0.005	0.208	0.001 ± 0.001	$0.010 \pm 0.003^{***}$	0.138	$-0.002 \pm 0.001^{*}$	$-0.037 \pm 0.002^{***}$
CART (m=5)	RXD	1923	0.055 ± 0.005	0.208	0.001 ± 0.001	$0.010 \pm 0.003^{***}$	0.143	0.003 ± 0.002	$0.063 \pm 0.010^{***}$
CART+Aux (m=1)	RXD	1923	0.055 ± 0.005	0.207	0.000 ± 0.002	0.003 ± 0.005	0.133	$-0.007 \pm 0.002^{***}$	$-0.125 \pm 0.006^{***}$
CART+Aux (m=5)	RXD	1923	0.055 ± 0.005	0.207	0.000 ± 0.002	0.003 ± 0.005	0.150	0.010 ± 0.007	$0.181 \pm 0.020^{***}$
GT	VIQ	1923	0.130 ± 0.005	0.193	0.000 ± 0.000	0.000 ± 0.000	0.189	0.000 ± 0.000	0.000 ± 0.000
Ignore	VIQ	1918.7	0.130 ± 0.005	0.209	$0.016 \pm 0.002^{***}$	$0.124 \pm 0.003^{***}$	0.207	$0.018 \pm 0.002^{***}$	$0.134 \pm 0.004^{***}$
Ignore (weighted)	VIQ	1918.7	0.130 ± 0.005	0.210	$0.018 \pm 0.007^{**}$	$0.134 \pm 0.003^{***}$	0.207	$0.018 \pm 0.002^{***}$	$0.136 \pm 0.003^{***}$
Ignore20	VIQ	1656.9	0.130 ± 0.005	0.214	$0.015 \pm 0.002^{***}$	$0.119 \pm 0.003^{***}$	0.208	$0.019 \pm 0.002^{***}$	$0.141 \pm 0.004^{***}$
kNN	VIQ	1923	0.130 ± 0.005	0.195	0.002 ± 0.002	$0.020 \pm 0.002^{***}$	0.186	-0.003 ± 0.002	$-0.023 \pm 0.002^{***}$
RF	VIQ	1923	0.130 ± 0.005	0.193	0.001 ± 0.002	$0.005 \pm 0.002^{**}$	0.185	$-0.003 \pm 0.002^{*}$	$-0.025 \pm 0.002^{***}$
Default (m=1)	VIQ	1923	0.130 ± 0.005	0.211	$0.019 \pm 0.002^{***}$	$0.141 \pm 0.002^{***}$	0.187	-0.001 ± 0.001	$-0.013 \pm 0.002^{***}$
Default (m=5)	VIQ	1923	0.130 ± 0.005	0.211	$0.019 \pm 0.003^{***}$	$0.141 \pm 0.002^{***}$	0.208	$0.019 \pm 0.008^{*}$	$0.156 \pm 0.012^{***}$
MICE RF (m=1)	VIQ	1923	0.130 ± 0.005	0.201	$0.008 \pm 0.001^{***}$	$0.065 \pm 0.002^{***}$	0.184	$-0.005 \pm 0.001^{***}$	$-0.037 \pm 0.001^{***}$
MICE RF (m=5)	VIQ	1923	0.130 ± 0.005	0.201	$0.008 \pm 0.002^{***}$	$0.065 \pm 0.002^{***}$	0.199	$0.010 \pm 0.004^{**}$	$0.075 \pm 0.007^{***}$
CART (m=1)	VIQ	1923	0.130 ± 0.005	0.208	$0.016 \pm 0.002^{***}$	$0.121 \pm 0.002^{***}$	0.185	$-0.003 \pm 0.001^{**}$	$-0.025 \pm 0.002^{***}$
CART (m=5)	VIQ	1923	0.130 ± 0.005	0.208	$0.016 \pm 0.002^{***}$	$0.121 \pm 0.002^{***}$	0.203	$0.014 \pm 0.005^{**}$	$0.112 \pm 0.008^{***}$
CART+Aux (m=1)	VIQ	1923	0.130 ± 0.005	0.193	-0.000 ± 0.002	0.000 ± 0.002	0.181	$-0.007 \pm 0.001^{***}$	$-0.057 \pm 0.003^{***}$
CART+Aux (m=5)	VIQ	1923	0.130 ± 0.005	0.193	-0.000 ± 0.003	0.000 ± 0.002	0.195	0.007 ± 0.005	$0.057 \pm 0.009^{***}$

Table A.13: Imputed FI Bias Summary by Block - 15% cMCAR (1/2)

Imputation	Block	N	Missingness	Mean FI	Bias	Bias Rate	FI SD	SD Bias	SD Bias Rate
GT	All	1923	0.150 ± 0.000	0.176	0.000 ± 0.000	0.000 ± 0.000	0.073	0.000 ± 0.000	0.000 ± 0.000
Ignore	All	1923	0.150 ± 0.000	0.176	0.000 ± 0.000	-0.000 ± 0.001	0.076	$0.002 \pm 0.000^{***}$	$0.014 \pm 0.001^{***}$
Ignore (weighted)	All	1923	0.150 ± 0.000	0.176	0.000 ± 0.002	-0.000 ± 0.001	0.075	$0.002 \pm 0.000^{***}$	$0.013 \pm 0.001^{***}$
Ignore20	All	1670.3	0.150 ± 0.000	0.176	0.000 ± 0.000	-0.000 ± 0.001	0.075	$0.002 \pm 0.001^{**}$	$0.013 \pm 0.001^{***}$
kNN	All	1923	0.150 ± 0.000	0.179	$0.002 \pm 0.000^{***}$	$0.012 \pm 0.001^{***}$	0.072	$-0.001 \pm 0.000^{***}$	$-0.009 \pm 0.000^{***}$
RF	All	1923	0.150 ± 0.000	0.177	$0.001 \pm 0.000^{*}$	$0.004 \pm 0.000^{***}$	0.074	$0.001 \pm 0.000^{**}$	$0.006 \pm 0.000^{***}$
Default (m=1)	All	1923	0.150 ± 0.000	0.188	$0.011 \pm 0.000^{***}$	$0.072 \pm 0.001^{***}$	0.076	$0.002 \pm 0.000^{***}$	$0.015 \pm 0.000^{***}$
Default (m=5)	All	1923	0.150 ± 0.000	0.193	$0.017 \pm 0.001^{***}$	$0.109 \pm 0.001^{***}$	0.078	$0.005 \pm 0.001^{***}$	$0.030 \pm 0.001^{***}$
MICE RF (m=1)	All	1923	0.150 ± 0.000	0.181	$0.004 \pm 0.000^{***}$	$0.029 \pm 0.000^{***}$	0.074	0.000 ± 0.000	$0.003 \pm 0.000^{***}$
MICE RF (m=5)	All	1923	0.150 ± 0.000	0.188	$0.011 \pm 0.001^{***}$	$0.073 \pm 0.001^{***}$	0.076	$0.002 \pm 0.001^{*}$	$0.017 \pm 0.002^{***}$
CART (m=1)	All	1923	0.150 ± 0.000	0.172	$-0.004 \pm 0.000^{***}$	$-0.029 \pm 0.000^{***}$	0.072	$-0.002 \pm 0.000^{***}$	$-0.012 \pm 0.000^{***}$
CART (m=5)	All	1923	0.150 ± 0.000	0.177	0.000 ± 0.001	$0.002 \pm 0.001^{**}$	0.073	-0.000 ± 0.001	0.002 ± 0.001
CART+Aux (m=1)	All	1923	0.150 ± 0.000	0.173	$-0.003 \pm 0.001^{***}$	$-0.022 \pm 0.000^{***}$	0.072	$-0.002 \pm 0.000^{***}$	$-0.010 \pm 0.000^{***}$
CART+Aux (m=5)	All	1923	0.150 ± 0.000	0.177	0.000 ± 0.001	-0.000 ± 0.001	0.074	0.000 ± 0.001	$0.004 \pm 0.001^{*}$
GT	BPX	1923	0.150 ± 0.003	0.289	0.000 ± 0.000	0.000 ± 0.000	0.263	0.000 ± 0.000	0.000 ± 0.000
Ignore	BPX	1923	0.150 ± 0.003	0.287	-0.001 ± 0.003	$-0.007 \pm 0.002^{**}$	0.277	$0.014 \pm 0.002^{***}$	$0.090 \pm 0.003^{***}$
Ignore (weighted)	BPX	1923	0.150 ± 0.003	0.287	-0.001 ± 0.009	$-0.007 \pm 0.003^{**}$	0.276	$0.014 \pm 0.002^{***}$	$0.088 \pm 0.003^{***}$
Ignore20	BPX	1670.3	0.150 ± 0.003	0.286	-0.001 ± 0.003	$-0.007 \pm 0.002^{**}$	0.275	$0.012 \pm 0.003^{***}$	$0.083 \pm 0.003^{***}$
kNN	BPX	1923	0.150 ± 0.003	0.299	$0.010 \pm 0.003^{***}$	$0.063 \pm 0.003^{***}$	0.256	$-0.007 \pm 0.001^{***}$	$-0.041 \pm 0.002^{***}$
RF	BPX	1923	0.150 ± 0.003	0.288	-0.001 ± 0.001	$-0.003 \pm 0.001^{*}$	0.263	0.000 ± 0.001	0.002 ± 0.001
Default (m=1)	BPX	1923	0.150 ± 0.003	0.291	0.002 ± 0.001	$0.015 \pm 0.001^{***}$	0.263	0.001 ± 0.001	$0.005 \pm 0.001^{***}$
Default (m=5)	BPX	1923	0.150 ± 0.003	0.295	$0.007 \pm 0.002^{***}$	$0.045 \pm 0.001^{***}$	0.268	$0.005 \pm 0.002^{*}$	$0.035 \pm 0.003^{***}$
MICE RF (m=1)	BPX	1923	0.150 ± 0.003	0.291	$0.003 \pm 0.001^{*}$	$0.020 \pm 0.001^{***}$	0.263	0.001 ± 0.001	$0.005 \pm 0.002^{**}$
MICE RF (m=5)	BPX	1923	0.150 ± 0.003	0.303	$0.014 \pm 0.002^{***}$	$0.094 \pm 0.001^{***}$	0.274	0.011 ± 0.007	$0.079 \pm 0.007^{***}$
CART (m=1)	BPX	1923	0.150 ± 0.003	0.282	$-0.006 \pm 0.002^{***}$	$-0.038 \pm 0.002^{***}$	0.260	$-0.003 \pm 0.000^{***}$	$-0.015 \pm 0.001^{***}$
CART (m=5)	BPX	1923	0.150 ± 0.003	0.287	-0.001 ± 0.002	$-0.008 \pm 0.002^{***}$	0.267	0.004 ± 0.004	$0.025 \pm 0.004^{***}$
CART+Aux (m=1)	BPX	1923	0.150 ± 0.003	0.283	$-0.006 \pm 0.002^{**}$	$-0.038 \pm 0.002^{***}$	0.260	$-0.002 \pm 0.001^{**}$	$-0.013 \pm 0.001^{***}$
CART+Aux (m=5)	BPX	1923	0.150 ± 0.003	0.287	-0.001 ± 0.002	$-0.009 \pm 0.002^{***}$	0.266	0.004 ± 0.003	$0.025 \pm 0.004^{***}$
GT	LB	1923	0.150 ± 0.001	0.147	0.000 ± 0.000	0.000 ± 0.000	0.087	0.000 ± 0.000	0.000 ± 0.000
Ignore	LB	1923	0.150 ± 0.001	0.147	-0.000 ± 0.001	-0.001 ± 0.001	0.092	$0.004 \pm 0.001^{***}$	$0.029 \pm 0.001^{***}$
Ignore (weighted)	LB	1923	0.150 ± 0.001	0.147	-0.000 ± 0.003	-0.000 ± 0.001	0.091	$0.004 \pm 0.001^{***}$	$0.028 \pm 0.001^{***}$
Ignore20	LB	1670.3	0.150 ± 0.001	0.147	0.000 ± 0.001	0.000 ± 0.001	0.091	$0.004 \pm 0.001^{***}$	$0.027 \pm 0.001^{***}$
kNN	LB	1923	0.150 ± 0.001	0.147	0.000 ± 0.001	0.001 ± 0.001	0.086	$-0.001 \pm 0.000^{***}$	$-0.008 \pm 0.001^{***}$
RF	LB	1923	0.150 ± 0.001	0.147	0.000 ± 0.001	0.001 ± 0.001	0.089	$0.002 \pm 0.000^{***}$	$0.012 \pm 0.001^{***}$
Default (m=1)	LB	1923	0.150 ± 0.001	0.152	$0.005 \pm 0.001^{***}$	$0.033 \pm 0.001^{***}$	0.089	$0.002 \pm 0.000^{***}$	$0.015 \pm 0.001^{***}$
Default (m=5)	LB	1923	0.150 ± 0.001	0.166	$0.019 \pm 0.001^{***}$	$0.121 \pm 0.001^{***}$	0.095	$0.008 \pm 0.003^{**}$	$0.050 \pm 0.004^{***}$
MICE RF (m=1)	LB	1923	0.150 ± 0.001	0.149	$0.002 \pm 0.001^{**}$	$0.014 \pm 0.001^{***}$	0.089	$0.001 \pm 0.001^{*}$	$0.008 \pm 0.001^{***}$
MICE RF (m=5)	LB	1923	0.150 ± 0.001	0.164	$0.017 \pm 0.001^{***}$	$0.112 \pm 0.001^{***}$	0.095	$0.008 \pm 0.003^{**}$	$0.053 \pm 0.004^{***}$
CART (m=1)	LB	1923	0.150 ± 0.001	0.137	$-0.010 \pm 0.001^{***}$	$-0.067 \pm 0.001^{***}$	0.085	$-0.002 \pm 0.001^{**}$	$-0.014 \pm 0.001^{***}$
CART (m=5)	LB	1923	0.150 ± 0.001	0.148	0.001 ± 0.001	$0.006 \pm 0.001^{***}$	0.091	0.004 ± 0.005	$0.027 \pm 0.005^{***}$
CART+Aux (m=1)	LB	1923	0.150 ± 0.001	0.140	$-0.007 \pm 0.001^{***}$	$-0.048 \pm 0.001^{***}$	0.086	$-0.001 \pm 0.000^{**}$	$-0.008 \pm 0.001^{***}$
CART+Aux (m=5)	LB	1923	0.150 ± 0.001	0.147	0.000 ± 0.001	0.001 ± 0.001	0.090	0.003 ± 0.002	$0.023 \pm 0.003^{***}$

Table A.14: Imputed FI Bias Summary by Block - 15% cMCAR (2/2)

Imputation	Block	N	Missingness	Mean FI	Bias	Bias Rate	FI SD	SD Bias	SD Bias Rate
GT	PFQ	1923	0.151 ± 0.002	0.095	0.000 ± 0.000	0.000 ± 0.000	0.125	0.000 ± 0.000	0.000 ± 0.000
Ignore	PFQ	1923	0.151 ± 0.002	0.095	0.000 ± 0.001	$0.002 \pm 0.001^{*}$	0.128	$0.003 \pm 0.001^{***}$	$0.016 \pm 0.001^{***}$
Ignore (weighted)	PFQ	1923	0.151 ± 0.002	0.095	0.000 ± 0.004	$0.001 \pm 0.001^{*}$	0.128	$0.003 \pm 0.001^{***}$	$0.016 \pm 0.001^{***}$
Ignore20	PFQ	1670.3	0.151 ± 0.002	0.095	0.000 ± 0.001	$0.002 \pm 0.001^{*}$	0.127	0.002 ± 0.002	$0.014 \pm 0.002^{***}$
kNN	PFQ	1923	0.151 ± 0.002	0.093	$-0.002 \pm 0.001^{***}$	$-0.012 \pm 0.000^{***}$	0.122	$-0.003 \pm 0.001^{***}$	$-0.020 \pm 0.001^{***}$
RF	PFQ	1923	0.151 ± 0.002	0.094	-0.001 ± 0.000	$-0.005 \pm 0.000^{***}$	0.125	-0.000 ± 0.000	-0.001 ± 0.001
Default (m=1)	PFQ	1923	0.151 ± 0.002	0.105	$0.010 \pm 0.001^{***}$	$0.065 \pm 0.001^{***}$	0.127	$0.002 \pm 0.001^{***}$	$0.015 \pm 0.001^{***}$
Default (m=5)	PFQ	1923	0.151 ± 0.002	0.105	$0.010 \pm 0.001^{***}$	$0.065 \pm 0.001^{***}$	0.131	$0.005 \pm 0.001^{***}$	$0.035 \pm 0.002^{***}$
MICE RF (m=1)	PFQ	1923	0.151 ± 0.002	0.097	$0.002 \pm 0.000^{***}$	$0.014 \pm 0.000^{***}$	0.125	-0.000 ± 0.000	$-0.002 \pm 0.001^{*}$
MICE RF (m=5)	PFQ	1923	0.151 ± 0.002	0.097	$0.002 \pm 0.001^{***}$	$0.014 \pm 0.000^{***}$	0.126	0.001 ± 0.001	$0.007 \pm 0.001^{***}$
CART (m=1)	PFQ	1923	0.151 ± 0.002	0.095	0.000 ± 0.001	0.000 ± 0.001	0.122	$-0.003 \pm 0.001^{***}$	$-0.022 \pm 0.001^{***}$
CART (m=5)	PFQ	1923	0.151 ± 0.002	0.095	0.000 ± 0.001	0.000 ± 0.001	0.125	-0.000 ± 0.003	-0.002 ± 0.003
CART+Aux (m=1)	PFQ	1923	0.151 ± 0.002	0.095	0.000 ± 0.001	-0.000 ± 0.001	0.122	$-0.003 \pm 0.001^{***}$	$-0.021 \pm 0.001^{***}$
CART+Aux (m=5)	PFQ	1923	0.151 ± 0.002	0.095	0.000 ± 0.001	-0.000 ± 0.001	0.125	-0.000 ± 0.002	-0.004 ± 0.002
GT	RXD	1923	0.147 ± 0.006	0.207	0.000 ± 0.000	0.000 ± 0.000	0.140	0.000 ± 0.000	0.000 ± 0.000
Ignore	RXD	1639.9	0.147 ± 0.006	0.207	0.000 ± 0.000	0.000 ± 0.000	0.140	-0.000 ± 0.001	0.000 ± 0.002
Ignore (weighted)	RXD	1639.9	0.147 ± 0.006	0.207	-0.001 ± 0.005	-0.002 ± 0.002	0.139	-0.001 ± 0.001	$-0.004 \pm 0.002^{*}$
Ignore20	RXD	1442.7	0.147 ± 0.006	0.206	0.000 ± 0.000	0.000 ± 0.000	0.140	-0.001 ± 0.002	-0.002 ± 0.002
kNN	RXD	1923	0.147 ± 0.006	0.205	-0.002 ± 0.001	$-0.013 \pm 0.001^{***}$	0.138	$-0.002 \pm 0.001^{*}$	$-0.013 \pm 0.001^{***}$
RF	RXD	1923	0.147 ± 0.006	0.206	-0.001 ± 0.001	$-0.008 \pm 0.001^{***}$	0.140	-0.000 ± 0.001	$-0.003 \pm 0.001^{***}$
Default (m=1)	RXD	1923	0.147 ± 0.006	0.208	0.001 ± 0.001	$0.008 \pm 0.001^{***}$	0.139	-0.001 ± 0.001	$-0.008 \pm 0.001^{***}$
Default (m=5)	RXD	1923	0.147 ± 0.006	0.208	0.001 ± 0.001	$0.008 \pm 0.001^{***}$	0.145	$0.005 \pm 0.001^{***}$	$0.034 \pm 0.002^{***}$
MICE RF (m=1)	RXD	1923	0.147 ± 0.006	0.207	-0.000 ± 0.001	-0.001 ± 0.001	0.138	$-0.002 \pm 0.001^{**}$	$-0.016 \pm 0.001^{***}$
MICE RF (m=5)	RXD	1923	0.147 ± 0.006	0.207	-0.000 ± 0.001	-0.001 ± 0.001	0.142	0.002 ± 0.001	$0.015 \pm 0.002^{***}$
CART (m=1)	RXD	1923	0.147 ± 0.006	0.207	-0.000 ± 0.002	-0.002 ± 0.002	0.133	$-0.007 \pm 0.001^{***}$	$-0.044 \pm 0.001^{***}$
CART (m=5)	RXD	1923	0.147 ± 0.006	0.207	-0.000 ± 0.002	-0.002 ± 0.002	0.150	0.010 ± 0.007	$0.074 \pm 0.009^{***}$
CART+Aux (m=1)	RXD	1923	0.147 ± 0.006	0.207	-0.000 ± 0.002	-0.003 ± 0.002	0.133	$-0.007 \pm 0.002^{***}$	$-0.045 \pm 0.002^{***}$
CART+Aux (m=5)	RXD	1923	0.147 ± 0.006	0.207	-0.000 ± 0.002	-0.003 ± 0.002	0.149	0.009 ± 0.006	$0.074 \pm 0.009^{***}$
GT	VIQ	1923	0.150 ± 0.005	0.193	0.000 ± 0.000	0.000 ± 0.000	0.189	0.000 ± 0.000	0.000 ± 0.000
Ignore	VIQ	1916.8	0.150 ± 0.005	0.193	0.000 ± 0.003	-0.002 ± 0.003	0.207	$0.018 \pm 0.004^{***}$	$0.120 \pm 0.005^{***}$
Ignore (weighted)	VIQ	1916.8	0.150 ± 0.005	0.193	0.000 ± 0.007	-0.002 ± 0.002	0.207	$0.019 \pm 0.005^{***}$	$0.121 \pm 0.005^{***}$
Ignore20	VIQ	1666.8	0.150 ± 0.005	0.193	0.000 ± 0.003	-0.000 ± 0.003	0.205	$0.016 \pm 0.004^{***}$	$0.112 \pm 0.004^{***}$
kNN	VIQ	1923	0.150 ± 0.005	0.195	0.002 ± 0.002	$0.017 \pm 0.002^{***}$	0.186	-0.003 ± 0.002	$-0.019 \pm 0.002^{***}$
RF	VIQ	1923	0.150 ± 0.005	0.193	0.001 ± 0.001	$0.004 \pm 0.001^{**}$	0.185	$-0.003 \pm 0.001^{*}$	$-0.022 \pm 0.002^{***}$
Default (m=1)	VIQ	1923	0.150 ± 0.005	0.211	$0.019 \pm 0.002^{***}$	$0.123 \pm 0.002^{***}$	0.187	-0.002 ± 0.001	$-0.011 \pm 0.002^{***}$
Default (m=5)	VIQ	1923	0.150 ± 0.005	0.211	$0.019 \pm 0.003^{***}$	$0.123 \pm 0.002^{***}$	0.208	$0.019 \pm 0.006^{***}$	$0.124 \pm 0.008^{***}$
MICE RF (m=1)	VIQ	1923	0.150 ± 0.005	0.201	$0.008 \pm 0.001^{***}$	$0.056 \pm 0.001^{***}$	0.184	$-0.005 \pm 0.001^{***}$	$-0.033 \pm 0.001^{***}$
MICE RF (m=5)	VIQ	1923	0.150 ± 0.005	0.201	$0.008 \pm 0.002^{***}$	$0.056 \pm 0.001^{***}$	0.201	0.012 ± 0.008	$0.071 \pm 0.009^{***}$
CART (m=1)	VIQ	1923	0.150 ± 0.005	0.193	0.001 ± 0.002	0.003 ± 0.002	0.179	$-0.010 \pm 0.002^{***}$	$-0.065 \pm 0.003^{***}$
CART (m=5)	VIQ	1923	0.150 ± 0.005	0.193	0.001 ± 0.003	0.003 ± 0.002	0.201	0.012 ± 0.010	$0.075 \pm 0.011^{***}$
CART+Aux (m=1)	VIQ	1923	0.150 ± 0.005	0.193	0.001 ± 0.002	0.002 ± 0.002	0.182	$-0.007 \pm 0.002^{***}$	$-0.046 \pm 0.002^{***}$
CART+Aux (m=5)	VIQ	1923	0.150 ± 0.005	0.193	0.001 ± 0.003	0.002 ± 0.002	0.197	0.009 ± 0.007	$0.058 \pm 0.008^{***}$

Table A.15: Imputed FI Prediction Summary by Block - 15% pMCAR (1/2)

Imputation	Block	N	Missingness	AUC	C-index	HR
GT	All	1923	0.157 ± 0.002	0.733 ± 0.036	0.654 ± 0.016	1.075 ± 0.007
Ignore	All	1923	0.157 ± 0.002	0.729 ± 0.037	0.648 ± 0.016	1.064 ± 0.007
Ignore (weighted)	All	1923	0.157 ± 0.002	–	–	–
Ignore20	All	1117	0.157 ± 0.002	0.733 ± 0.050	0.655 ± 0.025	1.073 ± 0.012
kNN	All	1923	0.157 ± 0.002	0.722 ± 0.038	0.644 ± 0.016	1.071 ± 0.008
RF	All	1923	0.157 ± 0.002	0.726 ± 0.038	0.647 ± 0.016	1.076 ± 0.008
Default (m=1)	All	1923	0.157 ± 0.002	$0.695 \pm 0.041^{***}$	0.630 ± 0.019	1.053 ± 0.007
Default (m=5)	All	1923	0.157 ± 0.002	$0.697 \pm 0.040^{***}$	0.631 ± 0.019	1.041 ± 0.007
MICE RF (m=1)	All	1923	0.157 ± 0.002	0.728 ± 0.037	0.648 ± 0.016	1.078 ± 0.008
MICE RF (m=5)	All	1923	0.157 ± 0.002	0.732 ± 0.037	0.651 ± 0.016	1.078 ± 0.008
CART (m=1)	All	1923	0.157 ± 0.002	0.730 ± 0.038	0.649 ± 0.016	1.077 ± 0.008
CART (m=5)	All	1923	0.157 ± 0.002	0.733 ± 0.037	0.652 ± 0.016	1.075 ± 0.008
CART+Aux (m=1)	All	1923	0.157 ± 0.002	0.732 ± 0.037	0.655 ± 0.016	1.078 ± 0.007
CART+Aux (m=5)	All	1923	0.157 ± 0.002	0.733 ± 0.037	0.655 ± 0.016	1.076 ± 0.008
GT	BPX	1923	0.064 ± 0.003	0.619 ± 0.040	0.544 ± 0.016	1.006 ± 0.002
Ignore	BPX	1832.4	0.064 ± 0.003	0.613 ± 0.042	0.540 ± 0.017	1.005 ± 0.002
Ignore (weighted)	BPX	1832.4	0.064 ± 0.003	–	–	–
Ignore20	BPX	1082.4	0.064 ± 0.003	0.617 ± 0.057	0.538 ± 0.025	1.005 ± 0.003
kNN	BPX	1923	0.064 ± 0.003	0.611 ± 0.041	0.540 ± 0.017	1.006 ± 0.002
RF	BPX	1923	0.064 ± 0.003	0.610 ± 0.041	0.542 ± 0.017	1.006 ± 0.002
Default (m=1)	BPX	1923	0.064 ± 0.003	0.609 ± 0.042	0.540 ± 0.017	1.006 ± 0.002
Default (m=5)	BPX	1923	0.064 ± 0.003	0.613 ± 0.041	0.541 ± 0.017	1.006 ± 0.002
MICE RF (m=1)	BPX	1923	0.064 ± 0.003	0.613 ± 0.041	0.541 ± 0.017	1.006 ± 0.002
MICE RF (m=5)	BPX	1923	0.064 ± 0.003	0.613 ± 0.041	0.541 ± 0.017	1.006 ± 0.002
CART (m=1)	BPX	1923	0.064 ± 0.003	0.611 ± 0.041	0.541 ± 0.018	1.006 ± 0.002
CART (m=5)	BPX	1923	0.064 ± 0.003	0.614 ± 0.041	0.541 ± 0.017	1.006 ± 0.002
CART+Aux (m=1)	BPX	1923	0.064 ± 0.003	0.612 ± 0.041	0.540 ± 0.018	1.006 ± 0.002
CART+Aux (m=5)	BPX	1923	0.064 ± 0.003	0.613 ± 0.041	0.540 ± 0.017	1.006 ± 0.002
GT	LB	1923	0.065 ± 0.003	0.689 ± 0.041	0.634 ± 0.016	1.046 ± 0.006
Ignore	LB	1821.8	0.065 ± 0.003	0.688 ± 0.043	0.635 ± 0.017	1.044 ± 0.006
Ignore (weighted)	LB	1821.8	0.065 ± 0.003	–	–	–
Ignore20	LB	1117	0.065 ± 0.003	0.680 ± 0.057	0.637 ± 0.025	1.047 ± 0.010
kNN	LB	1923	0.065 ± 0.003	0.674 ± 0.045	0.623 ± 0.018	1.043 ± 0.006
RF	LB	1923	0.065 ± 0.003	0.672 ± 0.045	0.623 ± 0.018	1.042 ± 0.006
Default (m=1)	LB	1923	0.065 ± 0.003	0.678 ± 0.044	0.629 ± 0.018	1.045 ± 0.006
Default (m=5)	LB	1923	0.065 ± 0.003	0.685 ± 0.042	0.634 ± 0.016	1.046 ± 0.006
MICE RF (m=1)	LB	1923	0.065 ± 0.003	0.674 ± 0.044	0.625 ± 0.017	1.043 ± 0.006
MICE RF (m=5)	LB	1923	0.065 ± 0.003	0.684 ± 0.042	0.633 ± 0.016	1.046 ± 0.006
CART (m=1)	LB	1923	0.065 ± 0.003	0.676 ± 0.044	0.625 ± 0.017	1.044 ± 0.006
CART (m=5)	LB	1923	0.065 ± 0.003	0.685 ± 0.041	0.632 ± 0.016	1.046 ± 0.006
CART+Aux (m=1)	LB	1923	0.065 ± 0.003	0.684 ± 0.042	0.632 ± 0.016	1.045 ± 0.006
CART+Aux (m=5)	LB	1923	0.065 ± 0.003	0.688 ± 0.041	0.636 ± 0.016	1.046 ± 0.006

AUC, C-index and HR were not calculated for Ignore (weighted) because the standard packages do not allow weights.

Table A.16: Imputed FI Prediction Summary by Block - 15% pMCAR (2/2)

Imputation	Block	N	Missingness	AUC	C-index	HR
GT	PFQ	1923	0.577 ± 0.010	0.607 ± 0.043	0.581 ± 0.016	1.027 ± 0.004
Ignore	PFQ	837.1	0.577 ± 0.010	0.609 ± 0.069	0.584 ± 0.030	1.029 ± 0.007
Ignore (weighted)	PFQ	837.1	0.577 ± 0.010	–	–	–
Ignore20	PFQ	777.8	0.577 ± 0.010	0.616 ± 0.073	0.589 ± 0.033	1.030 ± 0.008
kNN	PFQ	1923	0.577 ± 0.010	0.598 ± 0.045	0.573 ± 0.019	1.024 ± 0.006
RF	PFQ	1923	0.577 ± 0.010	0.453 ± 0.039	0.538 ± 0.017	1.023 ± 0.006
Default (m=1)	PFQ	1923	0.577 ± 0.010	0.560 ± 0.047	0.549 ± 0.024	1.007 ± 0.003
Default (m=5)	PFQ	1923	0.577 ± 0.010	0.560 ± 0.047	0.549 ± 0.024	1.007 ± 0.003
MICE RF (m=1)	PFQ	1923	0.577 ± 0.010	0.577 ± 0.045	0.558 ± 0.020	1.026 ± 0.006
MICE RF (m=5)	PFQ	1923	0.577 ± 0.010	0.577 ± 0.045	0.558 ± 0.020	1.026 ± 0.006
CART (m=1)	PFQ	1923	0.577 ± 0.010	0.608 ± 0.045	0.579 ± 0.020	1.032 ± 0.006
CART (m=5)	PFQ	1923	0.577 ± 0.010	0.608 ± 0.045	0.579 ± 0.020	1.032 ± 0.006
CART+Aux (m=1)	PFQ	1923	0.577 ± 0.010	0.619 ± 0.043	0.591 ± 0.018	1.034 ± 0.005
CART+Aux (m=5)	PFQ	1923	0.577 ± 0.010	0.619 ± 0.043	0.591 ± 0.018	1.034 ± 0.005
GT	RXD	1923	0.469 ± 0.006	0.630 ± 0.042	0.590 ± 0.017	1.021 ± 0.003
Ignore	RXD	1021.6	0.469 ± 0.006	0.624 ± 0.065	0.589 ± 0.029	1.022 ± 0.006
Ignore (weighted)	RXD	1021.6	0.469 ± 0.006	–	–	–
Ignore20	RXD	801.6	0.469 ± 0.006	0.635 ± 0.076	0.598 ± 0.036	1.025 ± 0.007
kNN	RXD	1923	0.469 ± 0.006	0.608 ± 0.044	0.579 ± 0.018	1.023 ± 0.004
RF	RXD	1923	0.469 ± 0.006	0.635 ± 0.044	0.595 ± 0.021	1.023 ± 0.004
Default (m=1)	RXD	1923	0.469 ± 0.006	0.629 ± 0.046	0.588 ± 0.021	1.023 ± 0.005
Default (m=5)	RXD	1923	0.469 ± 0.006	0.629 ± 0.046	0.588 ± 0.021	1.023 ± 0.005
MICE RF (m=1)	RXD	1923	0.469 ± 0.006	0.617 ± 0.045	0.584 ± 0.021	1.024 ± 0.005
MICE RF (m=5)	RXD	1923	0.469 ± 0.006	0.617 ± 0.045	0.584 ± 0.021	1.024 ± 0.005
CART (m=1)	RXD	1923	0.469 ± 0.006	0.615 ± 0.045	0.580 ± 0.020	1.022 ± 0.005
CART (m=5)	RXD	1923	0.469 ± 0.006	0.614 ± 0.045	0.580 ± 0.020	1.022 ± 0.005
CART+Aux (m=1)	RXD	1923	0.469 ± 0.006	0.622 ± 0.046	0.587 ± 0.019	1.024 ± 0.005
CART+Aux (m=5)	RXD	1923	0.469 ± 0.006	0.622 ± 0.046	0.587 ± 0.019	1.024 ± 0.005
GT	VIQ	1923	0.278 ± 0.011	0.642 ± 0.041	0.560 ± 0.016	1.011 ± 0.003
Ignore	VIQ	1406.1	0.278 ± 0.011	0.643 ± 0.049	0.566 ± 0.021	1.011 ± 0.003
Ignore (weighted)	VIQ	1406.1	0.278 ± 0.011	–	–	–
Ignore20	VIQ	1026.3	0.278 ± 0.011	0.649 ± 0.058	0.565 ± 0.026	1.011 ± 0.004
kNN	VIQ	1923	0.278 ± 0.011	0.619 ± 0.044	0.555 ± 0.019	1.010 ± 0.003
RF	VIQ	1923	0.278 ± 0.011	0.556 ± 0.115	0.556 ± 0.019	1.010 ± 0.004
Default (m=1)	VIQ	1923	0.278 ± 0.011	0.636 ± 0.045	0.565 ± 0.020	1.013 ± 0.003
Default (m=5)	VIQ	1923	0.278 ± 0.011	0.636 ± 0.045	0.565 ± 0.020	1.013 ± 0.003
MICE RF (m=1)	VIQ	1923	0.278 ± 0.011	0.630 ± 0.044	0.561 ± 0.019	1.011 ± 0.004
MICE RF (m=5)	VIQ	1923	0.278 ± 0.011	0.630 ± 0.044	0.561 ± 0.019	1.011 ± 0.004
CART (m=1)	VIQ	1923	0.278 ± 0.011	0.639 ± 0.043	0.567 ± 0.019	1.012 ± 0.003
CART (m=5)	VIQ	1923	0.278 ± 0.011	0.639 ± 0.044	0.567 ± 0.019	1.012 ± 0.003
CART+Aux (m=1)	VIQ	1923	0.278 ± 0.011	0.645 ± 0.043	0.565 ± 0.019	1.012 ± 0.003
CART+Aux (m=5)	VIQ	1923	0.278 ± 0.011	0.646 ± 0.043	0.565 ± 0.019	1.012 ± 0.003

AUC, C-index and HR were not calculated for Ignore (weighted) because the standard packages do not allow weights.

Table A.17: Imputed FI Prediction Summary by Block - 15% pMAR (1/2)

Imputation	Block	N	Missingness	AUC	C-index	HR
GT	All	1923	0.157 ± 0.003	0.733 ± 0.036	0.654 ± 0.016	1.075 ± 0.007
Ignore	All	1923	0.157 ± 0.003	0.732 ± 0.037	0.649 ± 0.016	1.070 ± 0.008
Ignore (weighted)	All	1923	0.157 ± 0.003	–	–	–
Ignore20	All	1117.4	0.157 ± 0.003	0.742 ± 0.047	0.666 ± 0.023	1.078 ± 0.010
kNN	All	1923	0.157 ± 0.003	0.721 ± 0.038	0.645 ± 0.016	1.071 ± 0.007
RF	All	1923	0.157 ± 0.003	0.728 ± 0.038	0.645 ± 0.017	1.070 ± 0.007
Default (m=1)	All	1923	0.157 ± 0.003	$0.695 \pm 0.042^{***}$	0.633 ± 0.019	1.065 ± 0.009
Default (m=5)	All	1923	0.157 ± 0.003	$0.697 \pm 0.041^{***}$	0.634 ± 0.018	1.046 ± 0.007
MICE RF (m=1)	All	1923	0.157 ± 0.003	0.728 ± 0.037	0.646 ± 0.017	1.073 ± 0.008
MICE RF (m=5)	All	1923	0.157 ± 0.003	0.732 ± 0.037	0.649 ± 0.016	1.074 ± 0.008
CART (m=1)	All	1923	0.157 ± 0.003	0.730 ± 0.037	0.650 ± 0.017	1.076 ± 0.008
CART (m=5)	All	1923	0.157 ± 0.003	0.733 ± 0.037	0.652 ± 0.016	1.075 ± 0.008
CART+Aux (m=1)	All	1923	0.157 ± 0.003	0.732 ± 0.037	0.654 ± 0.016	1.076 ± 0.007
CART+Aux (m=5)	All	1923	0.157 ± 0.003	0.734 ± 0.037	0.655 ± 0.016	1.075 ± 0.007
GT	BPX	1923	0.063 ± 0.006	0.619 ± 0.040	0.544 ± 0.016	1.006 ± 0.002
Ignore	BPX	1832.7	0.063 ± 0.006	0.618 ± 0.041	0.543 ± 0.017	1.005 ± 0.002
Ignore (weighted)	BPX	1832.7	0.063 ± 0.006	–	–	–
Ignore20	BPX	1085.6	0.063 ± 0.006	0.622 ± 0.055	0.553 ± 0.023	1.007 ± 0.003
kNN	BPX	1923	0.063 ± 0.006	0.613 ± 0.040	0.542 ± 0.016	1.006 ± 0.002
RF	BPX	1923	0.063 ± 0.006	0.613 ± 0.041	0.542 ± 0.016	1.005 ± 0.002
Default (m=1)	BPX	1923	0.063 ± 0.006	0.613 ± 0.041	0.543 ± 0.016	1.006 ± 0.002
Default (m=5)	BPX	1923	0.063 ± 0.006	0.616 ± 0.040	0.544 ± 0.016	1.006 ± 0.002
MICE RF (m=1)	BPX	1923	0.063 ± 0.006	0.613 ± 0.040	0.541 ± 0.016	1.006 ± 0.002
MICE RF (m=5)	BPX	1923	0.063 ± 0.006	0.615 ± 0.040	0.542 ± 0.016	1.006 ± 0.002
CART (m=1)	BPX	1923	0.063 ± 0.006	0.614 ± 0.041	0.543 ± 0.017	1.006 ± 0.002
CART (m=5)	BPX	1923	0.063 ± 0.006	0.616 ± 0.040	0.543 ± 0.016	1.006 ± 0.002
CART+Aux (m=1)	BPX	1923	0.063 ± 0.006	0.614 ± 0.040	0.542 ± 0.016	1.005 ± 0.002
CART+Aux (m=5)	BPX	1923	0.063 ± 0.006	0.615 ± 0.040	0.542 ± 0.016	1.006 ± 0.002
GT	LB	1923	0.065 ± 0.005	0.689 ± 0.041	0.634 ± 0.016	1.046 ± 0.006
Ignore	LB	1823	0.065 ± 0.005	0.690 ± 0.042	0.634 ± 0.017	1.043 ± 0.007
Ignore (weighted)	LB	1823	0.065 ± 0.005	–	–	–
Ignore20	LB	1117.4	0.065 ± 0.005	0.700 ± 0.052	0.643 ± 0.022	1.048 ± 0.008
kNN	LB	1923	0.065 ± 0.005	0.676 ± 0.043	0.624 ± 0.017	1.043 ± 0.006
RF	LB	1923	0.065 ± 0.005	0.678 ± 0.043	0.626 ± 0.018	1.043 ± 0.006
Default (m=1)	LB	1923	0.065 ± 0.005	0.681 ± 0.043	0.629 ± 0.017	1.044 ± 0.006
Default (m=5)	LB	1923	0.065 ± 0.005	0.688 ± 0.041	0.633 ± 0.016	1.046 ± 0.006
MICE RF (m=1)	LB	1923	0.065 ± 0.005	0.678 ± 0.043	0.627 ± 0.017	1.044 ± 0.006
MICE RF (m=5)	LB	1923	0.065 ± 0.005	0.686 ± 0.041	0.632 ± 0.016	1.046 ± 0.006
CART (m=1)	LB	1923	0.065 ± 0.005	0.680 ± 0.043	0.628 ± 0.017	1.044 ± 0.006
CART (m=5)	LB	1923	0.065 ± 0.005	0.687 ± 0.041	0.632 ± 0.016	1.046 ± 0.006
CART+Aux (m=1)	LB	1923	0.065 ± 0.005	0.685 ± 0.041	0.633 ± 0.016	1.045 ± 0.006
CART+Aux (m=5)	LB	1923	0.065 ± 0.005	0.689 ± 0.041	0.635 ± 0.016	1.046 ± 0.006

AUC, C-index and HR were not calculated for Ignore (weighted) because the standard packages do not allow weights.

Table A.18: Imputed FI Prediction Summary by Block - 15% pMAR (2/2)

Imputation	Block	N	Missingness	AUC	C-index	HR
GT	PFQ	1923	0.583 ± 0.012	0.607 ± 0.043	0.581 ± 0.016	1.027 ± 0.004
Ignore	PFQ	824.5	0.583 ± 0.012	0.629 ± 0.062	0.594 ± 0.030	1.031 ± 0.006
Ignore (weighted)	PFQ	824.5	0.583 ± 0.012	–	–	–
Ignore20	PFQ	763.5	0.583 ± 0.012	0.628 ± 0.064	0.593 ± 0.031	1.030 ± 0.007
kNN	PFQ	1923	0.583 ± 0.012	0.593 ± 0.046	0.568 ± 0.019	1.022 ± 0.006
RF	PFQ	1923	0.583 ± 0.012	0.411 ± 0.041	0.554 ± 0.018	1.027 ± 0.006
Default (m=1)	PFQ	1923	0.583 ± 0.012	0.516 ± 0.045	0.523 ± 0.021	1.003 ± 0.003
Default (m=5)	PFQ	1923	0.583 ± 0.012	0.516 ± 0.045	0.523 ± 0.021	1.003 ± 0.003
MICE RF (m=1)	PFQ	1923	0.583 ± 0.012	0.595 ± 0.048	0.566 ± 0.021	1.030 ± 0.006
MICE RF (m=5)	PFQ	1923	0.583 ± 0.012	0.595 ± 0.048	0.566 ± 0.021	1.030 ± 0.006
CART (m=1)	PFQ	1923	0.583 ± 0.012	0.603 ± 0.047	0.576 ± 0.022	1.034 ± 0.006
CART (m=5)	PFQ	1923	0.583 ± 0.012	0.603 ± 0.047	0.576 ± 0.022	1.034 ± 0.006
CART+Aux (m=1)	PFQ	1923	0.583 ± 0.012	0.612 ± 0.045	0.589 ± 0.019	1.034 ± 0.005
CART+Aux (m=5)	PFQ	1923	0.583 ± 0.012	0.612 ± 0.045	0.589 ± 0.019	1.034 ± 0.005
GT	RXD	1923	0.472 ± 0.011	0.630 ± 0.042	0.590 ± 0.017	1.021 ± 0.003
Ignore	RXD	1014.7	0.472 ± 0.011	0.614 ± 0.056	0.590 ± 0.025	1.020 ± 0.005
Ignore (weighted)	RXD	1014.7	0.472 ± 0.011	–	–	–
Ignore20	RXD	800.9	0.472 ± 0.011	0.622 ± 0.060	0.600 ± 0.026	1.021 ± 0.005
kNN	RXD	1923	0.472 ± 0.011	0.611 ± 0.044	0.581 ± 0.019	1.022 ± 0.004
RF	RXD	1923	0.472 ± 0.011	0.631 ± 0.043	0.591 ± 0.019	1.022 ± 0.004
Default (m=1)	RXD	1923	0.472 ± 0.011	0.616 ± 0.044	0.584 ± 0.019	1.022 ± 0.004
Default (m=5)	RXD	1923	0.472 ± 0.011	0.616 ± 0.044	0.583 ± 0.019	1.022 ± 0.004
MICE RF (m=1)	RXD	1923	0.472 ± 0.011	0.615 ± 0.045	0.589 ± 0.019	1.023 ± 0.005
MICE RF (m=5)	RXD	1923	0.472 ± 0.011	0.615 ± 0.045	0.589 ± 0.019	1.023 ± 0.005
CART (m=1)	RXD	1923	0.472 ± 0.011	0.614 ± 0.044	0.587 ± 0.019	1.022 ± 0.005
CART (m=5)	RXD	1923	0.472 ± 0.011	0.614 ± 0.044	0.587 ± 0.019	1.022 ± 0.005
CART+Aux (m=1)	RXD	1923	0.472 ± 0.011	0.609 ± 0.045	0.581 ± 0.019	1.021 ± 0.005
CART+Aux (m=5)	RXD	1923	0.472 ± 0.011	0.609 ± 0.045	0.581 ± 0.019	1.021 ± 0.005
GT	VIQ	1923	0.274 ± 0.013	0.642 ± 0.041	0.560 ± 0.016	1.011 ± 0.003
Ignore	VIQ	1415.1	0.274 ± 0.013	0.653 ± 0.048	0.565 ± 0.020	1.011 ± 0.003
Ignore (weighted)	VIQ	1415.1	0.274 ± 0.013	–	–	–
Ignore20	VIQ	1034	0.274 ± 0.013	0.653 ± 0.055	0.567 ± 0.024	1.012 ± 0.004
kNN	VIQ	1923	0.274 ± 0.013	0.631 ± 0.042	0.561 ± 0.017	1.011 ± 0.003
RF	VIQ	1923	0.274 ± 0.013	0.633 ± 0.041	0.560 ± 0.017	1.011 ± 0.003
Default (m=1)	VIQ	1923	0.274 ± 0.013	0.637 ± 0.044	0.559 ± 0.019	1.012 ± 0.003
Default (m=5)	VIQ	1923	0.274 ± 0.013	0.637 ± 0.044	0.559 ± 0.019	1.012 ± 0.003
MICE RF (m=1)	VIQ	1923	0.274 ± 0.013	0.639 ± 0.042	0.564 ± 0.018	1.012 ± 0.003
MICE RF (m=5)	VIQ	1923	0.274 ± 0.013	0.639 ± 0.042	0.564 ± 0.018	1.012 ± 0.003
CART (m=1)	VIQ	1923	0.274 ± 0.013	0.641 ± 0.043	0.565 ± 0.019	1.012 ± 0.003
CART (m=5)	VIQ	1923	0.274 ± 0.013	0.641 ± 0.043	0.565 ± 0.019	1.012 ± 0.003
CART+Aux (m=1)	VIQ	1923	0.274 ± 0.013	0.653 ± 0.043	0.562 ± 0.018	1.013 ± 0.003
CART+Aux (m=5)	VIQ	1923	0.274 ± 0.013	0.654 ± 0.043	0.562 ± 0.018	1.013 ± 0.003

AUC, C-index and HR were not calculated for Ignore (weighted) because the standard packages do not allow weights.

Table A.19: Imputed FI Prediction Summary by Block - 15% cMCAR (1/2)

Imputation	Block	N	Missingness	AUC	C-index	HR
GT	All	1923	0.150 ± 0.000	0.733 ± 0.036	0.654 ± 0.016	1.075 ± 0.007
Ignore	All	1923	0.150 ± 0.000	0.728 ± 0.037	0.648 ± 0.016	1.070 ± 0.007
Ignore (weighted)	All	1923	0.150 ± 0.000	–	–	–
Ignore20	All	1670.3	0.150 ± 0.000	0.729 ± 0.041	0.648 ± 0.020	1.071 ± 0.009
kNN	All	1923	0.150 ± 0.000	0.730 ± 0.037	0.652 ± 0.016	1.076 ± 0.007
RF	All	1923	0.150 ± 0.000	0.732 ± 0.037	0.653 ± 0.016	1.074 ± 0.007
Default (m=1)	All	1923	0.150 ± 0.000	0.734 ± 0.037	0.654 ± 0.016	1.073 ± 0.007
Default (m=5)	All	1923	0.150 ± 0.000	0.734 ± 0.037	0.654 ± 0.016	1.071 ± 0.007
MICE RF (m=1)	All	1923	0.150 ± 0.000	0.733 ± 0.037	0.653 ± 0.016	1.074 ± 0.007
MICE RF (m=5)	All	1923	0.150 ± 0.000	0.734 ± 0.037	0.654 ± 0.016	1.073 ± 0.007
CART (m=1)	All	1923	0.150 ± 0.000	0.730 ± 0.037	0.651 ± 0.016	1.076 ± 0.007
CART (m=5)	All	1923	0.150 ± 0.000	0.732 ± 0.037	0.652 ± 0.016	1.075 ± 0.008
CART+Aux (m=1)	All	1923	0.150 ± 0.000	0.733 ± 0.037	0.653 ± 0.016	1.076 ± 0.007
CART+Aux (m=5)	All	1923	0.150 ± 0.000	0.735 ± 0.037	0.654 ± 0.016	1.076 ± 0.008
GT	BPX	1923	0.150 ± 0.003	0.619 ± 0.040	0.544 ± 0.016	1.006 ± 0.002
Ignore	BPX	1923	0.150 ± 0.003	0.614 ± 0.040	0.542 ± 0.016	1.005 ± 0.002
Ignore (weighted)	BPX	1923	0.150 ± 0.003	–	–	–
Ignore20	BPX	1670.3	0.150 ± 0.003	0.618 ± 0.046	0.542 ± 0.019	1.005 ± 0.002
kNN	BPX	1923	0.150 ± 0.003	0.616 ± 0.040	0.543 ± 0.017	1.006 ± 0.002
RF	BPX	1923	0.150 ± 0.003	0.619 ± 0.040	0.545 ± 0.016	1.006 ± 0.002
Default (m=1)	BPX	1923	0.150 ± 0.003	0.620 ± 0.040	0.544 ± 0.016	1.006 ± 0.002
Default (m=5)	BPX	1923	0.150 ± 0.003	0.621 ± 0.041	0.545 ± 0.016	1.006 ± 0.002
MICE RF (m=1)	BPX	1923	0.150 ± 0.003	0.620 ± 0.040	0.545 ± 0.016	1.006 ± 0.002
MICE RF (m=5)	BPX	1923	0.150 ± 0.003	0.624 ± 0.040	0.546 ± 0.016	1.006 ± 0.002
CART (m=1)	BPX	1923	0.150 ± 0.003	0.619 ± 0.040	0.544 ± 0.016	1.006 ± 0.002
CART (m=5)	BPX	1923	0.150 ± 0.003	0.623 ± 0.040	0.547 ± 0.016	1.006 ± 0.002
CART+Aux (m=1)	BPX	1923	0.150 ± 0.003	0.619 ± 0.040	0.543 ± 0.016	1.006 ± 0.002
CART+Aux (m=5)	BPX	1923	0.150 ± 0.003	0.622 ± 0.041	0.544 ± 0.017	1.006 ± 0.002
GT	LB	1923	0.150 ± 0.001	0.689 ± 0.041	0.634 ± 0.016	1.046 ± 0.006
Ignore	LB	1923	0.150 ± 0.001	0.683 ± 0.042	0.628 ± 0.017	1.041 ± 0.006
Ignore (weighted)	LB	1923	0.150 ± 0.001	–	–	–
Ignore20	LB	1670.3	0.150 ± 0.001	0.685 ± 0.045	0.632 ± 0.019	1.042 ± 0.007
kNN	LB	1923	0.150 ± 0.001	0.685 ± 0.041	0.630 ± 0.016	1.046 ± 0.006
RF	LB	1923	0.150 ± 0.001	0.688 ± 0.041	0.633 ± 0.016	1.045 ± 0.006
Default (m=1)	LB	1923	0.150 ± 0.001	0.688 ± 0.041	0.632 ± 0.016	1.045 ± 0.006
Default (m=5)	LB	1923	0.150 ± 0.001	0.691 ± 0.041	0.635 ± 0.017	1.046 ± 0.006
MICE RF (m=1)	LB	1923	0.150 ± 0.001	0.687 ± 0.041	0.631 ± 0.016	1.044 ± 0.006
MICE RF (m=5)	LB	1923	0.150 ± 0.001	0.690 ± 0.041	0.635 ± 0.016	1.046 ± 0.006
CART (m=1)	LB	1923	0.150 ± 0.001	0.682 ± 0.041	0.629 ± 0.016	1.045 ± 0.006
CART (m=5)	LB	1923	0.150 ± 0.001	0.687 ± 0.042	0.633 ± 0.016	1.047 ± 0.006
CART+Aux (m=1)	LB	1923	0.150 ± 0.001	0.684 ± 0.041	0.631 ± 0.017	1.045 ± 0.006
CART+Aux (m=5)	LB	1923	0.150 ± 0.001	0.691 ± 0.041	0.636 ± 0.016	1.047 ± 0.006

AUC, C-index and HR were not calculated for Ignore (weighted) because the standard packages do not allow weights.

Table A.20: Imputed FI Prediction Summary by Block - 15% cMCAR (2/2)

Imputation	Block	N	Missingness	AUC	C-index	HR
GT	PFQ	1923	0.151 ± 0.002	0.607 ± 0.043	0.581 ± 0.016	1.027 ± 0.004
Ignore	PFQ	1923	0.151 ± 0.002	0.606 ± 0.043	0.579 ± 0.017	1.025 ± 0.004
Ignore (weighted)	PFQ	1923	0.151 ± 0.002	–	–	–
Ignore20	PFQ	1670.3	0.151 ± 0.002	0.605 ± 0.047	0.578 ± 0.021	1.026 ± 0.004
kNN	PFQ	1923	0.151 ± 0.002	0.607 ± 0.043	0.580 ± 0.017	1.028 ± 0.004
RF	PFQ	1923	0.151 ± 0.002	0.609 ± 0.043	0.582 ± 0.017	1.027 ± 0.004
Default (m=1)	PFQ	1923	0.151 ± 0.002	0.608 ± 0.044	0.584 ± 0.017	1.027 ± 0.004
Default (m=5)	PFQ	1923	0.151 ± 0.002	0.608 ± 0.044	0.584 ± 0.017	1.027 ± 0.004
MICE RF (m=1)	PFQ	1923	0.151 ± 0.002	0.609 ± 0.043	0.584 ± 0.017	1.027 ± 0.004
MICE RF (m=5)	PFQ	1923	0.151 ± 0.002	0.609 ± 0.043	0.584 ± 0.017	1.027 ± 0.004
CART (m=1)	PFQ	1923	0.151 ± 0.002	0.611 ± 0.043	0.582 ± 0.017	1.028 ± 0.004
CART (m=5)	PFQ	1923	0.151 ± 0.002	0.611 ± 0.043	0.582 ± 0.017	1.028 ± 0.004
CART+Aux (m=1)	PFQ	1923	0.151 ± 0.002	0.613 ± 0.043	0.585 ± 0.017	1.028 ± 0.004
CART+Aux (m=5)	PFQ	1923	0.151 ± 0.002	0.613 ± 0.043	0.586 ± 0.017	1.028 ± 0.004
GT	RXD	1923	0.147 ± 0.006	0.630 ± 0.042	0.590 ± 0.017	1.021 ± 0.003
Ignore	RXD	1639.9	0.147 ± 0.006	0.623 ± 0.047	0.587 ± 0.019	1.020 ± 0.004
Ignore (weighted)	RXD	1639.9	0.147 ± 0.006	–	–	–
Ignore20	RXD	1442.7	0.147 ± 0.006	0.622 ± 0.050	0.588 ± 0.021	1.020 ± 0.004
kNN	RXD	1923	0.147 ± 0.006	0.628 ± 0.043	0.589 ± 0.018	1.022 ± 0.004
RF	RXD	1923	0.147 ± 0.006	0.633 ± 0.042	0.592 ± 0.018	1.022 ± 0.004
Default (m=1)	RXD	1923	0.147 ± 0.006	0.630 ± 0.042	0.589 ± 0.017	1.022 ± 0.004
Default (m=5)	RXD	1923	0.147 ± 0.006	0.630 ± 0.042	0.589 ± 0.017	1.022 ± 0.004
MICE RF (m=1)	RXD	1923	0.147 ± 0.006	0.630 ± 0.043	0.590 ± 0.018	1.022 ± 0.004
MICE RF (m=5)	RXD	1923	0.147 ± 0.006	0.630 ± 0.043	0.590 ± 0.018	1.022 ± 0.004
CART (m=1)	RXD	1923	0.147 ± 0.006	0.625 ± 0.042	0.588 ± 0.017	1.021 ± 0.004
CART (m=5)	RXD	1923	0.147 ± 0.006	0.624 ± 0.043	0.588 ± 0.017	1.021 ± 0.004
CART+Aux (m=1)	RXD	1923	0.147 ± 0.006	0.624 ± 0.042	0.586 ± 0.017	1.021 ± 0.004
CART+Aux (m=5)	RXD	1923	0.147 ± 0.006	0.624 ± 0.042	0.586 ± 0.017	1.021 ± 0.004
GT	VIQ	1923	0.150 ± 0.005	0.642 ± 0.041	0.560 ± 0.016	1.011 ± 0.003
Ignore	VIQ	1916.8	0.150 ± 0.005	0.627 ± 0.043	0.553 ± 0.018	1.009 ± 0.003
Ignore (weighted)	VIQ	1916.8	0.150 ± 0.005	–	–	–
Ignore20	VIQ	1666.8	0.150 ± 0.005	0.627 ± 0.047	0.552 ± 0.022	1.008 ± 0.003
kNN	VIQ	1923	0.150 ± 0.005	0.638 ± 0.041	0.559 ± 0.016	1.011 ± 0.003
RF	VIQ	1923	0.150 ± 0.005	0.640 ± 0.041	0.558 ± 0.016	1.011 ± 0.003
Default (m=1)	VIQ	1923	0.150 ± 0.005	0.642 ± 0.042	0.559 ± 0.017	1.011 ± 0.003
Default (m=5)	VIQ	1923	0.150 ± 0.005	0.642 ± 0.042	0.559 ± 0.017	1.011 ± 0.003
MICE RF (m=1)	VIQ	1923	0.150 ± 0.005	0.645 ± 0.042	0.562 ± 0.017	1.011 ± 0.003
MICE RF (m=5)	VIQ	1923	0.150 ± 0.005	0.645 ± 0.042	0.562 ± 0.017	1.011 ± 0.003
CART (m=1)	VIQ	1923	0.150 ± 0.005	0.637 ± 0.042	0.560 ± 0.018	1.012 ± 0.003
CART (m=5)	VIQ	1923	0.150 ± 0.005	0.637 ± 0.042	0.560 ± 0.018	1.012 ± 0.003
CART+Aux (m=1)	VIQ	1923	0.150 ± 0.005	$0.649 \pm 0.042^{**}$	0.564 ± 0.017	1.012 ± 0.003
CART+Aux (m=5)	VIQ	1923	0.150 ± 0.005	$0.649 \pm 0.042^{**}$	0.564 ± 0.017	1.012 ± 0.003

AUC, C-index and HR were not calculated for Ignore (weighted) because the standard packages do not allow weights.

Table A.21: Imputed FI Prediction Summary by Block - 15% cMNAR (1/2)

Imputation	Block	N	Missingness	AUC	C-index	HR
GT	All	1923	0.150 ± 0.000	0.733 ± 0.036	0.654 ± 0.016	1.075 ± 0.007
Ignore	All	1923	0.150 ± 0.000	0.732 ± 0.037	0.653 ± 0.016	1.069 ± 0.007
Ignore (weighted)	All	1923	0.150 ± 0.000	–	–	–
Ignore20	All	1658.6	0.150 ± 0.000	0.735 ± 0.039	0.654 ± 0.017	1.069 ± 0.007
kNN	All	1923	0.150 ± 0.000	0.730 ± 0.037	0.652 ± 0.016	1.076 ± 0.007
RF	All	1923	0.150 ± 0.000	0.733 ± 0.037	0.653 ± 0.016	1.073 ± 0.007
Default (m=1)	All	1923	0.150 ± 0.000	0.733 ± 0.037	0.653 ± 0.016	1.073 ± 0.007
Default (m=5)	All	1923	0.150 ± 0.000	0.733 ± 0.037	0.654 ± 0.016	1.071 ± 0.007
MICE RF (m=1)	All	1923	0.150 ± 0.000	0.733 ± 0.037	0.653 ± 0.016	1.074 ± 0.007
MICE RF (m=5)	All	1923	0.150 ± 0.000	0.734 ± 0.037	0.654 ± 0.016	1.073 ± 0.007
CART (m=1)	All	1923	0.150 ± 0.000	0.734 ± 0.037	0.654 ± 0.016	1.073 ± 0.007
CART (m=5)	All	1923	0.150 ± 0.000	0.735 ± 0.037	0.655 ± 0.016	1.072 ± 0.007
CART+Aux (m=1)	All	1923	0.150 ± 0.000	0.730 ± 0.037	0.651 ± 0.016	1.076 ± 0.007
CART+Aux (m=5)	All	1923	0.150 ± 0.000	0.731 ± 0.037	0.652 ± 0.016	1.075 ± 0.007
GT	BPX	1923	0.140 ± 0.003	0.619 ± 0.040	0.544 ± 0.016	1.006 ± 0.002
Ignore	BPX	1923	0.140 ± 0.003	0.617 ± 0.040	0.543 ± 0.017	1.005 ± 0.002
Ignore (weighted)	BPX	1923	0.140 ± 0.003	–	–	–
Ignore20	BPX	1658.6	0.140 ± 0.003	0.614 ± 0.044	0.543 ± 0.019	1.005 ± 0.002
kNN	BPX	1923	0.140 ± 0.003	0.616 ± 0.040	0.543 ± 0.017	1.006 ± 0.002
RF	BPX	1923	0.140 ± 0.003	0.620 ± 0.040	0.546 ± 0.016	1.006 ± 0.002
Default (m=1)	BPX	1923	0.140 ± 0.003	0.620 ± 0.040	0.544 ± 0.016	1.006 ± 0.002
Default (m=5)	BPX	1923	0.140 ± 0.003	0.621 ± 0.040	0.546 ± 0.016	1.006 ± 0.002
MICE RF (m=1)	BPX	1923	0.140 ± 0.003	0.618 ± 0.040	0.543 ± 0.016	1.006 ± 0.002
MICE RF (m=5)	BPX	1923	0.140 ± 0.003	0.622 ± 0.040	0.546 ± 0.016	1.006 ± 0.002
CART (m=1)	BPX	1923	0.140 ± 0.003	0.619 ± 0.040	0.544 ± 0.016	1.006 ± 0.002
CART (m=5)	BPX	1923	0.140 ± 0.003	0.621 ± 0.040	0.544 ± 0.016	1.006 ± 0.002
CART+Aux (m=1)	BPX	1923	0.140 ± 0.003	0.618 ± 0.040	0.544 ± 0.016	1.006 ± 0.002
CART+Aux (m=5)	BPX	1923	0.140 ± 0.003	0.621 ± 0.041	0.545 ± 0.017	1.006 ± 0.002
GT	LB	1923	0.162 ± 0.001	0.689 ± 0.041	0.634 ± 0.016	1.046 ± 0.006
Ignore	LB	1923	0.162 ± 0.001	0.688 ± 0.041	0.633 ± 0.016	1.040 ± 0.005
Ignore (weighted)	LB	1923	0.162 ± 0.001	–	–	–
Ignore20	LB	1658.6	0.162 ± 0.001	0.693 ± 0.044	0.634 ± 0.018	1.040 ± 0.006
kNN	LB	1923	0.162 ± 0.001	0.685 ± 0.041	0.630 ± 0.016	1.046 ± 0.006
RF	LB	1923	0.162 ± 0.001	0.689 ± 0.041	0.633 ± 0.016	1.045 ± 0.006
Default (m=1)	LB	1923	0.162 ± 0.001	0.686 ± 0.041	0.632 ± 0.017	1.045 ± 0.006
Default (m=5)	LB	1923	0.162 ± 0.001	0.689 ± 0.042	0.635 ± 0.017	1.046 ± 0.006
MICE RF (m=1)	LB	1923	0.162 ± 0.001	0.686 ± 0.041	0.630 ± 0.016	1.044 ± 0.006
MICE RF (m=5)	LB	1923	0.162 ± 0.001	0.692 ± 0.041	0.634 ± 0.016	1.046 ± 0.006
CART (m=1)	LB	1923	0.162 ± 0.001	0.686 ± 0.041	0.631 ± 0.016	1.044 ± 0.006
CART (m=5)	LB	1923	0.162 ± 0.001	0.691 ± 0.041	0.634 ± 0.016	1.046 ± 0.006
CART+Aux (m=1)	LB	1923	0.162 ± 0.001	0.683 ± 0.042	0.630 ± 0.017	1.045 ± 0.006
CART+Aux (m=5)	LB	1923	0.162 ± 0.001	0.686 ± 0.042	0.633 ± 0.017	1.047 ± 0.006

AUC, C-index and HR were not calculated for Ignore (weighted) because the standard packages do not allow weights.

Table A.22: Imputed FI Prediction Summary by Block - 15% cMNAR (2/2)

Imputation	Block	N	Missingness	AUC	C-index	HR
GT	PFQ	1923	0.158 ± 0.002	0.607 ± 0.043	0.581 ± 0.016	1.027 ± 0.004
Ignore	PFQ	1923	0.158 ± 0.002	0.608 ± 0.043	0.581 ± 0.017	1.026 ± 0.003
Ignore (weighted)	PFQ	1923	0.158 ± 0.002	–	–	–
Ignore20	PFQ	1658.6	0.158 ± 0.002	0.611 ± 0.046	0.583 ± 0.018	1.026 ± 0.004
kNN	PFQ	1923	0.158 ± 0.002	0.607 ± 0.043	0.580 ± 0.017	1.028 ± 0.004
RF	PFQ	1923	0.158 ± 0.002	0.609 ± 0.043	0.582 ± 0.017	1.027 ± 0.004
Default (m=1)	PFQ	1923	0.158 ± 0.002	0.610 ± 0.044	0.586 ± 0.017	1.027 ± 0.004
Default (m=5)	PFQ	1923	0.158 ± 0.002	0.610 ± 0.044	0.586 ± 0.017	1.027 ± 0.004
MICE RF (m=1)	PFQ	1923	0.158 ± 0.002	0.608 ± 0.044	0.583 ± 0.017	1.028 ± 0.004
MICE RF (m=5)	PFQ	1923	0.158 ± 0.002	0.608 ± 0.044	0.583 ± 0.017	1.028 ± 0.004
CART (m=1)	PFQ	1923	0.158 ± 0.002	0.611 ± 0.044	0.585 ± 0.018	1.028 ± 0.004
CART (m=5)	PFQ	1923	0.158 ± 0.002	0.611 ± 0.044	0.585 ± 0.018	1.028 ± 0.004
CART+Aux (m=1)	PFQ	1923	0.158 ± 0.002	0.610 ± 0.043	0.586 ± 0.017	1.028 ± 0.004
CART+Aux (m=5)	PFQ	1923	0.158 ± 0.002	0.610 ± 0.043	0.586 ± 0.017	1.028 ± 0.004
GT	RXD	1923	0.055 ± 0.005	0.630 ± 0.042	0.590 ± 0.017	1.021 ± 0.003
Ignore	RXD	1816.4	0.055 ± 0.005	0.630 ± 0.043	0.590 ± 0.018	1.021 ± 0.004
Ignore (weighted)	RXD	1816.4	0.055 ± 0.005	–	–	–
Ignore20	RXD	1576.6	0.055 ± 0.005	0.633 ± 0.046	0.593 ± 0.021	1.021 ± 0.004
kNN	RXD	1923	0.055 ± 0.005	0.628 ± 0.043	0.589 ± 0.018	1.022 ± 0.004
RF	RXD	1923	0.055 ± 0.005	0.632 ± 0.042	0.591 ± 0.017	1.022 ± 0.004
Default (m=1)	RXD	1923	0.055 ± 0.005	0.630 ± 0.043	0.589 ± 0.017	1.021 ± 0.004
Default (m=5)	RXD	1923	0.055 ± 0.005	0.630 ± 0.043	0.589 ± 0.017	1.021 ± 0.004
MICE RF (m=1)	RXD	1923	0.055 ± 0.005	0.631 ± 0.042	0.590 ± 0.017	1.022 ± 0.004
MICE RF (m=5)	RXD	1923	0.055 ± 0.005	0.631 ± 0.042	0.590 ± 0.017	1.022 ± 0.004
CART (m=1)	RXD	1923	0.055 ± 0.005	0.630 ± 0.043	0.590 ± 0.018	1.022 ± 0.004
CART (m=5)	RXD	1923	0.055 ± 0.005	0.630 ± 0.043	0.590 ± 0.018	1.022 ± 0.004
CART+Aux (m=1)	RXD	1923	0.055 ± 0.005	0.630 ± 0.043	0.588 ± 0.018	1.022 ± 0.004
CART+Aux (m=5)	RXD	1923	0.055 ± 0.005	0.630 ± 0.043	0.588 ± 0.018	1.022 ± 0.004
GT	VIQ	1923	0.130 ± 0.005	0.642 ± 0.041	0.560 ± 0.016	1.011 ± 0.003
Ignore	VIQ	1918.7	0.130 ± 0.005	0.635 ± 0.042	0.555 ± 0.017	1.009 ± 0.003
Ignore (weighted)	VIQ	1918.7	0.130 ± 0.005	–	–	–
Ignore20	VIQ	1656.9	0.130 ± 0.005	0.635 ± 0.046	0.554 ± 0.020	1.009 ± 0.003
kNN	VIQ	1923	0.130 ± 0.005	0.638 ± 0.041	0.559 ± 0.016	1.011 ± 0.003
RF	VIQ	1923	0.130 ± 0.005	0.640 ± 0.041	0.559 ± 0.016	1.011 ± 0.003
Default (m=1)	VIQ	1923	0.130 ± 0.005	0.645 ± 0.042	0.561 ± 0.017	1.011 ± 0.003
Default (m=5)	VIQ	1923	0.130 ± 0.005	0.645 ± 0.042	0.561 ± 0.017	1.011 ± 0.003
MICE RF (m=1)	VIQ	1923	0.130 ± 0.005	0.643 ± 0.042	0.561 ± 0.017	1.011 ± 0.003
MICE RF (m=5)	VIQ	1923	0.130 ± 0.005	0.642 ± 0.042	0.561 ± 0.017	1.011 ± 0.003
CART (m=1)	VIQ	1923	0.130 ± 0.005	0.647 ± 0.042	0.562 ± 0.017	1.012 ± 0.003
CART (m=5)	VIQ	1923	0.130 ± 0.005	0.647 ± 0.042	0.562 ± 0.017	1.012 ± 0.003
CART+Aux (m=1)	VIQ	1923	0.130 ± 0.005	0.646 ± 0.042	0.561 ± 0.017	1.012 ± 0.003
CART+Aux (m=5)	VIQ	1923	0.130 ± 0.005	0.646 ± 0.042	0.561 ± 0.017	1.012 ± 0.003

AUC, C-index and HR were not calculated for Ignore (weighted) because the standard packages do not allow weights.

Table A.23: Lab FI variables.

Quantity	Units	Low (M)[1]	High (M)[1]	Low (F)[2]	High (F)[2]	Name	File (2003/2005)
Albumin	g/L	32	45	32	45	LBDSALSI	L40/BIOPRO
Alkaline phosphotase	U/L	20	130	20	130	LBXSAPSI	L40/BIOPRO
Bicarbonate	mmol/L	21	28	21	28	LBXSC3SI	L40/BIOPRO
Bilirubin, total	umol/L	2	21	2	21	LBDSTBSI	L40/BIOPRO
Blood pressure- diastolic	mmHG	60	90	60	90	BPXDI	BPX/BPX
Blood pressure- systolic	mmHG	90	140	90	140	BPXSY	BPX/BPX
Blood urea nitrogen	mmol/L	2.9	8.2	2.9	8.2	LBDSBUSI	L40/BIOPRO
C-reactive protein	mg/dL	0	1	0	1	LBXCRP	L11/CRP
Creatinine	umol/L	60	110	45	90	LBDSCRSI	L40/BIOPRO
Direct HDL-Cholesterol	mmol/L	1.3	∞	1.3	∞	LBDHDDSI	L13/HDL
Folate, RBC	nmol/L	376	1450	376	1450	LBDRBFSI	L06NB/FOLATE
Glucose, serum	mmol/L	3.9	6.1	3.9	6.1	LBDSGLSI	L40/BIOPRO
Glycohemoglobin levels	%	0	5.7	0	5.7	LBXGH	L10/GHB
Hemoglobin	g/dL	13.5	18	12	16	LBXHGB	L25/CBC
Iron, refigerated	umol/L	10.7	26.9	10.7	26.9	LBDSIRSI	L40/BIOPRO
Lactate dehydrogenase LDH	U/L	100	190	100	190	LBXSLDSI	L40/BIOPRO
Mean arterial pressure	mmHg	70	105	70	105	BPXMAP[3]	BPX/BPX
Mean cell volume	fL	80	96	80	96	LBXMCVSI	L25/CBC
Phosphorus	mmol/L	0.74	1.52	0.74	1.52	LBDSPHSI	L40/BIOPRO
Platelet count SI	1000 cells/uL	150	450	150	450	LBXPLTSI	L25/CBC
Protein, total	g/L	60	78	60	78	LBDSTPSI	L40/BIOPRO
Pulse	bpm	60	99	60	99	BPXPLS	BPX/BPX
Pulse pressure	mmHg	30	65	30	65	BPXPP[4]	BPX/BPX
Red cell distribution width	%	11.6	14.6	11.6	14.6	LBXRDW	L25/CBC
Segmented neutrophils percent (40-80%)	%	40	80	40	80	LBXNEPCT	L25/CBC
Sodium	mmol/L	136	142	136	142	LBXSNASI	L40/BIOPRO
Total calcium	mmol/L	2.3	2.74	2.3	2.74	LBDSCASI	L40/BIOPRO
Total Cholesterol	mmol/L	3.88	6.47	3.88	6.47	LBDSCHSI	L40/BIOPRO
Triglyceride	mmol/L	0.11	2.74	0.11	2.74	LBDSTRSI	L40/BIOPRO
Uric acid	umol/L	240	510	160	430	LBDSUASI	L40/BIOPRO
Vitamin B12, serum	pmol/L	118	701	118	701	LBDB12SI	L06NB/B12
Vitamin D	nmol/L	12	50	12	50	LBDVIDMS	VID/VID

[1] M: male.
[2] F: female.
[3] BPXMAP: BPXMeanArterialPressure.
[4] BPXPP: BPXPulsePressure.
Dichotomization converted lab values within the sex-specific ranges to 0, or 1 if out of range.

Table A.24: Self-reported Health FI variables.

Measurement	Description	Name	File
Angina/angina pectoris	Ever told you had angina/angina pectoris	MCQ160D	MCQ
Broken hip	Doctor told you hip broke/fracture	OSQ010A	OSQ
Cancer	Ever told you had cancer or malignancy	MCQ220	MCQ
Cataract operation	Ever had a cataract operation	VIQ071	VIQ
Confusion or inability to remember things	Experience confusion/memory problems	PFQ057	PFQ
Cough regularly	Do you cough most days for 3 months+?	RDQ031	RDQ
Difficulty attending social events	Difficulty attending social events	PFQ061R	PFQ
Difficulty dressing yourself	Difficulty dressing yourself	PFQ061L	PFQ
Difficulty getting in and out of bed	Difficulty getting in and out of bed	PFQ061J	PFQ
Difficulty grasping/holding small objects	Difficulty grasping/holding small objects	PFQ061P	PFQ
Difficulty lifting or carrying	Difficulty lifting or carrying	PFQ061E	PFQ
Difficulty managing money	Difficulty managing money	PFQ061A	PFQ
Difficulty preparing meals	Difficulty preparing meals	PFQ061G	PFQ
Difficulty pushing or pulling large objects	Push or pull large objects difficulty	PFQ061T	PFQ
Difficulty seeing in dim light	Difficulty seeing steps/curbs in dim light	VIQ051C	VIQ
Difficulty standing up from armless chair	Difficulty standing up from armless chair	PFQ061I	PFQ
Difficulty stooping, crouching, kneeling	Difficulty stooping, crouching, kneeling	PFQ061D	PFQ
Difficulty using fork and knife	Using fork, knife, drinking from cup	PFQ061K	PFQ
Difficulty walking between rooms	Difficulty walking between rooms on same floor	PFQ061H	PFQ
Arthritis	Doctor ever said you had arthritis	MCQ160A	MCQ
Diabetes	Doctor told you have diabetes	DIQ010	DIQ
High blood pressure	Ever told you had high blood pressure	BPQ020	BPQ
Frequency of healthcare use	#Times receive healthcare over past year	HUQ050	HUQ
General hearing	General condition of hearing	AUQ130/131[1]	AUQ
General vision	General condition of eyesight	VIQ031	VIQ
Health compared to 1 year ago	Health compared to 1 year ago	HUQ020	HUQ
Heart attack	Ever told you had heart attack	MCQ160E	MCQ
Heart disease	Ever told you had coronary heart disease	MCQ160C	MCQ
Leaked/lost control of urine	Leak urine during nonphysical activities	KIQ046	KIQ_U
Medications	Number of prescription medicines taken	RXDCOUNT	RXQ_RX
Osteoporosis	Ever told had osteoporosis/brittle bones	OSQ060	OSQ
Overnight hospital stays	Overnight hospital patient in last year	HUQ071	HUQ
Self-reported health	Would you say your health in general is...	HUQ010	HUQ
Stroke	Ever told you had a stroke	MCQ160F	MCQ
Thyroid condition	Ever told you had a thyroid problem	MCQ160M	MCQ
Weak/failing kidneys	Ever told you had weak/failing kidneys	KIQ022	KIQ_U

[1] 2003-04: AUQ130, 2004-05: AUQ131.

Table A.25: Auxiliary Lab Variables

Quantity	Description	Type	Name	File (2003/2005)
LBXPT21	Parathyroid Hormone	continuous	LBXPT21	L11/PTH
LBXWBCSI	White blood cell count	continuous	LBXWBCSI	L25/CBC
LBDLYMNO	Lymphocyte number	continuous	LBDLYMNO	L25/CBC
LBDMONO	Monocyte number	continuous	LBDMONO	L25/CBC
LBDEONO	Eosinophils number	continuous	LBDEONO	L25/CBC
LBDBANO	Basophils number	continuous	LBDBANO	L25/CBC
LBXRBCSI	Red blood cell count (million cells/uL)	continuous	LBXRBCSI	L25/CBC
LBXHCT	Hematocrit (%)	continuous	LBXHCT	L25/CBC
LBXMCHSI	Mean cell hemoglobin (pg)	continuous	LBXMCHSI	L25/CBC
LBXMC	Mean cell hemoglobin concentration (g/dL)	continuous	LBXMC	L25/CBC
LBXMPSI	Mean platelet volume (fL)	continuous	LBXMPSI	L25/CBC
LBXSATSI	Alanine aminotransferase ALT (U/L)	continuous	LBXSATSI	L40/BIOPRO
LBXSASSI	Aspartate aminotransferase AST (U/L)	continuous	LBXSASSI	L40/BIOPRO
LBXSGTSI	Gamma glutamyl transferase (U/L)	continuous	LBXSGTSI	L40/BIOPRO
LBXSKSI	Potassium (mmol/L)	continuous	LBXSKSI	L40/BIOPRO
LBXSCLSI	Chloride (mmol/L)	continuous	LBXSCLSI	L40/BIOPRO
LBXSOSSI	Osmolality (mmol/Kg)	continuous	LBXSOSSI	L40/BIOPRO
LBXSGB	Globulin (g/dL)	continuous	LBXSGB	L40/BIOPRO
PEASCST1	Blood pressure status	ordinal	PEASCST1	BPX/BPX
PEASCTM1	Blood pressure time in seconds	continuous	PEASCTM1	BPX/BPX
BPQ150A	Had food in the past 30 minutes?	categorical	BPQ150A	BPX/BPX
BPQ150C	Had coffee in the past 30 minutes?	categorical	BPQ150C	BPX/BPX
BPQ150D	Had cigarettes in the past 30 minutes?	categorical	BPQ150D	BPX/BPX
BPAARM	Arm selected	categorical	BPAARM	BPX/BPX
BPACSZ	Coded cuff size	ordinal	BPACSZ	BPX/BPX
BPXPULS	Pulse regular or irregular?	categorical	BPXPULS	BPX/BPX
BPXPTY	Pulse type	categorical	BPXPTY	BPX/BPX

Table A.26: Auxiliary Self-reported Health Variables (1/2)

Measurement	Description	Type	Name	File
AUQ150	Ever worn a hearing aid	categorical	AUQ150	AUQ
BPQ060	Ever had blood cholesterol checked	categorical	BPQ060	BPQ
DIQ050	Taking insulin now	categorical	DIQ050	DIQ
HUQ030	Routine place to go for healthcare	categorical	HUQ030	HUQ
HUQ060	How long since last healthcare visit	ordinal	HUQ060	HUQ
HUQ090	Seen mental health professional - past yr[1]	categorical	HUQ090	HUQ
KIQ042	Leak urine during physical activities	categorical	KIQ042	KIQ_U
KIQ044	Urinated before reaching the toilet	categorical	KIQ044	KIQ_U
MCQ010	Ever been told you have asthma	categorical	MCQ010	MCQ
MCQ053	Taking treatment for anemia - past 3 mos[2]	categorical	MCQ053	MCQ
MCQ080	Doctor ever said you were overweight	categorical	MCQ080	MCQ
MCQ092	Ever receive blood transfusion	categorical	MCQ092	MCQ
MCQ140	Trouble seeing even with glass/contacts	categorical	MCQ140	MCQ
MCQ160B	Ever told had congestive heart failure	categorical	MCQ160B	MCQ
MCQ160G	Ever told you had emphysema	categorical	MCQ160G	MCQ
MCQ160K	Ever told you had chronic bronchitis	categorical	MCQ160K	MCQ
MCQ160L	Ever told you had any liver condition	categorical	MCQ160L	MCQ
MCQ245A	Work days missed for illness/maternity	categorical	MCQ245A	MCQ
MCQ265	Blood relative have/had prostate cancer	categorical	MCQ265	MCQ
OSQ010B	Broken or fractured a wrist	categorical	OSQ010B	OSQ
OSQ010C	Broken or fractured spine	categorical	OSQ010C	OSQ
PFQ049	Limitations keeping you from working	categorical	PFQ049	PFQ
PFQ051	Limited in amount of work you can do	categorical	PFQ051	PFQ
PFQ054	Need special equipment to walk	categorical	PFQ054	PFQ
PFQ059	Physical, mental, emotional limitations	categorical	PFQ059	PFQ
PFQ061B	Walking for a quarter mile difficulty	categorical	PFQ061B	PFQ
PFQ061C	Walking up ten steps difficulty	categorical	PFQ061C	PFQ
PFQ061F	House chore difficulty	categorical	PFQ061F	PFQ
PFQ061M	Standing for long periods difficulty	categorical	PFQ061M	PFQ
PFQ061N	Sitting for long periods difficulty	categorical	PFQ061N	PFQ
PFQ061O	Reaching up over head difficulty	categorical	PFQ061O	PFQ
PFQ061Q	Going out to movies, events difficulty	categorical	PFQ061Q	PFQ
PFQ061S	Leisure activity at home difficulty	categorical	PFQ061S	PFQ
PFQ090	Require special healthcare equipment	categorical	PFQ090	PFQ
RDQ050	Bring up phlegm most days - 3 mo period	categorical	RDQ050	RDQ
RDQ070	Wheezing or whistling in chest - past yr	categorical	RDQ070	RDQ
RDQ140	Had dry cough at night in past year	categorical	RDQ140	RDQ
VIQ041	Time worrying about eyesight	ordinal	VIQ041	VIQ
VIQ051A	Difficulty reading ordinary newsprint	ordinal	VIQ051A	VIQ
VIQ051B	Difficulty with up close work or chores	ordinal	VIQ051B	VIQ
VIQ051D	Difficulty noticing objects to side	ordinal	VIQ051D	VIQ
VIQ051E	Difficulty finding object on crowded shelf	ordinal	VIQ051E	VIQ
VIQ056	Difficulty driving daytime-familiar place	ordinal	VIQ056	VIQ
VIQ061	Vision limits how long can do activities	ordinal	VIQ061	VIQ

[1] yr: year.
[2] mos: months.

Table A.27: Auxiliary Self-reported Health Variables (2/2)

Measurement	Description	Type	Name	File
RIDEXMON	Six month time period of exam	categorical	RIDEXMON	DEMO
RIAGENDR	Gender	categorical	RIAGENDR	DEMO
RIDAGEYR	Age at screeening adjudicated	continuous	RIDAGEYR	DEMO
RIDRETH1	Race/Ethnicity - recode	categorical	RIDRETH1	DEMO
DMQMILIT	Veteran/Military status	categorical	DMQMILIT	DEMO
DMDBORN	Country of birth - recode	categorical	DMDBORN	DEMO
DMDCITZN	Citizenship status	categorical	DMDCITZN	DEMO
DMDYRSUS	Length of time in US	continuous	DMDYRSUS	DEMO
DMDMARTL	Marital status	categorical	DMDMARTL	DEMO
DMDHHSIZ	Total number of people in the household	continuous	DMDHHSIZ	DEMO
INDHHINC	Income but with inconvenient coding	categorical	INDHHINC	DEMO
INDFMPIR	PIR truncated at 5	continuous	INDFMPIR	DEMO
RIDEXPRG	Pregnancy Status at Exam - recode	categorical	RIDEXPRG	DEMO
DMDHRGND	HH ref person gender	categorical	DMDHRGND	DEMO
DMDHRAGE	Age of reference person	continuous	DMDHRAGE	DEMO
DMDHRBRN	HH ref person country of birth	categorical	DMDHRBRN	DEMO
DMDHREDU	HH ref person education level	ordinal	DMDHREDU	DEMO
DMDHRMAR	HH ref person marital status	categorical	DMDHRMAR	DEMO
DMDHSEDU	HH ref person's spouse education level	ordinal	DMDHSEDU	DEMO
SIALANG	Language of SP interview	categorical	SIALANG	DEMO
SIAPROXY	Proxy used in SP interview?	categorical	SIAPROXY	DEMO
SIAINTRP	Interpreter used in SP interview?	categorical	SIAINTRP	DEMO
FIALANG	Language of family interview	categorical	FIALANG	DEMO
FIAPROXY	Proxy used in family interview?	categorical	FIAPROXY	DEMO
FIAINTRP	Interpreter used in family interview?	categorical	FIAINTRP	DEMO
MIALANG	Language of MEC interview	categorical	MIALANG	DEMO
MIAPROXY	Proxy used in MEC interview?	categorical	MIAPROXY	DEMO
MIAINTRP	Interpreter used in MEC interview?	categorical	MIAINTRP	DEMO
AIALANG	Language of ACASI interview	categorical	AIALANG	DEMO

Table A.28: Imputed FI Statistics for Real Missingness - Summary by Block

Imputation	Block	N	Mean FI	'Bias'[1,2]	FI SD	SD 'Bias'[1]	C-index	HR	AUC
Ignore	All	9307	0.1442	0.0000 ± 0.0000	0.0782	0.0000	0.683 ± 0.009	1.077 ± 0.004	0.832 ± 0.017
Ignore (weighted)	All	9307	0.1477	$-0.0036 \pm 0.0011^{**}$	0.0785	−0.0004	–	–	–
Ignore20	All	5701	0.1611	0.0000 ± 0.0000	0.0801	−0.0019	0.677 ± 0.010	1.078 ± 0.004	$0.792 \pm 0.021^{**}$
kNN	All	9307	0.1601	$-0.0160 \pm 0.0004^{***}$	0.0710	0.0071	0.664 ± 0.009	1.073 ± 0.004	$0.773 \pm 0.021^{***}$
Default (m=1)	All	9307	0.1437	$0.0005 \pm 0.0002^{*}$	0.0739	0.0042	0.678 ± 0.009	1.076 ± 0.004	$0.819 \pm 0.018^{***}$
Default (m=5)	All	9307	0.1466	$-0.0024 \pm 0.0005^{***}$	0.0877	−0.0095	0.684 ± 0.009	1.077 ± 0.004	0.829 ± 0.018
CART (m=1)	All	9307	0.1382	$0.0059 \pm 0.0002^{***}$	0.0751	0.0031	0.675 ± 0.009	1.077 ± 0.004	$0.826 \pm 0.018^{*}$
CART (m=5)	All	9307	0.1412	$0.0029 \pm 0.0004^{***}$	0.0816	−0.0034	0.684 ± 0.009	1.079 ± 0.004	$0.839 \pm 0.017^{***}$
CART+Aux (m=1)	All	9307	0.1381	$0.0061 \pm 0.0002^{***}$	0.0745	0.0037	0.679 ± 0.009	1.077 ± 0.004	0.832 ± 0.017
CART+Aux (m=5)	All	9307	0.1410	$0.0031 \pm 0.0003^{***}$	0.0784	−0.0003	0.686 ± 0.009	1.079 ± 0.004	$0.841 \pm 0.017^{***}$
Ignore	BPX	8889	0.2046	0.0000 ± 0.0000	0.2465	0.0000	0.524 ± 0.010	1.003 ± 0.001	0.667 ± 0.024
Ignore (weighted)	BPX	8889	0.2123	$-0.0078 \pm 0.0037^{*}$	0.2477	−0.0012	–	–	–
Ignore20	BPX	5544	0.2380	0.0000 ± 0.0000	0.2613	−0.0148	0.515 ± 0.011	1.003 ± 0.001	0.637 ± 0.026
kNN	BPX	9307	0.1983	$0.0012 \pm 0.0005^{*}$	0.2389	0.0077	0.522 ± 0.010	1.003 ± 0.001	$0.666 \pm 0.023^{*}$
Default (m=1)	BPX	9307	0.1978	$0.0010 \pm 0.0005^{*}$	0.2407	0.0058	0.513 ± 0.010	1.002 ± 0.001	0.647 ± 0.024
Default (m=5)	BPX	9307	0.2064	-0.0011 ± 0.0007	0.2572	−0.0107	0.522 ± 0.010	1.003 ± 0.001	$0.670 \pm 0.022^{**}$
CART (m=1)	BPX	9307	0.1969	$0.0013 \pm 0.0005^{*}$	0.2405	0.0060	0.513 ± 0.010	1.002 ± 0.001	0.650 ± 0.024
CART (m=5)	BPX	9307	0.2059	-0.0008 ± 0.0007	0.2571	−0.0106	0.523 ± 0.009	1.003 ± 0.001	$0.673 \pm 0.022^{***}$
CART+Aux (m=1)	BPX	9307	0.1964	$0.0014 \pm 0.0005^{**}$	0.2407	0.0058	0.512 ± 0.010	1.001 ± 0.001	0.645 ± 0.024
CART+Aux (m=5)	BPX	9307	0.2056	-0.0008 ± 0.0007	0.2514	−0.0049	0.523 ± 0.010	1.003 ± 0.001	$0.668 \pm 0.022^{**}$
Ignore	LB	8862	0.1379	0.0000 ± 0.0000	0.0869	0.0000	0.654 ± 0.010	1.054 ± 0.003	0.720 ± 0.024
Ignore (weighted)	LB	8862	0.1391	-0.0011 ± 0.0013	0.0863	0.0005	–	–	–
Ignore20	LB	5701	0.1425	0.0000 ± 0.0000	0.0889	−0.0021	0.655 ± 0.010	1.056 ± 0.003	0.716 ± 0.025
kNN	LB	9307	0.1322	$0.0013 \pm 0.0002^{***}$	0.0861	0.0008	0.627 ± 0.010	1.048 ± 0.003	$0.677 \pm 0.025^{**}$
Default (m=1)	LB	9307	0.1327	$0.0011 \pm 0.0001^{***}$	0.0865	0.0003	0.642 ± 0.010	1.052 ± 0.003	0.697 ± 0.024
Default (m=5)	LB	9307	0.1383	0.0000 ± 0.0002	0.0888	−0.0019	0.658 ± 0.009	1.057 ± 0.003	0.721 ± 0.023
CART (m=1)	LB	9307	0.1324	$0.0011 \pm 0.0001^{***}$	0.0864	0.0005	0.633 ± 0.010	1.049 ± 0.003	$0.683 \pm 0.025^{*}$
CART (m=5)	LB	9307	0.1383	0.0000 ± 0.0002	0.0912	−0.0043	0.657 ± 0.009	1.056 ± 0.003	0.721 ± 0.022
CART+Aux (m=1)	LB	9307	0.1328	$0.0012 \pm 0.0001^{***}$	0.0861	0.0007	0.634 ± 0.010	1.049 ± 0.003	$0.685 \pm 0.024^{*}$
CART+Aux (m=5)	LB	9307	0.1385	0.0001 ± 0.0002	0.0882	−0.0013	0.654 ± 0.009	1.055 ± 0.003	0.716 ± 0.023
Ignore	PFQ	4425	0.1199	0.0000 ± 0.0000	0.1599	0.0000	0.613 ± 0.010	1.023 ± 0.002	0.649 ± 0.026
Ignore (weighted)	PFQ	4425	0.1180	0.0019 ± 0.0030	0.1599	0.0001	–	–	–
Ignore20	PFQ	4156	0.1173	0.0000 ± 0.0000	0.1566	0.0033	0.611 ± 0.011	1.023 ± 0.002	0.647 ± 0.028
kNN	PFQ	9307	0.2276	$0.0012 \pm 0.0002^{***}$	0.1982	−0.0382	0.573 ± 0.010	1.018 ± 0.002	$0.541 \pm 0.025^{*}$
Default (m=1)	PFQ	9307	0.1304	$-0.0016 \pm 0.0002^{***}$	0.1375	0.0224	0.610 ± 0.010	1.023 ± 0.002	$0.594 \pm 0.028^{**}$
Default (m=5)	PFQ	9307	0.1304	$-0.0016 \pm 0.0003^{***}$	0.2649	−0.1049	0.610 ± 0.010	1.023 ± 0.002	$0.593 \pm 0.028^{**}$
CART (m=1)	PFQ	9307	0.1008	$-0.0008 \pm 0.0002^{***}$	0.1183	0.0416	0.619 ± 0.010	1.025 ± 0.002	$0.643 \pm 0.029^{*}$
CART (m=5)	PFQ	9307	0.1008	$-0.0008 \pm 0.0003^{**}$	0.2022	−0.0423	0.619 ± 0.010	1.025 ± 0.002	$0.643 \pm 0.029^{*}$
CART+Aux (m=1)	PFQ	9307	0.1003	$-0.0012 \pm 0.0002^{***}$	0.1161	0.0439	0.629 ± 0.010	1.025 ± 0.002	$0.651 \pm 0.029^{**}$
CART+Aux (m=5)	PFQ	9307	0.1003	$-0.0012 \pm 0.0003^{***}$	0.1584	0.0015	0.629 ± 0.010	1.025 ± 0.002	$0.651 \pm 0.029^{**}$
Ignore	RXD	5269	0.1796	0.0000 ± 0.0000[3]	0.1379	0.0000	0.608 ± 0.010	1.026 ± 0.002	0.688 ± 0.025
Ignore (weighted)	RXD	5269	0.1838	-0.0042 ± 0.0024	0.1370	0.0009	–	–	–
Ignore20	RXD	4282	0.1910	0.0000 ± 0.0000	0.1402	−0.0023	0.601 ± 0.011	1.024 ± 0.002	0.663 ± 0.027
kNN	RXD	9307	0.1351	0.0000 ± 0.0000	0.1185	0.0194	0.607 ± 0.010	1.027 ± 0.002	0.743 ± 0.021
Default (m=1)	RXD	9307	0.1435	0.0000 ± 0.0000	0.1154	0.0225	0.604 ± 0.010	1.026 ± 0.002	0.725 ± 0.024
Default (m=5)	RXD	9307	0.1435	0.0000 ± 0.0000	0.1340	0.0038	0.604 ± 0.010	1.026 ± 0.002	0.725 ± 0.024
CART (m=1)	RXD	9307	0.1448	0.0000 ± 0.0000	0.1149	0.0229	0.609 ± 0.010	1.027 ± 0.002	0.728 ± 0.023
CART (m=5)	RXD	9307	0.1448	0.0000 ± 0.0000	0.1454	−0.0075	0.609 ± 0.010	1.027 ± 0.002	0.728 ± 0.023
CART+Aux (m=1)	RXD	9307	0.1433	0.0000 ± 0.0000	0.1157	0.0222	0.601 ± 0.010	1.026 ± 0.002	0.729 ± 0.023
CART+Aux (m=5)	RXD	9307	0.1433	0.0000 ± 0.0000	0.1248	0.0130	0.601 ± 0.010	1.026 ± 0.002	0.729 ± 0.023
Ignore	VIQ	6993	0.1517	0.0000 ± 0.0000	0.1788	0.0000	0.562 ± 0.010	1.011 ± 0.001	0.717 ± 0.024
Ignore (weighted)	VIQ	6993	0.1579	$-0.0061 \pm 0.0029^{*}$	0.1868	−0.0080	–	–	–
Ignore20	VIQ	5303	0.1706	0.0000 ± 0.0000	0.1879	−0.0091	0.558 ± 0.011	1.010 ± 0.002	0.696 ± 0.026
kNN	VIQ	9307	0.1334	$0.0025 \pm 0.0003^{***}$	0.1540	0.0248	0.560 ± 0.010	1.011 ± 0.002	0.728 ± 0.024
Default (m=1)	VIQ	9307	0.1372	$0.0010 \pm 0.0003^{***}$	0.1566	0.0222	0.569 ± 0.010	1.012 ± 0.002	0.729 ± 0.024
Default (m=5)	VIQ	9307	0.1372	$0.0010 \pm 0.0004^{**}$	0.1725	0.0063	0.569 ± 0.010	1.012 ± 0.002	0.730 ± 0.024
CART (m=1)	VIQ	9307	0.1366	$0.0010 \pm 0.0003^{***}$	0.1568	0.0220	0.571 ± 0.010	1.012 ± 0.002	0.734 ± 0.024
CART (m=5)	VIQ	9307	0.1366	$0.0010 \pm 0.0004^{*}$	0.1998	−0.0210	0.571 ± 0.010	1.012 ± 0.002	0.735 ± 0.023
CART+Aux (m=1)	VIQ	9307	0.1333	0.0005 ± 0.0003	0.1574	0.0214	0.565 ± 0.010	1.012 ± 0.002	0.737 ± 0.023
CART+Aux (m=5)	VIQ	9307	0.1333	0.0005 ± 0.0006	0.1674	0.0114	0.565 ± 0.010	1.012 ± 0.002	0.738 ± 0.023

[1] This is the bias proxy: Ignore − Value. NAs were excluded.

[2] NA exclusions may cause the average of the bias (this column) to differ from the bias of the averages.

[3] RXD has only 1 variable and hence the bias must be 0.

Table A.29: Imputed FI Statistics for Real Missingness, RI-based imputations - Summary by Block (1/2)

Imputation	Block	N	Mean FI	'Bias'[1,2]	FI SD	SD 'Bias'[1]	C-index	HR	AUC
Ignore	All	9307	0.1442	0.0000 ± 0.0000	0.0782	0.0000	0.683 ± 0.009	1.077 ± 0.004	0.832 ± 0.017
Ignore+RI	All	9307	0.1330	$0.0112 \pm 0.0001^{***}$	0.0803	−0.0021	0.686 ± 0.009	1.077 ± 0.003	$0.851 \pm 0.016^{***}$
Ignore (weighted) + RI	All	9307	0.1330	$0.0112 \pm 0.0001^{***}$	0.0803	−0.0021	0.686 ± 0.009	1.078 ± 0.004	$0.851 \pm 0.016^{***}$
Ignore20 + RI	All	8728	0.1327	$0.0108 \pm 0.0001^{***}$	0.0774	0.0008	0.680 ± 0.010	1.079 ± 0.004	0.848 ± 0.017
kNN + RI	All	9307	0.1302	$0.0140 \pm 0.0002^{***}$	0.0771	0.0011	0.677 ± 0.009	1.076 ± 0.004	$0.841 \pm 0.016^{***}$
Default+RI (m=1)	All	9307	0.1309	$0.0133 \pm 0.0002^{***}$	0.0781	0.0001	0.678 ± 0.009	1.076 ± 0.004	$0.842 \pm 0.016^{***}$
Default+RI (m=5)	All	9307	0.1338	$0.0104 \pm 0.0002^{***}$	0.0790	−0.0009	0.684 ± 0.009	1.077 ± 0.004	$0.850 \pm 0.016^{***}$
CART+RI (m=1)	All	9307	0.1307	$0.0135 \pm 0.0002^{***}$	0.0776	0.0005	0.678 ± 0.009	1.077 ± 0.004	$0.841 \pm 0.016^{***}$
CART+RI (m=5)	All	9307	0.1336	$0.0106 \pm 0.0002^{***}$	0.0789	−0.0007	0.685 ± 0.009	1.079 ± 0.004	$0.851 \pm 0.016^{***}$
CART+Aux+RI (m=1)	All	9307	0.1305	$0.0137 \pm 0.0002^{***}$	0.0777	0.0005	0.679 ± 0.009	1.077 ± 0.004	$0.843 \pm 0.016^{***}$
CART+Aux+RI (m=5)	All	9307	0.1334	$0.0107 \pm 0.0002^{***}$	0.0786	−0.0005	0.686 ± 0.009	1.079 ± 0.004	$0.852 \pm 0.016^{***}$
Ignore	BPX	8889	0.2046	0.0000 ± 0.0000	0.2465	0.0000	0.524 ± 0.010	1.003 ± 0.001	0.667 ± 0.024
Ignore+RI	BPX	8889	0.2046	0.0000 ± 0.0000	0.2465	0.0000	0.524 ± 0.010	1.003 ± 0.001	0.667 ± 0.024
Ignore (weighted) + RI	BPX	8889	0.2046	0.0000 ± 0.0000	0.2465	0.0000	0.524 ± 0.010	1.003 ± 0.001	0.667 ± 0.024
Ignore20 + RI	BPX	8410	0.2033	0.0000 ± 0.0000	0.2458	0.0007	0.520 ± 0.010	1.003 ± 0.001	0.665 ± 0.026
kNN + RI	BPX	9307	0.2003	$0.0011 \pm 0.0005^{*}$	0.2397	0.0068	0.521 ± 0.009	1.003 ± 0.001	0.661 ± 0.023
Default+RI (m=1)	BPX	9307	0.1980	$0.0012 \pm 0.0005^{*}$	0.2408	0.0057	0.515 ± 0.010	1.001 ± 0.001	0.652 ± 0.024
Default+RI (m=5)	BPX	9307	0.2069	-0.0011 ± 0.0007	0.2558	−0.0093	0.523 ± 0.010	1.003 ± 0.001	$0.674 \pm 0.022^{**}$
CART+RI (m=1)	BPX	9307	0.1974	$0.0013 \pm 0.0005^{*}$	0.2405	0.0060	0.510 ± 0.010	1.001 ± 0.001	0.647 ± 0.024
CART+RI (m=5)	BPX	9307	0.2059	-0.0009 ± 0.0006	0.2539	−0.0073	0.523 ± 0.009	1.003 ± 0.001	$0.672 \pm 0.022^{***}$
CART+Aux+RI (m=1)	BPX	9307	0.1971	$0.0011 \pm 0.0005^{*}$	0.2407	0.0058	0.512 ± 0.009	1.001 ± 0.001	0.655 ± 0.023
CART+Aux+RI (m=5)	BPX	9307	0.2059	-0.0010 ± 0.0008	0.2644	−0.0179	0.522 ± 0.009	1.003 ± 0.001	$0.675 \pm 0.022^{**}$

[1] This is the bias proxy: Ignore − Value. NAs were excluded.

[2] NA exclusions may cause the average of the bias (this column) to differ from the bias of the averages.

Table A.30: Imputed FI Statistics for Real Missingness, RI-based imputations - Summary by Block (2/2)

Imputation	Block	N	Mean FI	'Bias'[1,2]	FI SD	SD 'Bias'[1]	C-index	HR	AUC
Ignore	LB	8862	0.1379	0.0000 ± 0.0000	0.0869	0.0000	0.654 ± 0.010	1.054 ± 0.003	0.720 ± 0.024
Ignore+RI	LB	8862	0.1379	0.0000 ± 0.0000	0.0869	0.0000	0.654 ± 0.010	1.054 ± 0.003	0.720 ± 0.024
Ignore (weighted) + RI	LB	8862	0.1379	0.0000 ± 0.0000	0.0869	0.0000	0.654 ± 0.010	1.054 ± 0.003	0.720 ± 0.024
Ignore20 + RI	LB	8728	0.1377	0.0000 ± 0.0000	0.0858	0.0010	0.655 ± 0.010	1.056 ± 0.003	0.720 ± 0.024
kNN + RI	LB	9307	0.1327	$0.0011 \pm 0.0001^{***}$	0.0858	0.0010	0.629 ± 0.010	1.048 ± 0.003	$0.680 \pm 0.025^{*}$
Default+RI (m=1)	LB	9307	0.1328	$0.0010 \pm 0.0001^{***}$	0.0866	0.0002	0.640 ± 0.010	1.052 ± 0.003	$0.695 \pm 0.024^{*}$
Default+RI (m=5)	LB	9307	0.1384	0.0000 ± 0.0002	0.0885	−0.0016	0.655 ± 0.009	1.056 ± 0.003	0.718 ± 0.023
CART+RI (m=1)	LB	9307	0.1327	$0.0011 \pm 0.0001^{***}$	0.0864	0.0005	0.635 ± 0.010	1.050 ± 0.003	$0.688 \pm 0.025^{*}$
CART+RI (m=5)	LB	9307	0.1385	0.0000 ± 0.0002	0.0914	−0.0046	0.655 ± 0.009	1.056 ± 0.003	0.720 ± 0.023
CART+Aux+RI (m=1)	LB	9307	0.1325	$0.0012 \pm 0.0001^{***}$	0.0864	0.0005	0.636 ± 0.010	1.050 ± 0.003	$0.689 \pm 0.024^{*}$
CART+Aux+RI (m=5)	LB	9307	0.1383	0.0001 ± 0.0002	0.0915	−0.0046	0.654 ± 0.009	1.055 ± 0.003	0.716 ± 0.023
Ignore	PFQ	4425	0.1199	0.0000 ± 0.0000	0.1599	0.0000	0.613 ± 0.010	1.023 ± 0.002	0.649 ± 0.026
Ignore+RI	PFQ	9306	0.0570	0.0000 ± 0.0000	0.1255	0.0344	0.627 ± 0.010	1.025 ± 0.002	0.770 ± 0.021
Ignore (weighted) + RI	PFQ	9306	0.0570	0.0000 ± 0.0000	0.1255	0.0344	0.627 ± 0.010	1.025 ± 0.002	0.770 ± 0.021
Ignore20 + RI	PFQ	8728	0.0559	0.0000 ± 0.0000	0.1229	0.0370	0.622 ± 0.010	1.025 ± 0.002	$0.769 \pm 0.022^{***}$
kNN + RI	PFQ	9307	0.0565	$0.0014 \pm 0.0002^{***}$	0.1245	0.0354	0.627 ± 0.010	1.025 ± 0.002	$0.769 \pm 0.021^{*}$
Default+RI (m=1)	PFQ	9307	0.0577	$-0.0014 \pm 0.0002^{***}$	0.1269	0.0330	0.628 ± 0.010	1.025 ± 0.002	$0.773 \pm 0.020^{**}$
Default+RI (m=5)	PFQ	9307	0.0577	$-0.0014 \pm 0.0003^{***}$	0.1273	0.0326	0.628 ± 0.010	1.025 ± 0.002	$0.773 \pm 0.020^{**}$
CART+RI (m=1)	PFQ	9307	0.0574	$-0.0008 \pm 0.0002^{***}$	0.1256	0.0343	0.628 ± 0.010	1.025 ± 0.002	$0.771 \pm 0.020^{*}$
CART+RI (m=5)	PFQ	9307	0.0574	$-0.0008 \pm 0.0003^{**}$	0.1262	0.0337	0.628 ± 0.010	1.025 ± 0.002	$0.771 \pm 0.020^{*}$
CART+Aux+RI (m=1)	PFQ	9307	0.0576	$-0.0011 \pm 0.0002^{***}$	0.1253	0.0347	0.631 ± 0.010	1.025 ± 0.002	$0.776 \pm 0.020^{**}$
CART+Aux+RI (m=5)	PFQ	9307	0.0576	$-0.0011 \pm 0.0004^{**}$	0.1261	0.0339	0.631 ± 0.010	1.025 ± 0.002	$0.776 \pm 0.020^{**}$
Ignore	RXD	5269	0.1796	0.0000 ± 0.0000[3]	0.1379	0.0000	0.608 ± 0.010	1.026 ± 0.002	0.688 ± 0.025
Ignore+RI	RXD	5269	0.1796	0.0000 ± 0.0000	0.1379	0.0000	0.608 ± 0.010	1.026 ± 0.002	0.688 ± 0.025
Ignore (weighted) + RI	RXD	5269	0.1796	0.0000 ± 0.0000	0.1379	0.0000	0.608 ± 0.010	1.026 ± 0.002	0.688 ± 0.025
Ignore20 + RI	RXD	4967	0.1783	0.0000 ± 0.0000	0.1366	0.0013	0.602 ± 0.011	1.024 ± 0.002	0.686 ± 0.027
kNN + RI	RXD	9307	0.1330	0.0000 ± 0.0000	0.1192	0.0187	0.607 ± 0.010	1.026 ± 0.002	0.743 ± 0.021
Default+RI (m=1)	RXD	9307	0.1439	0.0000 ± 0.0000	0.1154	0.0225	0.606 ± 0.010	1.026 ± 0.002	0.730 ± 0.023
Default+RI (m=5)	RXD	9307	0.1439	0.0000 ± 0.0000	0.1411	−0.0032	0.606 ± 0.010	1.026 ± 0.002	0.730 ± 0.023
CART+RI (m=1)	RXD	9307	0.1446	0.0000 ± 0.0000	0.1154	0.0225	0.605 ± 0.010	1.027 ± 0.002	0.730 ± 0.023
CART+RI (m=5)	RXD	9307	0.1446	0.0000 ± 0.0000	0.1652	−0.0273	0.605 ± 0.010	1.027 ± 0.002	0.731 ± 0.023
CART+Aux+RI (m=1)	RXD	9307	0.1435	0.0000 ± 0.0000	0.1161	0.0218	0.607 ± 0.010	1.027 ± 0.002	0.737 ± 0.023
CART+Aux+RI (m=5)	RXD	9307	0.1435	0.0000 ± 0.0000	0.1323	0.0056	0.607 ± 0.010	1.027 ± 0.002	0.737 ± 0.023
Ignore	VIQ	6993	0.1517	0.0000 ± 0.0000	0.1788	0.0000	0.562 ± 0.010	1.011 ± 0.001	0.717 ± 0.024
Ignore+RI	VIQ	6993	0.1517	0.0000 ± 0.0000	0.1788	0.0000	0.562 ± 0.010	1.011 ± 0.001	0.717 ± 0.024
Ignore (weighted) + RI	VIQ	6993	0.1517	0.0000 ± 0.0000	0.1788	0.0000	0.562 ± 0.010	1.011 ± 0.001	0.717 ± 0.024
Ignore20 + RI	VIQ	6549	0.1508	0.0000 ± 0.0000	0.1772	0.0016	0.557 ± 0.010	1.010 ± 0.002	0.718 ± 0.025
kNN + RI	VIQ	9307	0.1336	$0.0014 \pm 0.0003^{***}$	0.1556	0.0232	0.561 ± 0.010	1.012 ± 0.002	0.733 ± 0.023
Default+RI (m=1)	VIQ	9307	0.1373	0.0005 ± 0.0003	0.1574	0.0214	0.568 ± 0.010	1.012 ± 0.002	0.731 ± 0.024
Default+RI (m=5)	VIQ	9307	0.1373	0.0005 ± 0.0003	0.1745	0.0043	0.568 ± 0.010	1.012 ± 0.002	0.732 ± 0.024
CART+RI (m=1)	VIQ	9307	0.1353	$0.0009 \pm 0.0003^{***}$	0.1575	0.0213	0.571 ± 0.010	1.012 ± 0.002	0.734 ± 0.024
CART+RI (m=5)	VIQ	9307	0.1353	$0.0009 \pm 0.0004^{*}$	0.1855	−0.0067	0.571 ± 0.010	1.012 ± 0.002	0.735 ± 0.023
CART+Aux+RI (m=1)	VIQ	9307	0.1328	0.0003 ± 0.0003	0.1578	0.0210	0.571 ± 0.010	1.012 ± 0.002	0.741 ± 0.023
CART+Aux+RI (m=5)	VIQ	9307	0.1328	0.0003 ± 0.0004	0.1756	0.0032	0.571 ± 0.010	1.012 ± 0.002	0.742 ± 0.023

[1] This is the bias proxy: Ignore − Value. NAs were excluded.
[2] NA exclusions may cause the average of the bias (this column) to differ from the bias of the averages.
[3] RXD has only 1 variable and hence the bias must be 0.

A.7 Figures

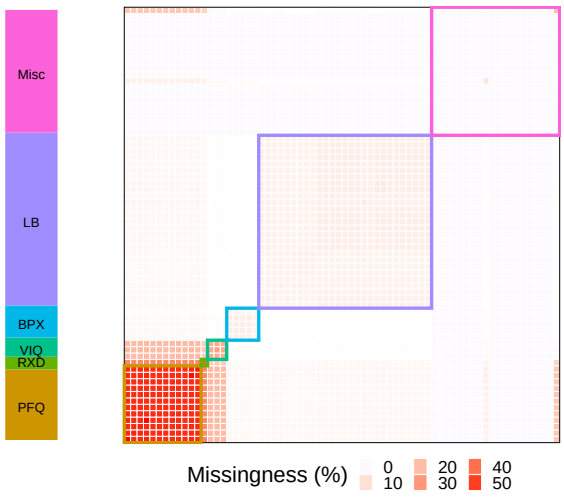

Figure A.1: Mutual missingness histogram of Full dataset. In contrast to Figure 4.2, young and old patients have not been separated. Note: variables are in the same order as Figure A.4.

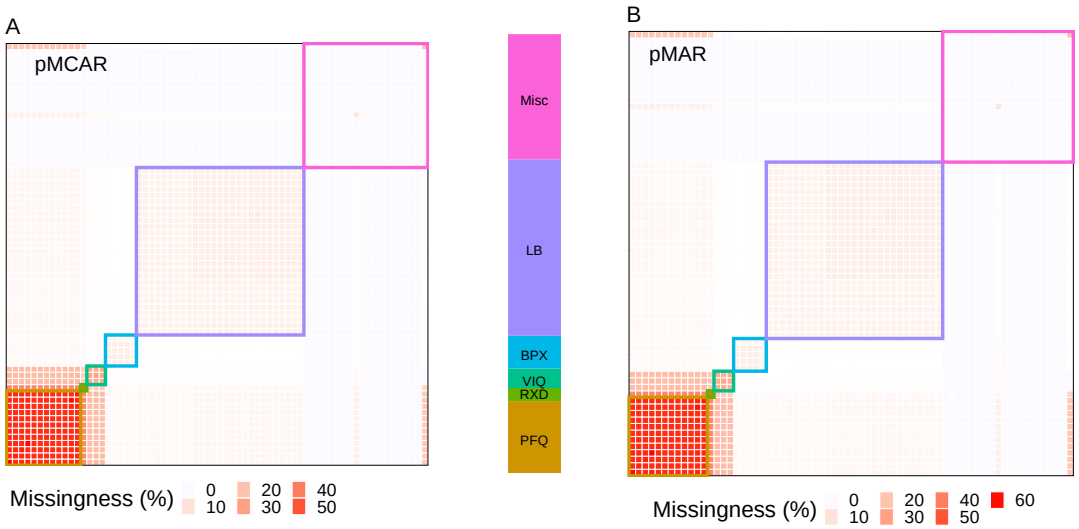

Figure A.2: Mutual missingness histogram of pMCAR and pMAR simulated data. A) pMCAR and B) pMAR. We see virtually identical results to Figure A.1, confirming our amputation preserved the patterns of missingness. Note: variables are in the same order as Figure A.4.

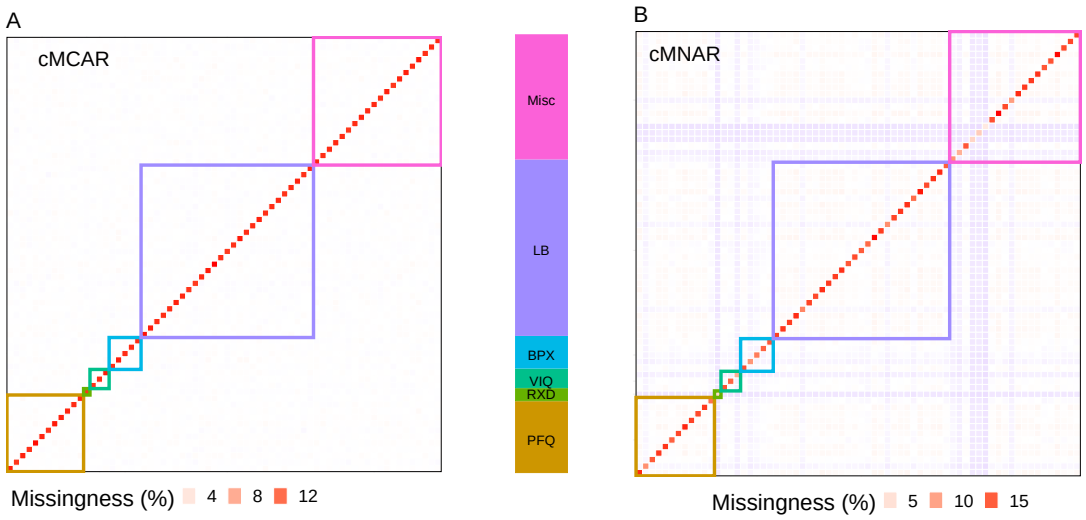

Figure A.3: Mutual missingness histogram of cMCAR and cMNAR simulated data. A) cMCAR and B) cMNAR. We see no patterns of missingness for cMCAR, as expected. For cMNAR we see some patterns of low missingness have begun to emerge for variables preferentially made not-missing. Note: variables are in the same order as Figure A.4.

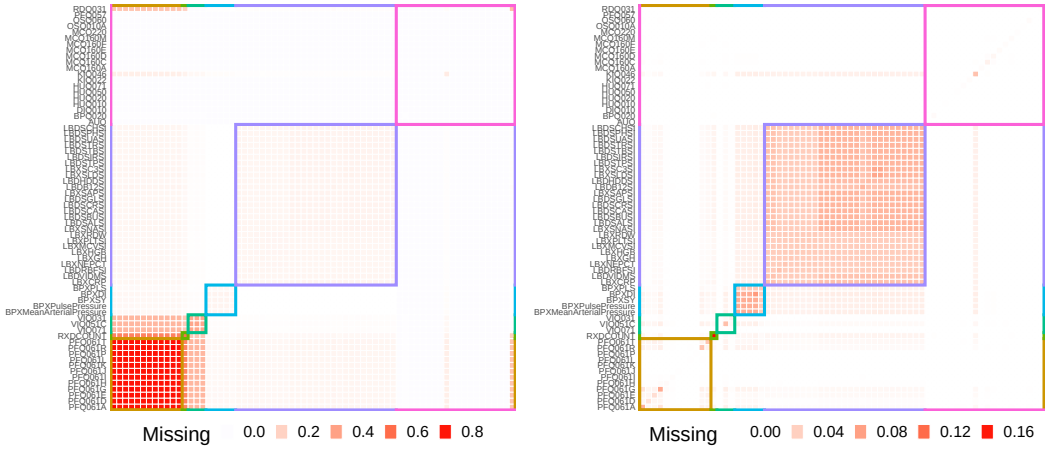

Figure A.4: Mutual missingness histograms with full variable names. The left-to-right x-axis is identical to the bottom-to-top y-axis. Left: missingness fraction of NHANES variables for young individuals (under 60). This histogram gives the mutual missingness fraction for (row, column) pairs of variables with the diagonal corresponding to each variables overall missingness. Right: missingness fraction of NHANES variables for older individuals (60+).

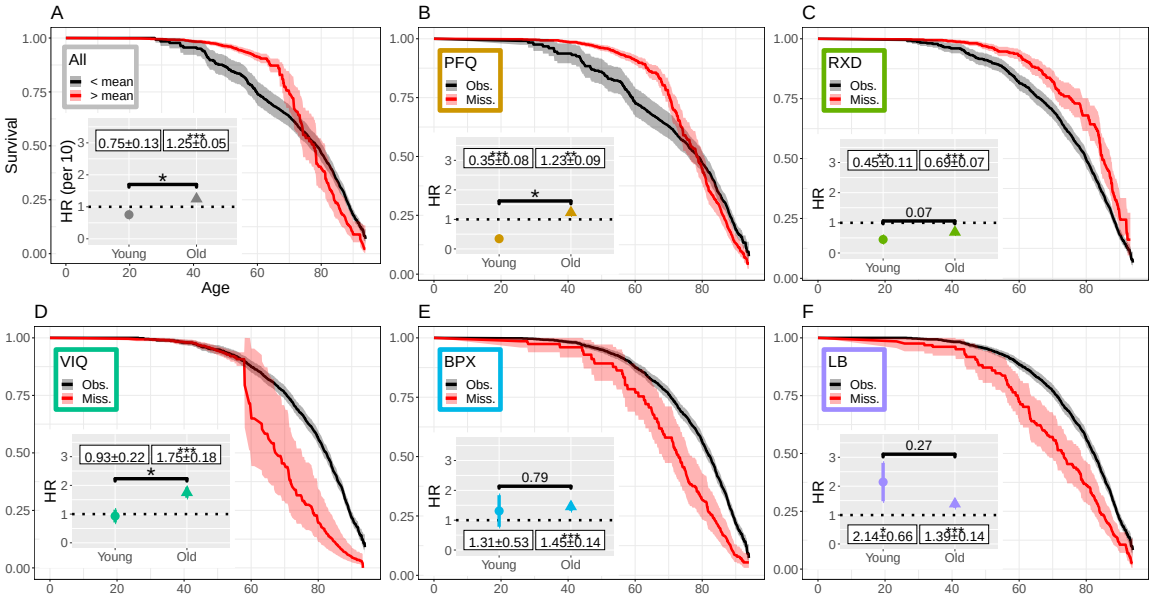

Figure A.5: Survival and missingness after moving cut to age 50 (instead of 60). A) all variables, B) personal fitness (PFQ), C) prescription drugs (RXD), D) vision (VIQ), E) blood pressure (BPX), and F) lab variables (LB). In A) the black line indicates the Kaplain-Meier survival curve for the subpopulation of individuals missing less than the mean (9.8 variables), the red line indicates individuals missing more than the mean. In B)-F), black lines indicate subpopulations without any of the variables in the block missing, red lines have at least one variable in the block missing. Shaded regions indicate 95% confidence intervals. Insets: hazard ratios (HRs) for Cox survival model for individuals stratified by young (< 50) or old (≥ 50), conditioned on age and sex. In A) the Cox model is HR per 10 deficits. In B)-F) each block Cox model was further conditioned on all other blocks (PFQ, RXD, VIQ, BPX and LB). p-values (log-rank test) are given in the caption of Figure A.6, they do not depend on the age cut because they consider all ages.

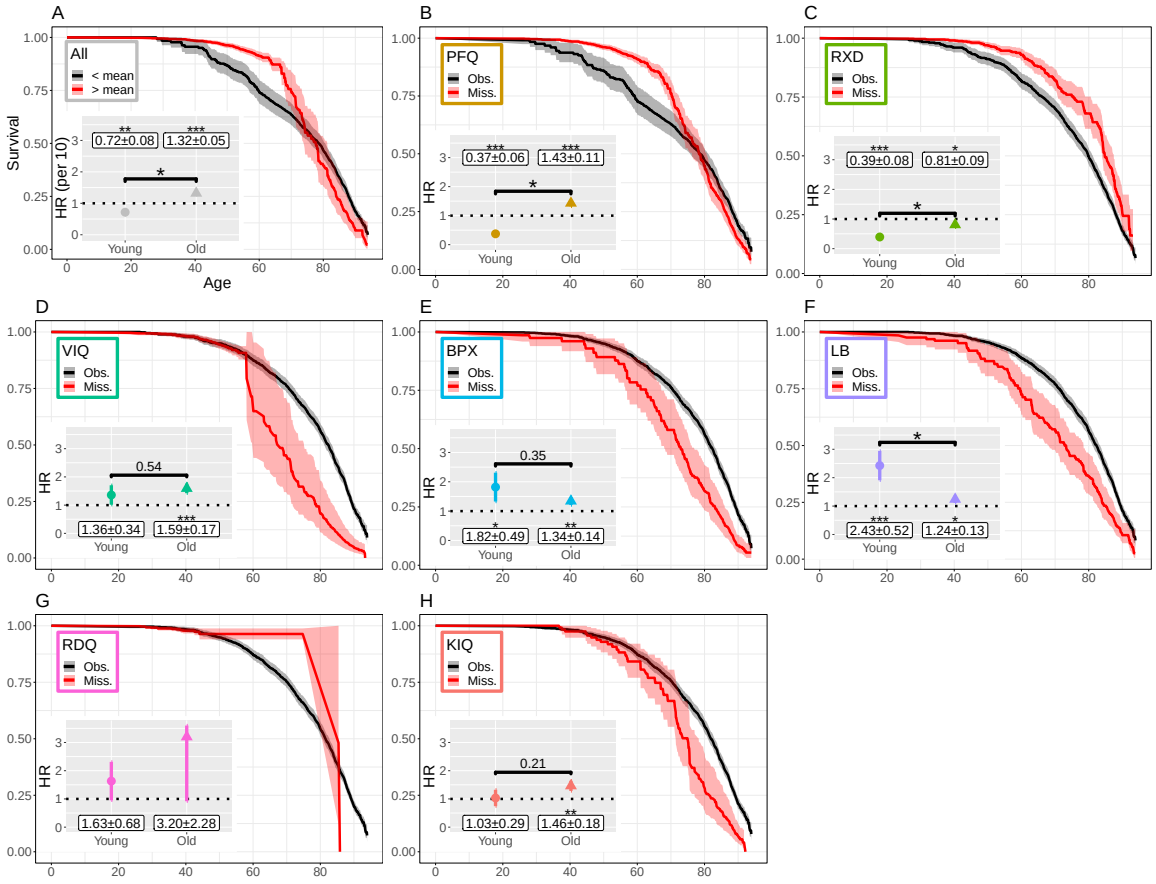

Figure A.6: Survival and missingness (extended). A) all variables, B) personal fitness (PFQ), C) prescription drugs (RXD), D) vision (VIQ), E) blood pressure (BPX), F) lab variables (LB), G) RDQ031: cough regularly (RDQ), and H) KIQ046: lost control of urine (KIQ). In A) the black line indicates the Kaplain-Meier survival curve for the subpopulation of individuals missing less than the mean (9.8 variables), the red line indicates individuals missing more than the mean. In B)-H), black lines indicate subpopulations without any of the variables in the block missing, red lines have at least one variable in the block missing. Shaded regions indicate 95% confidence intervals. Insets: hazard ratios (HRs) for Cox survival model for individuals stratified by young (< 60) or old (≥ 60), conditioned on age and sex. In A) the Cox model is HR per 10 deficits. In B)-H) each block Cox model was further conditioned on all other blocks (PFQ, RXD, VIQ, BPX, LB, RDQ and KIQ). Log-rank test p-values (overall effect): 0.37 (All), 0.016 (PFQ), $6.3 \cdot 10^{-6}$ (RXD), $3.3 \cdot 10^{-8}$ (VIQ), $8.3 \cdot 10^{-7}$ (BPX), $1.9 \cdot 10^{-5}$ (LB), 0.26 (RDQ) and $1.4 \cdot 10^{-6}$ (KIQ).

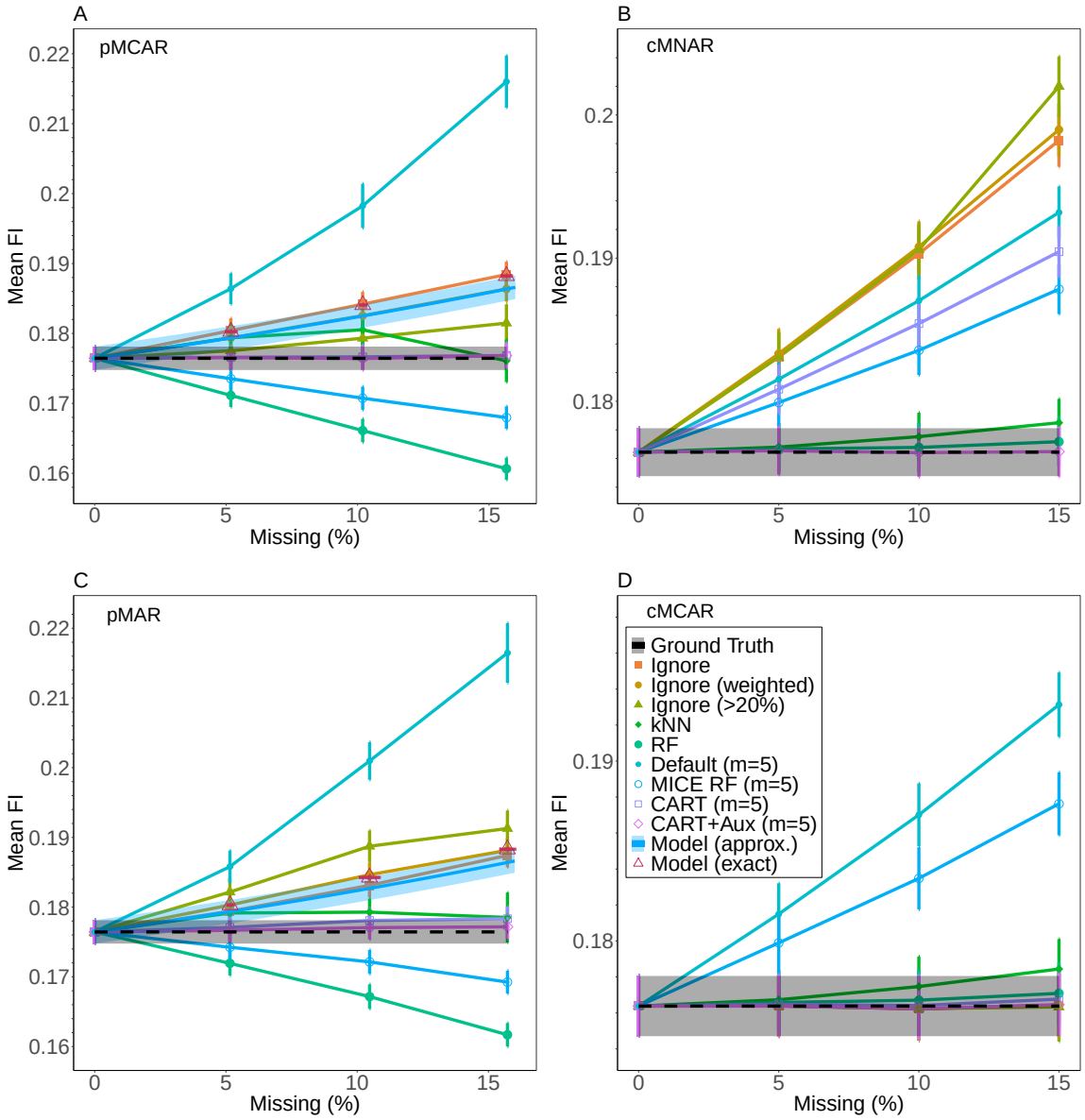

Figure A.7: Missingness biases the FI with most imputation strategies — extended. Using different percentages of simulated missingness and four mechanisms: A) pMCAR, B) cMNAR, C) pMAR, D) cMCAR. We show the mean FI calculated using different imputation strategies, as indicated by the legend. cMCAR had no bias for the ignore-based methods, whereas MICE RF and Default did. Note the similarity of pMCAR and pMAR, where only Ignore (> 20%) (i.e. Ignore20) performed differently. The Default method (teal circles) showed the largest bias compared to the ground truth (black dashed) for pMCAR/pMAR. Observe that the imputation strategies are all approximately linear, justifying the use of a linear bias rate.

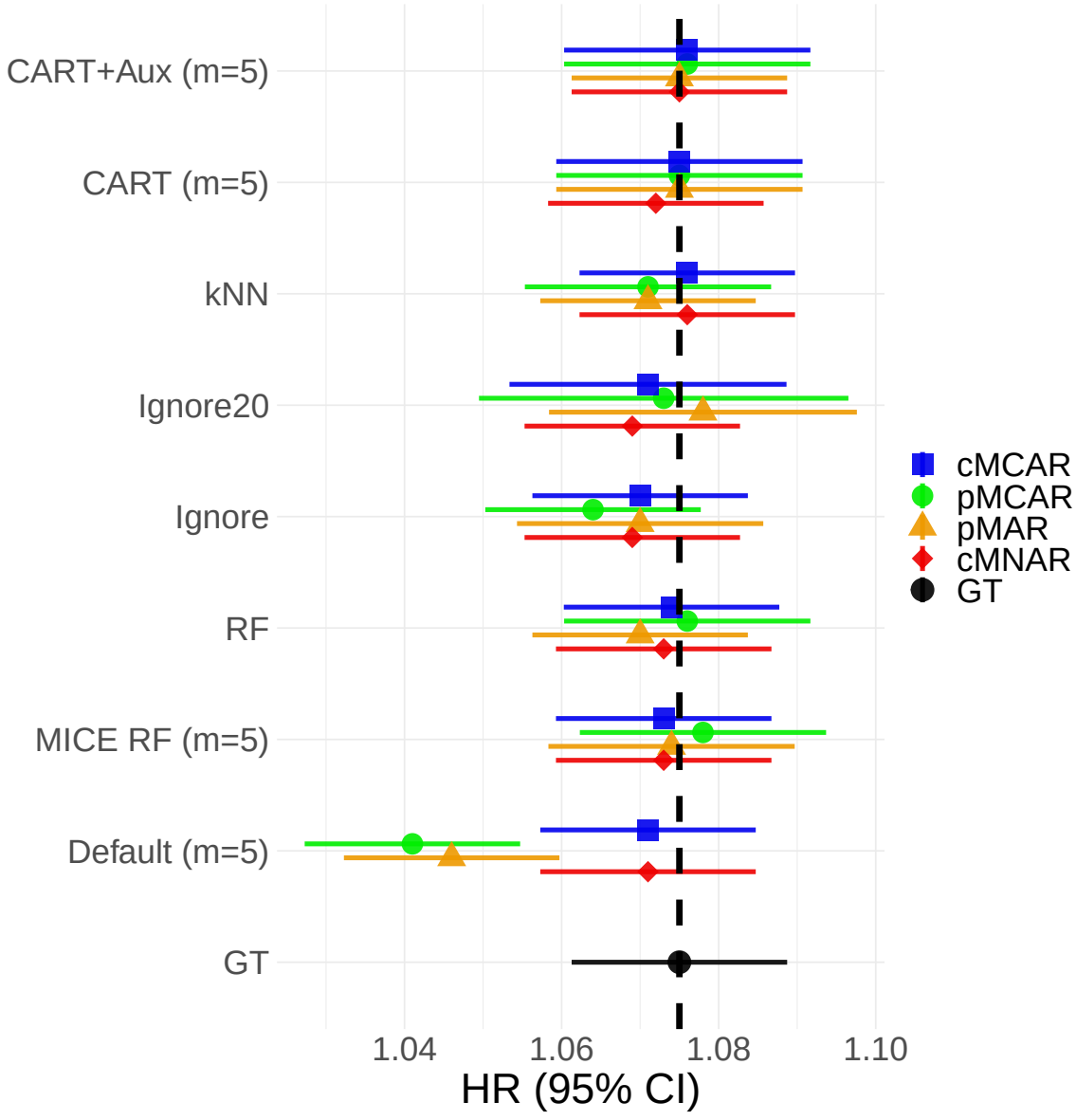

Figure A.8: Forest plot for HRs calculated from data with simulated 15% missingness. GT: ground truth.

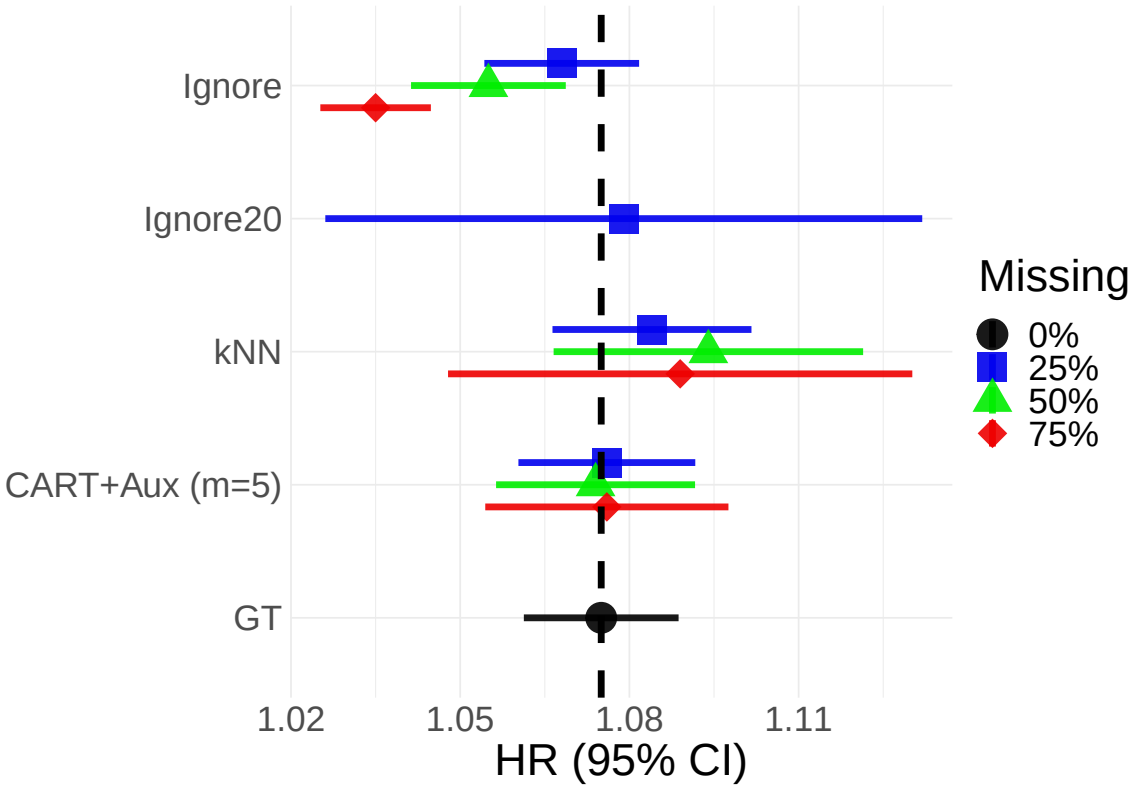

Figure A.9: Forest plot for HRs calculated from data with simulated cMCAR. Note that for 50% and 75% missingness there were not enough individuals to calculate a HR for Ignore20. GT: ground truth.

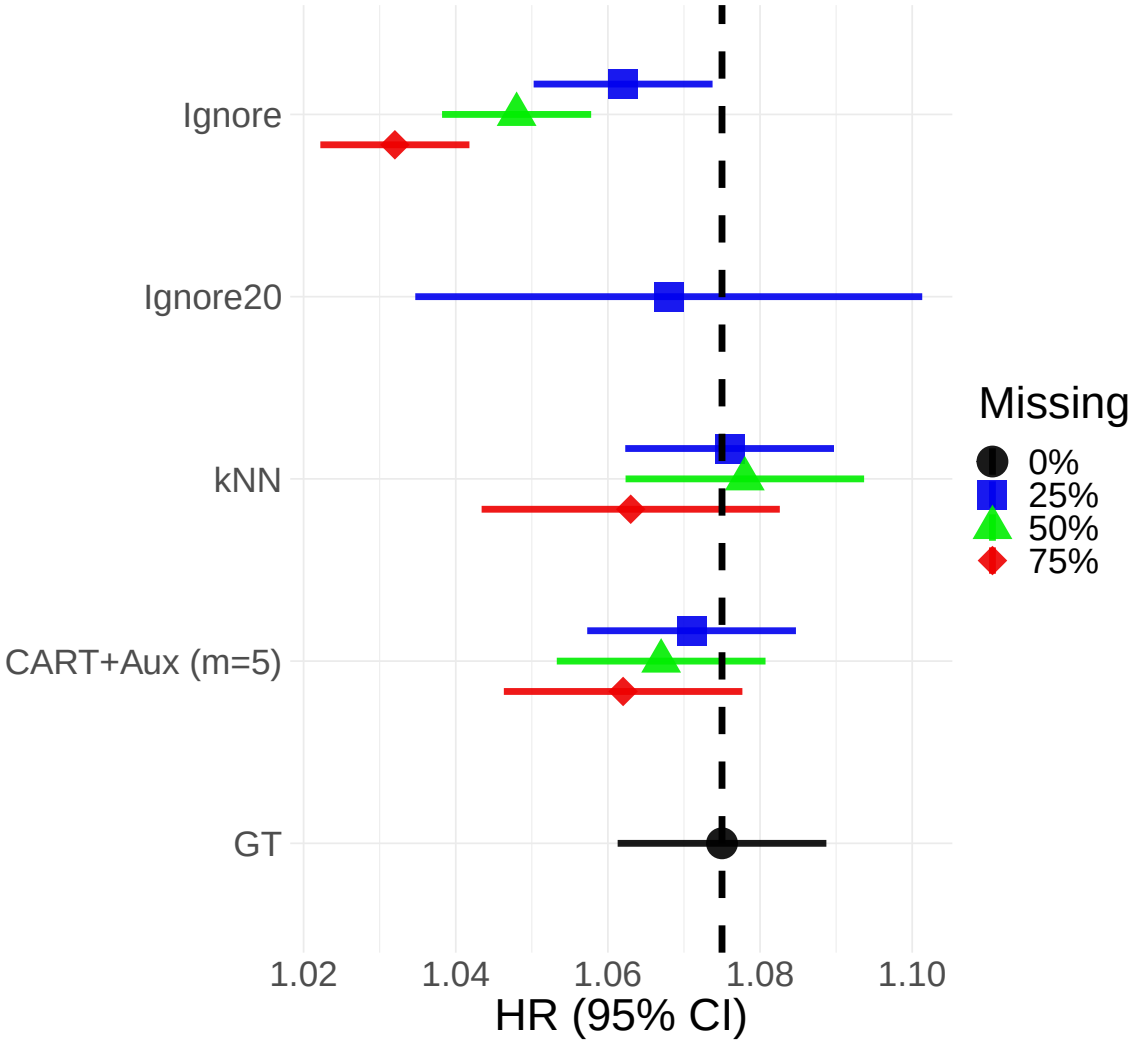

Figure A.10: Forest plot for HRs calculated from data with simulated cMNAR. Note that for 50% and 75% missingness there were not enough individuals to calculate a HR for Ignore20. GT: ground truth.

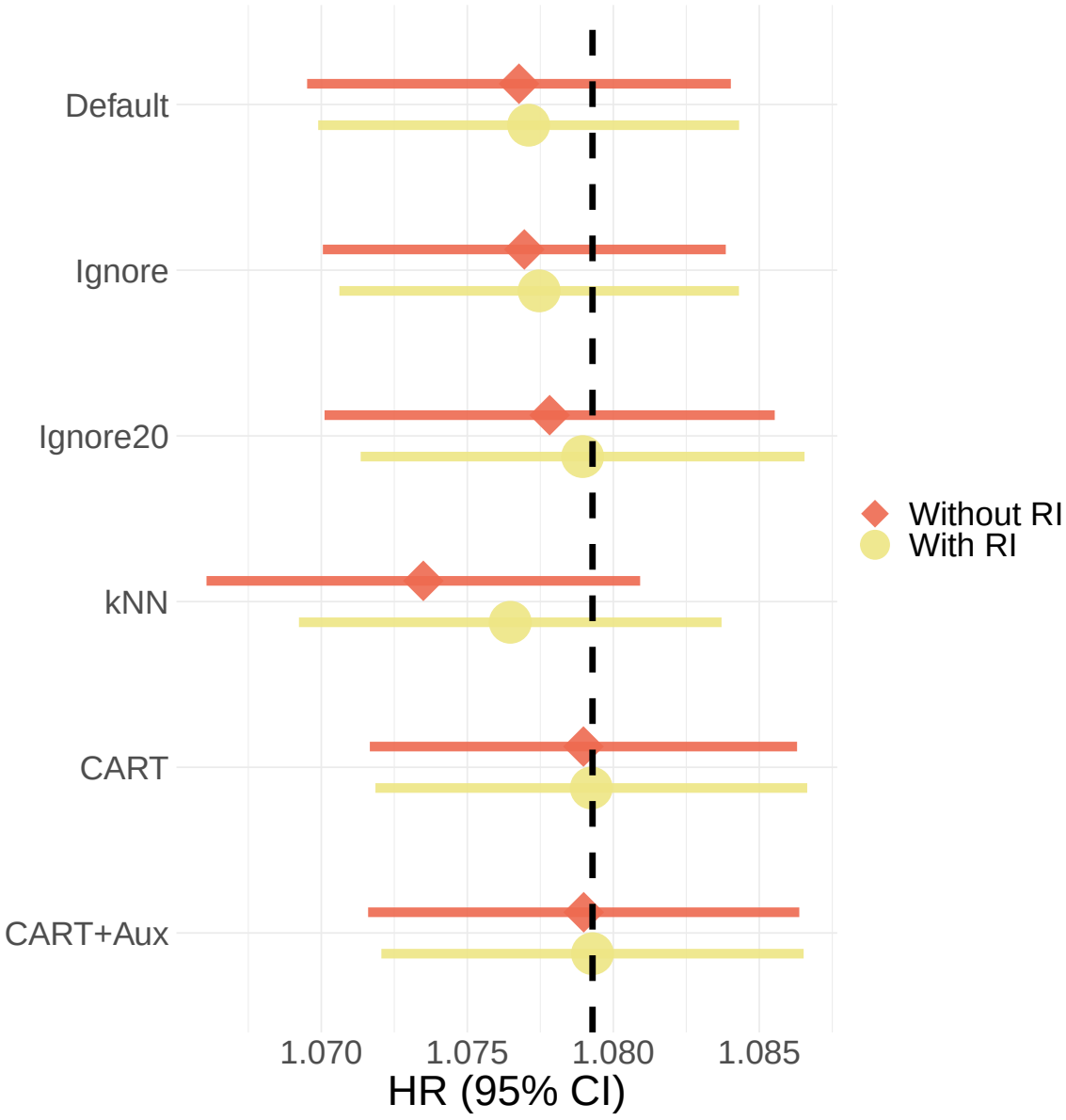

Figure A.11: Forest plot for HRs calculated from Full dataset (real missingness). RI: rule-based imputation.

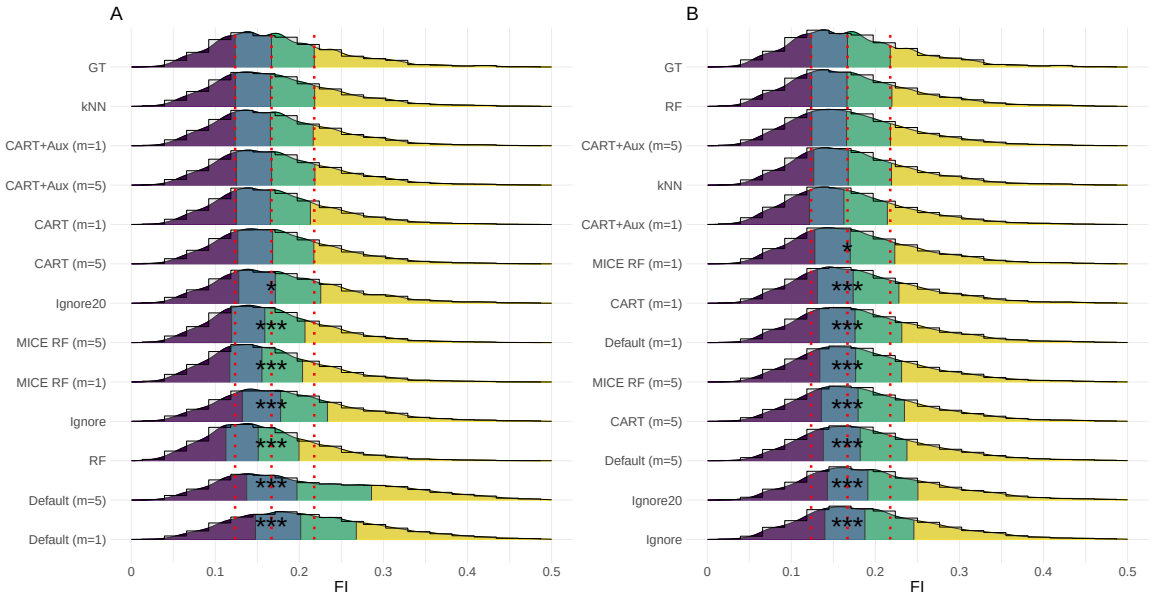

Figure A.12: FI distributions by imputation type for simulated 15% missingness. A: pMCAR, B: cMNAR. Colours: quartiles. Vertical lines: GT quartiles. Stars: KS-test significance (vs GT). Distributions are sorted by KS-similarity to GT from most (top) to least (bottom) similar.

Appendix B

Supplemental Information for Efficient representations of binarized health deficit data: the frailty index and beyond

This supplemental includes additional details for methods including imputation (Section B.1), validation and consistency checks that support our analysis (Section B.2), along with additional results for the imputed data that were not included in the paper for want of space (Section B.3). Finally, we have included the complete case results from a parallel analysis of the data (Section B.4).

B.1 Methods

B.1.1 Variables Used

Our choice of predictors, i.e. health deficit variables, was based on another NHANES study comparing the lab and clinical frailty index (FI) [19]. We have made appropriate modifications for our current study. We excluded alkaline phosphatase (ALP) and lactate dehydrogenase as they were rarely measured (only 84 individuals). We moved c-reactive protein (CRP) into outcomes as a biomarker of inflammation [60]. We moved all questionnaire health condition variables into outcomes. The clinical predictors used are listed in Table B.1. All predictors were binarized using the rules in the summary tables (Tables B.1 and B.2). Binarizing ordinal deficit data appears to have little impact on the FI [139], and we therefore expect it to have little impact on other summary health measures, such as principal components. Lab variables, Table B.2, were binarized according to normal ranges (sex-dependent).

Outcomes included IADL/ADL disability, morbidity, health care utilization, survival, and several specific outcomes of interest. In the latter case, we included mean telomere length, body mass index (BMI), CRP, chronological age, FI, frailty phenotype (FP), gait speed (time to walk 20 feet in seconds) and microalbuminuria (albumin/creatinine ratio [208]; a biomarker of kidney disease). The IADL/ADL variables were split into disability outcomes (high ordinal scales) or dysfunction input (low ordinal scales; see Tables B.1 and B.3 for exact cuts). Individuals with disabled IADL/ADL also had dysfunctional IADL/ADL. Similarly, albumin and creatinine were used both to calculate microalbuminuria (an outcome), then they were binarized and used as input variables, meaning that we expected albumin and creatinine deficits to be good predictors of microalbuminuria. Table B.3 summarizes the outcomes used. The demographics for the population, including statistics for the outcomes used are summarized in Table B.4.

We used the NHANES-adapted frailty phenotype (FP) [208, 47, 183]. FP was defined as 3 or more out of 5: low BMI, bottom 20% for gait speed (sex-adjusted), and self-reported: weakness, exhaustion, and low physical activity [208]. Consistent with other researchers [88], we defined common morbidities by responses to the medical condition questionnaire, e.g. "has a doctor or other health professional ever told (you that you have) arthritis?"

We inspected the distribution of each continuous-valued variable. To reduce the skewness, we log transformed: BMI, gait speed, CRP concentration (as others have done [142, 67]) and mean telomere length (done previously [75]). All continuous outcomes were scaled to zero mean and unit variance.

When pooling to calculate aggregate performance measures (e.g. overall Youden index), we reduced the outcomes to a non-redundant subset to prevent strongly related outcomes from dominating the pooled measures. In particular, non-specific ADL and IADL disability were included rather than the specific disabilities. The complete FI was included, not the FI CLINIC nor FI LAB. Only 10 year survival was included since all survival outcomes had similar patterns (Figure 5.6) and 10 years was the most balanced outcome (58% survived, 42% died). More balanced outcomes required less dramatic weighting for the GLM, and should therefore provide more reliable fits.

We used 7 demographical variables based on other researcher's work [73, 137], including: age (top-coded at 85), sex (female: 1, male: 0), race (white: 1, non: 0), family income (above poverty level: 1, below: 0) education level (greater than high school: 1, not: 0), smoker status (smoker: 1, non: 0) and partner status (married or living with partner: 1, not: 0). The demographic covariates we used are summarized in Table B.5.

All variables were imputed according to the procedure delineated in the next section.

B.1.2 Imputation

This section delineates our imputation procedure and characterises the missing data, and the individuals with missing data.

Imputation Details

Imputation was performed using MICE (multivariate imputation by chained equations) version 3.10.0 [196] via the CART (classification and regression tree) method. Imputation was performed 15 times reflecting the 15% overall missingness in the dataset [205]. We treated all variables encoded as "refused", "don't know", "missing", or "does not do this for other reasons" as NA ("not available" i.e. missing [152]).

Some variables were gated, meaning that the missing values can be inferred based on the response to a previous question ("gated imputation"). For example, SMQ040 (do you now smoke) was gated by SMQ020 (have you smoked at least 100 cigarettes in your life); we assumed SMQ040 was "no" if SMQ020 was "no". We assumed the "currently have" medical conditions (MCQ170K: chronic bronchitis, MCQ170L: liver condition and MCQ170M: thyroid disease) were "no" if the individual reported having never had the associated condition. Individuals whom had not taken a prescription medicine in the past month (RXD030) were assumed to be taking 0 prescription medications and therefore did not have polypharmacy.

Mean arterial pressure (MAP) and pulse pressure (PP) were computed using the systolic and diastolic blood pressure measurements (Table B.2), and were thus passively imputed, meaning that they were computed using imputed systolic and diastolic values [196].

We included all convenient auxiliary variables from the loaded files (mostly blood pressure related), the full list is given in Table B.6. Using auxiliary variables, without recoding, with the present method can improve imputation without risk of overfitting [145].

Missingness Statistics

We characterized the missingness in terms of the magnitude of the effect, the presence of "patterned missingness" [145] and looked for evidence of informative missingness. If data are not missing completely at random then imputation is needed to avoid biased conclusions [180].

Missingness frequency is one minus the frequency with which any particular variable was measured. In Figure B.1 we observed that lab variables tended to have much higher missingness than clinical variables.

The missingness frequency of binary outcomes and demographic covariates are given in Figure B.2. The missingness was very low except for family income, which was missing almost 10% of the time.

The missingness frequency of continuous outcomes and demographic covariates are given in Figure B.3.

Looking closer at the predictors, we see the joint missingness histogram has a large block of mutually-missing lab variables, Figure B.4. It is clear that some predictor variables tended to be reported together, especially lab variables. For lab variables, we can see that large blocks of variables were taken together and hence were either all reported or all missing. Blocks of missing variables, "patterned missingness", can cause problems with some common imputation algorithms, but not the one used here [145].

Individuals missing data had significantly worse survival (Hazard ratio = 1.6(1), $p = 7 \cdot 10^{-13}$, log-rank test), as shown in Figure B.5. This means that the data were not missing complete at random, which means we cannot simply ignore the missingness [180].

Sanity Checks

We performed sanity checks on the imputed values to ensure that they are reasonable and logically consistent. Logical inconsistencies would indicate a poor choice of imputation algorithm. In short, we found the imputed values to be consistent and we conclude that the imputation was sound.

We expect that the frequency of deficits should be higher in the imputed data because of the worse survival of individuals missing data. In Figure B.6 we observed that this was the case, imputed values were almost always more likely to be deficit than the measured values. We have superimposed the missingness frequencies for each variable since it indicates the relative importance of having a good imputation (if the missingness frequency is very low then the imputed values are unlikely to affect our final results). Similarly, we investigated the imputation of outcomes in Figure B.7. We've included demographical covariates as outcomes, although clearly some of them are not "deficit" (e.g. sex).

Next we looked for significant differences in the distribution of data after imputing values. Grossly, the joint 2D frequency histograms showed no change for predictors (Figure B.8) or outcomes (Figure B.9). Comparing the available case predictors to post-imputation predictors we see that the major change in lab variables — due to block missingness — seems to have not affected the histogram structure. It does look, however, that correlations between predictors were stronger in the imputed values which could indicate individuals with many deficits. This is expected since they had higher risk of mortality.

Many (39%) of individuals missing any data were missing all lab variables (see Figure B.4 for visualization). We propose a novel validation for the imputed lab variables. We compare the FI LAB of individuals with their lab variables imputed vs those without imputation. We know that the FI LAB should be higher in the imputed individuals because they had worse survival, and we can quantify the effect. Given the hazard ratio (HR) of missing the lab data, l, and the HR of the FI LAB, f, we expect,

$$\Delta FI = \frac{\ln l}{\ln f} \tag{B.1}$$

where ΔFI is the change in FI between the imputed and complete individuals. We conditioned on the FI CLINIC. The HR of missing the lab data was $l = 1.266 \pm 0.081$ and the HR of the FI LAB using the complete data was $f = 1.0299 \pm 0.0032$, yielding an estimated shift of $\Delta FI = 0.083$. The actual shift in the median FI LAB was 0.034, as shown in Figure B.10. The results are consistent: imputed patients had higher FI LAB and the estimate is within the interquartile range of the distribution.

Imputed Outcomes

We imputed all outcomes and included them in our prediction models. This was a convenience justified by the relative low missingness and the use of multiple imputations, which propagates uncertainty. In this section we verify that our conclusions are unaffected by this choice.

In summary, most outcomes had too few missing values to make a difference in their performance metrics. Of the few with non-trivial missingness, only gait speed prediction was significantly different when imputed (R^2 increased by about 0.05 when imputed values were included). We suspect that imputed individuals were missing not at random, that is, had they been measured they would have had very low gait speeds, which explains why they were easy to predict. Regardless, the shapes of the curves are unchanged and therefore so too are the study conclusion, which are based on relative performance.

Only 6 outcomes had enough missing values to meaningfully affect the final performance metrics, the remaining outcomes had missingness below 3% (Figures B.2 and B.3). We compared the prediction of our generalized linear models (GLMs) with and without including the imputed values, Figure B.11. Clearly the overall patterns are unaffected by the inclusion of imputed values: the same number of principal components are needed regardless of imputation status. The absolute differences were small and in all cases — except for gait — the performance measures overlapped within error. Also, sometimes imputed values were easier to predict and sometimes they were harder to predict, depending on the variable, indicating no global bias to imputed values.

Only gait showed a change in performance when imputed values were included. Gait was better predicted when imputed values were included. This likely reflects the underlying missingness mechanism, as discussed above we observed that the individuals missing data tended to be older and have higher risk of mortality. This suggests that individuals missing gait may have had very slow gait speeds, perhaps being completely unable to perform the task, hence their exclusion. We confirmed this by reading the study documentation, for the 236 (out of 444) individuals with a reason for missingness reported, they all indicated conditions that would likely slow gait speed. Of the 236 individuals with a reason for missing gait speed reported: 10 (4%) recently had chest/abdomen surgery, 7 (3%) recently had a heart attack, 87 (37%) had an aneurysm or stroke, 19 (8%) had severe neck or back pain on the day of the test, 55 (23%) had difficulty moving their knee, and 58 (25%) had a knee or hip replacement. These are all reasons that would *slow* gait speed, and are mostly related to poor health.

It should be no surprise that the imputed gait speeds were significantly slower, with the time to walk 20 feet (in $\log_e$-seconds): 2.22 ± 0.03 for imputed vs 1.94 ± 0.01 ($p = 2 \cdot 10^{-15}$), Figure B.12. These individuals may have simply been easier to predict because other biomarkers captured their relatively poor health. Regardless of whether or not these individuals should or shouldn't be easy to predict our key results are unaffected by their inclusion since we focus on relative performance. The absolute performance of our gait models would have been a little lower, $\Delta R^2 \sim 0.05$, if imputed values were excluded.

B.2 Analysis Validation

We performed a number of internal studies to confirm the validity of our analysis methodology. In particular, we show that the generalized linear models (GLMs) for binary outcomes, i.e. logistic regression, were optimized when we including a simple set of weights. This has a satisfying interpretation within the context of the logistic regression literature.

We have also provided a simulation study that validated our error estimates using cross-validation.

B.2.1 Unbalanced Data

Many of the binary outcomes of interest are uncommon or rare, e.g. only 3.3% died within the first year of followup. This poses a challenge when modelling, since high accuracies can be achieved by models with no real predictive power: they simply return the most common class, e.g. 'everybody survives the first year' is 96.7% accurate. Among other problems, this invalidates any measure of feature importance since the model isn't actually doing any discriminating. We found that this problem was ameliorated by weighting the individuals in the minority group more heavily.

We tested a simple weighting scheme where individuals in the majority group were given weight 1 whereas individuals in the minority group were given weight w, where w was selected through hyperparameter optimization. Each outcome was optimized for w using

$$w_i = \begin{cases} w, & \text{if } i \text{ is in minority.} \\ 1, & \text{if } i \text{ is in the majority.} \end{cases} \tag{B.2}$$

We used Bayesian optimization [211] as it is well-suited for optimizing models that are slow to train. Bayesian optimization uses statistical inference to predict the best set of parameters to try at each iteration. We used Youden's index [215] as the objective function (Youden index $\equiv$ sensitivity + specificity $-$ 1). We initialized with an 11-point, $\log_{10}$, grid search from $\log_{10} w \in [0, 2]$, then 10 iterations of Bayesian optimization. Finally, we selected the weight, w, with the highest Youden index. For predictors we used the 7 covariates: age, sex, ethnicity, family income, education level, marital status, smoker status and the FI. We used the GLM logistic regression model to predict 36 of the binary outcomes (everything except survival) and the 6 binary demographical variables for a total of 42 data points.

We investigated the 42 optimized prediction models to see if the degree of balance, b, was a strong predictor of the optimal w where,

$$b \equiv \frac{\text{Frequency of minority class}}{\text{Frequency of majority class}} \tag{B.3}$$

We observed power law behaviour in the optimal weights as a function of balance, b, Figure B.13. The best fit line yielded the relationship $w \sim b^{-1.12}$, with the simpler model $w \sim b^{-1}$ also fitting well. Note: no correlation between Youden index and either balance or weight was observed (not shown), demonstrating the power law relationship was real and not an artifact of the optimization algorithm.

We observed that the intuitive model fits well,

$$w = \frac{\text{Frequency of majority class}}{\text{Frequency of minority class}} = \frac{1}{b}. \tag{B.4}$$

Hence the optimal weights for each individual were:

$$w_i = \begin{cases} \frac{\text{Frequency of majority class}}{\text{Frequency of minority class}}, & \text{if } i \text{ is in minority.} \\ 1, & \text{if } i \text{ is in the majority.} \end{cases} \tag{B.5}$$

We implemented Eq. B.5 when fitting the GLMs in the present study.

Assuming a balanced prior, we can show that Eq. B.5 is equivalent to the weighted exogenous sampling method [96]. The exogenous method weights are defined by [96]

$$w_1 = \frac{p_1}{s_1} \tag{B.6a}$$

$$w_0 = \frac{p_0}{s_0} = \frac{1 - p_1}{1 - s_1}, \tag{B.6b}$$

where p_1/p_0 is the population frequency of class 1/0 and s_1/s_0 is the sample frequency of class 1/0. Choosing which class is 1 is arbitrary, so we pick class 1 to be the majority class, without loss of generality. The weighted likelihood [96] is insensitive to an overall scale—i.e. the minimum doesn't change—so we are free to divide both weights by w_1, yielding

$$\frac{w_0}{w_1} = \frac{s_1/p_1 - s_1}{1 - s_1} = \frac{s_1}{s_0}\left(\frac{1}{p_1} - 1\right) = \tilde{w}_0 \tag{B.7a}$$

$$\tilde{w}_1 = 1, \tag{B.7b}$$

where $\tilde{w}_0$ and $\tilde{w}_1$ are the rescaled weights. If we further set $p_1 = 1/2$, that is the population frequency of the event is exactly 50%, then we have

$$\tilde{w}_0 = \frac{s_1}{s_0} \equiv \frac{\text{Frequency of majority class}}{\text{Frequency of minority class}} \tag{B.8a}$$

$$\tilde{w}_1 = 1, \tag{B.8b}$$

which is identical to our heuristic scheme, Eq. B.5. Hence our empirical weights are optimal if we assume that the 'true' underlying population has as many cases as controls ($p_1 = 1/2$, i.e. a balanced study design).

The Cox proportional hazards model was not weighted since the survival data were already nearly balanced (43% survived to censorship, 57% died). We compared C-indexes for the Cox model with and without weights, 0.6638 ± 0.0010 vs 0.6644 ± 0.0012, confirming weights were unnecessary.

B.2.2 Out-of-sample Errors

Prior research has shown that under extreme conditions naïve cross-validation errors may be too narrow [11]. Nevertheless, we use naïve cross-validation because it is simple and easy. Here we use a simulation study to validate that the out-of-sample error estimates were reasonable proxies for the true errors. Note that these are standard errors, not confidence intervals.

Using a simulation study allows us to have a ground truth for the exact error intervals which we compared to those estimated via naïve cross-validation. To produce realistic data, we trained a model to generate predictors and outcomes that were similar to the complete case data. This generative model was subsequently used for the simulation study.

We trained the generative model as follows:

1. Combine predictors and outcomes, keeping only complete case data (no missingness, even in outcomes).

2. Compute complete PCA of the concatenated dataset of predictors and outcomes.

3. Generate random sample in the latent space i.e. generate a Gaussian random variable with covariance equal to the covariance of the PCA scores (diagonal covariance).

4. Use the PCA rotation to map the latent space into an observed space.

5. Binarize any binary-type data using a hard-cut at 0.5.

6. Split predictors and outcomes.

7. Drop outcomes with no case data.

8. (This is the ground truth data.)

9. Compute the cross-validation error estimate for the ground truth data using 10-fold cross-validation.

10. Iteratively generate $N = 100$ samples from this model and compute accuracy metrics.

11. Compare the simulated accuracy metrics to the cross-validation metric estimates. We know both the ground truth accuracy and error bars, so we can test the validity of the cross-validation estimate.

Figure B.14 compares the measures from the simulated data to the cross-validation estimate of the synthetic ground truth. We see good agreement in error bar magnitudes and reasonable agreement in point estimates of each accuracy metric for each variable.

B.3 Additional Results

This section contains additional results to the main text. Data have been pre-processed and imputed according to the rules outlined therein.

B.3.1 Youden Index vs AUC

The Youden index [215] is a relatively uncommon metric, so we compared it to a more common metric: the AUC (area-under-the-ROC-curve), in this section. We hope that this will help to develop our reader's intuition about the scale of the Youden index. Recall from the Section 5.2.2 of the main text that a Youden index (AUC) of 0 (0.5) indicates a useless test, i.e. no better than a guess, and 1 (1) indicates a perfect test, i.e. always correct. The Youden index is preferable since it provides a definite accuracy at a specific threshold, such as we would see in medical diagnosis [214]. In contrast, the AUC considers all possible thresholds.

We pooled the GLM predictions for all binary outcomes and number of PCs and plotted the Youden index vs AUC in Figure B.15. The two were strongly correlated and were well fit by the relationship [214]:

$$\text{AUC} = \Phi\left[\sqrt{2}\Phi^{-1}\left(\frac{\text{Youden} + 1}{2}\right)\right] \tag{B.9}$$

where Φ is the normal cumulative distribution function. This model is derived by assuming the two classes have normal statistics and the same variance, but different means. In general, the AUC is sensitive only to location parameter, e.g. mean, whereas the Youden index is sensitive to location and shape parameters, e.g. mean and variance [214]. Using Eq. B.9, we see that Youden indexes of 0.2, 0.4, 0.6 and 0.8 indicate similar performance to AUCs of 0.64, 0.77, 0.88 and 0.97, respectively.

By inspection of Figure B.15B, it is clear that only three outcomes deviated from Eq. B.9: ever had a liver condition (liver_con; green triangles), still have a liver condition (have_liver_con; red circles), and significant difficulty using a knife/fork (adl_knifeDIS; blue squares). These are the three rarest outcomes: 1.2%: have_liver_con, 1.9%: adl_knifeDIS, and 3.0% liver_con. These were similarly rare to the rarest input deficits, the rarest being phosphorous at a rate of 2.2%. The outcomes were predicted using weighted GLMs whereas the compression measures used unweighted PCA. This implicates the weighting scheme as the primary culprit in causing the deviation of these three (rarest) outcomes from Eq. B.9 (Section B.2.1).

We see that for the present study, the AUC saturates at 1 much faster than the Youden index, suggesting the Youden index is a more conservative metric.

B.3.2 Input Compression

We illustrated the action of PCA on the 2D joint deficit histogram in Figure 5.2. The raw, unscaled version of Figure 5.2 is shown in Figure B.16.

Our main compression result was comparing data reconstruction accuracy as a function of latent dimension size. In Figure 5.3 we plotted the Youden index as the measure of accuracy. In Figure B.17 we reproduce the plot using the AUC instead of the Youden index. The AUC shows the same trend and relative performance for each algorithm, however, the algorithms reach approximately perfect reconstruction faster with the AUC than the Youden index. For example, LSVD reaches perfect AUC near 20 features versus 30 for the Youden index. This reflects the Youden index being a more sensitive measure (Figure B.15).

We observed strong correlations between the first feature of each algorithm in Figure 5.4. The extended Spearman correlations including centered PCA, LPCA and LSVD are given in Figure B.18. Centering means that the mean has been subtracted, which should reduce the correlation with the FI. With or without centering, the correlations between the FI and PC1/LPC1/LSV1 are very strong.

We also compared the sex and age dependence of the 1D features in Figure B.19. The overall age and sex dependencies were markedly similar: further evidence that they describe the same underlying phenomenon.

B.3.3 Generalized Linear Models (GLMs)

We explored which PCs were needed to optimally predict the pooled outcomes by stepwise adding another PC: starting with covariate demographical information, then PC1 and so forth. Figure B.20 gives the pooled predictive power for the GLM using the AUC instead of Youden index for comparison to Figure 5.7. The scale is different (AUC vs Youden) but the results are nearly identical.

We then explored specific outcomes, plotted up to PC6 in Figure 5.9; the complete results are plotted in Figure B.21. Limiting up to PC6 did not exclude any meaningful improvements in performance for most outcomes – though the 'high-dimensional outcomes' in Figure 5.10 continue to improve to a high number of PCs. Figure 5.10 is also truncated, which misses the improvement in prediction for some ADL/IADL disabilities, e.g. "difficulty managing money" (iadl_moneyDIS) from PC21-PC25 and "difficulty preparing meals" (iadl_mealDIS) from PC47-PC49 (Figure B.21). These improvements could have been caused by our choice to include ADL/IADL dysfunction/difficulties as input variables (see Section B.1.1). Higher PCs tend to become increasingly specific towards the input variables (e.g. Figures 5.5 and 5.11), so each ADL/IADL dysfuction may have a specific PC associated with it, as those associated PCs are included the ADL/IADL disability prediction could improve without risk of overfitting (because everybody with disability had dysfunction).

We next consider variations of Figures 5.9 and 5.10 to explore robustness of those results.

We first consider the possibility of quadratic behaviour and interactions within the GLM in Figure B.22. We included a linear and quadratic term for each PC as well as all possible interactions between the PCs. We observe that the maximum predictive performance was the same or lower than the linear model, Figures 5.9 and 5.10, in all cases. We also observe severe overfitting, presumably due to the greatly increased number of parameters due to interactions.

We next use the AUC instead of Youden index in Figure B.23 for comparison to Figures 5.9 and 5.10. The results look nearly identical.

Finally, instead of using PCA, we used LPCA in Figure B.24 and LSVD in Figure B.25, where we have grouped by outcome type. These are for comparison with PCA (Figures 5.9 and 5.10). We can summarize all three feature sets — PCA, LPCA and LSVD — similarly: most outcomes required 1 or 2 features ("low-dimensional") with a few requiring many more ("high-dimensional"), with agreement on which outcomes are low vs high–dimensional. It is, however, clear by inspection that the LSVD accuracy metrics tend to be lower than either PCA or LPCA (the latter two are roughly the same). This is consistent with our observations in Figures 5.7 and 5.8 that LSVD features were typically worse predictors.

Now we move on to feature importance. We estimated feature importance by using a feature selection algorithm, LASSO, then computed the selection frequency for each feature. LASSO picked the optimal performing model, minimum mean-squared error (continuous outcomes) or deviance (binary outcomes) [58]. All 55 PCs and 7 demographical covariates were available for the selection algorithm. Features selected more often are, ostensibly, more important. We cross-validated each outcome and pooled the selection frequencies together. In Figure B.26, left, we present the aggregated feature selection probabilities for the GLMs. Selection probability decreased with increasing PC number. We observe an elbow past 7 PCs in the selection probability, implying that the first 7 PCs were consistently better predictors whereas the higher PCs were less useful and could therefore be vulnerable to overfitting. Similarly, income, education, partner status and smoker status seem to be relatively unimportant variables for prediction; conversely: sex, race, and age seem important. PC feature importance appeared to satisfy a power law, as displayed in Figure B.26, right (linearity on a log-log scale confirms a power law). The best fit was generated using weighted linear regression, where each weight was $1/\sigma^2$ (σ= standard error).

The observation that feature importance drops with increasing PC number justifies truncating at a particular PC. Together with our observations from Figure 5.12, that implied features below the bilinear trend cause overfitting, this suggests a good strategy for feature selection is to use a hard cut as PCs start to drop below the bilinear trend.

B.3.4 Robustness Analysis

Here we further examine how robust our results were to changes in the dataset.

The exact values for the PCA rotation coefficients (the matrix U in Appendix 5.6) from our full dataset are reported in Table B.7 up to PC7. They can be used directly to generate features for new datasets: each PC is simply a linear transformation using the coefficients in the respective column.

We tested the robustness of the PCA rotation to choice of variables. We randomly sampled a subset of 30 variables then computed the PC rotation coefficients and repeated the process 2000 times. Figure B.27 illustrates the result. The first 3 PCs were robust to choice of variables, beyond which were not statistically significant. These PCs all have strong 'domain' signals: clinic, lab, blood pressure and metabolism, each domain being composed of several associated variables. We infer that the first 3 PCs were robust to choice of variables and that this robustness comes from having distinct domains of like-variables.

We expect that the rotation is robust to choice of similar variables, for example, if we replaced the clinical variables with a different set of clinical variables we expect that the rotation wouldn't change very much. This relates to the coarse graining action of PCA, which relies on domains. Substituting a variable within a particular domain for another variable in the same domain is unlikely to affect the resulting PCs.

B.3.5 Age Stratification

Recall from the main text that we stratified by age to further test the stability of our results (in the Section 5.3.5). In this section we include additional results related to age stratification and to other demographic variables.

The feature associations between primary features and demographical variables are given in Figure B.28, analogous to Figure 5.6. We observed strong associations between age and the first latent features (FI, PC1, LPC1 and LSV1).

The deficit frequency increased with age, Figure B.19, but did the frequency of joint deficits change? We stratified by age quartile and then normalized the joint histogram by the mean deficit frequency, effectively conditioning on the probability of a deficit occurring in that quartile. The resulting histogram appears to become slightly more saturated with increasing age, Figure B.29. The increase in saturation implies that PC1, which tends to be a large block, is likely to be more important with increasing age (confirmed in the main text, Section 5.3.5). The histogram is relatively stable with age, however, suggesting that the PCs past PC1 change only slightly with age. Looking at Figure B.28 we can confirm that PC1 is strongly correlated with age whereas the remaining PCs are not.

We can show that the eigenvectors do not change when the histogram is simply scaled.

Proof: Consider two joint histograms of different ages, $H^{(y)}$ and $H^{(o)}$, say young and old,

$$H_{ij}^{(y)} \equiv \frac{1}{N_{young}} \sum_{young} x_i x_j \tag{B.10a}$$

$$H_{ij}^{(o)} \equiv \frac{1}{N_{old}} \sum_{old} x_i x_j. \tag{B.10b}$$

Suppose that the two histograms are identical except for an overall scale, i.e. $H^{(o)} = \alpha H^{(y)}$. In our case α would be the ratio of mean FIs for the age groups. The eigenvalues of $H^{(y)}$ are, by definition,

$$H^{(y)} = \sum_i \lambda_i \left(\vec{\phi}_i \otimes \vec{\phi}_i\right) \tag{B.11}$$

where λ_i and $\vec{\phi}_i$ are the ith eigenvalue and eigenvector pair, respectively. By assumption $H^{(o)}$ is just scaled so,

$$H^{(o)} = aH^{(y)} = \sum_i (a\lambda_i)\left(\vec{\phi}_i \otimes \vec{\phi}_i\right) \tag{B.12}$$

thus the eigenvectors are the same but the overall eigenvalues are different. If we normalize the eigenvalues by their sum then they are also the same, $(a\lambda_j)/\sum_i(a\lambda_i) = \lambda_j/\sum_i \lambda_i$. **QED**

Hence, our observation in Figure B.29 that the main age-related change to the histogram is a change in scale should neither affect the eigenvectors nor the PC rotation. The overall scale does, however, affect the spacing between eigenvalues, which could be important for PC ranks. For example, if spacing between eigenvalues is small then PCs are more likely to swap order, making it harder to robustly estimate them across samples.

We compared the PC second moment (eigenvalue) spectrum ("eigenspectrum") as a function of age, Figure B.30. The first eigenvalue clearly increases with age, consistent with PC1 $\sim$ FI (the FI is known to increase exponentially with age [124]). There may also be a bilinear structure in the spectrum on a log-log scale, although the first line only has three data points. This PCA structure has been observed previously for fractal time series data [61], where the slope is proportional to the fractal dimension – which is a measure of complexity [111].

We split at the median age, and saw little difference in compression between the young and old cohorts (Figure B.31). We then compared prediction of outcomes. We excluded demographical variables to better facilitate comparison, because the baseline model that used only demographical variables did not perform equally well for each cohort. In Figure B.32 we observed that older cohorts tended to have better predictions than younger cohorts. Performance was nevertheless better with the full population, possibly because it contained more individuals for training.

We then used GLM modelling to summarize the differences between the young and old cohorts. In Figure B.33 we performed a compression test: using the GLM — with PCs as features — to predict the baseline predictor variables. In Figure B.34 we performed a prediction test using the GLM to predict outcomes as normal. Of note, there was a much stronger focus on predicting/compressing creatinine, BUN, microalbuminuria, and to a lesser extend gait, in the older cohort than in the younger cohort. The younger cohort tended to prioritize other input variables e.g. calcium, HDL and iron. The difference in higher PCs between young and old cohorts reflects the lack or robustness of the higher PCs that we observed in Figure 5.11.

B.3.6 Benchmarks

We benchmarked each of the dimensionality reduction techniques using a personal computer (i7-10750H CPU @ 2.60 GHz (typical clock: 4.4 GHz), 16 GB of RAM). Computations were performed using RStudio with R version 4.0.1 [152]. Sample data were generated from the complete case predictors. We shuffled the individuals once then selected incrementally larger sets of individuals and predictor.

Benchmarks were computed using the median of 100 repeats via the `microbenchmark` package [120]. The resulting benchmarks are reported in Figure B.35. Grossly, PCA was 10x faster than LPCA which was 10x faster than LSVD. PCA, LPCA and LSVD all scaled similarly. The scaling with respect to number of individuals was sublinear. For the 100-individual sample, the scaling was exponentially with increasing number of predictors; the 1000-individual sample was slower than exponential. FI scaled much slower, showing essentially no change with increasing dimension. It is surprising that PCA was faster than the FI, but this could reflect the FI having a large fixed computational cost due to its implementation, consistent with the trivial scaling with increasing dimension.

B.4 Complete Case Results

Before imputation, we performed an initial analysis on the complete case predictor data. That is, individuals were required to have all 55 predictors and 7 covariates reported, although missing values were allowed in the outcomes (which were then dropped one at a time when assessing accuracy). We used complete case (no missing values allow) instead of available case (missing values ignored) because PCA, LPCA and LSVD are impractical (PCA) or impossible (LPCA and LSVD) when an individual has missing values. For example, if a PC has a non-zero contribution from each input variable then that PC will be NA for all individuals except for those with complete data. Demanding complete case reduced our dataset from $N = 1872$ individuals to $N_{cc} = 1123$ individuals. As described in the main text and Section B.1.2, there were significant differences between the two datasets with the complete case data being younger: median (IQR) 71 (65-78) vs 76 (67-83) and had significantly better survival, implying they were also healthier (hazard ratio for individuals missing data: 1.6(1)). Some differences between the datasets are thus expected. The purpose of this section is to point out any salient differences between the complete case results and the imputed results reported in the main paper.

Mostly we observed the same salient results in both the complete case and imputed datasets. The major differences appear to be due to poor data quantity for the complete case data which had far fewer adverse outcome cases, particularly in ADL/IADL disability, e.g. "difficulty using knife" (adl_knifeDIS) had only 4 cases. Ostensibly individuals with disabilities were more likely to have missing data, perhaps due to physical limitations affecting data collection. This is consistent with our other observations that the complete case individuals were in better overall health. Minor differences between the imputed results and complete case could be attributed to the relatively poor health of the individuals with missing data. When using only complete case data our study population is biased towards a subpopulation with healthier, younger individuals.

We start with the eigen-decomposition of the joint 2D histogram, Figure B.36. Visually, Figure B.36 is similar to Figure 5.2, reaffirming the two major results: the first PC has nearly uniform weights, akin to the FI, and the subsequent PCs tend to block out like domains. The primary difference is the increased saturation of Figure 5.2 which shows that the complete case had lower deficit rates, consistent with those individuals being healthier.

B.4.1 Complete Case: Input Compression

Investigating the compression ability of each algorithm, we observed nearly identical results to the imputed data (Figure B.37). We also see the same order for relative performance of the algorithms (FI $\lesssim$ PCA $<$ LPCA $<$ LSVD), the same difference between clinical and lab data, and the same scale of overall performance. As with the imputed data, LSVD was the most efficient and saturated at approximately 30 dimensions, implying that the dataset can be fully represented by a smaller set of ≤ 30 features.

We again computed the correlation matrix between the first features from each algorithm, Figure B.38. The correlations were weaker but showed the same trends: the first feature from PCA, LPCA and LSVD all correlated strongly with each other and the FI. They also correlated more strongly with the FI CLINIC than the FI LAB.

B.4.2 Complete Case: Feature Associations

Compression generated new features in a latent space of reduced dimension that we then interpreted. We associated the features with both input predictors and outcomes to infer how information was aggregated by the compression algorithms.

In Figure B.39 we see a nearly identical pattern to the imputed data, Figure 5.5. Notably, the associations seem to be stronger in the complete case data (with smaller CIs), e.g. the association between PC8 and iron. This could be because the complete case data is a more homogeneous population with lower deficit frequencies. It could also be that imputation over-estimated the endemic stochasticity.

Our major observations are unchanged. PC1, LPC1, LSV1 and the FI all have nearly identical prediction patterns. The higher PCs tend to be of weaker significance and more specific to a particular variable or domain of variables.

We can infer the information content of each feature through the association scores — a higher score implies more information related to a particular input variable. Similarly, in Figure B.40 we score the predictive power of each feature for each outcome. In both figures the inner colour indicates the lower limit of the 95% confidence interval (CI): lighter values are less significant (white is non-significant). Consistent with the compression observations, we see nearly identical patterns between the FI and the first latent dimension: PC1, LPC1 and LSV1; note also the similarity to the FI CLINIC. As with the imputed data, we observe higher PCA dimensions tend to be weaker, but also more specific predictors.

B.4.3 Complete Case: Generalized Linear Models (GLMs)

By using a GLM we were able to infer conditional associations with outcomes given other features and demographical covariates. In Figure B.41 we plot the pooled outcome accuracies — Youden indexes — for predicting binary outcomes using each algorithm as a dimensionality reduction step. In contrast to the imputed data, Figure 5.7, we see that the complete case data, Figure B.41, is much more sensitive to overfitting. We see a rapid drop in predictive power for all algorithms past about 10 PCs. This may be because binary outcomes were much less common in the complete case data. The three least common outcomes for the complete case dataset were "difficulty using knife" (adl_knifeDIS): 4 cases, "difficulty walking between rooms" (exhaustionDIS): 7 cases, and "have a liver condition" (have_liver_con): 11 cases. We suspect the associated Youden indexes are unreliable. Modelling outcomes with such low case rates required very large weights which would make the fits very sensitive to a small number of data points (causing a leverage problem [86]). What's more, cross-validation estimates may be unreliable, e.g. for adl_knifeDIS there would be, on average, less than one case in each test set. In contrast, the full dataset had 36, 123 and 21 cases, respectively (before imputing the outcomes). The least common outcome in the full dataset had 21 cases vs 4 for the complete case data. Looking at Figure B.43, we confirm that adl_knifeDIS and exhaustionDIS have volatile accuracies and are very sensitive to the number of features and, presumably, overfitting (other cases were also rare, e.g. FP: 17 cases and survival_1year: 11 cases). This supports our hypothesis that overfitting of the complete case data was caused by case rarity.

Comparatively, the continuous outcomes showed much lower sensitivity to overfitting, Figure B.42. This is further evidence that the rare case data is the culprit in the overfitting sensitivity observed in Figure B.41. The continuous outcomes showed a slight tendency to overfit, having worse performance as the number of PCs becomes large. In contrast, we saw no evidence of overfitting in the imputed dataset, Figure 5.8.

We break down the predictions by specific outcomes in Figure B.43. The key observations from the imputed data, Figure 5.9, hold: most outcomes are near-optimally predicted using just PC1, with a few specific outcomes taking far more. The complete plots without truncation are given in Figure B.44. In contrast to the imputed data, we observe that the fits are more volatile and the ADL/IADL performance is notably lower, ostensibly due to the poor performance when predicting "difficulty using knife" (adl_knifeDIS) and "difficulty walking between rooms" (exhaustionDIS). As discussed above, this is likely a manifestation of the rarity of those outcomes, having just 4 and 7 cases, respectively.

Focusing on the outcomes that appeared to respond to more than 2 PCs, we replotted these outcomes in Figure B.45. These are the "high-dimensional" outcomes that we visually observed to improve with increasing number of PCs (using Figure 5.9), as well as the frailty phenotype (FP), which is theoretically high dimensional, having been proposed to emerge from a complex interplay of several biological systems [57]. In contrast to the imputed data, Figure 5.10, the discrete outcomes are quite noisy — probably due to lack of case data. Interestingly, PCA using only the clinical variables (PCACLINIC) yielded a prediction of age that appears to be much lower dimensional. This may relate the lower deficit frequencies in the complete case population (compare Figure B.36 to Figure 5.2). We preserve strong similarities with the imputed data, however, with similar curves for the continuous outcomes and age. Both imputed and complete case analyses indicate that there is something unique about these outcomes in requiring far more dimensions (PCs) to optimally predict than most other outcomes.

B.4.4 Complete Case: Robustness Analysis

We tested the robustness of PCA by bootstrapping the sample population. The PCA rotation for the first 10 PCs is given in Figure B.46 for the complete case data. The results are very similar to the imputed data, Figure 5.11, with PC1 being nearly uniform weights across variables, PC2 was a contrast term between clinical and lab variables (i.e. lab had opposing signs to clinical, weights were roughly equal), PC3 was a contrast term between heart and metabolism, and PC4 was primarily metabolic (glucose and glycohemoglobin with hearing). The significance appeared to be lower in the complete case data, with PC5 being almost entirely non-significant.

Finally, looking at the eigenvalues (second moments) for each PC we observed Figure B.47. We observe the same bilinear structure spanning the same PCs as we did with the imputed data, Figure 5.12.

B.5 Tables

Table B.1: Predictors Used — Clinical Variables

Variable	Encoding	Description	NHANES Code
adl_bed	0: no difficulty, 1: otherwise	Difficulty getting in/out of bed	PFQ060J
adl_dress	0: no difficulty, 1: otherwise	Difficulty dressing self	PFQ060L
adl_knife	0: no difficulty, 1: otherwise	Difficulty using fork, knife, cup	PFQ060K
broken_hip	0: no, 1: yes	Broken or fractured a hip	OSQ010A
confusion	0: no, 1: yes	Experience confusion/memory problems	PFQ056
cough	0: no, 1: yes	Coughing most days - over 3 mo period	RDD030
gpa_crouch	0: no difficulty, 1: otherwise	Difficulty kneeling/crouching	PFQ060D
gpa_grasp	0: no difficulty, 1: otherwise	Difficulty grasping/holding small objects	PFQ060P
gpa_overhead	0: no difficulty, 1: otherwise	Difficulty reaching over head	PFQ060O
gpa_sit	0: no difficulty, 1: otherwise	Difficulty sitting for long periods	PFQ060N
gpa_stand	0: no difficulty, 1: otherwise	Difficulty standing for long periods	PFQ060M
gpa_standup	0: no difficulty, 1: otherwise	Difficulty standing from armless chair	PFQ060I
hear	0: good, 1: otherwise	General condition of hearing	AUQ130
iadl_chores	0: no difficulty, 1: otherwise	Household chore difficulty	PFQ060F
iadl_meals	0: no difficulty, 1: otherwise	Difficulty preparing meals	PFQ060G
iadl_money	0: no difficulty, 1: otherwise	Difficulty managing money	PFQ060A
leaked	0: no, 1: yes	Leak urine during nonphysical activities	KIQ046
lem_10steps	0: no difficulty, 1: otherwise	Difficulty walking 10 steps	PFQ060C
lem_qmile	0: no difficulty, 1: otherwise	Difficulty walking a quarter mile	PFQ060B
lsa_homeleis	0: no difficulty, 1: otherwise	Difficulty leisuring at home	PFQ060S
lsa_movies	0: no difficulty, 1: otherwise	Difficulty going to movies, events	PFQ060Q
lsa_socials	0: no difficulty, 1: otherwise	Difficulty attending social events	PFQ060R
sight_dim	0: no difficulty, 1: any difficulty	Difficulty seeing steps/curbs-dim light	VIQ050C
sight_gen	0: excellent or good, 1: fair, poor or very poor	General condition of eyesight	VIQ030
srh	0: excellent, very good or good, 1: fair or poor	General health condition	HUQ010
srh_change	0: better or same, 1: worse	Health now compared with 1 year ago	HUQ020

Table B.2: Predictors Used — Lab Variables

Variable	Normal Range[1]	Description	Units	NHANES Code
albumin	M: [32,45]; F: [32,45]	Albumin	g/L	LBDSALSI
B12	M: [118,701]; F: [118,701]	Vitamin B12, serum	pmol/L	LBDB12SI
bicarb	M: [21,28]; F: [21,28]	Bicarbonate	mmol/L	LBXSC3SI
bilirubin	M: [2,21]; F: [2,21]	Bilirubin, total	umol/L	LBDSTBSI
BPdi	M: [60,90]; F: [60,90]	Blood pressure, diastolic	mmHG	BPXDI
BPsy	M: [90,140]; F: [90,140]	Blood pressure, systolic	mmHG	BPXSY
BUN	M: [2.9,8.2]; F: [2.9,8.2]	Blood urea nitrogen	mmol/L	LBDSBUSI
calcium	M: [2.3,2.74]; F: [2.3,2.74]	Total calcium	mmol/L	LBDSCASI
chol	M: [3.88,6.47]; F: [3.88,6.47]	Total cholesterol	mmol/L	LBDSCHSI
creatinine	M: [60,110]; F: [45,90]	Creatinine	umol/L	LBDSCRSI
folate	M: [376,1450]; F: [376,1450]	Folate, RBC	nmol/L	LBDRBFSI
glucose	M: [3.9,6.1]; F: [3.9,6.1]	Glucose, serum	mmol/L	LBDSGLSI
glycohemo	M: [0,5.7]; F: [0,5.7]	Glycohemoglobin	%	LBXGH
HDL	M: [1.3,∞); F: [1.3,∞)	Direct HDL-Cholesterol	mmol/L	LBDHDLSI
hemo	M: [13.5,18]; F: [12,16]	Hemoglobin	g/dL	LBXHGB
iron	M: [10.7,26.9]; F: [10.7,26.9]	Iron, refrigerated	umol/L	LBDSIRSI
MAP	M: [70,105]; F: [70,105]	Mean arterial pressure $\equiv$ Bpsy/3 + 2 $\cdot$ BPdi/3	mmHg	–
MCV	M: [80,96]; F: [80,96]	Mean cell volume	fL	LBXMCVSI
neutrophils	M: [40,80]; F: [40,80]	Segmented neutrophils percent (40-80%)	%	LBXNEPCT
phosophorus	M: [0.74,1.52]; F: [0.74,1.52]	Phosphorus	mmol/L	LBDSPHSI
platelet	M: [150,450]; F: [150,450]	Platelet count SI	1000 cells/uL	LBXPLTSI
PP	M: [30,65]; F: [30,65]	Pulse pressure $\equiv$ BPsy $-$ BPdi	mmHg	–
protein	M: [60,78]; F: [60,78]	Protein, total	g/L	LBDSTPSI
pulse	M: [60,99]; F: [60,99]	Pulse	bpm	BPXPLS
RCBW	M: [11.6,14.6]; F: [11.6,14.6]	Red cell distribution width	%	LBXRDW
sodium	M: [136,142]; F: [136,142]	Sodium	mmol/L	LBXSNASI
trigly	M: [0.11,2.74]; F: [0.11,2.74]	Triglyceride	mmol/L	LBDSTRSI
uric	M: [240,510]; F: [160,430]	Uric acid	umol/L	LBDSUASI
vitd	M: [12,50]; F: [12,50]	Vitamin D	ng/mL	LBDVIDMS

[1] 0: normal; 1: abnormal, outside range (M: male, F: female).

Table B.3: Outcomes Used

Variable	Encoding	Description	NHANES Code
ADLDIS	0: healthy, 1: disabled	Had any ADL disability	ADLDIS
Age	Continuous	Age in years, top-coded at 85	RIDAGEYR
arthritis	0: no, 1: yes	Told had arthritis	MCQ160A
BMI[1]	Continuous	Body mass index	BMXBMI
bronchitis	0: no, 1: yes	Told had chronic bronchitis	MCQ160K
cancer	0: no, 1: yes	Told had cancer or malignancy	MCQ220
cataracts	0: no, 1: yes	Had a cataract operation	VIQ070
CRP[1]	Continuous	Inflammation biomarker	LBXCRP
diabetes	0: no, 1: yes	Told you have diabetes	DIQ010
emphysema	0: no, 1: yes	Told had emphysema	MCQ160G
exhaustion	0: no difficulty, 1: otherwise	Difficulty walking between rooms	PFQ060H
FI	Continuous	Frailty index	–
FP	0: < 3 diagnostics, 1: ≥ 3 [208]	Frailty phenotype	–
gait[1]	Continuous	Time to walk 20 feet	MSXW20TM
h_angina	0: no, 1: yes	Told had angina	MCQ160D
h_attack	0: no, 1: yes	Told had heart attack	MCQ160E
h_disease	0: no, 1: yes	Told had coronary heart disease	MCQ160C
h_failure	0: no, 1: yes	Told had congestive heart failure	MCQ160B
have_bronchitis	0: no, 1: yes	Still have chronic bronchitis	MCQ170K
have_liver_con	0: no, 1: yes	Still have a liver condition	MCQ170L
have_thyroid_dis	0: no, 1: yes	Still have a thyroid problem	MCD170M
hu_hosp	0: no, 1: yes	Overnight hospital patient, past year	HUD070
hu_times	0: ≤ 3, 1: ≥ 4[16]	#Times receive healthcare, past year	HUQ050
hypertension	0: no, 1: yes	Told had high blood pressure	BPQ020
IADLDIS	0: healthy, 1: disabled	Had any IADL disability	IADLDIS
liver_con	0: no, 1: yes	Told had any liver condition	MCQ160L
microalb.	Continuous	microalbuminuria = albumin/creatinine	–
osteoporosis	0: no, 1: yes	Told had osteoporosis/brittle bones	OSQ060
overweight	0: no, 1: yes	Told you were overweight	MCQ160J
physical activity	0: more active/same, 1: less active	Compare activity w/others same age	PAQ520
polypharmacy	0: < 4, 1: ≥ 4 [169]	Number of prescription medicines	RXD295
stroke	0: no, 1: yes	Told you had a stroke	MCQ160F
survival_10year	0: died, 1: survived	Survived 10+ years post-study	survival_10year
telomere[1]	Continuous	Mean telomere length	TELOMEAN
thyroid	0: no, 1: yes	Told had a thyroid problem	MCD160M
weak_kidneys	0: no, 1: yes	Told had weak/failing kidneys	KIQ022
weakness	0: no difficulty, 1: otherwise	Difficulty lifting or carrying	PFQ060E
adl_bedDIS[2]	0: no/some difficulty, 1: much/unable	Difficulty getting in/out of bed	PFQ060JDIS
adl_dressDIS[2]	0: no/some difficulty, 1: much/unable	Difficulty dressing self	PFQ060LDIS
adl_knifeDIS[2]	0: no/some difficulty, 1: much/unable	Difficulty using fork, knife, cup	PFQ060KDIS
exhaustionDIS[2]	0: no/some difficulty, 1: much/unable	Difficulty walking between rooms	PFQ060HDIS
FICLINIC[2]	Continuous	Frailty index, clinical data only	–
FILAB[2]	Continuous	Frailty index, lab data only	–
iadl_choresDIS[2]	0: no/some difficulty, 1: much/unable	Household chore difficulty	PFQ060FDIS
iadl_mealsDIS[2]	0: no/some difficulty, 1: much/unable	Difficulty preparing meals	PFQ060GDIS
iadl_moneyDIS[2]	0: no/some difficulty, 1: much/unable	Difficulty managing money	PFQ060ADIS
survival_1year[2]	0: died, 1: survived	Survived 1+ year post-study	survival_1year
survival_5year[2]	0: died, 1: survived	Survived 5+ years post-study	survival_5year

[1] Log-scaled.
[2] Excluded during pooling.

Table B.4: Demographics / Outcome Statistics

Outcome	Mean or Prevalence[1]
age	72.8 (sd=8.2,N=1872)
FI	0.216 (sd=0.135,N=1872)
FP	6.8% (127/1872)
TELOMEAN	0.916 (sd=0.216,N=1382)
BMI	28.175 (sd=5.478,N=1434)
CRP	0.551 (sd=1.044,N=1532)
gait	7.480 (sd=3.561,N=1428)
microalbuminuria	0.509 (sd=0.144,N=1518)
hu_times	52.3% (978/1869)
hu_hosp	18.1% (338/1872)
survival_1year	96.7% (1811/1872)
survival_5year	79.4% (1486/1872)
survival_10year	57.5% (1077/1872)
ADL Disability	10.9% (203/1870)
adl_dressDIS	5.7% (107/1870)
adl_bedDIS	5.8% (108/1870)
adl_knifeDIS	1.9% (36/1870)
exhaustionDIS	6.6% (123/1870)
IADL Disability	19.0% (355/1864)
iadl_moneyDIS	7.3% (137/1868)
iadl_choresDIS	15.6% (291/1865)
iadl_mealsDIS	10.3% (192/1868)
arthritis	46.8% (874/1868)
h_failure	7.9% (146/1851)
h_disease	10.3% (190/1836)
h_angina	7.9% (145/1846)
h_attack	11.3% (210/1861)
stroke	8.9% (165/1864)
emphysema	4.0% (74/1864)
overweight	32.9% (614/1869)
bronchitis	6.5% (122/1866)
liver_con	3.0% (56/1864)
thyroid	14.2% (265/1864)
cancer	20.8% (389/1868)
have_bronchitis	3.5% (66/1862)
have_liver_con	1.1% (21/1861)
have_thyroid_dis	9.4% (174/1856)
diabetes	20.0% (374/1871)
weak_kidneys	5.0% (93/1859)
osteoporosis	14.1% (262/1853)
cataracts	26.2% (486/1858)
hypertension	52.5% (978/1863)
polypharmacy	45.2% (688/1521)

[1] Before any imputation (prevalence/number non-missing).

Table B.5: Covariates Used

Variable	Encoding	Description	NHANES Code
age	Continuous, top-coded at 85	Age in years, top-coded at 85	RIDAGEYR
education	0: no post-secondary, 1: post-secondary	Education level	DMDEDUC2
has_partner	0: no partner, 1: cohabitating or married	Partner status	DMDMARTL
income	0: below poverty line, 1: above	Family poverty income ratio	INDFMPIR
race	0: non-white, 1: white	Race	RIDRETH1
sex	0: male, 1: female	Sex	RIAGENDR
smoker	0: non-smoker, 1: smoker	Smoker status	SMQ040

Table B.6: Auxiliary Variables Used For Imputation

Description	NHANES Code
MSDEXCLU	Exclusion criteria for muscle strength
TELOSTD	Standard deviation of TELOMEAN
PEASCST1	Blood Pressure Status
PEASCTM1	Blood Pressure Time in Seconds
PEASCCT1	Blood Pressure Comment
BPXCHR	Heart rate (2x30 second)
BPQ150A	Had food in the past 30 minutes?
BPQ150B	Had alcohol in the past 30 minutes?
BPQ150C	Had coffee in the past 30 minutes?
BPQ150D	Had cigarettes in the past 30 minutes?
BPAARM	Arm selected
BPACSZ	Coded cuff size
BPXDB	# of dropped beats in 30 seconds
BPXPULS	Pulse regular or irregular?
BPXPTY	Pulse type
BPXML1	MIL: maximum inflation levels (mm Hg)
BPAEN1	Enhancement used first reading
BPAEN2	Enhancement used second reading
BPAEN3	Enhancement used third reading
BPAEN4	Enhancement used fourth reading
BPXSAR	BPsy average reported to examinee
BPXDAR	BPdi average reported to examinee

Table B.7: PCA Rotation Coefficients, Bootstrapped (N=2000)

Input	PC1	PC2	PC3	PC4	PC5	PC6	PC7
BPdi	-0.120(6)	0.153(16)	0.064(34)	-0.021(71)	-0.245(88)	0.059(152)	-0.057(169)
BPsy	-0.204(7)	0.377(15)	0.430(25)	0.184(50)	0.121(62)	-0.042(88)	0.057(87)
pulse	-0.078(6)	0.111(15)	0.017(27)	-0.073(41)	0.008(72)	-0.077(99)	0.074(113)
PP	-0.255(7)	0.389(15)	0.398(21)	0.116(37)	0.049(45)	0.021(56)	0.034(49)
MAP	-0.101(5)	0.193(15)	0.239(24)	0.084(39)	-0.056(53)	-0.003(68)	-0.005(79)
albumin	-0.027(3)	0.023(8)	-0.011(13)	-0.025(17)	-0.009(22)	0.003(23)	-0.004(25)
bicarb	-0.037(4)	0.029(9)	-0.022(14)	-0.027(20)	-0.041(24)	0.001(29)	-0.004(28)
bilirubin	-0.017(3)	0.020(7)	-0.015(10)	-0.033(13)	-0.020(21)	-0.028(21)	0.003(27)
BUN	-0.102(6)	0.043(15)	-0.068(31)	-0.108(67)	-0.256(73)	0.027(126)	-0.022(126)
creatinine	-0.121(6)	0.073(16)	-0.080(35)	-0.120(84)	-0.337(86)	0.000(158)	-0.039(148)
HDL	-0.213(7)	0.265(19)	-0.466(30)	-0.138(72)	0.056(153)	-0.131(308)	0.338(223)
folate	-0.053(5)	0.045(11)	0.018(18)	-0.012(29)	-0.072(33)	-0.015(46)	0.014(45)
glucose	-0.124(6)	0.128(18)	-0.323(32)	0.321(45)	0.036(89)	0.074(122)	-0.104(127)
glycohemo	-0.183(7)	0.205(20)	-0.388(38)	0.436(57)	0.136(110)	0.084(139)	-0.121(124)
hemo	-0.075(5)	0.052(13)	-0.041(25)	-0.047(56)	-0.234(52)	-0.023(94)	-0.004(69)
iron	-0.120(6)	0.072(16)	-0.061(35)	-0.078(77)	-0.269(92)	0.009(157)	0.005(172)
MCV	-0.081(6)	0.067(14)	0.017(28)	-0.147(50)	-0.144(68)	-0.041(98)	-0.042(90)
phosophorus	-0.011(2)	0.004(4)	-0.006(7)	-0.009(10)	-0.021(12)	0.005(14)	-0.007(13)
platelet	-0.025(3)	0.018(7)	-0.017(12)	-0.030(19)	-0.050(20)	-0.007(26)	0.000(23)
protein	-0.064(5)	0.069(12)	-0.029(22)	0.053(35)	-0.042(45)	-0.017(60)	-0.032(63)
RDW	-0.046(4)	0.034(10)	-0.018(18)	-0.005(38)	-0.146(36)	0.006(61)	-0.005(48)
neutrophils	-0.016(2)	0.016(7)	-0.012(8)	-0.001(12)	-0.018(13)	-0.001(16)	-0.006(16)
sodium	-0.095(6)	0.079(16)	-0.006(27)	-0.000(45)	-0.037(62)	0.002(91)	-0.047(95)
calcium	-0.094(6)	0.104(17)	-0.041(32)	-0.087(56)	-0.124(91)	-0.079(135)	0.082(162)
chol	-0.096(6)	0.099(15)	-0.043(29)	-0.024(48)	0.019(70)	-0.045(82)	0.005(97)
trigly	-0.065(5)	0.078(13)	-0.168(21)	0.095(34)	0.063(50)	-0.007(88)	0.088(59)
uric	-0.057(4)	0.034(11)	-0.040(19)	-0.059(35)	-0.118(40)	-0.003(61)	-0.019(56)
B12	-0.038(4)	0.027(9)	0.004(16)	-0.032(20)	-0.019(24)	-0.013(30)	-0.018(29)
vitd	-0.032(4)	0.021(8)	0.002(13)	0.037(19)	-0.051(22)	-0.003(30)	0.001(27)
adl_dress	-0.120(5)	-0.157(8)	0.057(19)	0.076(27)	0.013(61)	-0.096(71)	0.060(95)
adl_bed	-0.134(5)	-0.174(7)	0.023(19)	0.088(25)	0.038(45)	-0.042(68)	0.065(60)
adl_knife	-0.051(4)	-0.075(8)	0.032(14)	0.038(21)	0.018(40)	-0.062(43)	0.027(62)
iadl_money	-0.095(5)	-0.103(9)	0.044(19)	0.024(34)	-0.038(78)	-0.144(72)	0.002(127)
iadl_chores	-0.203(4)	-0.215(7)	0.038(16)	0.045(21)	-0.028(32)	0.022(44)	0.033(42)
iadl_meals	-0.120(5)	-0.172(7)	0.044(19)	0.083(33)	-0.030(80)	-0.149(80)	0.045(138)
lsa_movies	-0.176(5)	-0.216(6)	0.056(18)	0.087(29)	-0.045(55)	-0.076(74)	0.064(89)
lsa_homeleis	-0.066(5)	-0.106(8)	0.009(15)	0.080(22)	0.022(49)	-0.090(45)	0.002(79)
lsa_socials	-0.158(5)	-0.208(6)	0.036(19)	0.104(30)	-0.051(65)	-0.107(74)	0.054(106)
lem_qmile	-0.237(4)	-0.159(10)	0.011(22)	-0.026(39)	-0.089(81)	0.150(82)	0.005(130)
lem_10steps	-0.204(5)	-0.170(9)	0.050(24)	0.012(38)	-0.051(83)	0.155(81)	0.032(127)
gpa_grasp	-0.112(5)	-0.117(9)	0.038(21)	0.047(31)	0.067(46)	-0.037(60)	0.033(66)
gpa_overhead	-0.129(5)	-0.122(9)	0.001(23)	0.111(28)	0.032(46)	-0.024(51)	0.018(57)
gpa_sit	-0.132(5)	-0.106(11)	0.011(26)	-0.014(47)	0.133(69)	0.088(97)	0.052(106)
gpa_stand	-0.266(4)	-0.133(11)	0.023(25)	-0.085(40)	0.022(96)	0.190(87)	-0.009(164)
gpa_standup	-0.195(4)	-0.187(8)	0.049(18)	0.058(25)	0.013(42)	0.027(68)	0.067(48)
gpa_crouch	-0.291(4)	-0.071(13)	-0.013(29)	-0.125(52)	0.096(127)	0.254(116)	-0.006(221)
srh	-0.181(5)	-0.026(14)	-0.105(30)	0.095(53)	-0.046(106)	-0.129(178)	-0.178(159)
srh_change	-0.087(5)	-0.057(10)	-0.018(20)	-0.003(29)	-0.036(44)	-0.049(58)	-0.042(60)
hear	-0.214(6)	0.127(18)	-0.062(56)	-0.563(82)	0.356(145)	0.010(181)	-0.071(176)
sight_gen	-0.156(5)	0.017(15)	-0.026(36)	-0.088(71)	0.042(166)	-0.232(295)	-0.317(227)
sight_dim	-0.191(5)	-0.082(13)	-0.002(33)	-0.112(61)	0.178(96)	-0.101(165)	-0.148(123)
broken_hip	-0.022(3)	-0.017(6)	0.013(9)	-0.004(9)	-0.007(12)	-0.005(13)	-0.001(15)
confusion	-0.097(5)	-0.069(10)	0.010(21)	-0.039(37)	0.026(80)	-0.152(77)	-0.022(129)
cough	-0.052(4)	0.012(9)	-0.002(17)	-0.051(22)	0.017(29)	0.005(36)	-0.026(31)
leaked	-0.097(5)	-0.020(11)	0.010(22)	-0.015(32)	-0.028(42)	0.001(51)	0.019(54)

B.6 Figures

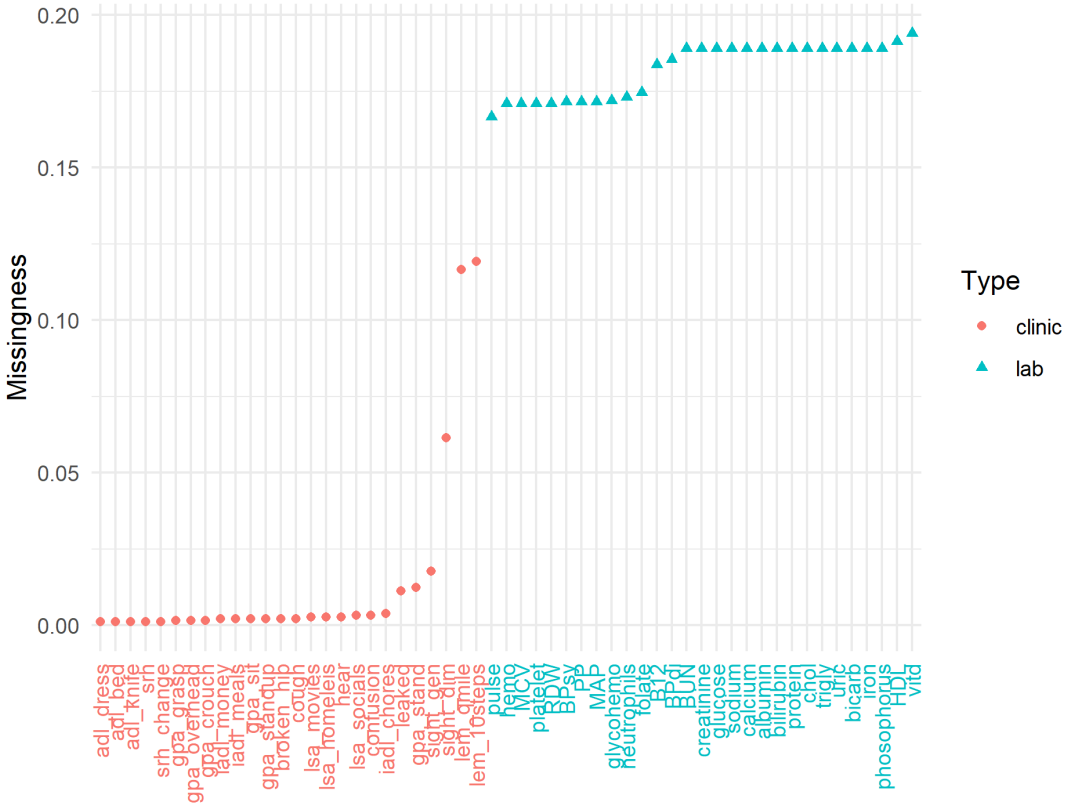

Figure B.1: Missingness frequencies of predictor variables (rank ordered). Note the clinical variables (red circles) have much lower missingness than lab variables (blue triangles). This is likely because clinical variables are self-reported. See also Figure B.4. Missingness is after gated imputation (Section B.1.2).

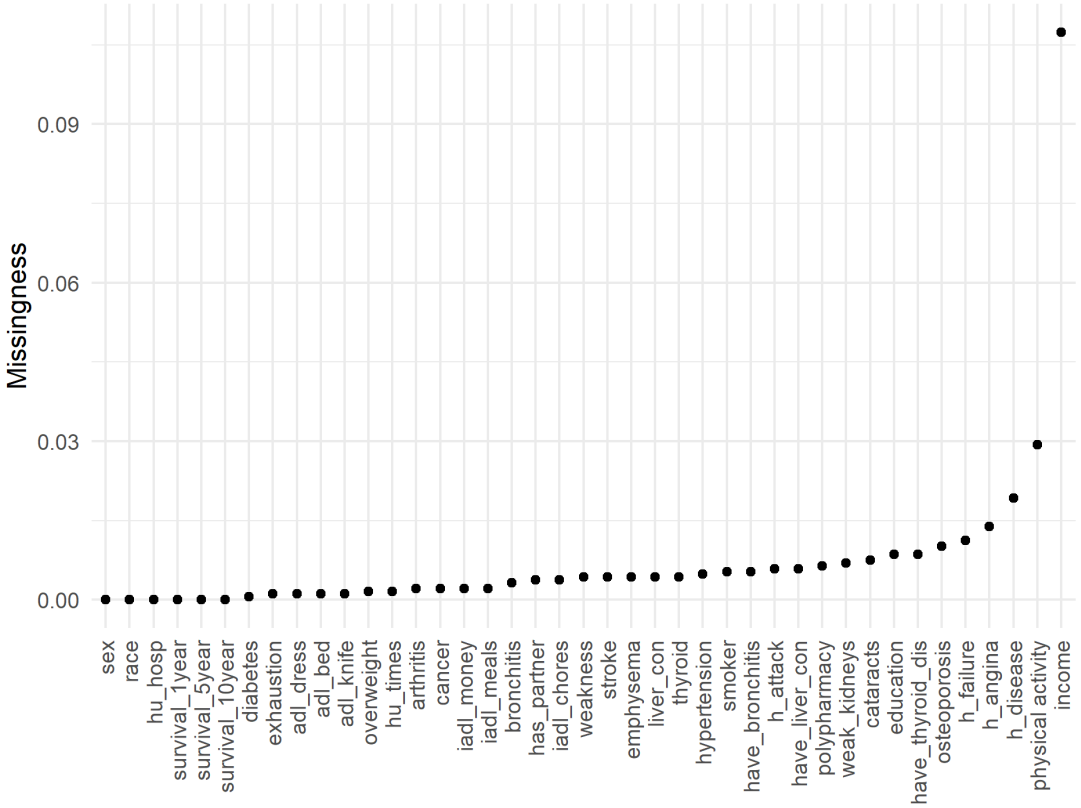

Figure B.2: Missingness frequencies of binary outcome variables and demographic covariates. The missingness was low in the binary outcomes and covariates likely because they are self-reported. We saw much higher missingness in the measured outcomes, Figure B.3. Missingness is after gated imputation (Section B.1.2).

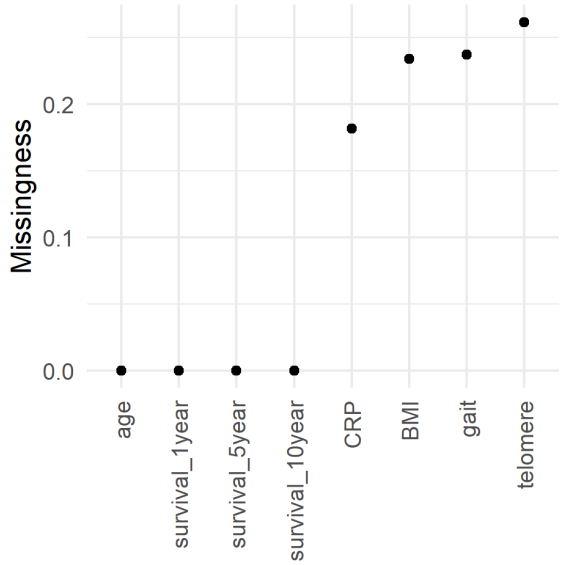

Figure B.3: Missingness frequencies of continuous outcome variables and demographic covariates. CRP, BMI, gait and telomere all had to be measured, which explains why they had much higher missingness rate than the other outcomes (here and Fig. B.2).

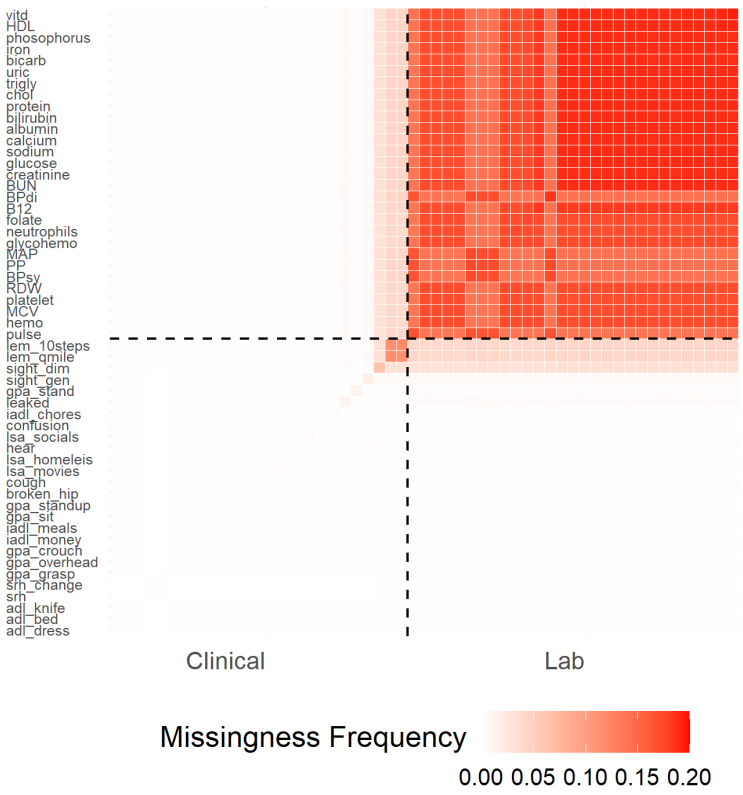

Figure B.4: Missingness joint frequency histogram (predictors). Diagonal is missingness frequency of each variable. Off-diagonal is mutual missingness frequency of variable pairs. Observe that the lab data tended to be mutually missing (top right), which can lead to serious problems with common imputation algorithms [145]. Imputation quality was validated in Section B.1.2.

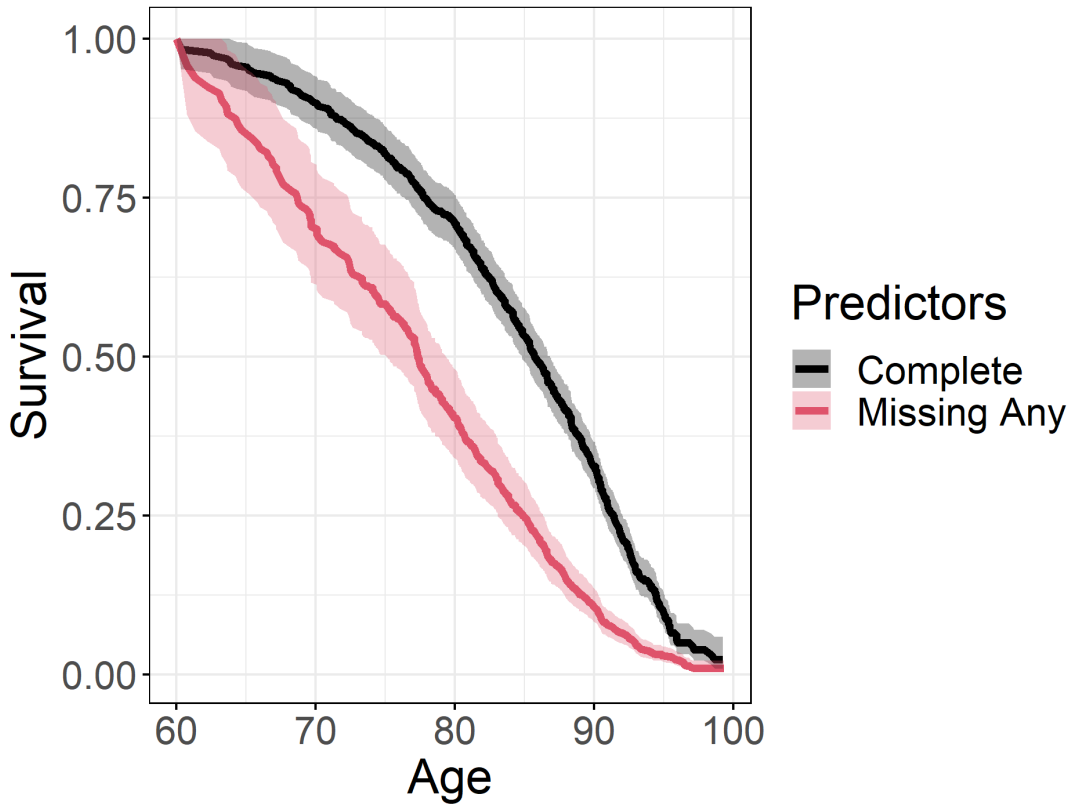

Figure B.5: Missingness survival effect. Individuals missing any predictor variable (red line) showed worse survival than individuals with all of their predictor variables reported (black line). This is an indication of informative censoring, meaning that the complete case analysis, Section B.4, could be biased [180]. Note: ages were top-coded at 85 which could cause distortions of the survival curves past age 85.

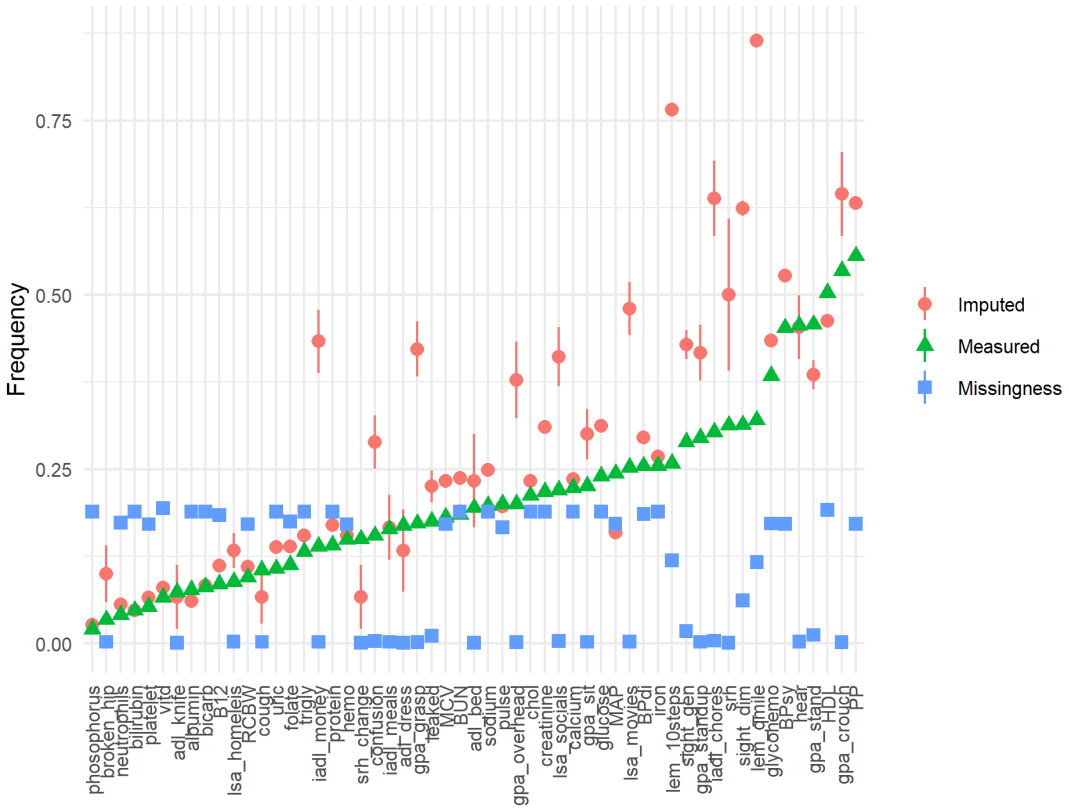

Figure B.6: Deficit frequencies for imputed versus measured health variables. Measured deficit frequency (green triangles) and imputed deficit frequency (red circles). The missingness frequency of the variable is given in blue (squares). Imputed variables (red circles) tended to be more frequently deficit. This is consistent with our other observations that individuals missing values tended to have worse overall health (e.g. worse survival, Figure B.5).

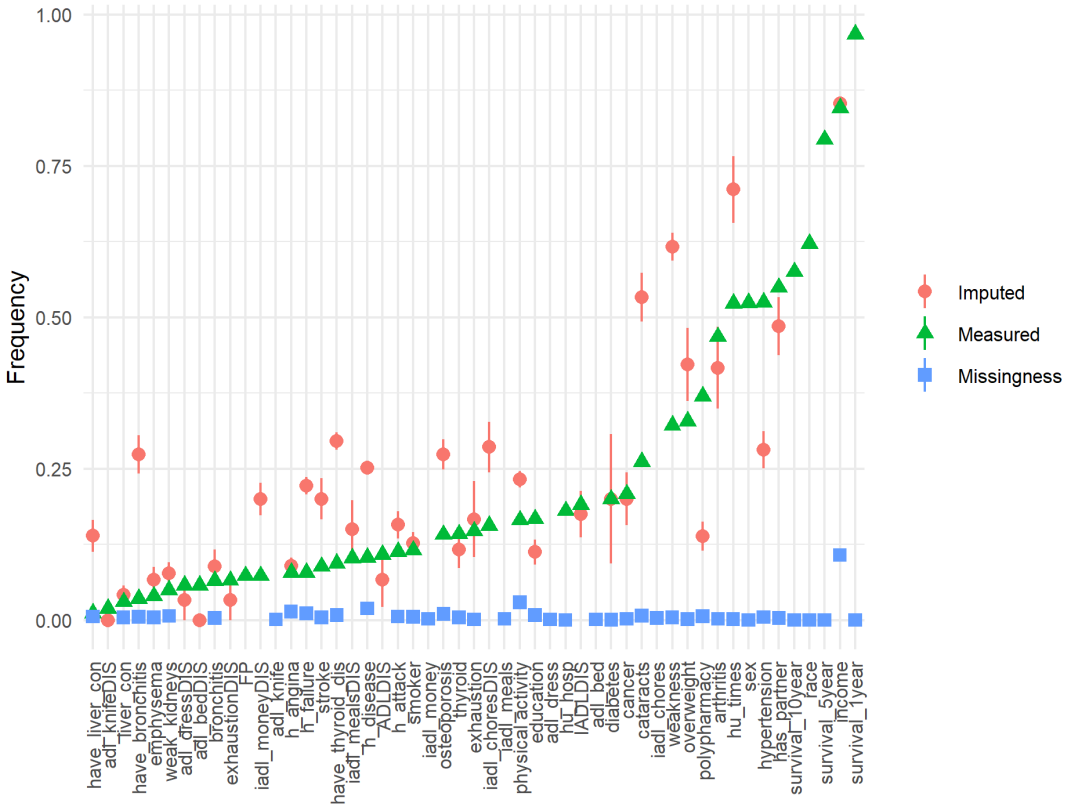

Figure B.7: Deficit frequencies for imputed versus measured binary outcomes. Measured frequency (green triangles) and imputed frequency (red circles). The missingness frequency of the variable is given in blue (squares). Imputed adverse outcomes tended to be more frequently deficit (red circles, excluding: income, race, survival, sex, education, smoker and partner). For the demographical covariates (income, race, sex, education, smoker and partner) and survival outcomes (1, 5 and 10 year), frequency indicates how often they were of value 1 (see Section B.1.1 for encoding rules). Where no red point is visible it is because there were no missing values and hence no imputed values (e.g. FP and survival). Imputed frequencies are clearly higher than measured frequencies, consistent with our other observations that individuals with missing values tended to have worse overall health (e.g. worse survival, Figure B.5, and more health deficits, Figure B.8).

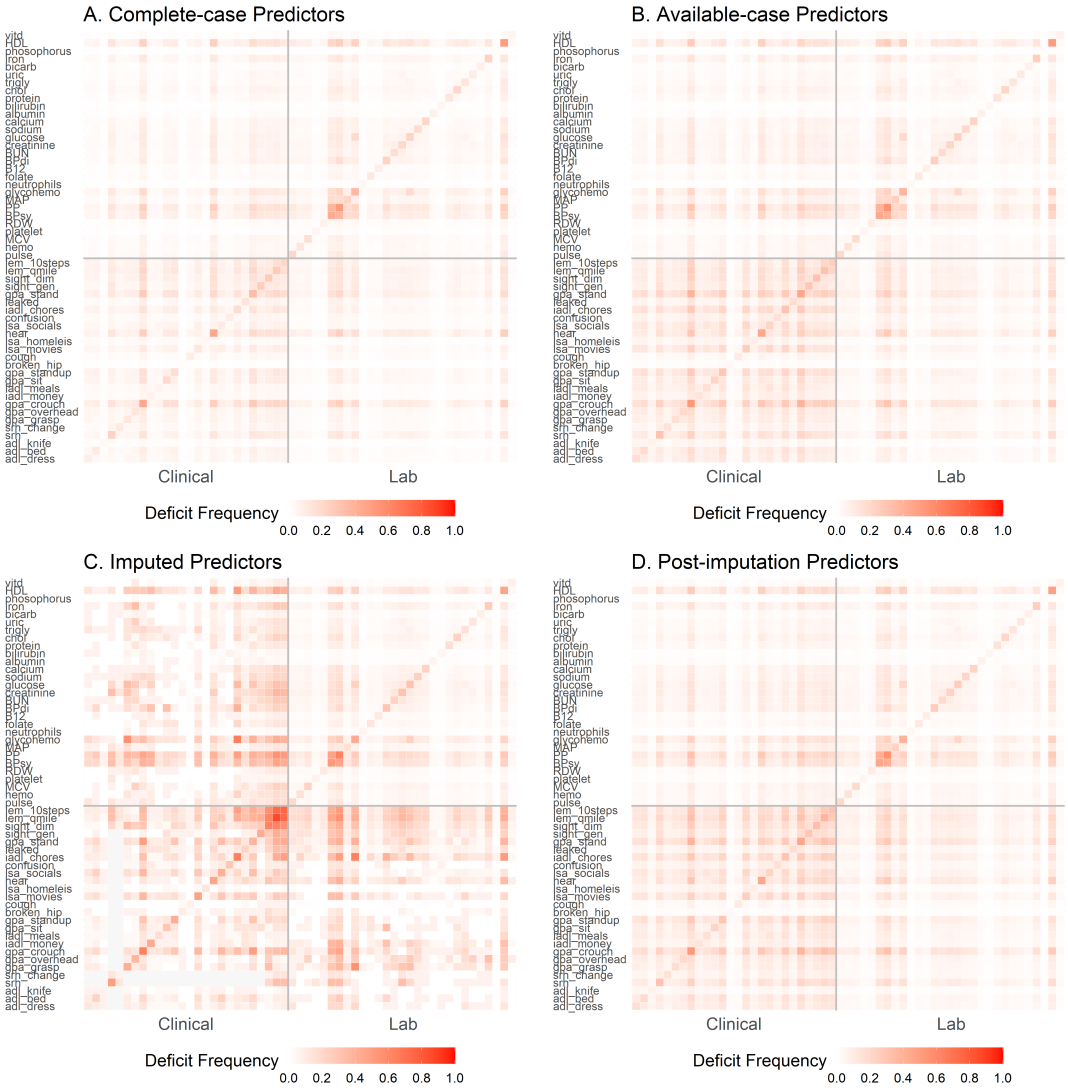

Figure B.8: Joint 2D frequency histogram for predictors with and without imputation. A: complete case data (individuals have no NA), B: available case data (NAs skipped), C: imputed predictors (imputed values only), and D: post-imputation predictors (all values, including imputed). The imputed values clearly have more deficits but the net effect on the post-imputation data is negligible relative to the available case data. We expected more deficits in the imputed values because individuals missing data had worse survival (Figure B.5). Individuals with complete data clearly had fewer deficits (A). Top values are lab variables, bottom are clinical. Tiles are grayed out if there were no values in respective variable pair.

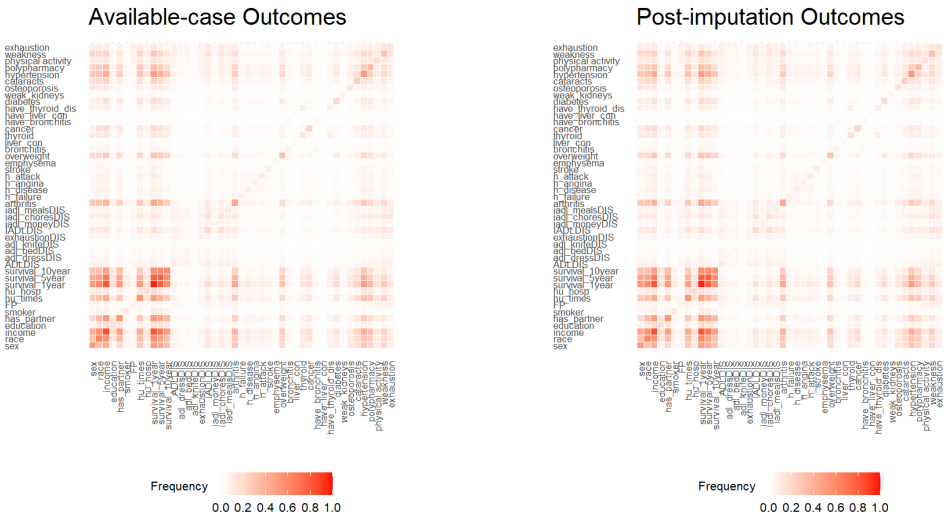

Figure B.9: Joint 2D frequency histogram for outcomes for available case or imputed data (frequency each binary outcome was '1'). We see no difference by eye. This could be because of the relatively low missingness for outcomes (Figures B.2 and B.3) Available case outcomes were included in the "complete case" dataset (only predictors were required to be complete case).

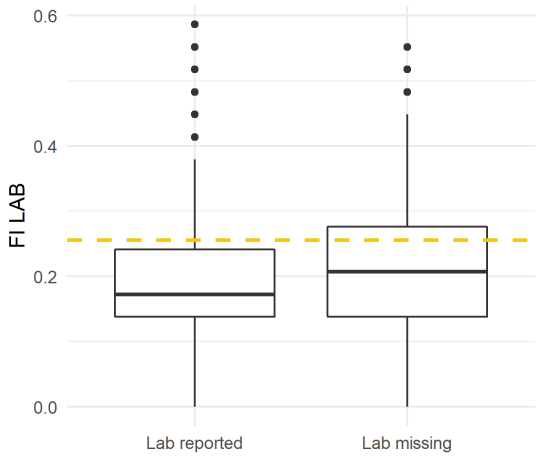

Figure B.10: Boxplot of individuals missing the lab block (right) versus those with lab variables measured (left). y-axis indicates the FI LAB value (after imputation). Solid black lines are the medians for each group. White boxes delineate the interquartile ranges (25% to 75% quantiles); the whiskers span the 95% confidence intervals (assuming asymptotic-normality) [206]. Patients missing data had worse survival, Figure B.5, and therefore we expect them to have a higher FI LAB. The expected shift, Eq B.1, of the median is the dashed yellow line (median of left $+ \Delta FI$). We expect this to be close to the median on the right side. The estimate is in the correct direction, and the line is within the interquartile range. This implies a good imputation.

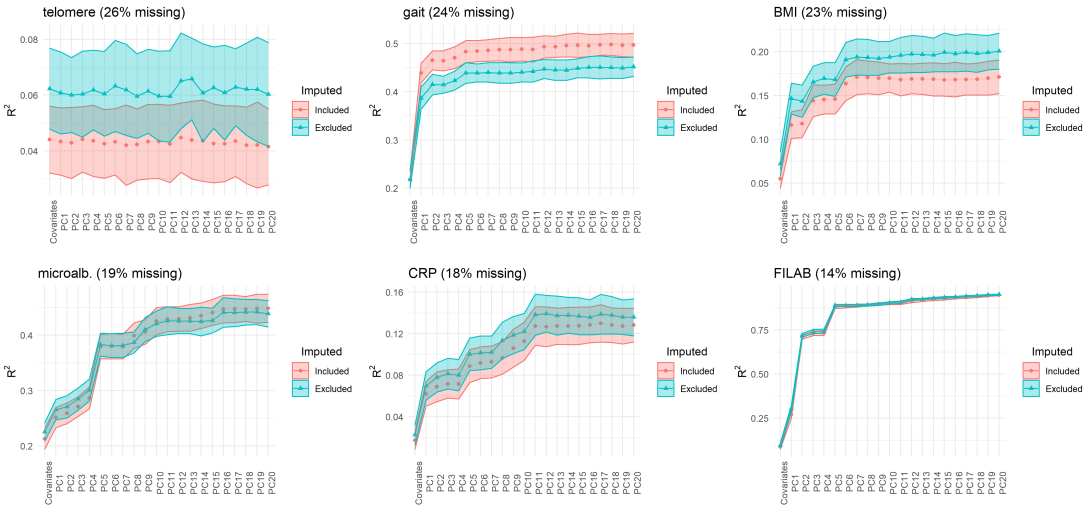

Figure B.11: Improvement in predictive power as more PCs are included, with and without imputed outcomes. Highest missingness outcomes. This figure allows us to surmise the effect that imputing the outcomes has had on the accuracy metrics. Red band (circles) is including imputed values, blue band (triangles) exclude imputed values. Bands tend to overlap, indicating non-significant differences. Band is the cross-validation error. In particular, the outcomes with the 4th-6th most missing data (bottom row) overlap heavily, implying that the remaining outcomes — which had much lower missingness ($< 3\%$) — would have a negligible difference due to the imputed values. In the outcomes with the 1st-3rd most missingness we see the same pattern with a global shift in accuracy, this does not affect our study conclusions which are based on the shape of the curves. In Figure B.12 we observed that imputed gait values tended to be slower than normal, which may explain why they were easier to predict. We do not think this is an indication of a poor imputation, since we expect those individuals to have low gait speeds (Section B.1.2).

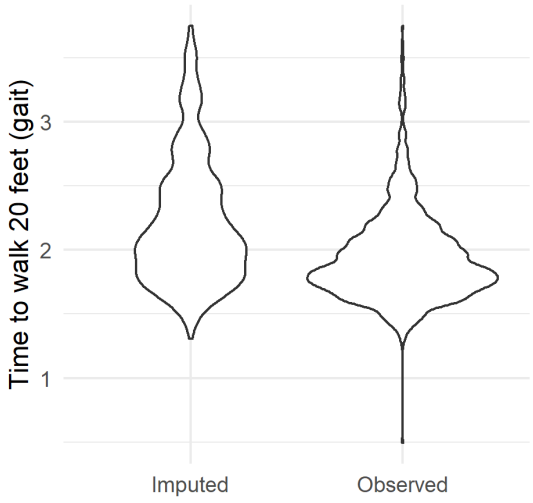

Figure B.12: Violin plot of imputed timed-gait values (log scale). Higher is worse. Outlines represent the distributions of imputed (left) and observed (right) values. Imputed values tended to be higher implying slower gait speeds. This is consistent both with worse mortality for those missing data, Figure B.5, and with the reasons for missingness of this particular variable (when it was reported). See Section B.1.2 for details.

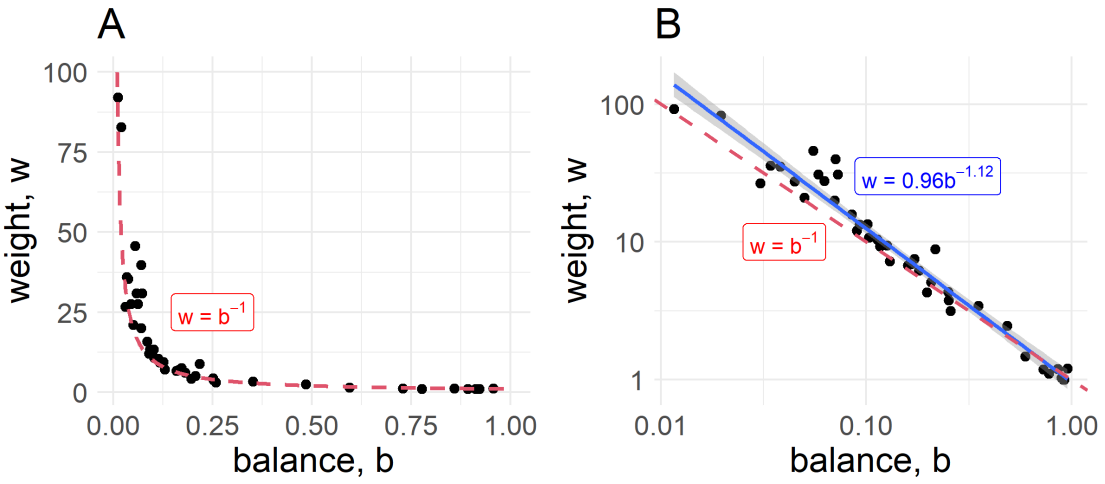

Figure B.13: GLM weight selection for binary outcomes on A: linear and B: log-log scale. Weight, w is the optimized parameter (Eq. B.2). Balance, b, is the ratio of minority over majority class frequencies (Eq. B.3). Each point represents an optimized binary GLM (logistic regression). There is clear power law behaviour — the log-log plot shows a linear relationship — with $w \sim b^{-1}$ being a good choice of the power (red dashed line, Eq. B.4). The solid blue line indicates the least-squares fit. See Section B.2.1 for complete discussion.

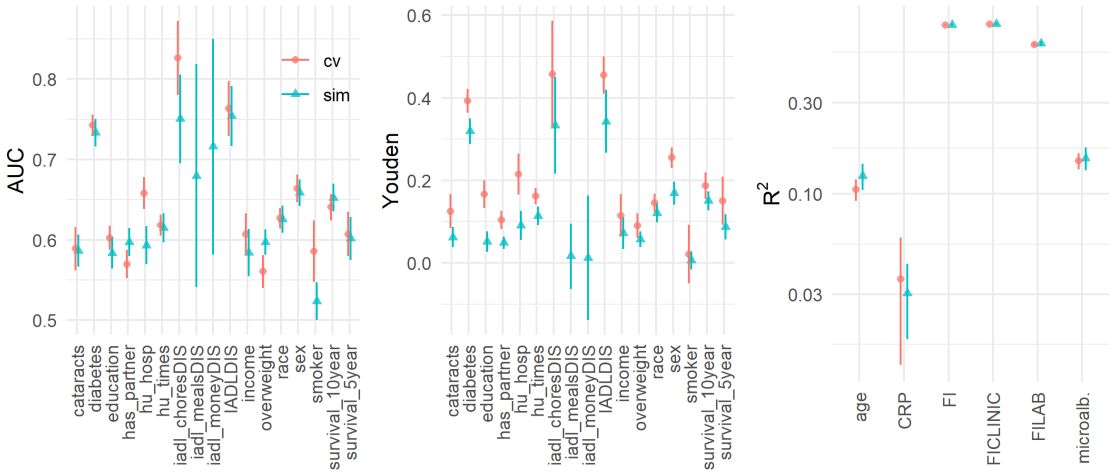

Figure B.14: Simulation study of cross-validated estimates. As described in Section B.2.2, we generated a synthetic dataset based on our study sample. Error bars were estimated directly from the synthetic dataset via cross-validation (red points) and compared to error bars generated by Monte Carlo sampling of the synthetic dataset distribution (blue triangles). The simulated values provide a ground truth, the cross-validation estimates show error bars of similar size, with roughly the correct coverage (point estimates are typically within one or two error bars of each other). This demonstrates our cross-validation procedure is correctly calibrated for our data. Note: missing data points are due to failed fitting of the ROC curve (due to insufficient case data in the cross-validation).

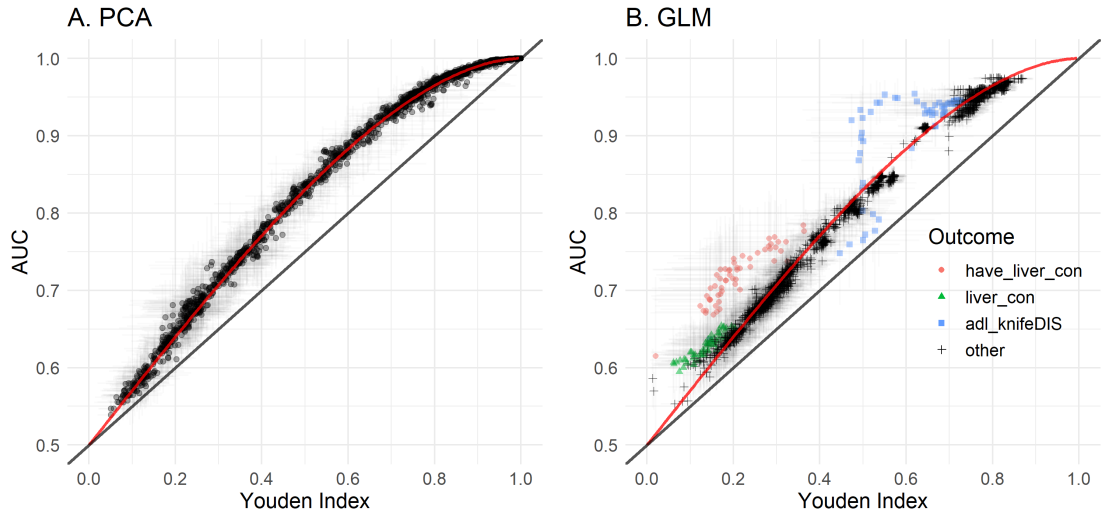

Figure B.15: AUC is strongly correlated with Youden index, closely following Eq. B.9 (red line). A. AUC vs Youden index for compression using PCA, each point is a unique input variable and a unique number of PCs (1-55), with cross-validation error (55 inputs $\times$ 55 PC options = 3025 points). Eq. B.9 fits excellently. The relationship is smooth, non-linear and saturating, with AUC reaching 1 before Youden index. The saturating of the AUC indicates that it is a less sensitive scale, and explains why we observed compression reaching unity faster on the AUC scale (Figure B.17) than the Youden scale (Figure 5.3). We considered also the scores from the binary GLMs in B., one point per model with each model having covariate information and between 0-55 PCs, with cross-validation error (39 outcomes $\times$ 56 PC options = 2184 points). Eq. B.9 fit both the GLM and compression well (red lines), although the compression scores clearly fit better. The GLM outcomes which fit poorly were: ever had a liver condition (liver_con; green triangles), still have a liver condition (have_liver_con; red circles), and significant difficulty using a knife/fork (adl_knifeDIS; blue squares). These happen to be the three rarest outcomes: 1.2%: have_liver_con, 1.9%: adl_knifeDIS, and 3.0% liver_con (the least common input deficit was phosphorous at a rate of 2.2%). The weighting scheme (Section B.2.1) may affect the relationship between AUC and Youden indexes when outcomes are very rare (PCA was not weighted). Diagonal black line is $y = x/2 + 1/2$, which illustrates that AUC $>$ Youden$/2 + 1/2$. The values in these figures are the same as used in Figure 5.3 (A) and Figures 5.7, 5.9 and 5.10 (B).

Figure B.16: Eigen-decomposition of the joint histogram, without scaling. The first column is the complete 2D joint deficit histogram, the remaining columns sum to the first column (Eq. A6). The first PC is clearly dominant and is dense, meaning it is nearly equal weights for each variable (akin to the FI). The eigen-decomposition naturally finds blocks of correlated variables. When it runs out of blocks it looks for strong diagonal terms. This causes PCA to naturally block out like-variables, e.g. lab vs clinical in PC2, similar to an expert choosing to create an FI out of variables from the same domain. Compared to the transformed scale, Figure 5.2, we see that the higher PCs are much dimmer, reflecting their minor contribution.

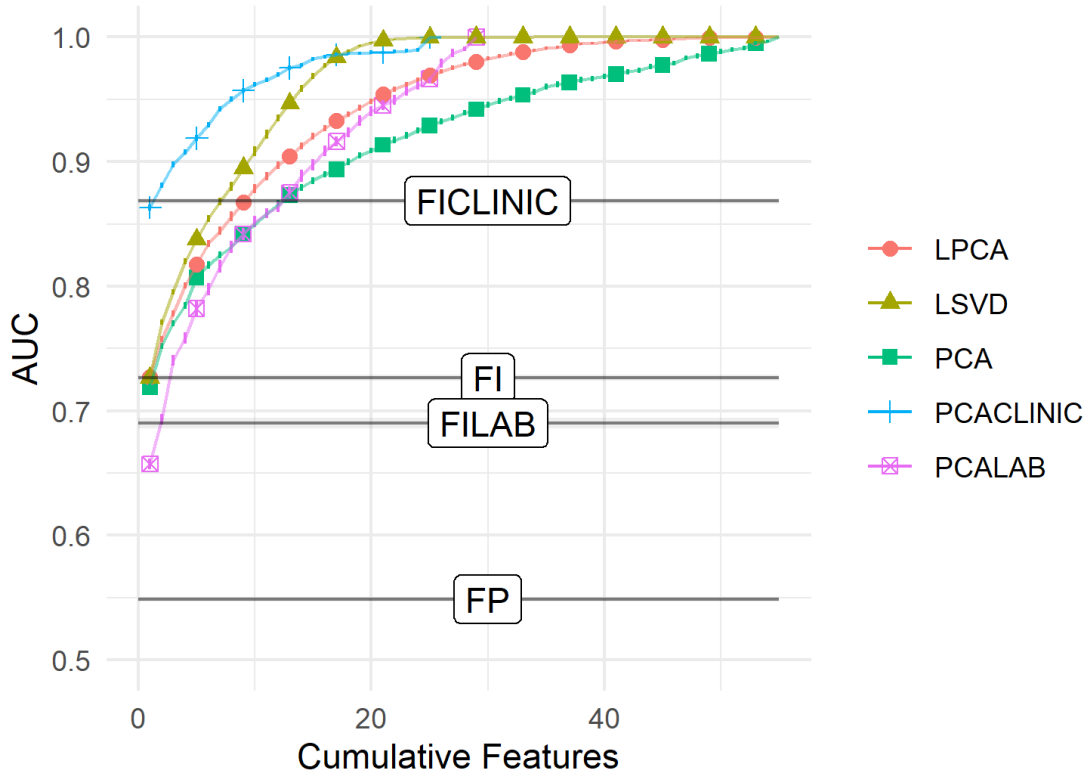

Figure B.17: Cumulative compression with AUC. The reader should look for two things: (1) the number of components (dimensions) needed to achieve a relatively high score, and (2) the slope of the curve, when it flattens we can expect the features are variable-specific or otherwise less important. We see the same relative importance as with the Youden index, Figure 5.3, but the AUC saturates much faster, with LSVD reaching perfect AUC near 20 features (vs 30 for Youden index). The faster saturation appears to be due to known differences between the AUC and Youden index (Eq. B.9 and Figure B.15). The Youden index is preferable since it provides a definite accuracy at a specific threshold, such as we would see in medical diagnosis [214].

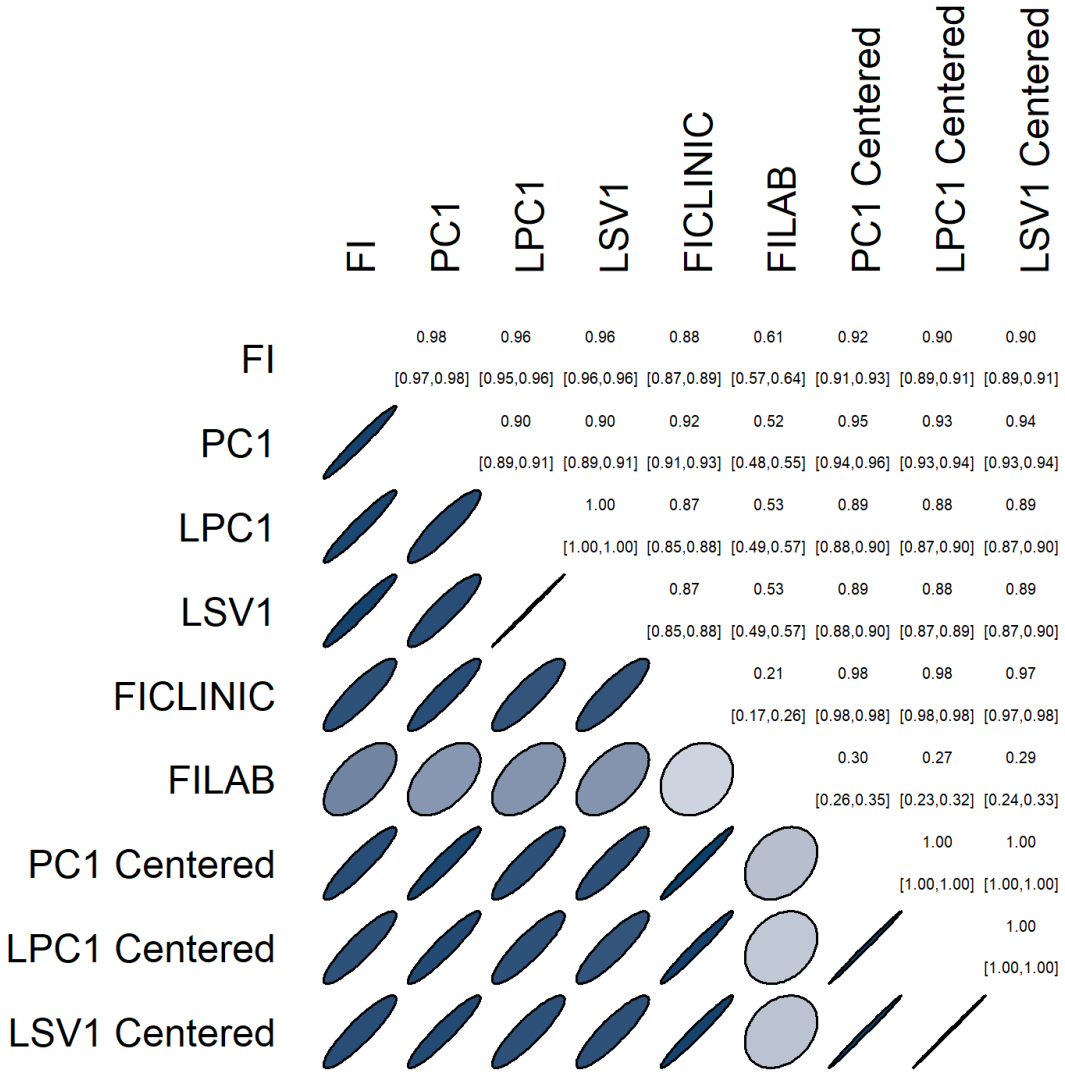

Figure B.18: Spearman correlation of features across algorithms, extended. This is an extension of Figure 5.4 to include centered features. Observe that the centered features show the same strong correlations as the uncentered features, illustrating that lack of centering is not the cause of the correlation. Upper triangle is correlation coefficient with 95% confidence interval; ellipses are equivalent Gaussian contours (for visualization) [129]. The first latent dimension for either PC, LPC or LSVD correlates strongly with the FI, even when centered. We also observe that the first latent dimension correlates more strongly with clinical than lab data.

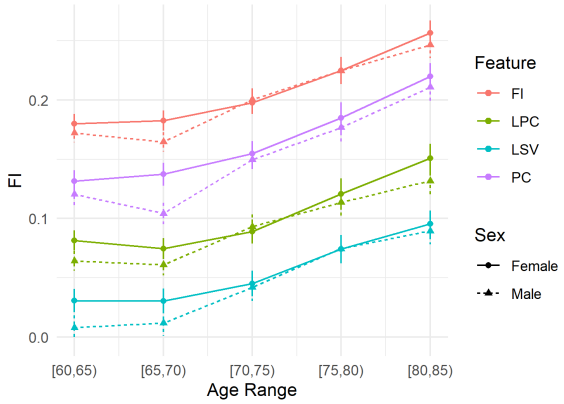

Figure B.19: Age and sex dependence of first latent feature (from PC1, LPC1, LSV1 and FI). All features show similar age and sex dependence. Females (solid, circles) increase approximately exponentially with age, males increase more linearly (dotted, triangles). Similar to the FI LAB in [21] we observed a strong sex-effect at younger ages that is smaller at older ages. Scale only applies to FI; PC1/LPC1/LSV1 have been globally scaled for visualization (linear scaling). Individuals age 85+ were excluded from this figure because age was top-coded at 85 (we don't know their true age). This is further evidence that all four algorithms are sensitive to the same underlying signal, see discussion in "The first latent dimension 'is' the frailty index".

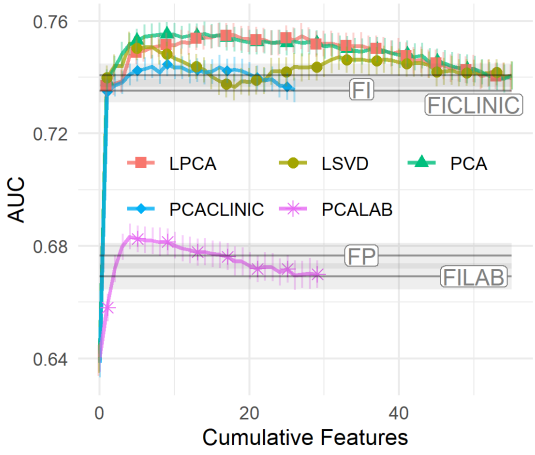

Figure B.20: Cumulative prediction plot for discrete outcomes, AUC (GLM). 0th dimension is demographic information. Prediction improved quickly, reaching a maximum at 5-10 features. Increasing the number of features initially improves prediction but eventually it gets worse due to overfitting. Results are qualitatively identical to the Youden index, Figure 5.7.

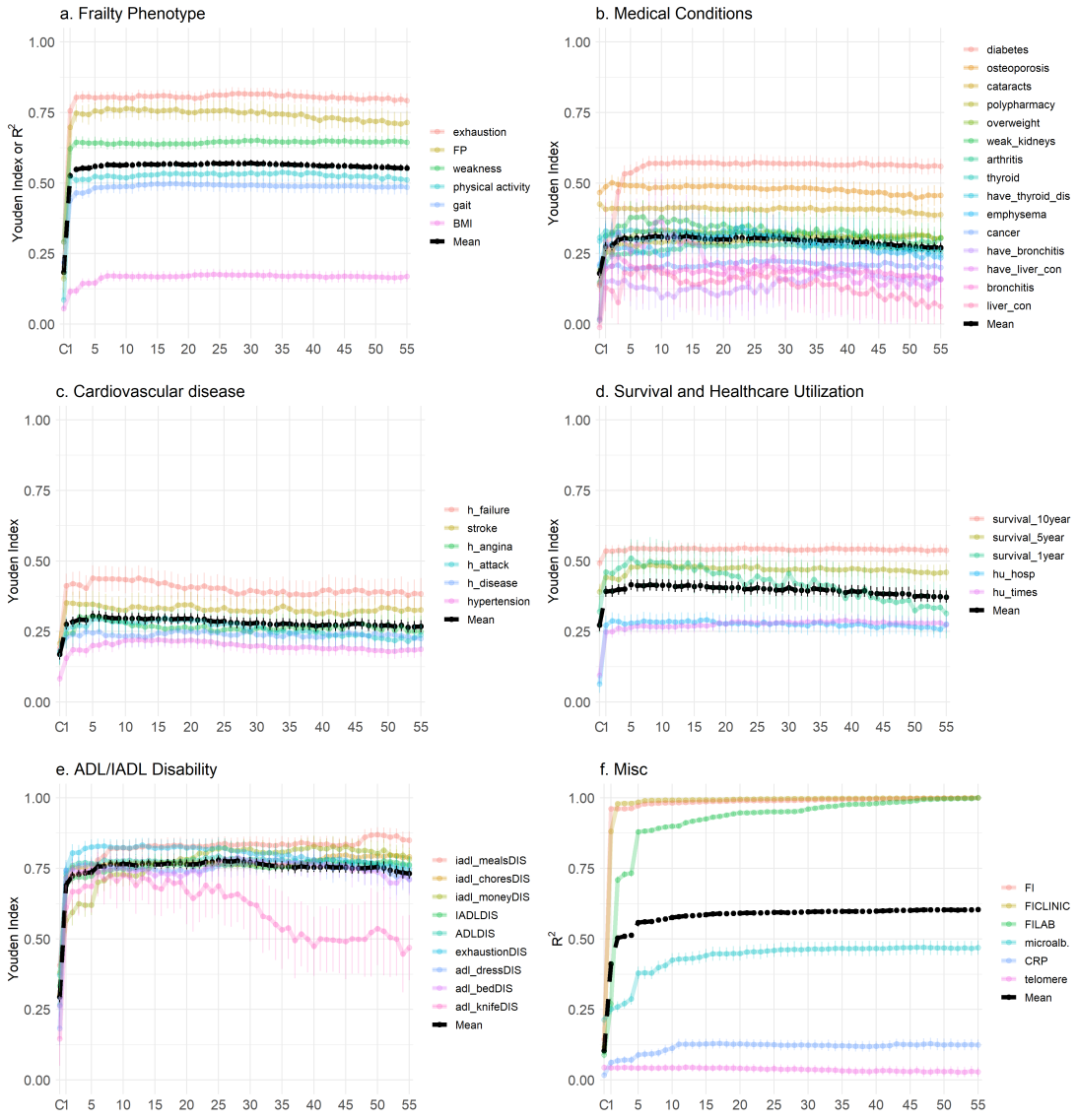

Figure B.21: Improvement in predictive power as more PCs are included, grouped by outcome type (GLM). This figure extends Figure 5.9 to include all PCs. X-axis labels indicate the cumulative number of PCs included, "C" means demographical covariates only. Coloured lines indicate specific outcomes, black line indicates the mean for each group. Scores saturate quickly, justifying truncating the plots. Several of the ADL/IADL disability appear to improve with high PCs, e.g. iadl_mealDIS from PC47-PC49, we suspect this is consequence of our choice of input variables (see Section B.3.3). Note: legends are sorted from top (best) to bottom (worst) performance for the PC55 model.

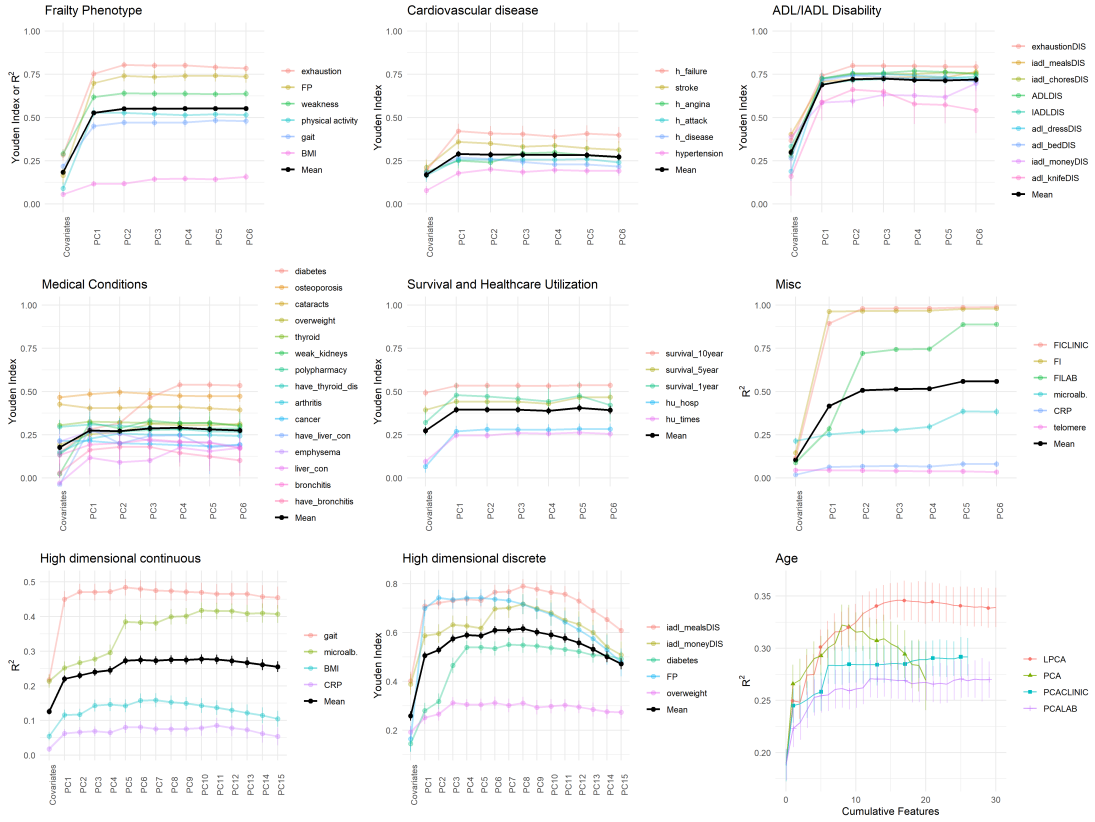

Figure B.22: Improvement in predictive power as more PCs are included, grouped by outcome type (with non-linear terms). Models included all cumulative linear, quadratic and interaction terms up to the indicated PC, starting with the model using only covariates. Coloured lines indicate specific outcomes, black line indicates the mean. The last row contains hand-picked variables based on their high-dimensional behaviour. Compare black lines to Figures 5.9 and 5.10, which used only linear terms. The linear models performed at least as well, e.g. ADL/IADL disability saturates at 0.75 both here (non-linear) and in Figure 5.9 (linear). The last row show a clear tendency to overfit (downward curving of performance with increasing number of predictors; compare to Figure 5.10).

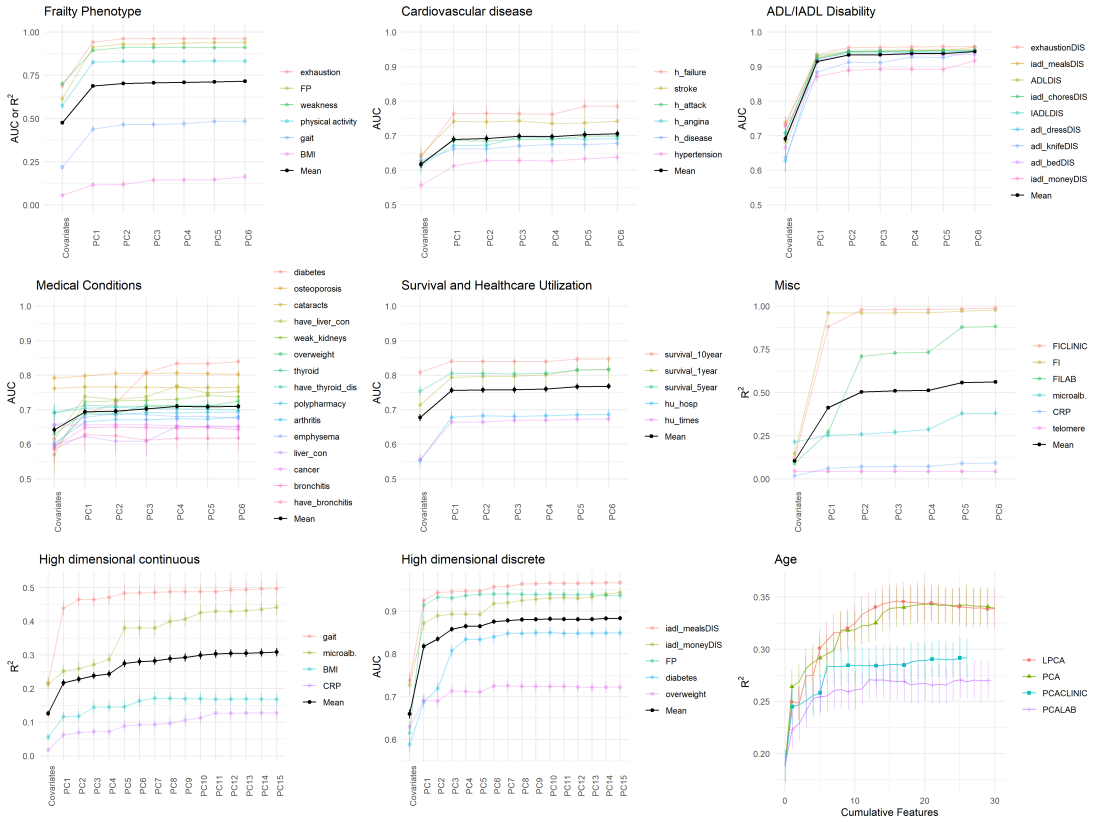

Figure B.23: Improvement in predictive power as more PCs are included, grouped by outcome type (with AUC; linear terms only). Coloured lines indicate specific outcomes, black line indicates the mean. The last row contains hand-picked variables based on their high-dimensional behaviour. We see little difference in the relative performances using the AUC versus the Youden index, Figures 5.9 and 5.10. This is not surprising given the strength of the correlation between AUC and Youden index, which is approximately linear for AUC $\lesssim 0.9$ (Figure B.15).

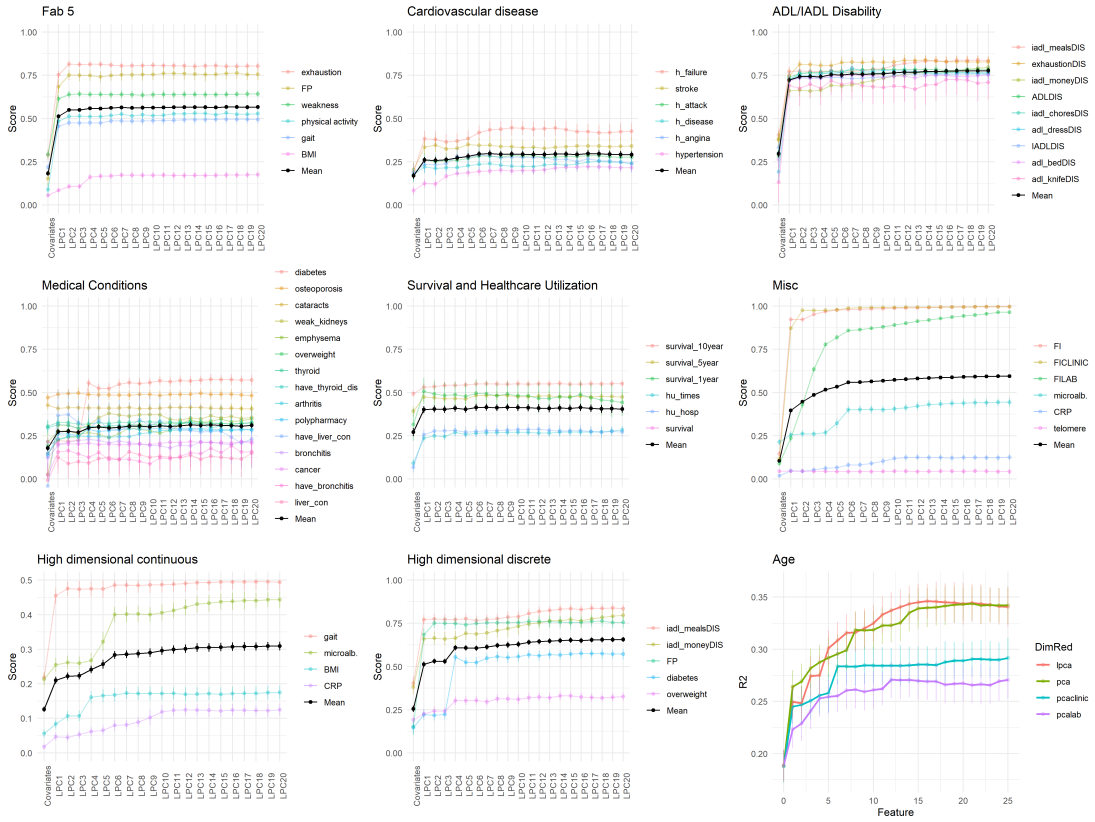

Figure B.24: Improvement in predictive power as more features are included, grouped by outcome type (using LPCA rather than PCA). Coloured lines indicate specific outcomes, black line indicates the mean. The last row contains hand-picked variables based on their high-dimensional behaviour. Score was Youden index for discrete outcomes and R^2 for continuous outcomes. Compare to Figures 5.9 and 5.10, which used PCA. Results are very similar to PCA, further evidence of the similarities between PCA and LPCA.

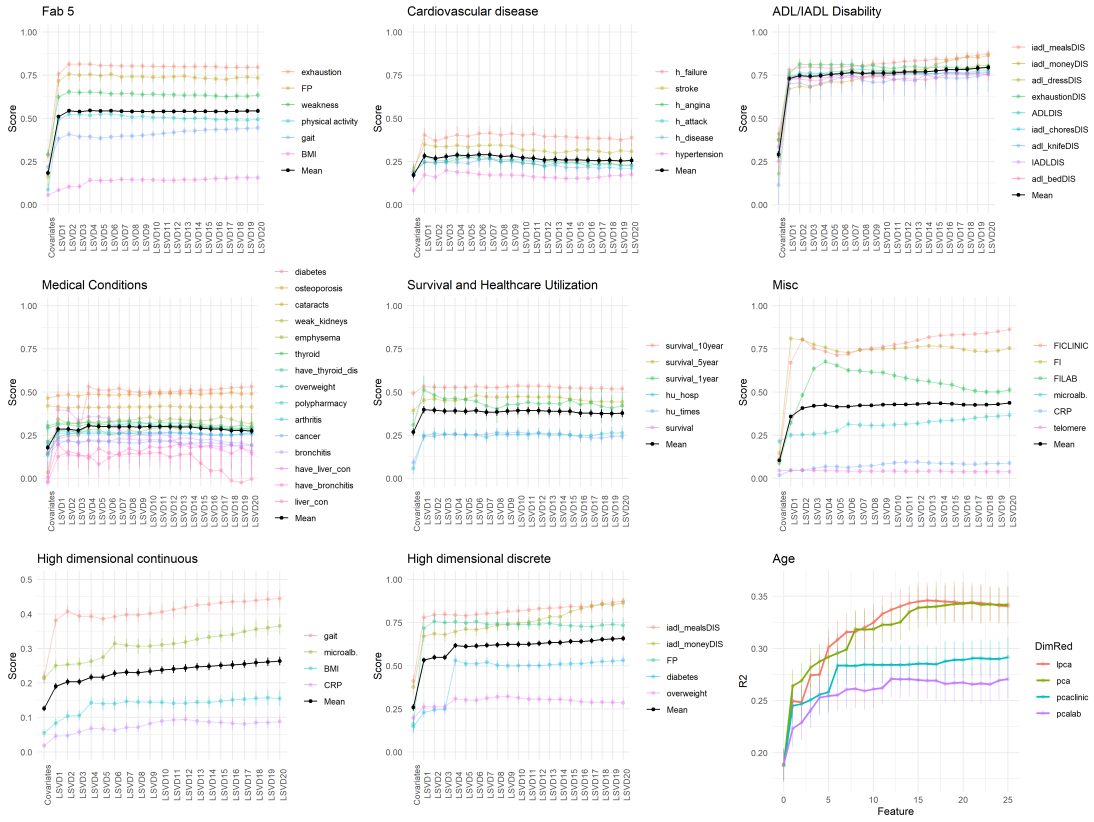

Figure B.25: Improvement in predictive power as more features are included, grouped by outcome type (using LSVD rather than PCA). Coloured lines indicate specific outcomes, black line indicates the mean. The last row contains hand-picked variables based on their high-dimensional behaviour. Score was Youden index for discrete outcomes and R^2 for continuous outcomes. Compare to Figures 5.9 and 5.10, which used PCA (or Figure B.24 which was very similar to PCA). Notice the overall scores are lower here than for PCA (e.g. look at the last row). This is consistent with our observations in Figures 5.7 and 5.8 which showed LSVD generally resulted in worse prediction scores.

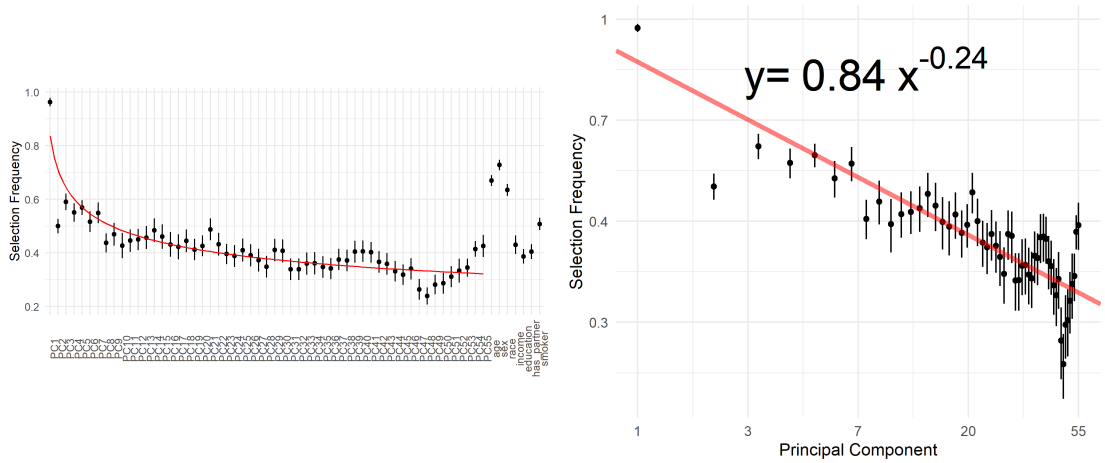

Figure B.26: GLM feature selection frequencies. Left: linear scale, right: log-log scale. GLM models were given all PCs and covariates and then LASSO picked the optimal subset for prediction (see Section B.3.3). We see a continuous drop in feature selection frequency with increasing PC number, suggesting less informative features. This helps explain why the prediction scores saturated at relative low PCs in Figure B.21. The linear behaviour on the log-log plot motivates a power law fit. The nth PC was selected with frequency $y = 0.84n^{-0.24}$ (red line). Results are pooled from 10-fold cross-validation of all outcomes, excluding the FP and FI (to prevent trivial self-prediction).

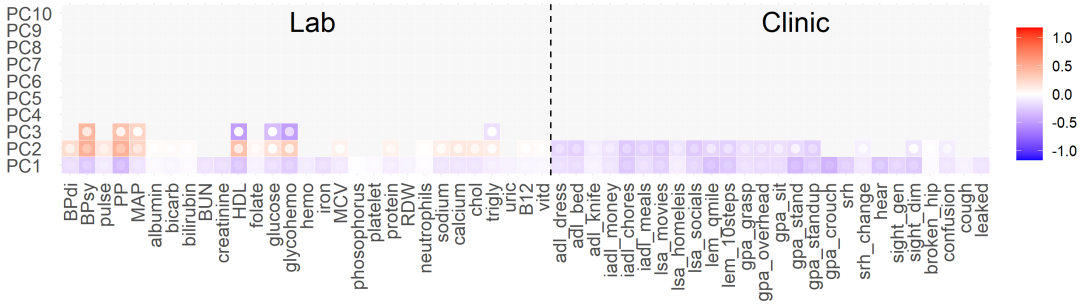

Figure B.27: PCA rotation sensitivity analysis. We randomly sampled subsets of 30 variables (out of 55), then performed PCA on the subset, and then aggregated the rotation coefficients. Left side are coefficients for the lab variables, right are clinical. The first three PCs are quantitatively robust. The remaining PCs were not robust (non-significant/grayed-out). It is worth comparing to Figure 5.11 which used all 55 variables and randomly sampled individuals (with replacement), and showed more robust PCs up to PC5 or PC6.

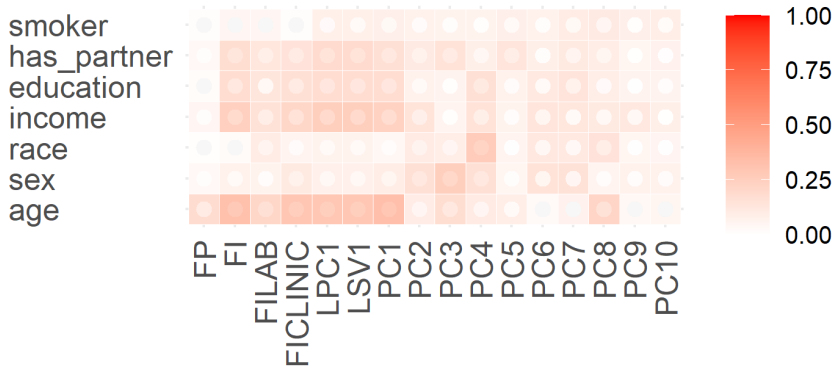

Figure B.28: Feature associations with demographical variables. Age vs FP, and sex, race, income, education, has_partner, and smoker vs all variables: Youden index (see the Section 5.2.2 for details). Age vs remaining features (FI, FILAB, ..., PC10): correlation coefficient (absolute value). The raw predictive power of each feature should flag any demographical-specific effects. Note the age effect for PC1, sex effect for PC3 and race effect for PC4. The age effect supports our claim in the "Age stratification" section of our results that PC1 becomes increasingly dominant with age. Inner circle fill colour is lower limit of 95% CI (white is non-significant).

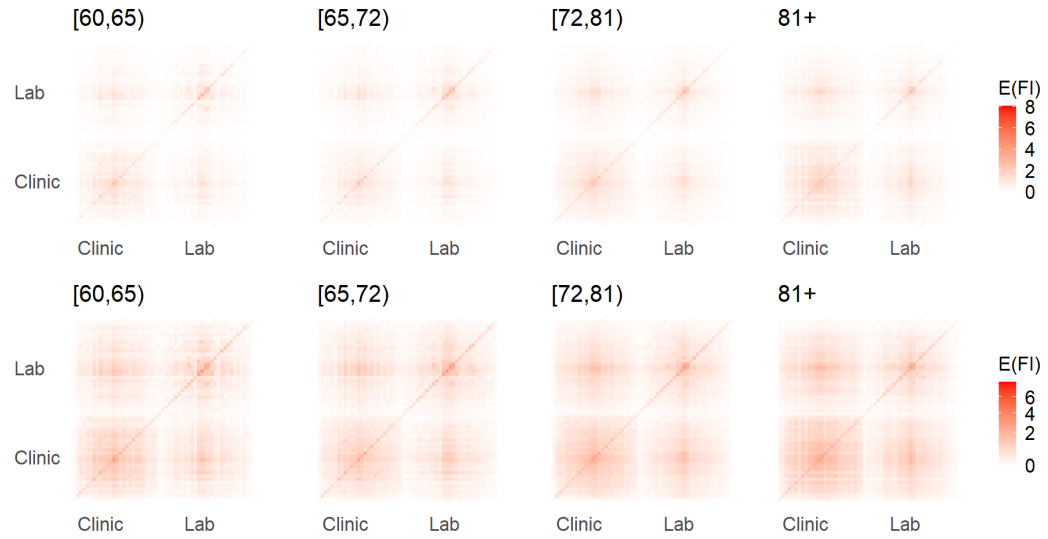

Figure B.29: 2D Histogram as a function of age, normalized. Top: linear fill scale, bottom: gamma transformed for visualization $(\text{sign}(x)|x|^{\gamma}, \gamma = 2/3)$. We have normalized by the probability of having a deficit at that age, i.e. the scale is in units of mean FI for that age range, E(FI). We see the 2D histogram structure is relatively stable with age, showing only an increase in saturation with age. See Section B.3.5 for context.

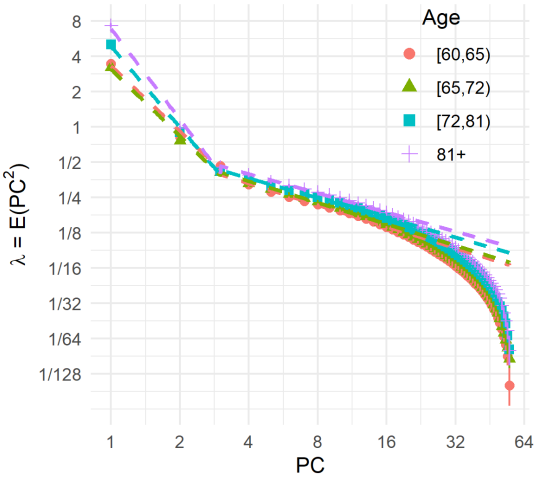

Figure B.30: PCA second moments (eigenvalues) with respect to age quartile, bootstrapped CI (N=2000). The first eigenvalue and the slope of the first line both increase with age. The increasing first eigenvalue is consistent with the increasing FI with age, a widely-reported phenomenon. The increasing slope is analogous to a decrease in fractal dimension with age [61]. Log-log scales. See Section B.3.5 for context.

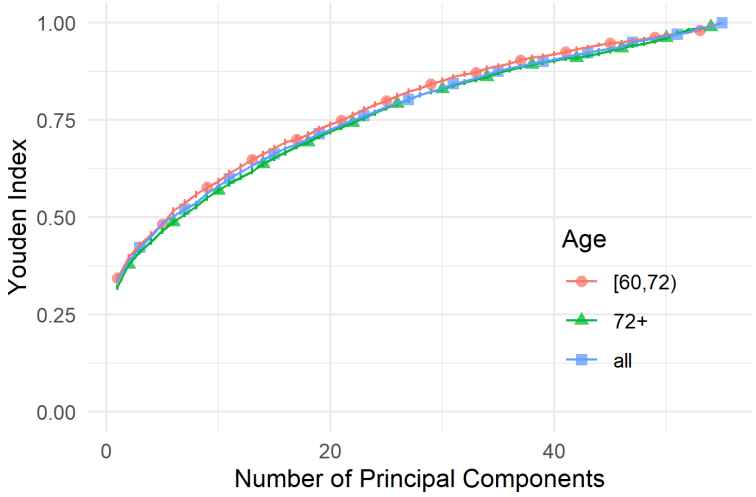

Figure B.31: Cumulative compression by age group using PCA. We see only a minor difference between the cohorts, with the young cohort compressing a little better. Note: age was top-coded at 85. For comparison with Figure 5.3.

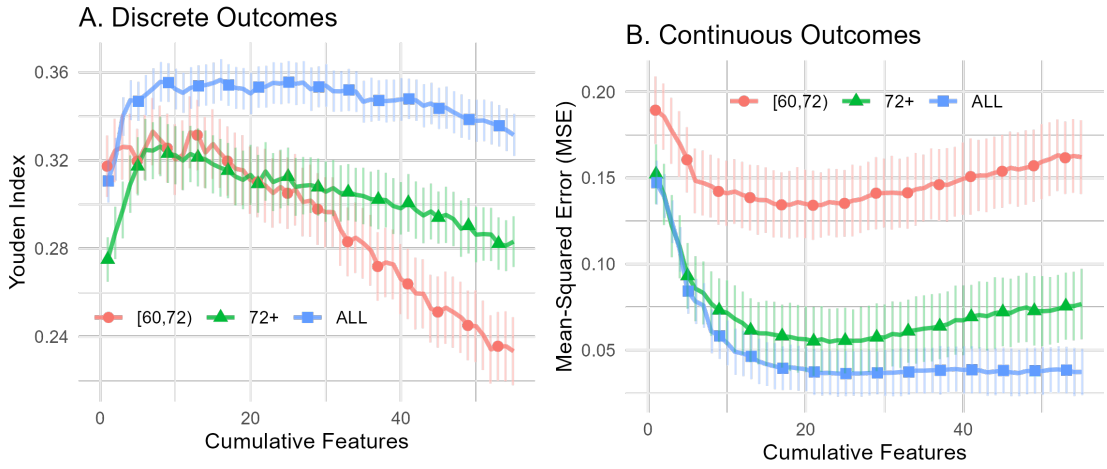

Figure B.32: Cumulative prediction plot stratified by age, but with no demographical variables (GLM). The older individuals (green triangles) clearly had better model performance — similar Youden index (A) and lower MSE (B) — than the younger individuals (red circles). We have included the full population for comparison (blue squares), which clearly performs the best (although it also has twice as much training data as the other two samples). In contrast to Figures 5.7 and 5.8, we have not included covariates as the 0th feature (see Section B.3.5).

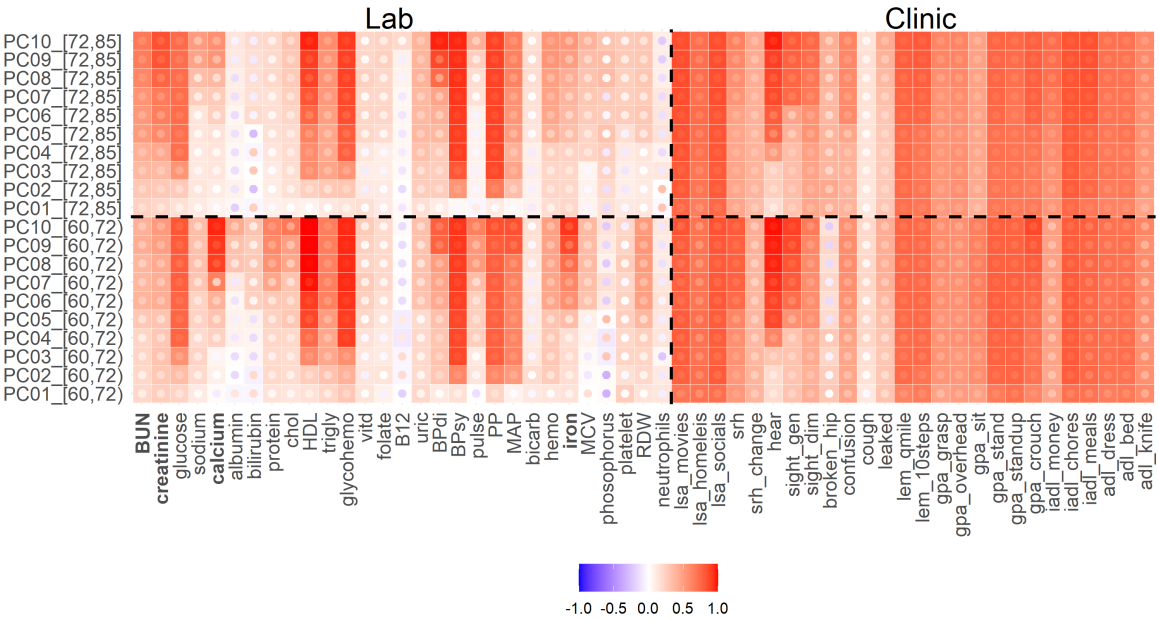

Figure B.33: GLM stepwise prediction of input variables, stratified by age. GLMs were trained using PCs to predict the input predictor variables. Age range is indicated in row name, top-coded at 85. The PC patterns are quite similar, indicating robustness with respect to age. Where they differ is of interest. Of note: BUN, creatinine, calcium, and iron (bolded). Youden index (higher is better). Inner circle fill colour is 95% CI limit closest to 0. GLMs were *not* conditioned on demographical variables (because we want to know everything that's in the PCs for comparison). Associated sections: Section B.3.5 and "Age stratification" in the main text.

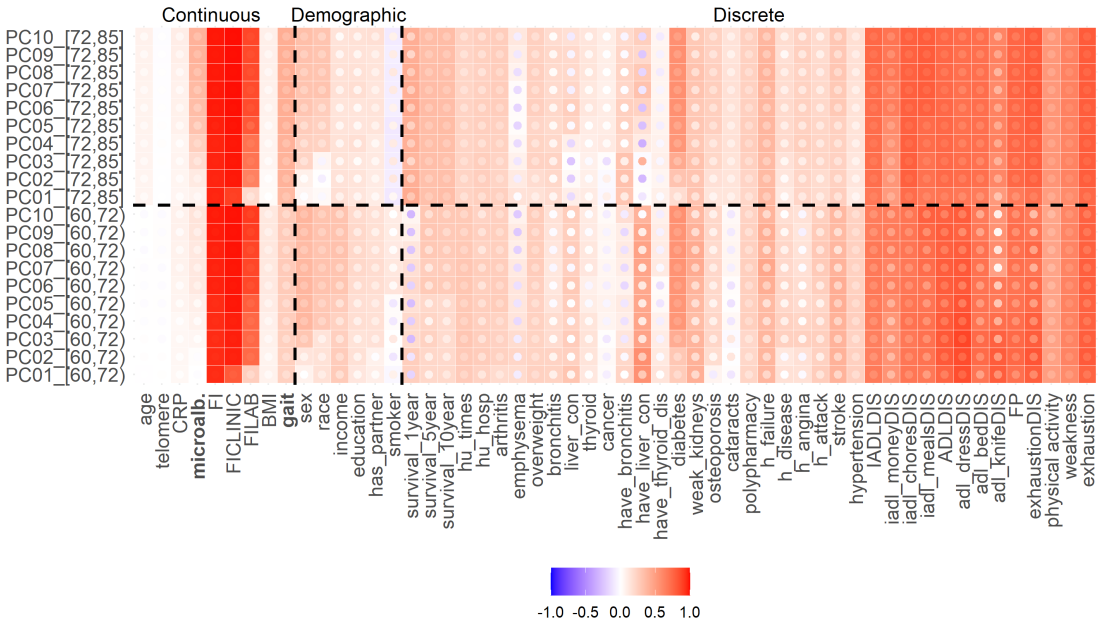

Figure B.34: GLM stepwise prediction, stratified by age. GLMs were used to predict outcomes and demographic covariates. Age range is indicated in row name, top-coded at 85. The PCs are quite similar, indicating robustness with respect to age. Where they differ is of interest. Of note: microalbuminuria and gait (bolded). Continuous score is R^2; demographic and discrete scores are both Youden (higher is better). Inner circle fill colour is 95% CI limit closest to 0. GLMs were *not* conditioned on demographical variables (because we want to know everything that's in the PCs for comparison). Associated sections: Section B.3.5 and "Age stratification" in the main text.

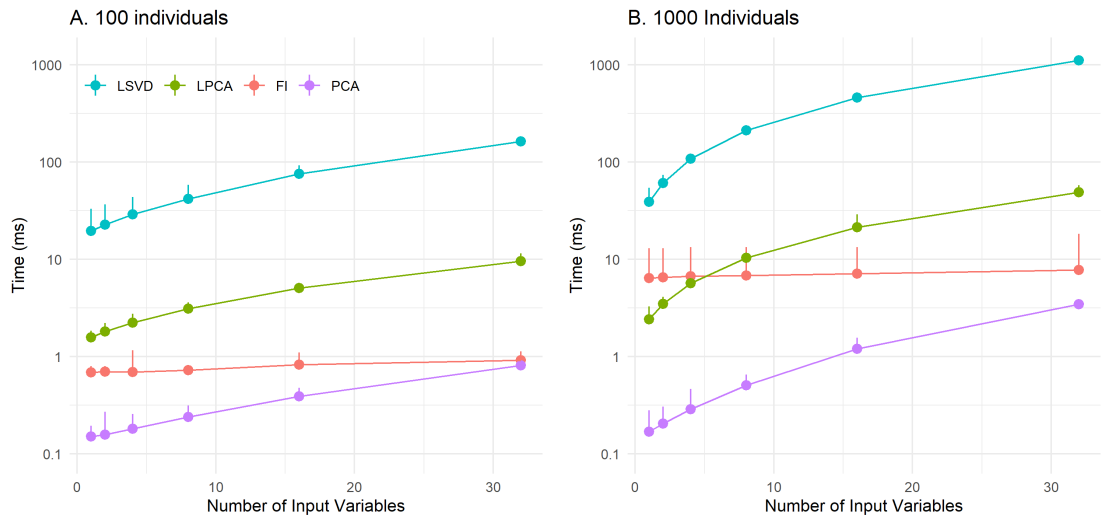

Figure B.35: Benchmarks for dimensionality reduction algorithms used. A. using a sample of 100 individuals, B. 1000. LSVD, LPCA and PCA all scaled similarly with increasing number of input variables, with PCA being about 10x faster than LPCA and LPCA being about 10x faster than LSVD. The FI, in comparison, scaled very well with increasing number of input variables, but had a high fixed computational cost (unlike the other algorithms, the FI code was not optimized). Increasing the number of individuals in the sample caused a sublinear increase in computation time (A vs B). See Section B.3.6 for details.

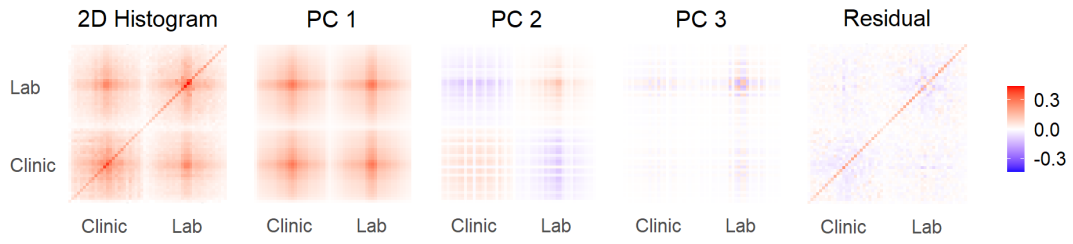

Figure B.36: Principal component analysis (PCA) of binary data is equivalent to eigen-decomposing the 2D joint deficit histogram, complete case data. The first column is the complete histogram, the remaining columns sum to the first column (Eq. A6). The first PC is clearly dominant and is dense, meaning it is nearly equal weights for each variable (akin to the FI). The eigen-decomposition naturally finds blocks of correlated variables. When it runs out of blocks it looks for strong diagonal terms. This causes PCA to naturally block out like-variables, e.g. lab vs clinical in PC2, similar to an expert choosing to create an FI out of variables from the same domain. Colour-scale has been transformed for visualization using $\text{sign}(x)|x|^{\gamma}$, $\gamma = 2/3$. Results are similar to imputed result, Figure 5.2, although the imputed histogram is clearly more saturated, reflecting the worse overall health of individuals with missing data (see Section B.1.2).

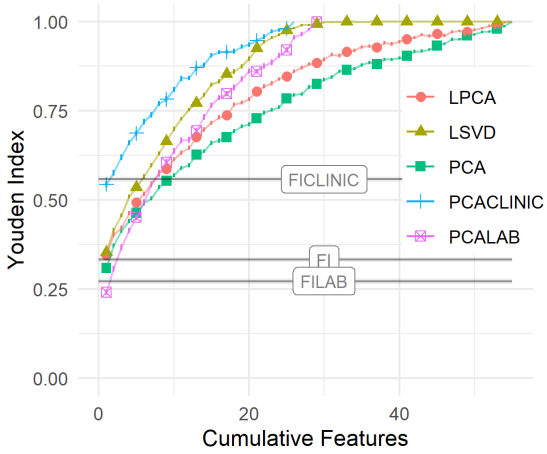

Figure B.37: Cumulative compression, complete case data. Tuning the size of the latent dimension bottleneck we inferred the maximum number of dimensions required to efficiently represent the input data. The reader should look for two things: (1) the number of components (dimensions) needed to achieve a relatively high score, and (2) the slope of the curve – when it flattens we can expect the features are noise, variable-specific or otherwise less important. Logistic SVD compresses the input most efficiently, saturating at around 30 features. Note the dramatic difference between lab and clinical compression both for PCA and the FI; the first PC of clinical data scores as well as 9 lab PCs. Results are similar to the imputed result, Figure 5.3.

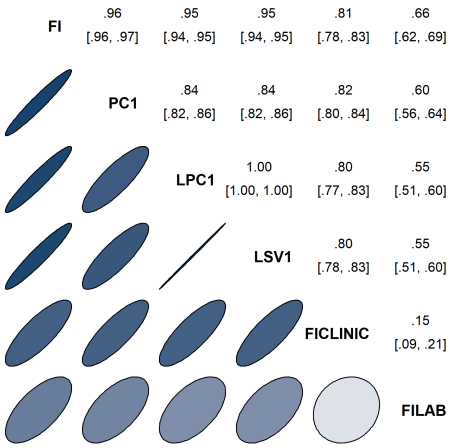

Figure B.38: Spearman correlation of primary features across algorithms, complete case data. The first latent dimension for either PC, LPC or LSVD correlated strongly with the FI and each other, and correlated more strongly with the FI CLINIC than FI LAB. This implies a strong mutual signal very close to the FI, especially the FI CLINIC. Upper triangle is correlation coefficient with 95% confidence interval. Ellipses indicate equivalent Gaussian contours [129]. Compared to the imputed result, Figure 5.4, we see somewhat smaller correlations between most features.

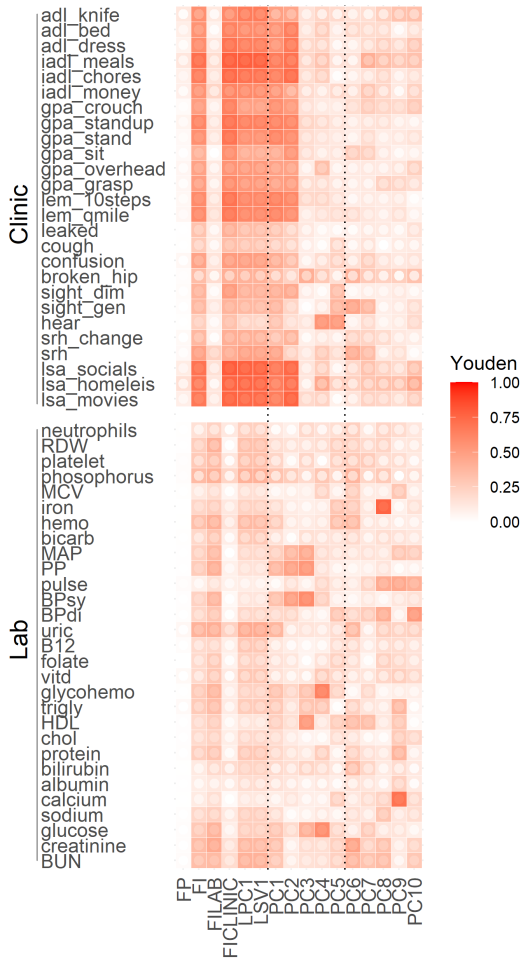

Figure B.39: Feature associations with individual input variables, i.e. what goes into each feature, complete case data. Association strength (fill colour) between features (x-axis) and adverse outcomes (y-axis); 0: no association, 1: perfect. Youden index. Inner circle fill colour is lower limit of 95% CI (white is non-significant). Compared to the imputed data, Figure 5.5, the higher PCs seem to have stronger signals/smaller confidence intervals, perhaps because the complete case data is more homogeneous.

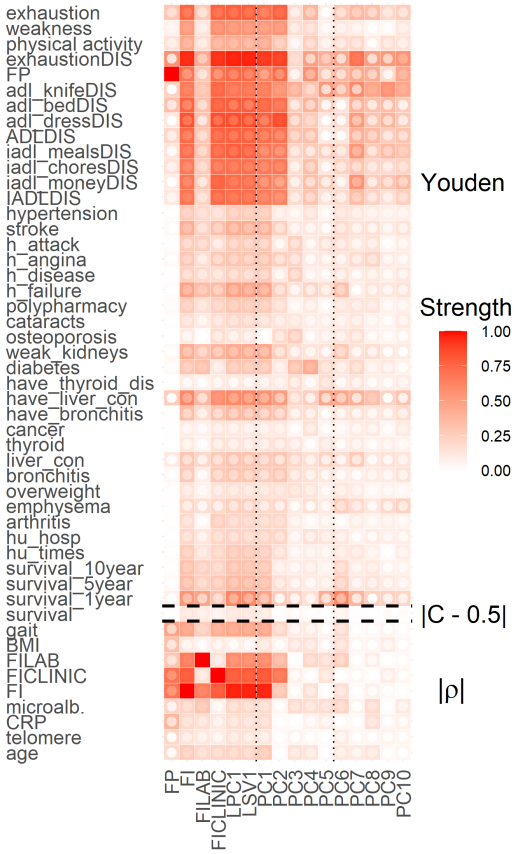

Figure B.40: Feature associations with individual outcomes, i.e. what we get out of each feature, complete case data. Association strength (fill colour) between features (x-axis) and adverse outcomes (y-axis); 0: no association, 1: perfect. Inner circle fill colour is lower limit of 95% CI (white is non-significant). Text on right denotes metric used. Compared to the imputed data, Figure 5.6, we see stronger signals in the higher PCs, perhaps because the complete case data is more homogeneous.

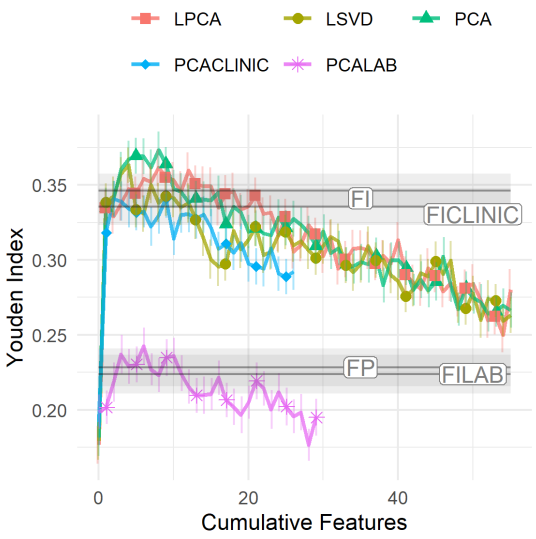

Figure B.41: Cumulative prediction plot for discrete outcomes (GLM), complete case data. 0th dimension is demographic information. Increasing the number of features initially improves prediction but quickly worsens, ostensibly due to overfitting. Youden index: higher is better. Compared to the imputed data, Figure 5.7, we see much stronger evidence of overfitting (decreasing score with increasing number of features). We suspect this is due to a lack of case data. For some outcomes, case data were rare enough that the scores could be unreliable (see Section B.4.3).

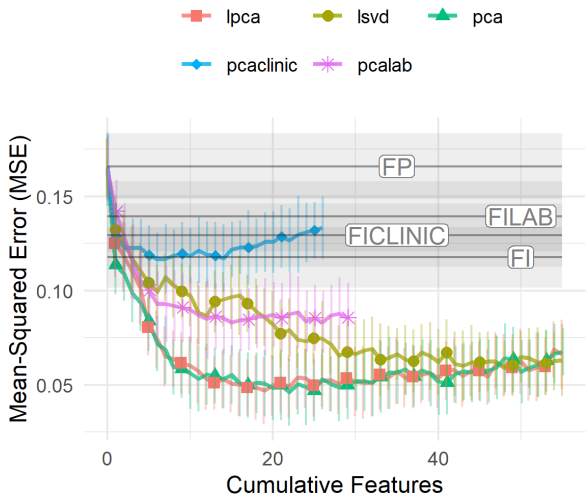

Figure B.42: Cumulative prediction plot for continuous outcomes (GLM), complete case data. 0th dimension is demographic information. Increasing the number of features improves prediction with a tendency to overfit as the number of PCs approaches the maximum. LSVD performs notably worse than PCA and LPCA. MSE is on standardized scale, therefore $R^2 = 1 - MSE$. MSE: lower is better. Compared to the imputed data, Figure 5.8, we see some evidence of overfitting (increasing error with increasing number of features).

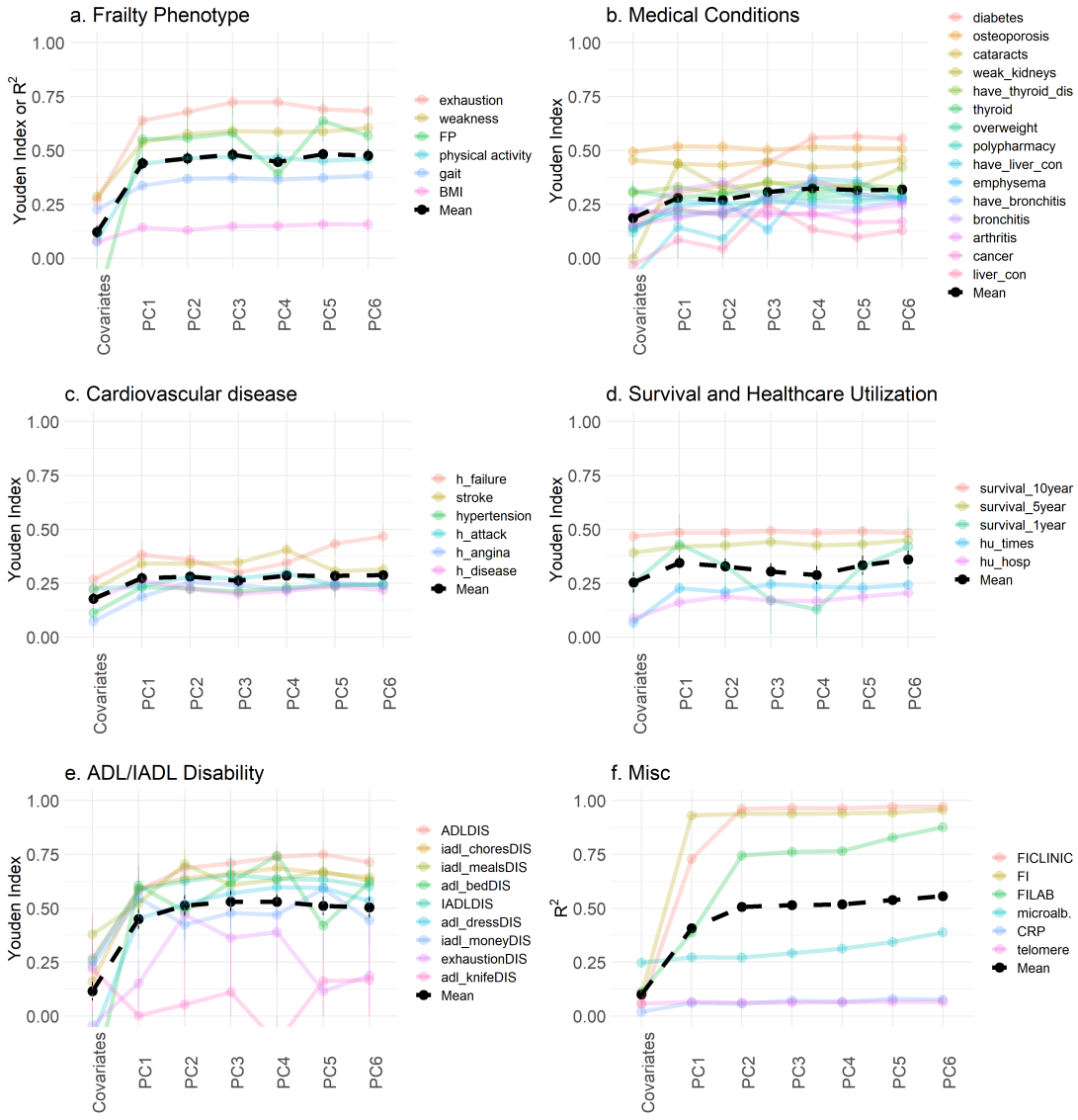

Figure B.43: Improvement in predictive power as more PCs are included, grouped by outcome type (GLM), complete case data. Coloured lines indicate specific outcomes, black line indicates the mean for each group. For most outcomes the performance stops improving after a few PCs, hence why we've truncated at PC6. The exceptions are explored in Figure B.45. Note: legend is sorted from best (top) to worse (bottom) performance of the PC6 model. See Figure B.44 for the complete plots without truncation. Compared to the imputed data, Figure 5.9, we see much more volatile fits and lower overall accuracies, particularly for ADL/IADL disability. Cases were rare for ADL/IADL disability, which could make the Youden index estimates unreliable (see Section B.4.3).

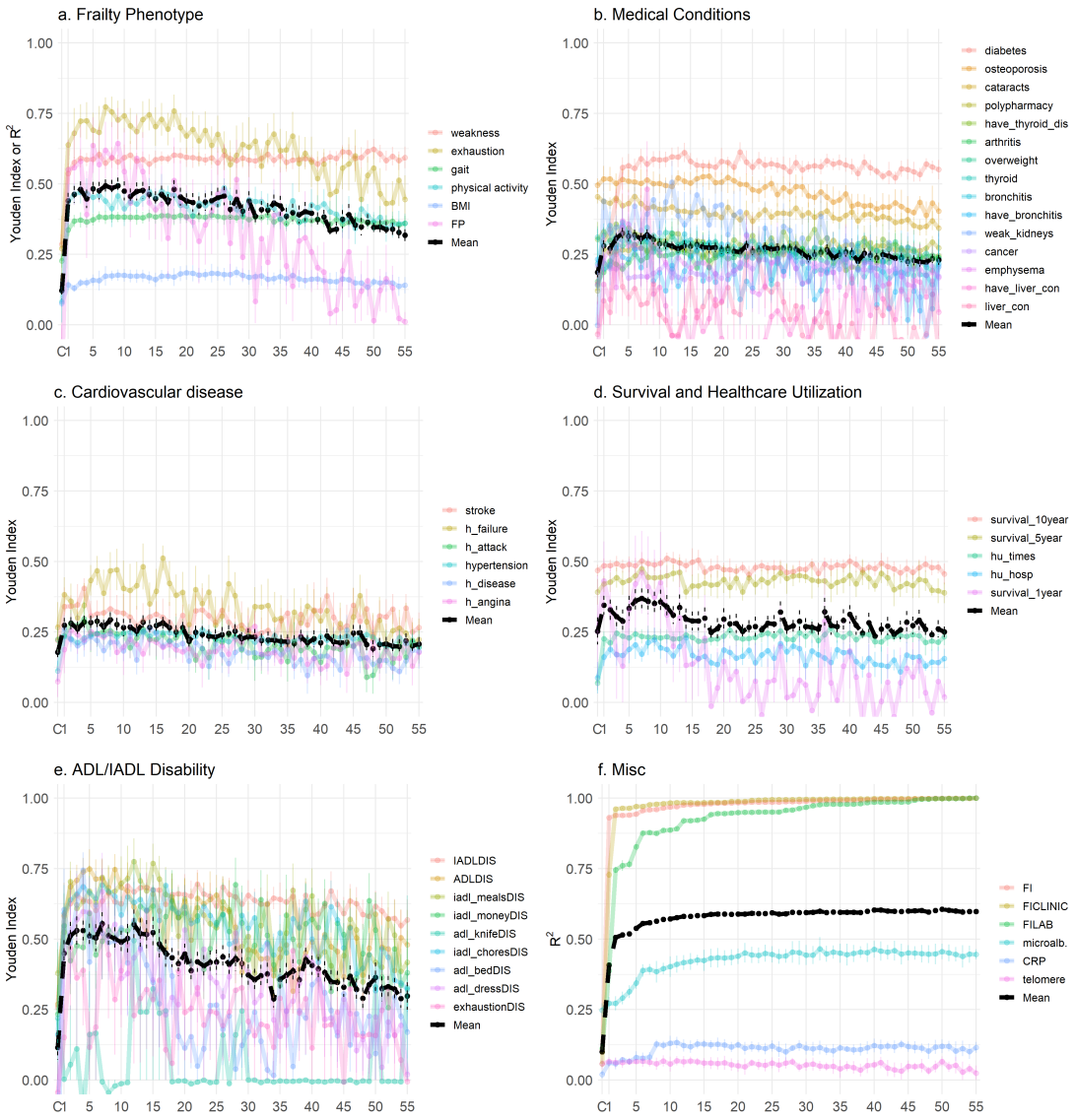

Figure B.44: Improvement in predictive power as more PCs are included, grouped by outcome type (GLM), complete case data, without truncation (all PCs present). X-axis labels indicate the cumulative number of PCs included, "C" means demographical covariates only. Coloured lines indicate specific outcomes, black line indicates the mean for each group. Note: legend is sorted from best (top) to worse (bottom) performance of the PC55 model. Cases were rare for ADL/IADL disability, which could make the Youden index estimates unreliable (see Section B.4.3). This is the extended version of Figure B.43.

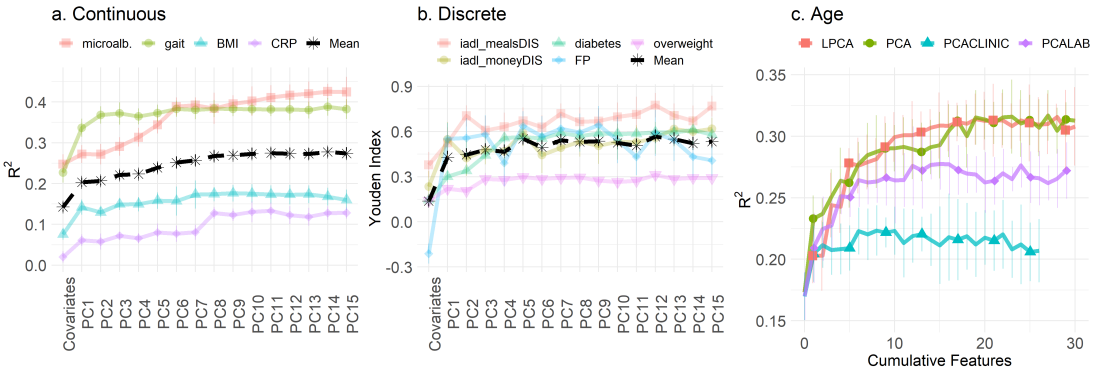

Figure B.45: Improvement in predictive power as more PCs are included, high-dimensional outcomes (GLM), complete case data. High-dimensional outcomes were identified by the imputed analysis (compare to Figure 5.10). We tend to see continual improvement for the discrete and continuous outcomes, excluding the FP (up to ~ 10). Age appeared to be the highest dimensional. Compared to the imputed data, Figure 5.10, we see more volatile curves, perhaps because of limited case data (see Section B.4.3); note the different coloured labels (labels are sorted by performance).

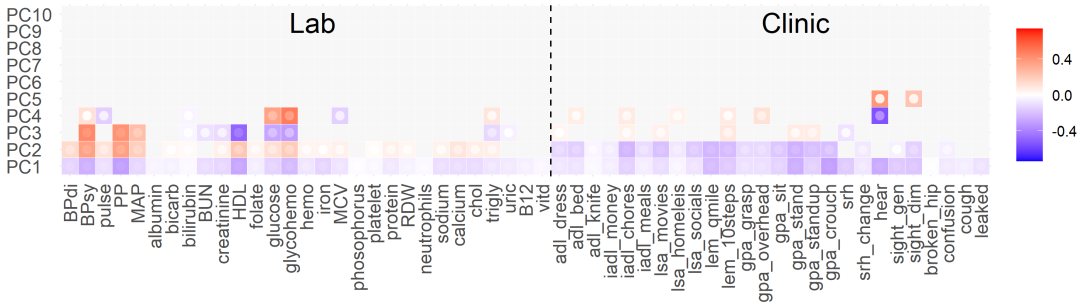

Figure B.46: PCA robustness, complete case data. Robustness of the PCA rotation was assessed by randomly sampling which individuals to include (i.e. bootstrapping, $N = 2000$). Left side are lab variables, right are clinical. Inner circle fill colour is 95% CI limit closest to 0. Grayed out tiles were non-significant. The first three PCs were quantitatively robust. We see the robustness drops with increasing PC number. The global sign for each PC were mutually aligned across replicates using the Pearson correlation between individual feature scores. Compared to the imputed data, Figure 5.11, we see that the PCs were a little less robust in the complete case data (lower significance), but otherwise similar.

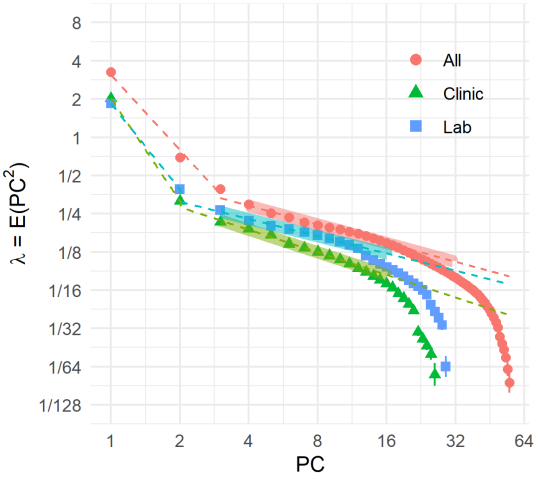

Figure B.47: PCA second moments (eigenvalues) with bootstrapped standard errors (N=2000), complete case data. Log-log scales. Note the bilinear structure. Banded region is optimal performance region (±1 error bar from best). Compared to the imputed data, Figure 5.12, the points have curved further away from the banded regions.

Appendix C

Supplemental Information for Network dynamical stability analysis reveals key mallostatic natural variables that erode homeostasis and drive age-related decline of health

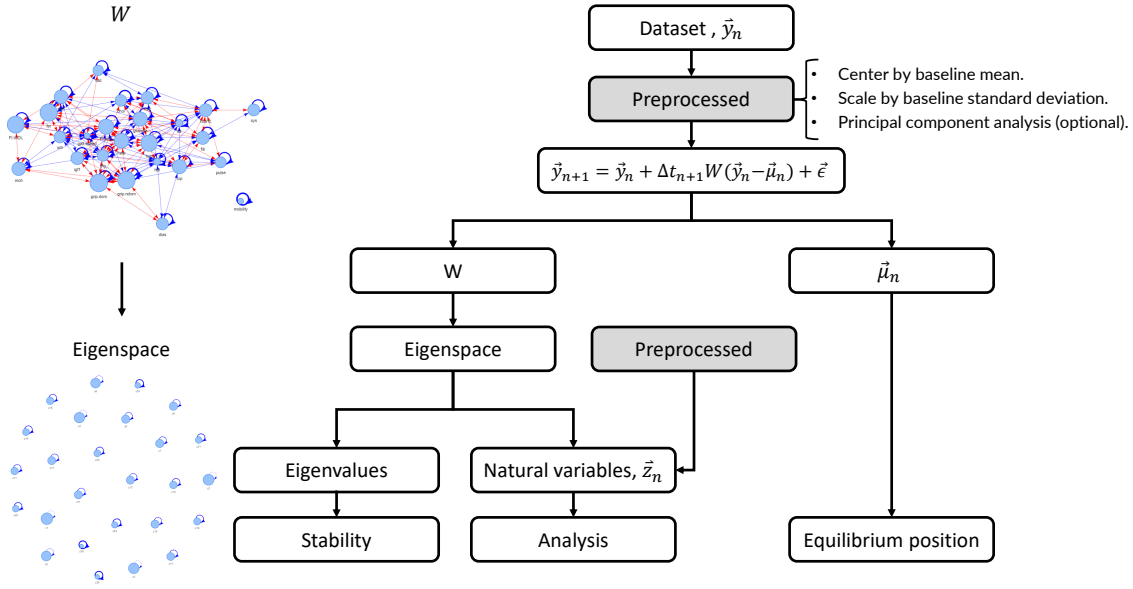

Figure C.1: Study pipeline. We analysed four datasets using our proposed model. We model the dynamics of biomarkers, $\vec{y}_n$, over time using equation (C.4). Our model extracts an interaction network, $\boldsymbol{W}$, and equilibrium positions $\vec{\mu}_n$, where the latter are allowed to depend on covariates (e.g. age and sex). The estimated network, $\boldsymbol{W}$, captures arbitrary linear interactions between biomarkers which can be removed by working with the natural variables, $\vec{z}_n$. Natural variables are defined by a linear mapping into the eigenspace of $\boldsymbol{W}$. The natural variables allowed us to analyse stability. We were also able to infer changes to the mean and variance of the observed variables based on changes in the natural variables.

C.1 Introduction

We modelled generic health biomarker data as a mean-reverting stochastic process. Our study pipeline is summarized in Figure C.1. Our model describes generic dynamics near an equilibrium solution (Section C.8.4). In this supplemental we provide additional information to support and validate both our methods and our conclusions. We provide a complete description of the data in Section C.2 and how we preprocessed it in Section C.3. We performed a number of consistency checks on missing data which were imputed according to Section C.4. We consider variations of our model in Section C.5 which demonstrates that our final model best describes the data. We provide the mathematics necessary to estimate model parameters, along with an iterative estimation algorithm in Section C.6. We then validate our algorithm using synthetic data in Section C.7. Additional mathematics useful for understanding our model and its connection to the literature are described in Section C.8. Finally, we include additional results in Section C.9 which support our conclusions.

Table C.1: Dataset Summary — Biomarkers

Dataset	Species	Primary Outcome	Variable	Description
SLAM (C57BL/6)	Mouse, C57BL/6	Death	bw	Body weight
SLAM (C57BL/6)	Mouse, C57BL/6	Death	fat	Total fat mass
SLAM (C57BL/6)	Mouse, C57BL/6	Death	lean	Total lean mass
SLAM (C57BL/6)	Mouse, C57BL/6	Death	fluid	Total fluid mass
SLAM (C57BL/6)	Mouse, C57BL/6	Death	glucose	Blood glucose (fasting)
SLAM (C57BL/6)	Mouse, C57BL/6	Death	lactate	
SLAM (Het3)	Mouse, Het 3	Death	bw	Body weight
SLAM (Het3)	Mouse, Het 3	Death	fat	Total fat mass
SLAM (Het3)	Mouse, Het 3	Death	lean	Total lean mass
SLAM (Het3)	Mouse, Het 3	Death	fluid	Total fluid mass
SLAM (Het3)	Mouse, Het 3	Death	glucose	Blood glucose (fasting)
SLAM (Het3)	Mouse, Het 3	Death	lactate	
Paquid	Human	Dementia	MMSE	Mini-mental state exam†
Paquid	Human	Dementia	BVRT	Benton visual retention test
Paquid	Human	Dementia	IST	Isaacs set test
Paquid	Human	Dementia	CESD$^{\diamond}$	depression scale‡
ELSA	Human	Death	vitd	Vitamin d
ELSA	Human	Death	srh	Self-reported health (higher: worse)
ELSA	Human	Death	eye	Self-reported (corrected) eyesight (higher: worse)
ELSA	Human	Death	hear	Self-reported (corrected) hearing (higher: worse)
ELSA	Human	Death	FI.ADL	Activities of Daily Living (ADL[46]) FI*
ELSA	Human	Death	FI.IADL	Instrumental ADL FI*
ELSA	Human	Death	gait.speed	Time to walk 8 feet (2.44 m)
ELSA	Human	Death	grip.ndom	Grip strength, non-dominant hand
ELSA	Human	Death	grip.dom	Grip strength, dominant hand
ELSA	Human	Death	crp	C-reactive protein*
ELSA	Human	Death	hba1c	Glycohaemoglobin
ELSA	Human	Death	glucose	Glucose
ELSA	Human	Death	hgb	Haemoglobin
ELSA	Human	Death	mch	Mean corpuscular haemoglobin
ELSA	Human	Death	fer	Ferritin*
ELSA	Human	Death	chol	Cholesterol
ELSA	Human	Death	ldl	Low density lipoprotein
ELSA	Human	Death	hdl	High density lipoprotein
ELSA	Human	Death	trig	Triglycerides*
ELSA	Human	Death	sys	Systolic blood pressure
ELSA	Human	Death	dias	Diastolic blood pressure
ELSA	Human	Death	pulse	Pulse
ELSA	Human	Death	fib	Fibrogen
ELSA	Human	Death	igf1	Insulin-like growth factor-1
ELSA	Human	Death	wbc	White blood cell count*

* FI: frailty index; defined as average number of health deficits[174].

† Transformed as $-\sqrt{\max(\text{MMSE}) - \text{MMSE}}$.

‡ Square root transformed for normality.

$^{\diamond}$ Center for Epidemiological Studies depression scale

* Log transformed for normality.

Table C.2: Covariate Summary

Dataset	Covariate	Description
SLAM (C57BL/6)	Age	Chronological age in weeks
SLAM (C57BL/6)	Sex	0: male, 1: female
SLAM (Het3)	Age	Chronological age in weeks
SLAM (Het3)	Sex	0: male, 1: female
Paquid	Age	Chronological age in years
Paquid	Sex	0: male, 1: female
Paquid	Education	0: did not complete primary school, 1: did
ELSA	Age	Chronological age in years
ELSA	Sex	0: male, 1: female

C.2 Materials

We analysed 4 datasets derived from 3 longitudinal studies. The datasets and predictors ("biomarkers") used are summarized in Table C.1. All predictor variables were continuous or, in the case of Paquid, ordinal with many scales (> 15).

We included covariates to reduce confounding effects and to look for allostasis, which depends on age. We included age (continuous) and binary variables. The covariates used are summarized in Table C.2.

C.3 Preprocessing

The data we analyzed were longitudinal with regular sampling rates. For this reason, data were conveniently stored as 3-dimensional arrays (individuals, biomarkers, time points), meaning that each individual had the same number of variables and measurements (although many of them missing). This means that some timepoints for some individuals had to be 'invented' (instantiated as NA) based on the sampling rate of the study in question.

The Study of Longitudinal Aging in Mice (SLAM) datasets were both processed using the same criteria. The initial data were downloaded and processed using the analysis script of another publication [135]. We then applied additional preprocessing as follows. Biomarkers were visually investigated for normality and deemed adequate. The sex-specific mean and standard deviation of the first measurement of each biomarker was used to center and scale all timepoints. Mice with less than 2 timepoints were excluded from analysis (about 1% of mice). Any observations made past the reported death age of each mouse were excluded from analysis (about 1% of observations). We excluded the first two timepoints from analysis because after encoding we found that approximately half of individuals had not yet had a body composition measurement (imputed values looked unrealistic). Missing timepoints — which occurred due to staggered data collection — were instantiated using a piecewise linear model between known observations. Data from SLAM and the other datasets were stored in 3-dimensional arrays, with missing values imputed according to Section C.4. The final arrays were size: (608, 6, 22) for SLAM C57/BL6, and (611, 6, 29) for SLAM Het3 (individuals, biomarkers, time points).

The Paquid dataset we used is available as part of a software package [150]. Biomarkers were visually investigated for normality. To improve normality we transformed CESD by the square-root and MMSE by $-\sqrt{30 - \text{MMSE}}$ where 30 is the maximum allowed score for the MMSE. All biomarkers were centered and scaled by their respective mean and standard deviation from the first timepoint. Missing timepoints were instantiated using a piecewise linear model between known observations. The final array was size (500, 4, 9); (individuals, biomarkers, time points).

The English Longitudinal Study of Ageing (ELSA) dataset is available from the UK data service (https://ukdataservice.ac.uk/). We analysed all of the waves which included lab work: 2, 4, 6 and 8 (i.e. the "nurse" waves). We included only individuals present in wave 2, thus excluding later recruits. Biomarkers were visually investigated for normality. We found that the log transformation improved normality for C-reactive protein, ferritin, triglycerides, and white blood-cell count. All biomarkers were centered and scaled by their respective mean and standard deviation from the first timepoint. Skipped timepoints were instantiated using linear interpolation of the available timepoints. Censored (or died) timepoints were instantiated using the mean followup time (which was uniform due to the study design). The final array was size (9330, 23, 4); (individuals, biomarkers, time points).

C.4 Missing data

We were presented with two forms of missing data for an individual at a particular timepoint. The entire timepoint could be missing or some subset of values could be missing. In either case the missingness could be informative; for example an individual may have temporarily left the study due to poor health and their biomarkers could have had abnormal values reflecting their poor health. In this way the population may appear abnormally healthy as it ages. Under such circumstances, failure to impute can lead to biased study conclusions [180], such as parameter estimates (Section C.4.1).

We considered three imputation approaches and selected the approach which gave the most reasonable values. First ("simplest"), we imputed a single value using either the individual's mean biomarker value, carry forward the previous value (and then carry back any skipped values), the conditional population mean (assuming multivariate Gaussian statistics), or the individual mean followed by the conditional mean for individuals whom did not have that variable reported. Second ("model mean"), we considered an iterative approach after applying one of the first methods wherein values were imputed according to the model mean (i.e. model prediction), equation (C.1) and equation (C.3). Third ("MICE"), we considered multivariate imputation using chained equations (MICE) [196]. MICE is a multiple imputation technique that uses a Gibbs' sampler along with a predictive model. We considered MICE using both classification and regression trees (CART) and 2-level modelling (normal for continuous variables and logistic for binary).

When imputing the model mean, at each iteration we estimated the model parameters then imputed the conditional mean for each missing value (Algorithm C.1). Let $\vec{y}$ denote the biomarker, $\vec{u}$ denote all unobserved y and $\vec{o}$ denote all observed y. The statistics are Gaussian (equation (C.4)), so we can use the factorization theorem [140] to compute the expectation value.

If $\vec{y}_{n+1}$ is known but a set of $\vec{y}_n$ are unknown then

$$\begin{aligned}
E(\vec{u}_n|\vec{o}_n,\vec{y}_{n+1}) &= \langle\vec{y}_{un}\rangle + \boldsymbol{\Sigma}_{uo}\boldsymbol{\Sigma}_{oo}^{-1}(\vec{o}_n - \langle\vec{y}_{on}\rangle) \quad \text{where,}\\
\langle\vec{y}_{un}\rangle &= (\boldsymbol{I} + \Delta t_{n+1}\boldsymbol{W})_{u\cdot}^{-1}\big(\vec{y}_{n+1} + \boldsymbol{W}\Delta t_{n+1}\vec{\mu}_n\big),\\
\langle\vec{y}_{on}\rangle &= (\boldsymbol{I} + \Delta t_{n+1}\boldsymbol{W})_{o\cdot}^{-1}\big(\vec{y}_{n+1} + \boldsymbol{W}\Delta t_{n+1}\vec{\mu}_n\big),\\
\Sigma_{uo} &= ((\boldsymbol{I} + \Delta t_{n+1}\boldsymbol{W})^T\boldsymbol{Q}(\boldsymbol{I} + \Delta t_{n+1}\boldsymbol{W}))_{uo}^{-1},\\
\Sigma_{oo}^{-1} &= \Big(((\boldsymbol{I} + \Delta t_{n+1}\boldsymbol{W})^T\boldsymbol{Q}(\boldsymbol{I} + \Delta t_{n+1}\boldsymbol{W}))_{oo}^{-1}\Big)^{-1},
\end{aligned} \tag{C.1}$$

where $E(x|y)$ denotes expectation value of x conditional on y, $\boldsymbol{Q}$ is the precision matrix defined below, and $\boldsymbol{I}$ is the identity matrix. Note that

$$\boldsymbol{\Sigma} \equiv \boldsymbol{Q}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{oo} & \boldsymbol{\Sigma}_{ou} \\ \boldsymbol{\Sigma}_{uo} & \boldsymbol{\Sigma}_{uu} \end{bmatrix} \tag{C.2}$$

is the block-decomposition of the noise covariance rearranged for observed (o) and unobserved (u) variables. If instead $\vec{y}_n$ is known but a set of $\vec{y}_{n+1}$ are unknown then

$$E(\vec{u}_{n+1}|\vec{o}_{n+1},\vec{y}_n) = \vec{y}_{un} + W_{u\cdot}\Delta t_{n+1}(\vec{y}_n - \vec{\mu}_n) + \mathbf{\Sigma}_{uo}\mathbf{\Sigma}_{oo}^{-1}(\vec{o}_{n+1} - \vec{y}_{on} - W_{o\cdot}\Delta t_{n+1}(\vec{y}_n - \vec{\mu}_n)). \tag{C.3}$$

We used the simplest approach as an initial imputation (e.g. carry forward/back). We then imputed $\vec{y}_1$ first using equation (C.1) then each subsequent timepoint using equation (C.3).

We compared the imputation quality and found that the model mean was both straightforward and effective, and therefore elected to use it for both SLAM datasets and Paquid. We initialized imputation with carry forward/back: that is, forward carrying previous values until the last timepoint was reached then backwards carrying to fill any values still missing. In rare cases a few data points were missing after imputation, these were simply ignored (we used all available case data). The ELSA dataset was sensitive to the model mean — perhaps due to the limited number of data points — hence we used a single imputation which combined first imputing the individual-specific variable mean followed by the conditional mean, assuming multivariate Gaussian statistics at each timepoint. Note that since we elected to use bootstrapping, we imputed each bootstrap replicate and then averaged to get an estimate for each missing value along with a standard error.

The final imputation was assessed for quality, Figure C.2 — and looked reasonable. When inspecting imputation quality we are looking for the same age-dependent pattern for both the imputed and observed values, both in terms of mean and dispersion. Informative censorship is possible, so for variables with survival effects we can expect that missing values should be at higher risk because they include individuals whom were censored due to poor health (or death). Risk can be inferred by the direction of drift with respect to age: data points which look 'older' are likely higher risk. Hence imputed values may look a little 'older' than observed values.

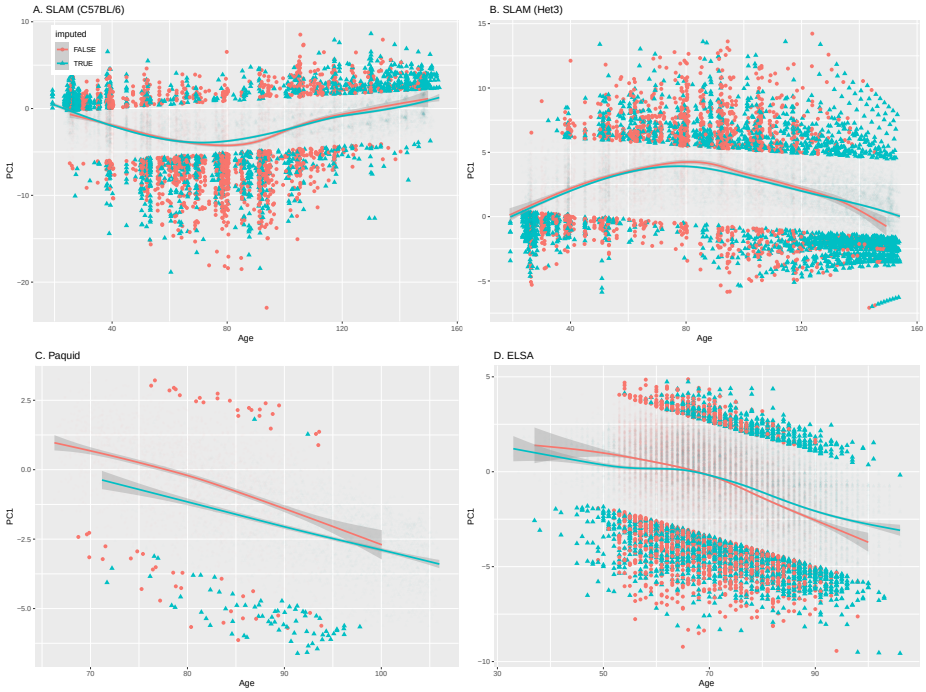

Figure C.2: Final imputation quality check, visualized using principal component 1. **A.** C57BL/6 mice (SLAM). **B.** Het3 mice (SLAM). **C.** Paquid (human, dementia). **D.** ELSA (human). Imputed values appear to be reasonable for each dataset. Principal component analysis (PCA) was applied to each dataset in the entirety, flatted across timepoints. Good quality imputation (blue triangles) should show the same trend and dispersion as the observed data (red points). Censored individuals likely have worse health, so imputed values may look a little 'older' than the observed. Age-dependence is indicated by the solid lines with confidence intervals (cubic spline; `geom_smooth` with defaults [206]). Outlying points are highlighted ($l \pm 3$ where l is the ordinary linear regression model). Data points were labelled as imputed (blue triangles) if the preponderance of the rotation weights were missing: $\sum_{i=missing} |U_{i1}|/(\sum_j |U_{j1}|) > 0.5$; where $\boldsymbol{U}$ is the PCA rotation matrix.

C.4.1 Informative censorship

Is it better to impute dropped individuals (dead, censored) or not? Dropped individuals may have abnormal biomarker values leading to their exclusion, i.e. informative censorship. There is "substantial" evidence that dropped individuals in longitudinal studies have worse health [71] and their health biomarkers will reflect this, leading to a potential survivorship bias. We used simulated data to test for potential bias and observed that — if done well — imputing values for dropped individuals can reduce this bias.

We simulated data from our model equation (C.4) using randomly generated parameters then imposed informative censorship. We simulated 100 times with 100 individuals in each simulation. Each simulation included 2 biomarkers. Parameters and biomarkers were draw from normal random variables. The diagonal of $\boldsymbol{W}$ was mean $-1/4$, the off-diagonals were mean 0 and the overall standard deviation was 0.1. The mean μ_0 was 0 and standard deviation was 0.1. No covariates were simulated. The noise was diagonal, $\boldsymbol{\Sigma} = 0.5\boldsymbol{I}$ (also used for instantiating the population). We censored using Gompertz statistics with proportional hazards for biomarker values [13] (shape: $\alpha = 0.1$, scale: $\lambda = 10^{-5}$). The proportional hazards coefficients were randomly sampled from a normal distribution with mean 1 and standard deviation 0.1, this ensured that large values of the biomarkers were preferentially censored.

The results of the simulation are shown in Figure C.3 for various imputation strategies, with the horizontal dashed line indicating unbiased results. We observed that a significant bias existed in both the diagonal and off-diagonal elements of $\boldsymbol{W}$, which were systematically over-estimated if the data were not imputed. Conversely, if we used only the simple carry forward/back imputation, an even worse bias ensued in the opposing direction. If we used the model mean imputation, however, we reduced the bias in $\boldsymbol{W}$ to nearly 0 without significantly increasing bias in the other parameters. For this reason, we imputed all dropped individuals: censored and dead.

C.5 Model selection

Our goal with model selection was two fold: (i) to find the optimal model(s) that best fit the data, and (ii) to test which model parameters were essential for fitting the data. This allows us to infer the existence of which model parameters are robustly supported by the data. Fit quality was measured using the root-mean squared error (RMSE) and mean absolute error (MAE). We used the 632 estimator for errors, which is a linear combination of 63.2% test error and 36.8% training error [74]. Test error was estimated via out-of-sample bootstrap replication, with 100 resamples. Each bootstrap selected a new dataset of the same size as the original by resampling individuals with replacement. The out-of-sample individuals are those whom were not selected. Estimation algorithms are reported in Section C.6.

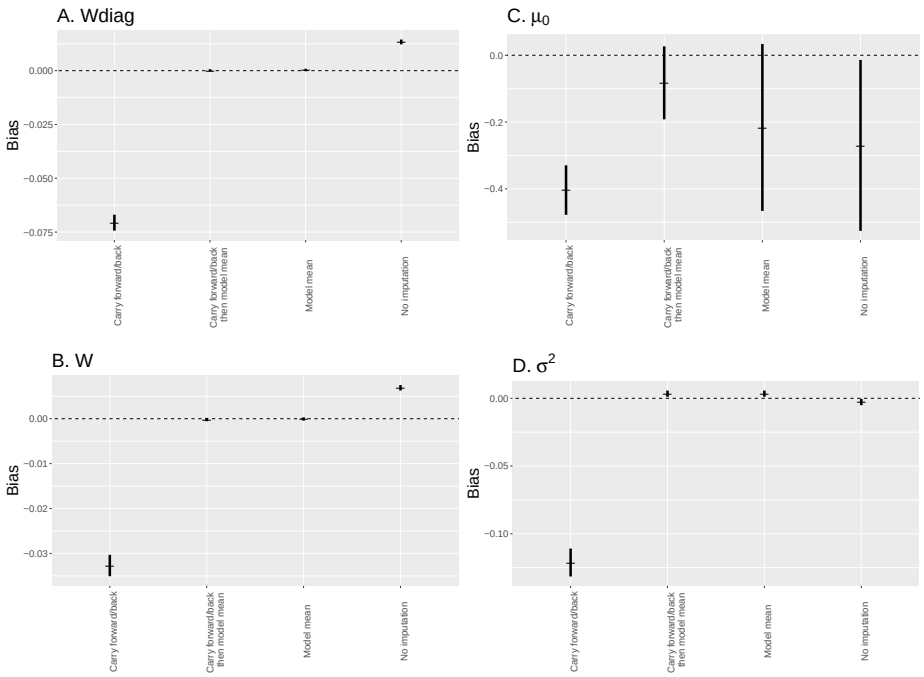

Figure C.3: Imputation of dropped individuals can reduce bias. We simulated informative censorship and here compare estimates from different missing data handling strategies. Observe that both the diagonal elements of $\boldsymbol{W}$ (A.) and all elements of $\boldsymbol{W}$ (B.) were biased high when data were not imputed. However, if we imputed using the model mean, the bias was greatly reduced. For μ_0 (C.) we also reduced the bias with the combined imputation strategy, which was the strategy employed on the real data. Imputation did introduce a small bias in the noise estimate (D.). The bias was largest if we used only the carry forward/back method.

We compare model performance in Figure C.4; we explain the model labels here. The general form of our model is ('full')

$$\begin{aligned}\vec{y}_{n+1} &= \vec{y}_n + \boldsymbol{W}\Delta t_{n+1}(\vec{y}_n - \vec{\mu}_n) + \vec{\epsilon},\\ \vec{\epsilon} &\sim \mathcal{N}(0, \Sigma|\Delta t|)\\ \vec{\mu}_n &\equiv \vec{\mu}_0 + \boldsymbol{\Lambda}\vec{x}_n + \vec{\mu}_{age}t_n,\end{aligned} \tag{C.4}$$

where t is the age. The error can be expressed in terms of the precision matrix,

$$\boldsymbol{Q} \equiv \boldsymbol{\Sigma}^{-1}. \tag{C.5}$$

We considered both using $\boldsymbol{Q} = \boldsymbol{I}$, the identity matrix (default), and estimating $\boldsymbol{Q}$ from the data using the log-likelihood ('Q'). Transforming into natural variables — wherein $\boldsymbol{W}$ is diagonal — we have

$$z_{jn+1} = z_{jn} + \lambda_j \Delta t_{n+1}(z_{jn} - \tilde{\mu}_{jn}) + \tilde{\epsilon}, \tag{C.6}$$

where $\vec{z}_n \equiv \boldsymbol{P}^{-1}\vec{y}_n$, $\lambda_i \equiv P_{i\cdot}^{-1}\boldsymbol{W}P_{\cdot i}$, $\tilde{\vec{\mu}}_n \equiv \boldsymbol{P}^{-1}\mu_n$ and $\tilde{\vec{\epsilon}} \equiv \boldsymbol{P}^{-1}\vec{\epsilon}$.

We considered ('pca') the possibility that principal component analysis (PCA) could be used as a preprocessing step to decouple the biomarkers such that we could fit equation (C.6), assuming independent noise between the z_j. Equation (C.49) states that in the steady-state the principal components are equivalent to the eigenvectors of $\boldsymbol{W}$, this self-consistency motivates using PCA. Prior work has also suggested that principal components don't change much during the aging process [146].

We considered that μ_n may be time-dependent and may also depend on other covariates ('covs'), which is discussed in Section C.2.

We considered simpler, nested forms of equation (C.4). Recall that the data were standard-deviation-scaled and mean-centered by the baseline value, which justifies some of the simplifications. Simplified forms allowed us to test whether $\boldsymbol{W}$ and $\vec{\mu}$ were necessary to fit the data. Removing these parameters leads to special cases of the model. The simplest model for the data is to simply carry forward the previous value. If recovery is small, $W\Delta t \to 0$ then we have ('carry'),

$$\vec{y}_{n+1} = \vec{y}_n + \tilde{\epsilon} \tag{C.7}$$

which corresponds to carrying the previous value forward ($\langle\vec{y}_{n+1}\rangle = \langle\vec{y}_n\rangle$). This model does not require any parameters to make predictions; it was used for the initial imputation of the Paquid and SLAM datasets (Section C.4).

If recovery is complete between each timepoint then $\boldsymbol{W}\Delta t \to -\boldsymbol{I}$ and we instead have the second simplest model ('fast'),

$$\vec{y}_{n+1} = \vec{\mu}_n + \vec{\epsilon} \tag{C.8}$$

which corresponds to biomarkers being randomly distributed about some mean value which depends on covariates ($\langle\vec{y}_{n+1}\rangle = \langle\vec{\mu}_n\rangle$). Alternatively, we could have $\vec{\mu}_n \equiv 0$ in which case we have ('noallo'),

$$\vec{y}_{n+1} = \vec{y}_n + W\Delta t_{n+1}\vec{y}_n + \vec{\epsilon} \tag{C.9}$$

which we refer to as the no allostasis model (it also implicitly sets homeostatic equilibrium to 0). In 1-dimension, equation (C.9) is simple exponential growth/decay in the mean (for small Δt).

While nonlinear behaviour can be captured by our model (Section C.8), we also directly investigated nonlinear behaviour by including a quadratic term ('quad'),

$$z_{jn+1} = z_{jn} + \lambda_j \Delta t_{n+1}(z_{jn} - \tilde{\mu}_{jn}) + \gamma \Delta t_{n+1}^2 z_{jn}^2 + \tilde{\epsilon}. \tag{C.10}$$

We only considered a quadratic term for the diagonal model, equation (C.6) with PCA preprocessing.

We compare model performance in Figure C.4. We found that the fast model, equation (C.8), fit very poorly, having error so large that it did not fit in the plot region. Within the plot region, the carry-forward model performed the worst, equation (C.7) ('carry'). Note the implication: biomarker recovery towards equilibrium is much closer to none ('carry') than complete ('fast'). Next worse was excluding μ, equation (C.9) ('noallo'). The remaining models typically performed similarly-well. The SLAM datasets both saw a noteworthy improvement in fit when age was included as a covariate (in μ_n). We observed no improvement with inclusion of a quadratic term, equation (C.10) ('quad'). Finally, and importantly, we found that a diagonal fit on principal components (PCs) yielded equivalent performance to the full model. This permitted a greatly simplified methodology since we were able estimate using weighted linear regression (Section C.6.1).

C.6 Estimation

We provide useful results for fitting equation (C.4) and its simplified form, equation (C.6). The latter can be solved using weighted linear regression.

C.6.1 (Weighted) Linear Regression

If the noise term is diagonal then the equations decouple and we have a set of linear equations which can be independently solved using linear regression. In the present study we used PCA (principal component analysis) as a preprocessing step prior to fitting a diagonal model. That is, we assumed the PCs do not interact with each other. We can rewrite equation (C.6) as

$$\begin{aligned} z_{ijn+1} - z_{ijn} &= \lambda_j(\Delta t_{in+1} z_{ijn}) + \vec{\beta}_j^{\,T}(\Delta t_{in+1}\vec{x}_{in}) + \epsilon_{ij}, \qquad \text{where} \\ \beta_{j0} &= \lambda_j \mu_0, \\ \beta_{jage} &= \lambda_j \mu_{jage}, \\ \beta_{jk} &= \lambda_j \Lambda_{k\cdot}, \qquad \text{and} \\ \epsilon_{ij} &\sim \mathcal{N}(0, \sigma_j^2 |\Delta t_{in+1}|). \end{aligned} \tag{C.11}$$

This is a weighted linear regression problem [210] where the predictors are $\Delta t_{in+1} z_{ijn}$, and $\Delta t_{in+1} x_{ijn}$; the weights are $|\Delta t_{in+1}|^{-1}$. During model selection, Section C.5, we found that equation (C.11) fit the data as well as the full model, equation (C.4).

C.6.2 Maximum likelihood estimators (MLEs)

We derive the MLEs for equation (C.4) in full generality. For convenience define

$$\hat{y}_{ibn+1} \equiv y_{ibn} + \Delta t_{in+1} \sum_k W_{bk}(y_{ikn} - \mu_{ikn}) = y_{ibn} + \Delta t_{in+1} \sum_k W_{bk}(y_{ikn} - \sum_j \Lambda_{kj} x_{ijn}) \tag{C.12}$$

where we have p variables, N individuals and $T+1$ timepoints. We index the N individuals with i and the T timepoint-pairs with n. For convenience we drop μ_0 and define the equivalent $\vec{\mu}_{in} \equiv \boldsymbol{\Lambda}\vec{x}_{in}$; where we use $x_{i0n} \equiv 1$ to recover μ_0. Estimators are denoted with a hat e.g. $\hat{\boldsymbol{W}}$ estimates $\boldsymbol{W}$.

The log-likelihood is,

$$\begin{aligned} l &= -\frac{1}{2}\sum_{i,n} \ln\left(\det\left|2\pi \boldsymbol{Q}^{-1} |\Delta t_{in+1}|\right|\right) - \frac{1}{2}\sum_{i,n} (\vec{y}_{in+1} - \hat{\vec{y}}_{in+1})^T \frac{\boldsymbol{Q}}{|\Delta t_{in+1}|} (\vec{y}_{in+1} - \hat{\vec{y}}_{in+1}) \\ &= \frac{1}{2} NT \ln\left(\det\left|\boldsymbol{Q}\right|\right) - \frac{p}{2}\sum_{i,n} \ln\left(2\pi |\Delta t_{in+1}|\right) - \frac{1}{2}\sum_{i,n} (\vec{y}_{in+1} - \hat{\vec{y}}_{in+1})^T \frac{\boldsymbol{Q}}{|\Delta t_{in+1}|} (\vec{y}_{in+1} - \hat{\vec{y}}_{in+1}). \end{aligned} \tag{C.13}$$

We derive analytical forms for the MLEs as well as providing derivatives for gradient-based optimization algorithms. We also report the curvature since this is used to estimate the asymptotic error via the inverse Fisher matrix [76]. We found that the asymptotic errors are well-calibrated for $\boldsymbol{W}$, but tend to be too small for $\vec{\mu}_n$ (Section C.7). In the present study, we report bootstrap errors.

Note that we find it useful to express the estimators in terms of the uncentered (cross)covariance,

$$\begin{aligned} \mathrm{Cov}_2(\vec{x}_{in}) &\equiv \langle \vec{x}\vec{x}^T \rangle_{i,n}, \text{ and} \\ \mathrm{Cov}_2(\vec{x}_{in}, \vec{y}_{in}) &\equiv \langle \vec{x}\vec{y}^T \rangle_{i,n} \end{aligned} \tag{C.14}$$

where the expectation value is taken over individuals, i, and timepoints, n. In general, $\langle f(x_{in}) \rangle_{i,n}$ denotes the average of $f(x_{in})$ over individuals, i, and timepoints n.

We start by considering $\boldsymbol{W}$. The derivatives are

$$\begin{aligned} \frac{\partial l}{\partial W_{\alpha\beta}} &= \sum_{i,n,a} \mathrm{sign}(\Delta t_{in+1}) Q_{a\alpha} (y_{ian+1} - y_{ian} - \Delta t_{in+1} \sum_k W_{ak}(y_{ikn} - \mu_{ikn}))(y_{i\beta n} - \mu_{i\beta n}) \\ \nabla_W l &= \sum_{i,n} \mathrm{sign}(\Delta t_{in+1}) \boldsymbol{Q} (\vec{y}_{in+1} - \vec{y}_{in} - \Delta t_{in+1} \boldsymbol{W}(\vec{y}_{in} - \vec{\mu}_{in}))(\vec{y}_{in} - \vec{\mu}_{in})^T \end{aligned} \tag{C.15}$$

where ∇_W denotes the gradient with respect to (vectorized) $\mathrm{vec}(\boldsymbol{W})$. The MLE is thus

$$\begin{aligned} \hat{\boldsymbol{W}} \langle |\Delta t_{in+1}| (\vec{y}_{in} - \vec{\mu}_{in})(\vec{y}_{in} - \vec{\mu}_{in})^T \rangle_{i,n} &= \langle \mathrm{sign}(\Delta t_{in+1}) (\vec{y}_{in+1} - \vec{y}_{in})(\vec{y}_{in} - \vec{\mu}_{in})^T \rangle_{i,n} \\ \mathrm{Cov}_2(\sqrt{|\Delta t_{in+1}|}(\vec{y}_{in} - \vec{\mu}_{in})) \hat{\boldsymbol{W}}^T &= \mathrm{Cov}_2(\mathrm{sign}(\Delta t_{in+1})(\vec{y}_{in} - \vec{\mu}_{in}), \vec{y}_{in+1} - \vec{y}_{in}). \end{aligned} \tag{C.16}$$

The latter equation is useful for linear algebra software packages. Alternatively, we can invert the uncentered covariance $\langle |\Delta t_{in+1}| (\vec{y}_{in} - \vec{\mu}_{in})(\vec{y}_{in} - \vec{\mu}_{in})^T \rangle_{i,n}$ which yields equation (12).

The curvature of $\boldsymbol{W}$ is

$$\frac{\partial^2 l}{\partial W_{\gamma\delta} W_{\alpha\beta}} = -NTQ_{\gamma\alpha} \langle |\Delta t_{in+1}| (y_{i\beta n} - \mu_{i\beta n})(y_{i\delta n} - \mu_{i\delta n}) \rangle_{i,n} \tag{C.17}$$

the Fisher information is the negative of this. The covariance of the MLE is given by the inverse Fisher information,

$$I^{-1}_{\alpha\beta\gamma\delta} = \frac{1}{NT} Q^{-1}_{\alpha\gamma} \langle |\Delta t_{in+1}| (\vec{y}_{in} - \vec{\mu}_{in})(\vec{y}_{in} - \vec{\mu}_{in}) \rangle^{-1}_{\beta\delta} \tag{C.18}$$

the standard errors are the square-roots of the diagonal elements,

$$\delta W_{\alpha\beta}^2 = \frac{Q_{\alpha\alpha}^{-1}}{NT}\langle|\Delta t_{in+1}|(\vec{y}_{in}-\vec{\mu}_{in})(\vec{y}_{in}-\vec{\mu}_{in})\rangle_{\beta\beta}^{-1}. \tag{C.19}$$

Next we consider $\vec{\mu}_n$. We condense all relevant parameters into $\mathbf{\Lambda}$ which has MLE,

$$\begin{aligned}
\frac{\partial l}{\partial \Lambda_{\alpha\beta}} &= -\sum_{i,n,a,b}\frac{Q_{ab}}{|\Delta t_{in+1}|}(y_{ian+1}-\hat{y}_{ian+1})(\Delta t_{in+1}W_{b\alpha}x_{i\beta n}) \\
&= -NTW_{\alpha\cdot}^T\boldsymbol{Q}\langle(\vec{y}_{in+1}-\vec{y}_n)x_{i\beta n}\text{sign}(\Delta t_{in+1})\rangle + NTW_{\alpha\cdot}^T\boldsymbol{QW}\langle|\Delta t_{in+1}|\vec{y}_{in}x_{i\beta n}\rangle \\
&\quad - W_{\alpha\cdot}^T\boldsymbol{QW\Lambda}\langle|\Delta t_{in+1}|\vec{x}_{in}x_{i\beta n}\rangle \\
\implies \frac{1}{NT}\nabla_\Lambda l &= -\boldsymbol{W}^T\boldsymbol{Q}\langle(\vec{y}_{in+1}-\vec{y}_n)\vec{x}_{in}^T\text{sign}(\Delta t_{in+1})\rangle + \boldsymbol{W}^T\boldsymbol{QW}\langle|\Delta t_{in+1}|\vec{y}_{in}\vec{x}_{in}^T\rangle \\
&\quad - \boldsymbol{W}^T\boldsymbol{QW\Lambda}\langle|\Delta t_{in+1}|\vec{x}_{in}\vec{x}_{in}^T\rangle
\end{aligned} \tag{C.20}$$

This implies

$$\boldsymbol{W}^T\boldsymbol{QW}\hat{\boldsymbol{\Lambda}}\langle|\Delta t_{in+1}|\vec{x}_{in}\vec{x}_{in}^T\rangle = \boldsymbol{W}^T\boldsymbol{QW}\langle|\Delta t_{in+1}|\vec{y}_{in}\vec{x}_{in}^T\rangle - \boldsymbol{W}^T\boldsymbol{Q}\langle\text{sign}(\Delta t_{in+1})(\vec{y}_{in+1}-\vec{y}_{in})\vec{x}_{in}^T\rangle \tag{C.21}$$

which we can write

$$\begin{aligned}
\boldsymbol{W}^T\boldsymbol{QW}\hat{\boldsymbol{\Lambda}} &= \boldsymbol{W}^T\boldsymbol{QW}\text{Cov}_2(|\Delta t_{in+1}|\vec{y}_{in},\vec{x}_{in})\big(\text{Cov}_2(\sqrt{|\Delta t_{in+1}|}\vec{x}_{in})^{-1}\big) \\
&\quad - \boldsymbol{W}^T\boldsymbol{Q}\text{Cov}_2(\text{sign}(\Delta t_{in+1})(\vec{y}_{in+1}-\vec{y}_{in}),\vec{x}_{in})\big(\text{Cov}_2(\sqrt{|\Delta t_{in+1}|}\vec{x}_{in})^{-1}\big).
\end{aligned} \tag{C.22}$$

We estimate from the general form, but note that equation (C.22) can be greatly simplified when $\boldsymbol{W}$ is invertible, which is expected because it empirically has strong diagonal elements. For invertible $\boldsymbol{W}$ we get equation (11).

The curvature is

$$\frac{\partial^2 l}{\partial\Lambda_{\gamma\delta}\partial\Lambda_{\alpha\beta}} = -(\boldsymbol{W}^T\boldsymbol{QW})_{\alpha\gamma}TN\langle|\Delta t_{in+1}|\vec{x}_{in}\vec{x}_{in}^T\rangle_{\beta\delta} \tag{C.23}$$

where the expectation is over times and individuals. The Fisher information is used to estimate the asymptotic error,

$$I_{\alpha\beta\gamma\delta}^{-1} = \frac{1}{NT}(\boldsymbol{W}^T\boldsymbol{QW})_{\alpha\gamma}^{-1}\langle|\Delta t_{in+1}|\vec{x}_{in}\vec{x}_{in}^T\rangle_{\beta\delta}^{-1} \tag{C.24}$$

the fit error is the square-root of the diagonal,

$$\big(\delta\Lambda_{\alpha\beta}\big)^2 = \frac{1}{NT}(\boldsymbol{W}^T\boldsymbol{QW})_{\alpha\alpha}^{-1}\langle|\Delta t_{in+1}|\vec{x}_{in}\vec{x}_{in}^T\rangle_{\beta\beta}^{-1}. \tag{C.25}$$

Finally, observe the equilibrium case where $\langle \vec{y}_{n+1} \rangle = \langle \vec{y}_n \rangle = \vec{\mu}_n$ and $\text{Cor}(\vec{y}_{n+1} - \vec{y}_n, \vec{x}) = \text{Cor}(\vec{y}_n - \vec{\mu}_n, \vec{x}) = 0$ (i.e. the fluctuations are random) then equation (C.20) becomes

$$\frac{1}{NT} \nabla_{\Lambda_{eq}} l = \boldsymbol{W}^T \boldsymbol{Q} \boldsymbol{W} \langle |\Delta t_{in+1}| \vec{y}_{in} \vec{x}_{in}^T \rangle - \boldsymbol{W}^T \boldsymbol{Q} \boldsymbol{W} \boldsymbol{\Lambda} \langle |\Delta t_{in+1}| \vec{x}_{in} \vec{x}_{in}^T \rangle \tag{C.26}$$

and we have

$$\begin{aligned} \hat{\boldsymbol{\Lambda}}_{eq} &= \left\langle |\Delta t_{in+1}| \vec{y}_{in+1} \vec{x}_{in}^T \right\rangle \left(\left\langle |\Delta t_{in+1}| \vec{x}_{in} \vec{x}_{in}^T \right\rangle \right)^{-1}, \\ &= \text{Cov}_2(|\Delta t_{in+1}| \vec{y}_{in+1}, \vec{x}_{in}) \big(\text{Cov}_2(\sqrt{|\Delta t_{in+1}|} \vec{x}_{in})^{-1} \big). \end{aligned} \tag{C.27}$$

equation (C.27) is useful for an initial $\boldsymbol{\Lambda}$ estimate as it does not depend on $\boldsymbol{W}$.

C.6.3 Noise estimator

We used a simple estimator for the noise, $\boldsymbol{\Sigma}$. For our model, equation (C.4), a simple estimator is derived by observing

$$\vec{y}_{n+1} - \langle \vec{y}_{n+1} \rangle = \vec{\epsilon} \tag{C.28}$$

which implies that

$$\langle (\vec{y}_{n+1} - \langle \vec{y}_{n+1} \rangle)(\vec{y}_{n+1} - \langle \vec{y}_{n+1} \rangle)^T \rangle = \langle \vec{\epsilon} \vec{\epsilon}^T \rangle = \boldsymbol{\Sigma} |\Delta t| \tag{C.29}$$

we conclude that

$$\hat{\boldsymbol{\Sigma}} = \left\langle \frac{1}{|\Delta t_{in+1}|} (\vec{y}_{in+1} - \langle \vec{y}_{in+1} \rangle)(\vec{y}_{in+1} - \langle \vec{y}_{in+1} \rangle)^T \right\rangle_{i,n}. \tag{C.30}$$

Where the expectation must be taken over individuals, i, and timepoints, n. Note that $\vec{y}_{n+1} - \langle \vec{y}_{n+1} \rangle$ is the model residual, which is easily computed after the model has been fit.

C.6.4 Iterative estimation

We found that a simple iterative scheme of alternating estimators from Section C.6.2 was able to correctly recovery true parameter values in a simulation study, Section C.7. The scheme proceeds according to Algorithm C.1 (see below). We defaulted to numIter = 5 iterations. Note that in the special, diagonal case, of equation (C.6) we simultaneously estimated $\boldsymbol{\Lambda}$ and $\boldsymbol{W}$ using weighted linear regression, Section C.6.1 ("PCA" case).

While we did estimate the asymptotic error, we found that the bootstrapped error had lower bias and hence we only report the latter (see Section C.7). To estimate the errors in parameters and prediction we bootstrapped Algorithm C.1 and took the standard deviation as the error estimate (100 resamples).

C.7 Validation

We used synthetic (simulated) data to validate: (i) Algorithm C.1, (ii) the parameter errorbars, and (iii) the prediction error estimator (RMSE). We used synthetic data based on the SLAM C57/BL6 dataset for validation. We fit the full model equation (C.4) to the dataset: all 6 predictors and 2 covariates (sex and age), as well as estimating the noise. We used the fit parameters to generate new data then tested to see if Algorithm C.1 recovered the true parameters and errors. Algorithm C.1 was bootstrapped 100 times, the prediction error was estimated using both in-sample (train) and out-of-sample (test). We simulated 1000 times for each synthetic dataset size: 10, 50, 100, 500 and 1000 individuals. Each dataset had 22 timepoints.

We confirmed that Algorithm C.1 is able to accurately reproduce true model parameters. In Figure C.5 we plot the parameter estimates versus the ground truth values. We see that the algorithm is accurate for $N \geq 50$. We see a bias-low for the diagonal elements of $\boldsymbol{W}$.

C.7.1 Parameter error

Here we test the calibration of our parameter errorbars. We compare both the bootstrap and asymptotic error estimates to the ground truth. Bootstrap errors were estimated using the standard deviation of bootstrap replicates. Asymptotic errors were estimated using the estimators in Section C.6.2. As we will demonstrate in this section, the asymptotic errorbars can be too small, whereas the bootstrap errors appeared to be correctly calibrated. For this reason, we always used the bootstrap estimates in the main text. The asymptotic error estimates are much faster to compute and are presented for posterity.

Algorithm C.1 Iterative estimator

if imputeFirst **then**

Impute missing $\vec{y}$ using simple algorithm (e.g. carry forward/back).

end if

if doPCA **then**

Estimate PCA rotation, $\boldsymbol{U}$, on first timepoint, $\vec{y}_1$, then apply to all timepoints.

end if

Estimate $\boldsymbol{\Lambda}$ using equation (C.27) which assumes $\langle \vec{y}_{n+1} \rangle = \langle \vec{y}_n \rangle = \vec{\mu}_n$ and $\text{Cor}(\vec{y}_n - \vec{\mu}_n, \vec{x}) = 0$.

Estimate $\boldsymbol{W}$ using equation (C.16).

for i in 1 to numIter **do**

if estimateNoise **then**

Estimate $\boldsymbol{\Sigma}$ using equation (C.30) and $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$.

end if

if imputeMean **then**

if doPCA **then**

Transform model parameters into observed space using $\boldsymbol{U}^{-1} = \boldsymbol{U}^T$.

end if

Impute $\vec{y}_1$ with the model mean using equation (C.1).

for n in 2 to numTimes **do**

Impute $\vec{y}_n$ with the model mean using equation (C.3).

end for

end if

Estimate $\boldsymbol{\Lambda}$ using equation (C.22).

Estimate $\boldsymbol{W}$ using equation (C.16).

end for

if doPCA **then**

Transform imputed values and parameters into observed space using $\boldsymbol{U}^{-1} = \boldsymbol{U}^T$.

end if

Estimate asymptotic errors.

Return $\boldsymbol{W}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\Sigma}$ and imputed values, $\vec{y}_{imp}$.

Where imputeFirst, doPCA, estimateNoise and imputeMean are Boolean user settings. numTimes is the number of observation timepoints in the dataset, $T + 1$.

In Figure C.6 we present the coverage of both error estimators. The coverage is the fraction of times that the true parameter value fell within the estimated error interval. The nominal coverage is 68.3% for a normal random variable. In Figure C.6A we present the coverage of the asymptotic error estimates and find that they are unsatisfactory for μ_{age} and μ_0 (errorbars were too small). These may be due to strong correlations between the two parameters, for example the parameters for body weight correlated across simulations at $\text{cor}(\mu_{age}, \mu_0) = -0.886$, which could make the asymptotic errors inaccurate. In Figure C.6B we observe that all of the bootstrap error parameter coverages were close to the nominal rate (dashed line), and were symmetrically distributed above and below. This indicates that our bootstrap parameter errorbars were properly calibrated.

C.7.2 Prediction error

Our primary measure of prediction error was the root-mean-squared error (RMSE). It is important that our measure is properly calibrated such that it estimates the correct RMSE, i.e. in a simulation study where the true error is known. We compare three RMSE estimators to the ground truth: (i) the testing error, which is the out-of-sample bootstrap error, (ii) the training error, which is the in-sample bootstrap error, and (iii) the 632 error which is a linear combination of 63.2% testing error and 36.8% training error [74]. The ground truth error is the error of the sample given the correct parameter values: this is the error of a single sample, not the distribution of possible values. The average ground truth error should be an unbiased estimate of the true, distribution error. In Figure C.7A we demonstrate that the 632 error is close to the ground truth error. In Figure C.7B we present the coverage of each estimator and find they are all close to the nominal rate. We conclude that the 632 error is a satisfactory estimator of the true model error.

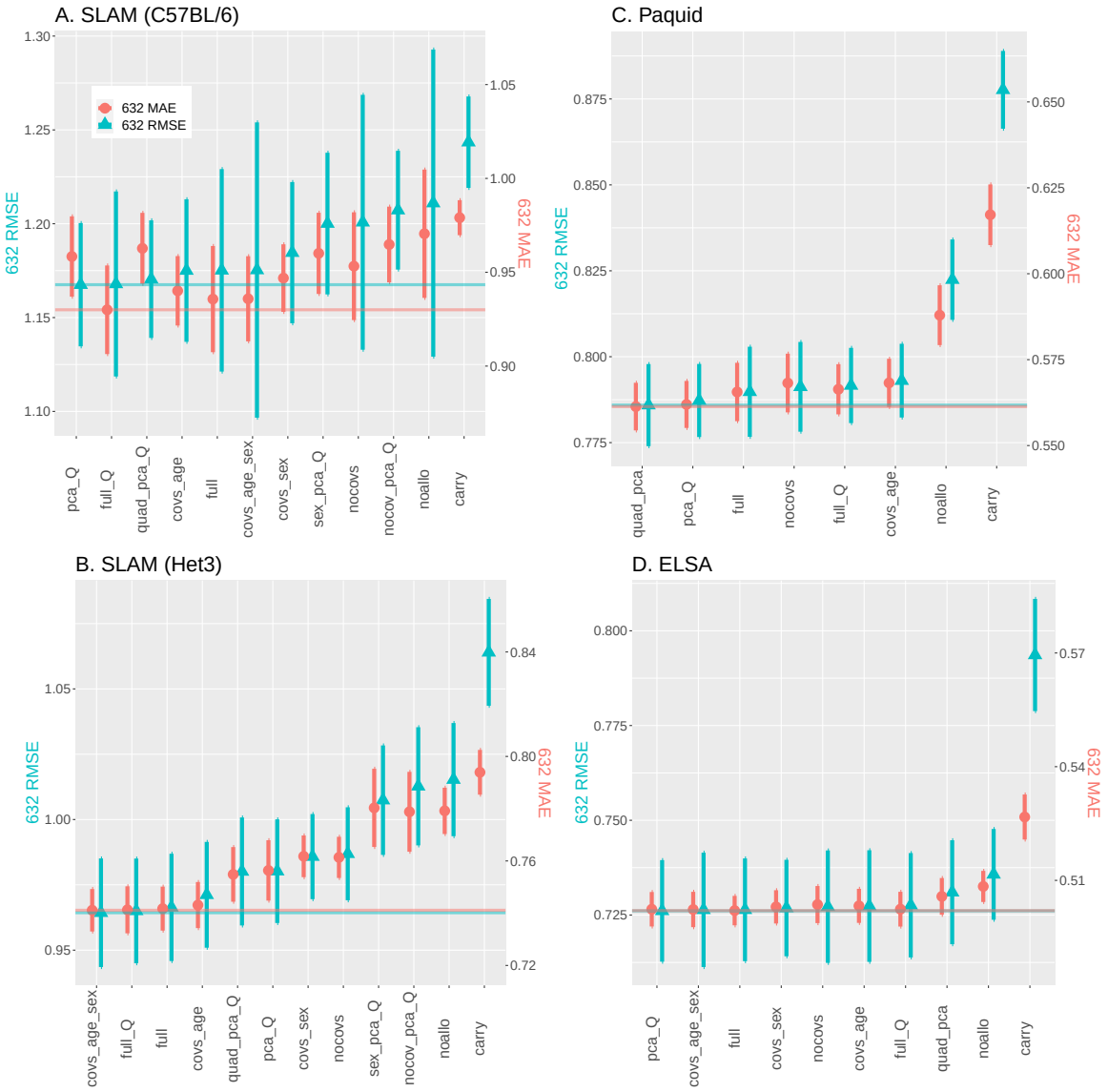

Figure C.4: Model selection. **A.** C57BL/6 mice (SLAM). **B.** Het3 mice (SLAM). **C.** Paquid (human, dementia). **D.** ELSA (human). Lower error is better. y-axis is 632-RMSE on left and 632-MAE on right. Horizontal lines indicate the best performing model. We are looking for the simplest model that consistently hits those lines across datasets. We considered models significantly worse if they do not have an error interval overlapping this line; prioritizing RMSE. Models: carry: equation (C.7); fast: equation (C.8); noallo: equation (C.9); quad: equation (C.10); full: equation (C.4). Additional parameters: pca: equation (C.6) with PCA preprocessing and diagonal noise; Q: the noise was estimated; covs: prefix, after which included covariates are listed; nocovs: no covariates were used. For example, `sex_pca_Q` included sex as a covariate (`sex`), used PCA as a preprocessing step and assumed diagonal $\boldsymbol{W}$ and $\boldsymbol{Q}$, and fit equation (C.6) (`pca`), and estimated Q from the data (`Q`). The fast model, equation (C.8), performed much worse for all datasets (points above plot region), 632-RMSE: 0.91(2) (Paquid), 0.92(1) (ELSA), 2.03(6) (SLAM C57) and 2.21(7) (SLAM HET3); 632-MAE: 0.68(1) (Paquid), 0.702(5) (ELSA), 1.32(3) (SLAM C57) and 1.47(3) (SLAM HET3).

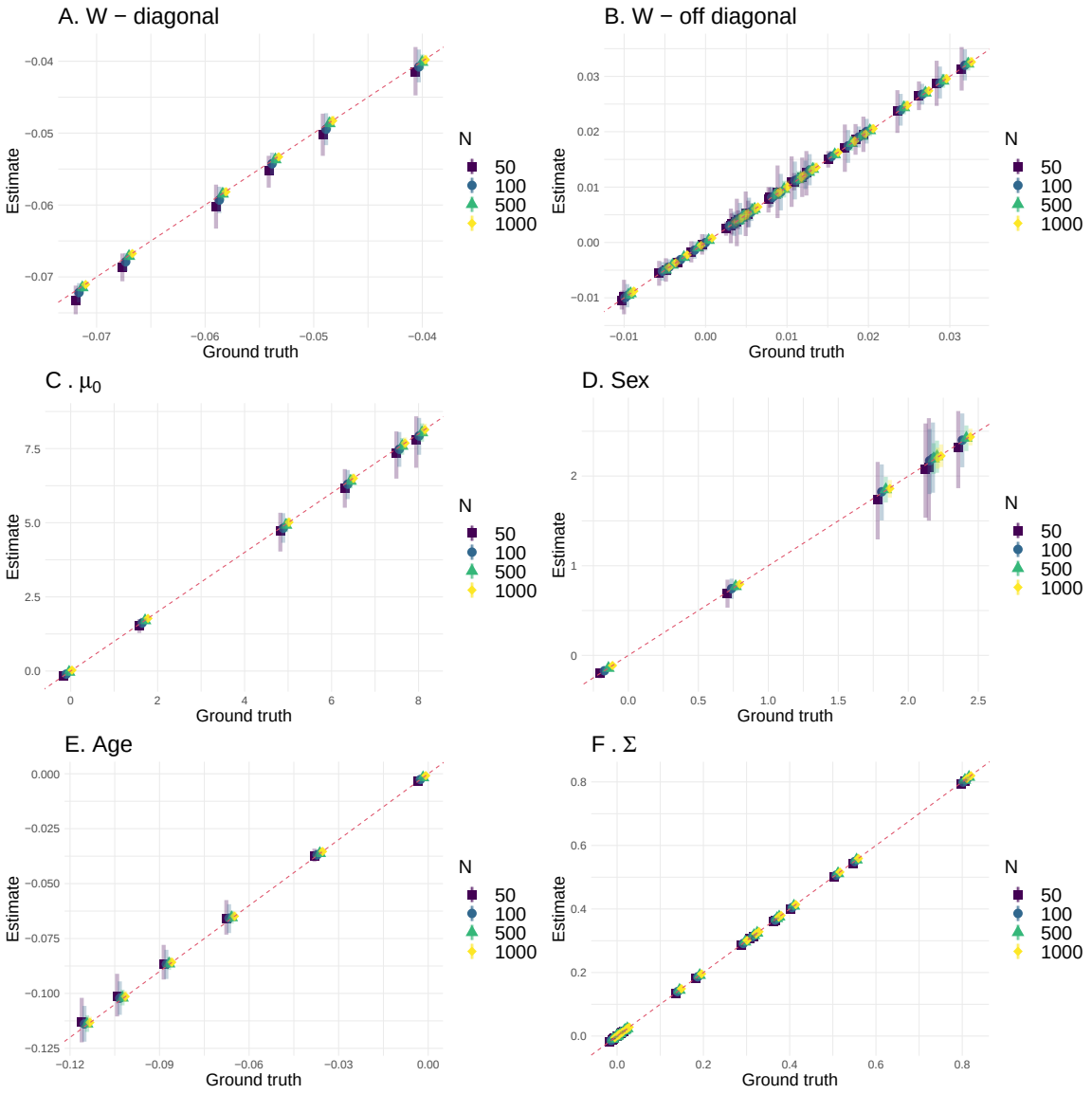

Figure C.5: Algorithm C.1 validation. For the indicated parameters in each measurement (A.-F.), the estimated value is plotted against the ground truth value for a variety of sample sizes (indicated by the legend). Points show mean; bands are the interquartile range (25th to 75th percentile). Bias is indicated by position of point relative to the red dashed line, $y = x$ (perfect estimator). Precision (and accuracy) are inferred by the dispersion (bands). As the number of individuals, N, is increased from 50 to 1000 we see the estimator becomes increasingly accurate and precise, with a small dispersion around the ground truth values for each parameter. Points are staggered for visualization. Note: $N = 10$ had large errors and hence was excluded for better visualization.

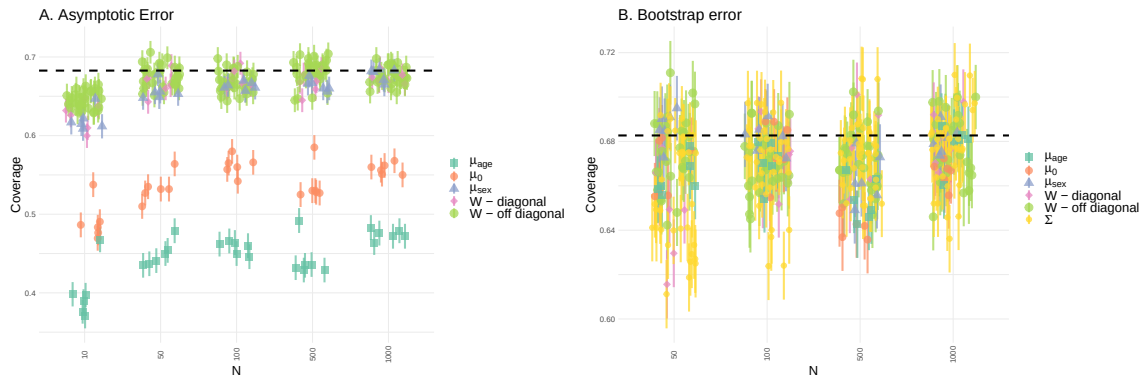

Figure C.6: Parameter errorbar validation (coverage). Asymptotic errorbars can be too small, whereas bootstrap errorbars appear to be valid. **A.** Asymptotic error clearly has abnormally low coverage for μ_0 and μ_{age}, perhaps due to strong correlations between the two parameters. Asymptotic error estimates for the other parameters look good. **B.** bootstrap error coverage looks good: parameters are close to the nominal rate (dashed line) and are (mostly) symmetrically distributed above and below. Note the scale. Errorbars are standard error in the mean. x-axis not to scale.

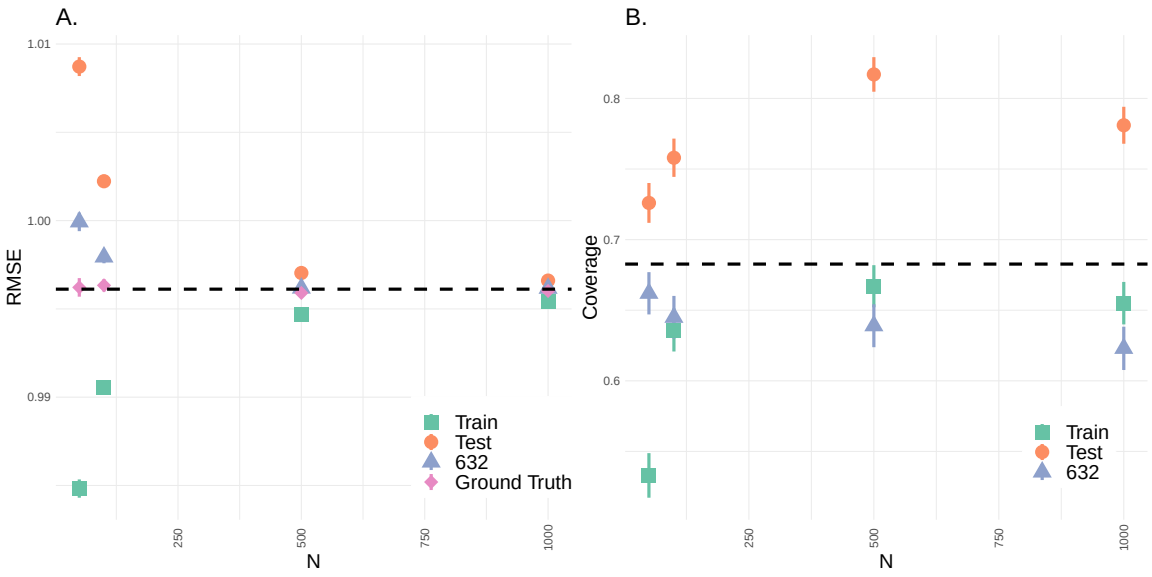

Figure C.7: Bootstrap error calibration. 632 error is a satisfactory estimator of the true error. **A.** Test error (out-of-sample) was biased high, training error (in-sample) was biased low, whereas 632 error was nearly unbiased relative to the ground truth. **B.** The coverage of the train and 632 error were close to the nominal rate, 68.3% (dashed line). The test error clearly had abnormally high coverage, indicating the errorbars on the test error are too large. Note: the true (stochastic) error is difficult to precisely estimate due to non-uniform sampling, so we used the average ground truth to estimate the true error. Errorbars are standard error.

C.8 Math

In this section we include supporting information for the model moments along with mappings to related approaches, i.e. other researcher's models.

C.8.1 Ordinary differential equation

We consider only 1-dimension since we found that we could transform our multivariate biomarkers into a set of decoupled, 1-dimensional equations using the $\boldsymbol{W}$-diagonalizing matrix, $\boldsymbol{P}$. That is,

$$\vec{z} \equiv \boldsymbol{P}^{-1}\vec{y} \tag{C.31}$$

decouples the z_j into a set of independent 1-dimensional equations. As needed, we can map back into $\vec{y}$ — which we do in Section C.8.5.

In the limit of small Δt our 1-dimensional model equation (C.6) becomes a modified Ornstein-Uhlenbeck process as follows,

$$\lim_{\Delta t \to 0} z_{jn+1} = z_{jn} + \lambda_j(z_{jn} - \tilde{\mu}_{jn})dt + \lim_{\Delta t \to 0} \tilde{\epsilon}, \tag{C.32}$$

with

$$\tilde{\epsilon} \sim \mathcal{N}(0, \tilde{\sigma}_j^2|\Delta t|). \tag{C.33}$$

A Wiener process, $d\xi$, has three criteria [77]: (i) independence, (ii) stationarity (statistics doesn't change over time), and (iii) $\mathcal{N}(0, |\Delta t|)$–distributed. These criteria are satisfied by $\tilde{\epsilon}$ once we scale out $\tilde{\sigma}_j$. Substituting t for timepoint n and $t + dt$ for timepoint $n + 1$ we have

$$z_j(t + dt) = z_j(t) + \lambda_j(z_j(t) - \tilde{\mu}_j(t))dt + \tilde{\sigma}_j d\xi(t), \tag{C.34}$$

which can be rewritten as

$$dz_j(t) = \lambda_j(z_j(t) - \tilde{\mu}_j(t))dt + \tilde{\sigma}_j d\xi(t), \tag{C.35}$$

which is an Ornstein-Uhlenbeck process with a non-constant equilibrium, $\tilde{\mu}_j(t)$, which depends on time through $\tilde{\mu}_{j,age}t$ [77]. Note that equation (C.35) holds for all $\tilde{\mu}_j(t)$ that can be Taylor expanded, since the nonlinear corrections go as $\mathcal{O}(\Delta t^2)$.

We rewrite equation (C.35) in a stripped down form as

$$dz = \lambda(z - \mu(t))dt + \sigma d\xi. \tag{C.36}$$

The solution is then

$$z(t) = z_0 e^{\lambda t} - \lambda e^{\lambda t} \int_0^t \mu(s) e^{-\lambda s} ds + \sigma e^{\lambda t} \int_0^t e^{-\lambda s} d\xi(s). \tag{C.37}$$

The integral is stochastic (Ito) and cannot be solved analytically, however its moments can be computed using two standard results [77]. The mean is

$$\langle \int_0^t f(s,\xi) d\xi \rangle = 0, \tag{C.38}$$

and the two-point correlations are

$$\langle \int_0^t f(s,\xi) d\xi(s) \int_0^{t'} g(u,\xi) d\xi(u) \rangle = \int_0^{\min(t,t')} \langle f(s,\xi) g(s,\xi) \rangle ds. \tag{C.39}$$

The statistics are Gaussian so all moments can be rewritten in terms of the mean and two-point correlations.

Moments

Here we provide additional details supporting the math in Box 1.

Let

$$\mu(t) = \mu_0 + \mu_{age} t, \tag{C.40}$$

that is, the only time-dependence is through the linear term $\mu_{age} t$ (μ_0 can still depend on covariates, but they can't vary in time).

Starting from equation (C.37)

$$\begin{aligned} z(t) &= z_0 e^{\lambda t} - \lambda e^{\lambda t} \int_0^t \mu(s) e^{-\lambda s} ds + \sigma e^{\lambda t} \int_0^t e^{-\lambda s} d\xi(s) \\ &= z_0 e^{\lambda t} - \lambda e^{\lambda t} \Big(- \frac{\mu_0}{\lambda} \big(e^{-\lambda t} - 1 \big) + \frac{\mu_{age}}{\lambda^2} \big(e^{-\lambda t} (-\lambda t - 1) + 1 \big) \Big) + \sigma e^{\lambda t} \int_0^t e^{-\lambda s} d\xi(s). \end{aligned} \tag{C.41}$$

The mean is easily computed as

$$\begin{aligned} \langle z(t) \rangle &= \langle z(0) \rangle e^{\lambda t} - \lambda e^{\lambda t} \Big(- \frac{\mu_0}{\lambda} \big(e^{-\lambda t} - 1 \big) + \frac{\mu_{age}}{\lambda^2} \big(e^{-\lambda t} (-\lambda t - 1) + 1 \big) \Big) \\ &= \langle z(0) \rangle e^{\lambda t} + (\mu_0 + \frac{\mu_{age}}{\lambda})(1 - e^{\lambda t}) + \mu_{age} t. \end{aligned} \tag{C.42}$$

The auto-covariance doesn't depend on $\mu(t)$, it is simply,

$$\left\langle (z(t+\tau) - \langle z(t+\tau)\rangle)(z(t) - \langle z(t)\rangle) \right\rangle = \sigma^2 e^{2\lambda t} e^{\lambda\tau} \left\langle \int_0^{t+\tau} e^{-\lambda s} d\xi(s) \int_0^t e^{-\lambda s} d\xi(s) \right\rangle$$
$$= -\frac{\sigma^2}{2\lambda}\left(e^{\lambda|\tau|} - e^{2\lambda t} e^{\lambda\tau} \right). \tag{C.43}$$

The variance is the special case $\tau = 0$,

$$\mathrm{Var}(z(t)) = -\frac{\sigma^2}{2\lambda}\left(1 - e^{2\lambda t}\right). \tag{C.44}$$

The mean and auto-covariance completely characterize Gaussian statistics, all other statistics can be calculated from them.

The above moments neglect the possibility that we may be unable to measure the system at $t = 0$. It is therefore useful to define the moments relative to a reference time, t_r. Doing some algebra we have

$$\begin{aligned}\langle z(t)\rangle &= \langle z(t_r)\rangle e^{\lambda(t-t_r)} + (\mu_0 + \frac{\mu_{age}}{\lambda} + \mu_{age} t_r)(1 - e^{\lambda(t-t_r)}) + \mu_{age}(t - t_r) \\ &= \langle z(t_r)\rangle e^{\lambda(t-t_r)} + (\mu(t_r) + \frac{\mu_{age}}{\lambda})(1 - e^{\lambda(t-t_r)}) + \mu_{age}(t - t_r)\end{aligned} \tag{C.45}$$

for the mean, where $\mu(t_r) \equiv \mu_0 + \mu_{age} t_r$. Note that it is convenient to write

$$\langle z(t)\rangle - \mu(t) = (\langle z(t_r)\rangle - \mu(t_r)) e^{\lambda(t-t_r)} + \frac{\mu_{age}}{\lambda}(1 - e^{\lambda(t-t_r)}). \tag{C.46}$$

For the variance we have

$$\mathrm{Var}(z(t)) = \mathrm{Var}(z(t_r)) e^{2\lambda(t-t_r)} - \frac{\sigma^2}{2\lambda}\left(1 - e^{2\lambda(t-t_r)}\right). \tag{C.47}$$

Note that if we wait a long time, $t - t_r \gg 1/\lambda$, we reach steady-state values (so long as $\lambda < 0$). For example, the steady-state variance is

$$\mathrm{Var}(z)_{ss} = -\frac{\sigma^2}{2\lambda}. \tag{C.48}$$

C.8.2 Biomarker Principal Components

The biomarkers, $\vec{y}$, are connected to the natural variables, $\vec{z}$, by the transformation, $\boldsymbol{P}^{-1}$, equation (C.31). $\boldsymbol{P}^{-1}$ is the (linear) diagonalizing transformation of $\boldsymbol{W}$. We can use this to calculate the steady-state principal components of $\vec{y}$,

$$\begin{aligned}\mathrm{Cov}(\vec{y}_{ss}, \vec{y}_{ss}) &= \langle (\vec{y}_{ss} - \langle \vec{y}_{ss}\rangle)(\vec{y}_{ss} - \langle \vec{y}_{ss}\rangle)^T \rangle \\ &= \boldsymbol{P} \langle (\vec{z}_{eq} - \langle \vec{z}_{ss}\rangle)(\vec{z}_{eq} - \langle \vec{z}_{ss}\rangle)^T \rangle \boldsymbol{P}^T \\ &= \boldsymbol{P}\mathrm{Cov}(\vec{z}_{ss}, \vec{z}_{ss})\boldsymbol{P}^T.\end{aligned} \tag{C.49}$$

If $\boldsymbol{P}$ is a rotation/orthogonal ($\boldsymbol{P}^{-1} = \boldsymbol{P}^T$) then, by definition [26], $\boldsymbol{P}$ is the diagonalizing transformation of $\mathrm{Cov}(\vec{y}_{ss}, \vec{y}_{ss})$, with eigenvalues equal to the steady-state variance of the z_j. Note: $\boldsymbol{P}$ is orthogonal if $\boldsymbol{W}$ is real and symmetric [26]. If we rank-order the $\mathrm{Var}(z_j)$ then we have exactly the principal components of $\vec{y}$ [146]. Hence, in the steady-state the principal components are exactly the same as the natural variables, $\vec{z}$, sorted in order of decreasing variance, equation (C.48).

C.8.3 Small Timesteps, Δt

Our model, equation (C.4), approximates an ordinary differential equation in the limit $|\lambda \Delta t| \ll 1$ (Sections C.8.1 and C.8.5). Sehl and Yates [176] found that most biomarkers decay linearly at a rate of $\lambda < 0.01$ year^{-1} with the fastest being about 0.03 year^{-1}. The frailty index — the average number of health deficits an individual has — accumulates at a similarly small rate of $0.025 - 0.04$ year^{-1} [123]. We observed typical rates in the range $0.025 - 0.05$ human-equivalent year^{-1} (Figure 6.2B), with sampling times Δt of 4 years for ELSA, 3 years for Paquid, 4.9 human-equivalent years for SLAM C57/BL6 and 3.6 human-equivalent years for SLAM Het3. This implies that we can expect $|\lambda \Delta t| \ll 1$ and therefore the small Δt approximation of equation (C.4) is likely fine. This means our model should behave similarly to an ordinary differential equation.

C.8.4 General dynamics

Linear and nonlinear dynamical models alike can be analysed for stability near an equilibrium position using the eigenvalues [106]. The system is linearized as

$$\frac{d}{dt}\vec{y} = \boldsymbol{W}\vec{y} + \vec{b}. \tag{C.50}$$

The system is stable if and only if the real parts of the eigenvalues are always negative (positive recovery). Observe that the mean of our model equation (C.4) can be written as

$$\begin{aligned} \frac{\langle \vec{y}_{n+1} - \vec{y}_n \rangle}{\langle \Delta t_{n+1} \rangle} &= \boldsymbol{W}\langle \vec{y}_n \rangle - \boldsymbol{W}\vec{\mu}_n \\ &= \boldsymbol{W}\langle \vec{y}_n \rangle + \vec{b} \end{aligned} \tag{C.51}$$

for $\vec{b} \equiv -\boldsymbol{W}\vec{\mu}_n$. Hence for small Δt we have (approximately)

$$\frac{d}{dt}\langle\vec{y}(t)\rangle = \boldsymbol{W}\langle\vec{y}(t)\rangle + \vec{b} \tag{C.52}$$

hence our approach probes the mean-stability of arbitrary linear or nonlinear dynamics.

C.8.5 Stochastic process model (SPM) approximation

Our model can be used to analyse arbitrary dynamical systems near equilibrium, as discussed in Section C.8.4. Here we show how a specific dynamical model — the stochastic process model (SPM) — is approximated by our model. Our model was motivated in part by earlier works which have shown that biomarker data can be modelled as a stochastic differential equation [213, 52]. The earlier work by Yashin *et al.* proposed the SPM as a generic framework for longitudinal aging biomarker data [213] where an individual's collection of biomarkers, $\vec{y}$, evolves over time as

$$d\vec{y} = \boldsymbol{A}(t)(\vec{y} - \vec{\mu}(t))dt + \boldsymbol{B}(t)d\xi_t \tag{C.53}$$

where $\vec{\mu}$ is the unknown equilibrium term ("functional state" of the organism), $\boldsymbol{A}$ is the interaction network and $d\xi_t$ is a Wiener noise term modified by the matrix $\boldsymbol{B}$. Subsequent work by Farrell *et al.* demonstrated that a deep neural network could be used to fit SPM and further that a time-independent linear interaction model was sufficient to describe the interaction network, $\boldsymbol{A}$, for ELSA data [52]. Our model, equation (C.4), is the appropriate approximation for equation (C.53) for small timesteps.

Proof: In Section C.8.1 we showed that our 1-dimensional model is equivalent to a Wiener process in the limit of $\Delta t \to 0$. Consider the SPM with constant regulation matrix, $\boldsymbol{A}$, and linear functional state, μ,

$$d\vec{y} = \boldsymbol{A}(\vec{y} - \vec{\mu}(t))dt + \boldsymbol{B}d\xi \tag{C.54}$$

Suppose $\boldsymbol{A}$ is diagonalizable then,

$$\begin{aligned} d\vec{y} &= \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}(\vec{y} - \vec{\mu})dt + \boldsymbol{B}d\xi, \\ \implies d(P^{-1}\vec{y}) &= \boldsymbol{D}(\boldsymbol{P}^{-1}(\vec{y} - \vec{\mu}))dt + \boldsymbol{P}^{-1}\boldsymbol{B}d\xi, \\ \implies d\vec{z} &= \boldsymbol{D}(\vec{z} - \tilde{\vec{\mu}})dt + \tilde{\boldsymbol{B}}d\xi \end{aligned} \tag{C.55}$$

for the latent space, $\vec{z} \equiv \boldsymbol{P}^{-1}\vec{y}$. By inspection, the latent space obeys Ornstein-Uhlenbeck dynamics with $D_{jj} = \lambda_j$ and hence we can approximate each z_j,

$$z_j(t + \Delta t) \approx z_j(t) + D_{jj}(z_j(t) - \tilde{\mu}_j)\Delta t + \tilde{\epsilon}_j, \quad \text{where}$$
$$\tilde{\vec{\epsilon}} \sim \mathcal{N}(0, \tilde{\boldsymbol{B}}\tilde{\boldsymbol{B}}^T|\Delta t|) \tag{C.56}$$

which we can map into the observed space using $\boldsymbol{P}$ to get,

$$\boxed{\vec{y}(t + \Delta t) \approx \vec{y}(t) + \boldsymbol{A}(\vec{y}(t) - \vec{\mu})\Delta t + \vec{\epsilon}} \tag{C.57}$$

which is equation (C.4) with $\boldsymbol{A} \equiv \boldsymbol{W}$. The transformed variance of $\tilde{\epsilon}_i$ is,

$$\begin{aligned}
\langle(\boldsymbol{P}\tilde{\vec{\epsilon}})(\boldsymbol{P}\tilde{\vec{\epsilon}})^T\rangle &= \boldsymbol{P}\langle\tilde{\vec{\epsilon}}\tilde{\vec{\epsilon}}^T\rangle\boldsymbol{P}^T \\
&= \boldsymbol{P}\langle\tilde{\boldsymbol{B}}\tilde{\boldsymbol{B}}^T\rangle\boldsymbol{P}^T|\Delta t| \\
&= \langle\boldsymbol{B}\boldsymbol{B}^T\rangle|\Delta t| \\
&\equiv \boldsymbol{\Sigma}|\Delta t|
\end{aligned} \tag{C.58}$$

QED.

C.8.6 Mapping to Sehl and Yates

Sehl and Yates performed a meta-analysis of 469 biomarkers across cross-sectional and longitudinal aging studies and observed that the vast majority of biomarkers decay linearly with age [176]. In the present section we demonstrate that their linear model describes the steady-state dynamics of our model. In other words, the long-time (old-age) behaviour of our model is consistent with their observations.

In the steady-state our model equation (7) becomes linear in time,

$$\langle z_{jn}\rangle_{ss} = \mu_{0j}(\vec{x}) + \mu_{age,j}t_n - \frac{\mu_{age,j}}{|\lambda_j|} \tag{C.59}$$

where we have included all time-independent covariates in $\mu_0(\vec{x})$ for convenience. The biomarkers have a one-to-one relationship with the natural variables through the orthogonal transformation $\boldsymbol{P}$ and thus evolve according to

$$\langle y_{ln}\rangle_{ss} = \sum_j P_{lj}\mu_{age,j}t_n + \sum_j P_{lj}\left(\mu_{0j}(\vec{x}) - \frac{\mu_{age,j}}{|\lambda_j|}\right). \tag{C.60}$$

The Sehl and Yates model [176] is

$$\frac{y_{ln}}{y_{l,30}} = 1 - k_l(t_n - 30) \tag{C.61}$$

for biomarker y_l with baseline value of $y_{l,30}$ at age 30; the age is in years and k_l is the rate in %-change per year. We can rewrite their model as

$$y_{ln} = -k_l y_{l,30} \mathrm{t}_n + (30k_l + 1)y_{l,30}, \tag{C.62}$$

which is exactly equation (C.60) with the substitutions

$$\begin{aligned} \sum_j P_{lj}\mu_{age,j} &\equiv -k_l y_{l,30}, \quad \text{and} \\ \sum_j P_{lj}\left(\mu_{0j} - \frac{\mu_{age,j}}{|\lambda_j|}\right) &\equiv (30k_l + 1)y_{l,30}, \end{aligned} \tag{C.63}$$

which can be mapped into $\vec{z}$ using $\boldsymbol{P}^{-1}$:

$$\begin{aligned} \mu_{age,j} &\equiv -\sum_l P_{jl}^{-1} k_l y_{l,30}, \quad \text{and} \\ \mu_{0j} - \frac{\mu_{age,j}}{|\lambda_j|} &\equiv \sum_l P_{jl}^{-1}(30k_l + 1)y_{l,30}. \end{aligned} \tag{C.64}$$

Observe that $\boldsymbol{P}$ permits the drift of only a few $\vec{z}$ to map into many observed biomarkers, $\vec{y}$. Together with our observation that many more biomarkers drift than do natural variables, Figure C.10, this implies that Sehl and Yates' observation that most biomarkers drift with age may be due to a only few underlying (allostatic) natural variables that are declining with age.

C.9 Additional Results

We restricted the main text to only our key results. Here we provide additional information to support our conclusions.

We included covariates, $\vec{x}$, in the equilibrium position, $\mu(\vec{x})$, to reduce confounding effects and to test for the presence of allostasis (which depends on age). Here we tested each parameter for significance using the bootstrap parameter error estimates. The z-score for each covariate is reported in Figure C.8; blue tiles are not significant, white and red are significant at $p \leq 0.05$. Most covariates were significant, particularly age for the human studies. In Section C.5 we found that the effect of covariates on prediction was typically small. This means that the effects of covariates were reliably estimated (small p) but did not explain much variation (minor effect on RMSE).

Our model estimates an interaction network, $\boldsymbol{W}$, together with equilibrium positions. In the main text we presented the ELSA network with suppressed diagonal (Figure 6.2). The complete networks for each dataset are provided in Figure C.9. The networks are all symmetrical because we used PCA as a preprocessing step. Relationships indicate how the y-axis variable will affect the x-axis variable during the next timestep.

Our model also estimates an equilibrium homeostatic position for each variable, μ. An important question is how strongly do variables adhere to homeostasis in the biomarkers, $\vec{y}$ versus the natural variables, $\vec{z}$? In the main text we presented the difference between the natural variable mean and the equilibrium position for each variable, $\langle z_j - \mu_j \rangle$. We reproduce that figure beside the observed biomarkers in Figure C.10. In Figure C.10B (and Figure 6.4A) we observed that the natural variables appear to be split into two groups: the majority group was close to μ, indicative of homeostasis, whereas the minority group was far from homeostasis. This latter group had a strong drift term, μ_{age}, which indicated that homeostasis was a moving target i.e. allostasis. In Figure C.10A we show that the observed biomarkers were much more likely to be far from homeostasis than the natural variables (B), implying that the natural variables are able to condense the effects of age-related drift (allostasis) into a few variables (see also Section C.8.6).

The natural variables appear to be efficient for representing age-related changes. What do the natural variables mean in terms of observable outcomes? In Figure C.11 we report the correlations between the accumulating/drifting natural variables and biomarkers. In Figure C.12 we report correlations with covariates. Together these give us an idea of what each natural variable represents and, by model implication, is controlling. For example, z_1 of Paquid is strongly correlated with the mental acuity scores: MMSE, BVRT and IST, implying it represents overall mental acuity. This may explain why z_1 was such a strong predictor of dementia (Figure C.14). In this way the age-related decline of mental acuity can be represented by changes to just one variable, z_1, but also observed across several biomarkers, MMSE, BVRT and IST.

The linear map, $\boldsymbol{P}$, allows a few natural variables to cause several biomarkers to drift. The effects of allostasis on observed biomarkers via the primary risk natural variables are illustrated in Figure C.13. The sign of each natural variable is arbitrary, due to idiosyncrasies in eigendecompositions [146]. The dominant survival dimension for the Het3 mouse data was z_2, which appears to capture a loss of body fat and muscle, and relative gain of fluid. The dominant z_2 dimension for the C57BL/6 was more specific to loss of fat (the z_1 signal for C57BL/6 was very similar to the z_2 signal for Het3, Figure C.11). z_1 was the dominant dementia-free-survival dimension for Paquid, and captured a system-wide drop in mental acuity scores (MMSE, BVRT and IST), which likely captures cognitive decline associated with dementia. z_1 for ELSA appears to be related to frailty [146], having its effects spread across many variables, especially those related to disability (eye, hear, FI ADL and FI IADL), physical condition (grip strength and gait speed), and self-reported health (SRH); note that higher is better for physical condition variables and worse for the other variables (eye, hear, SRH, FI, etc). In all cases the effects are strongest in the natural variables, which is ensured by the orthogonality of $\boldsymbol{P}$. This means that the effects of the natural variable drift must be diluted across the observed biomarkers (e.g. Figure C.10), potentially hiding them within healthy variation.

The drift rate of the natural variables, μ_{age}, was correlated with the risk of adverse outcome (Figure 6.4A). We named this phenomenon "mallostasis": the tendency of an aging system towards an ever-worsening equilibrium. Here we consider the role of confounding variables by constructing complete survival models for each natural variable and adverse outcome (mortality or dementia onset). We constructed a (time-dependent) survival model for each natural variable, with age, sex and the natural variable as predictors. We then recorded the Cox proportional hazards coefficients, which represent the (conditional) log–hazard ratios per unit increase for each natural variable. We observed that the Cox coefficients correlated with the drift rate, μ_{age}: Figure C.14. This provides more robust support for mallostasis: that the steady-state behaviour of aging mice and humans is declining natural variables and commensurately declining health.

As an illustration of mallostasis, we consider a simple composite health measure, $b \equiv \vec{\mu}_{age}^{T}\vec{z}$. Figure C.14 demonstrates that Cox coefficient is proportional to μ_{age} therefore b is proportional to the hazard. This is confirmed in Figure C.15 which demonstrates that b for each dataset is a good predictor of survival (or dementia onset).

The natural variables are connected to survival via mallostasis, but how do they relate to the observed phenotype? That is, how do the changes in the natural variables with age affect the observed biomarkers? The total mean and variance are conserved between the biomarkers and the natural variables by Parseval's theorem. This means that natural variables with large means and variances will dominate the means and variances of the observed biomarkers, thus controlling the major changes we see. The steady-state mean grows indefinitely proportional to μ_{age}, what about the variance? The equilibrium (steady-state) dispersion, equation (C.48), for the natural variables are plotted in Figure C.16. Smaller eigenvalues are associated with higher variance. Some dimensions (e.g. z_1 for the C57BL/6) can contain as much as 10x more variance than the next highest dimension. These dimensions will dominate the observed variance in the steady-state. The model predicts that these dimensions will eventually become the dominant principal components (PCs), equation (C.49), implying they would dominate the observed phenotype in the steady-state. Hence what we observe will be dominated by natural variables with small λ and large μ_{age}, such as z_1 of ELSA, which appears to be closely related to frailty.

We used PCA (principal component analysis) as a preprocessing step, which allowed us to fit a diagonal model, equation C.6. This simplified analysis and yielded equivalent performance to the full model (Section C.5). Here we test the self-consistency of the approach. By using PCA, at each bootstrap the eigenvectors of $\boldsymbol{W}$ are principal components, possibly reordered (because we fit a diagonal model for $\boldsymbol{W}$). Averaging over multiple bootstrap replicates removes this equivalence — although in the steady-state the model predicts that the principal components and eigenvectors of $\boldsymbol{W}$ will coincide, equation (C.49). Here we test the similarity of the PCA rotation and the eigenvector rotation: if they coincide then the principal components are eigenvectors. In Figure C.17 we present the inner product between these matrices, which varies from -1 to 1, with ± 1 representing perfect similarity. We observed strong similarities between the transformations, indicating that the principal components and natural variables will be strongly correlated. This suggests that PCA may be a useful shortcut for approximating the eigenvectors of $\boldsymbol{W}$.

Finally, we include a survival summary for each dimension in terms of conditional Cox regression and the C-index in Figure C.18. The values are identical to those used in Figures C.14 (Cox coefficient) and 6.4 (C-index). This permits the reader to investigate the relative importance of each dimension. Comparing to the correlates of each dimension, Figure C.11, one can infer potential mechanisms. For example, z_2 of C57BL/6 has a strong survival effect (low is bad) and shows increasing glucose and fat, which could indicate metabolic dysfunction, which C57BL/6 are prone to [121].

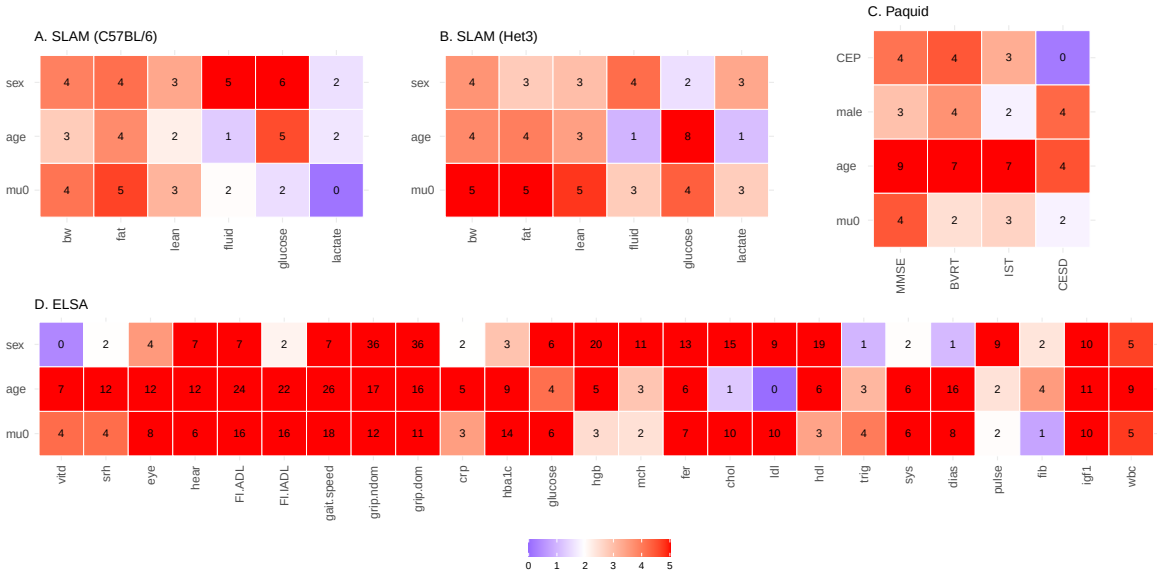

Figure C.8: Covariate significance (z-scores). **A.** C57BL/6 mice (SLAM). **B.** Het3 mice (SLAM). **C.** Paquid (human, dementia). **D.** ELSA (human). The equilibrium term, μ, was a linear function of these covariates. Most covariates were significant (red or white). Only the blue tiles were not significant at 95% ($z = 1.96$). Tile number is z-score. Colour scale is truncated at $z = 5$ ($p = 6 \cdot 10^{-7}$). See Figures C.11 and C.12 for the directions of the covariate effects.

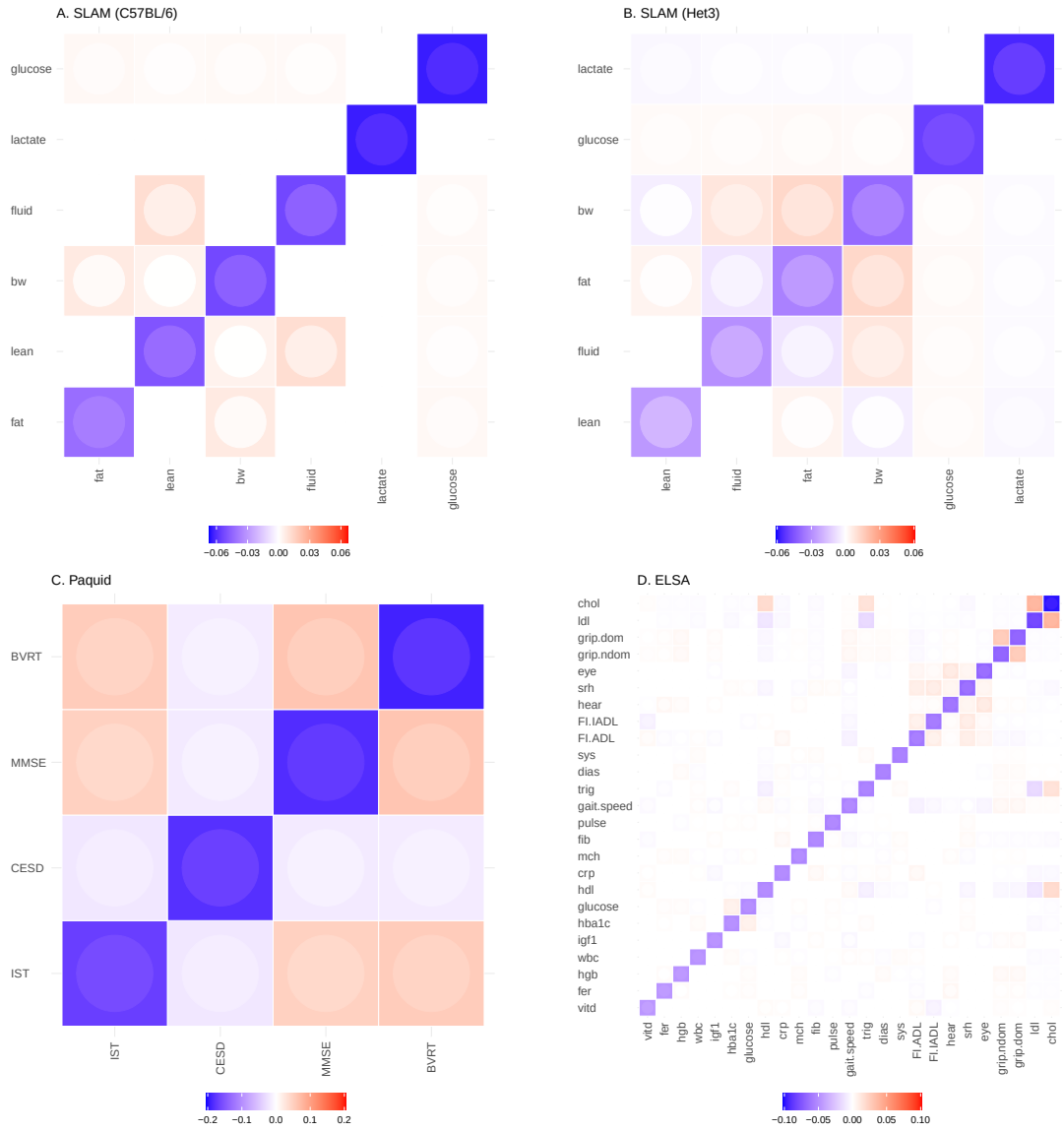

Figure C.9: Interaction networks for all datasets. **A.** C57BL/6 mice (SLAM). **B.** Het3 mice (SLAM). **C.** Paquid (human, dementia). **D.** ELSA (human). Tile colour indicates interaction strength (saturation) and direction (colour) of the interaction from the y-axis variable to the x-axis variable. Inner colour indicates the limit of 68% confidence interval (CI) closest to zero (i.e. standard error). Non-significant interactions, at 68%, have been whited-out. Variables are sorted by diagonal strength (increasing rate). The matrices are real and symmetric because the data were diagonalized by an orthogonal matrix (PCA).

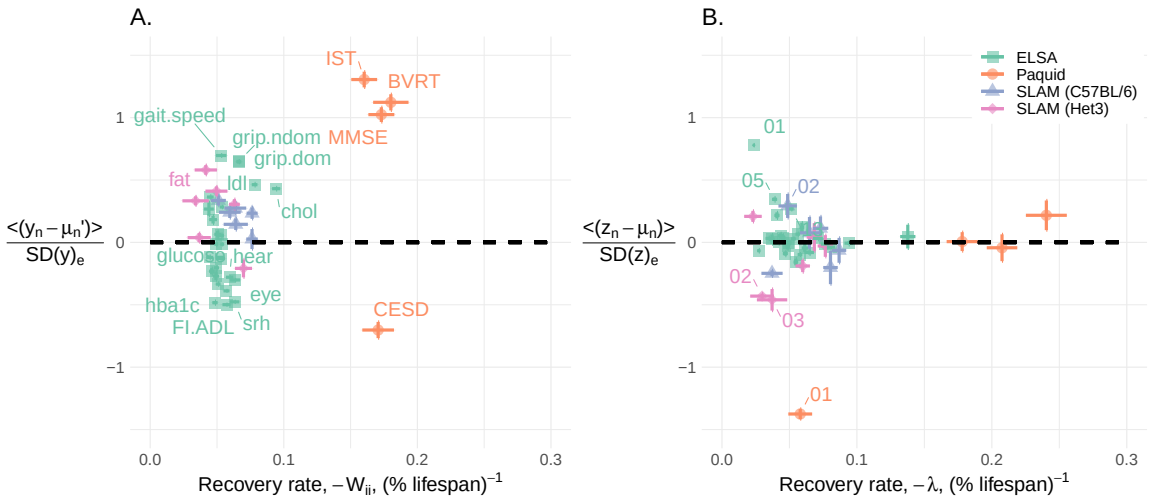

Figure C.10: Homeostasis of biomarkers vs natural variables. The dysruption of homeostasis seems to be diffuse across biomarkers whereas it is concentrated into a few natural variables. **A.** Observed biomarkers were typically far from equilibrium (dotted line). **B.** In contrast, most natural variables were close to equilibrium. We inferred that variables close to equilibrium were in homeostasis whereas those far from equilibrium were allostatic. Together these plots suggest that the natural variables were able to condense the effects of allostasis into a few major variables.

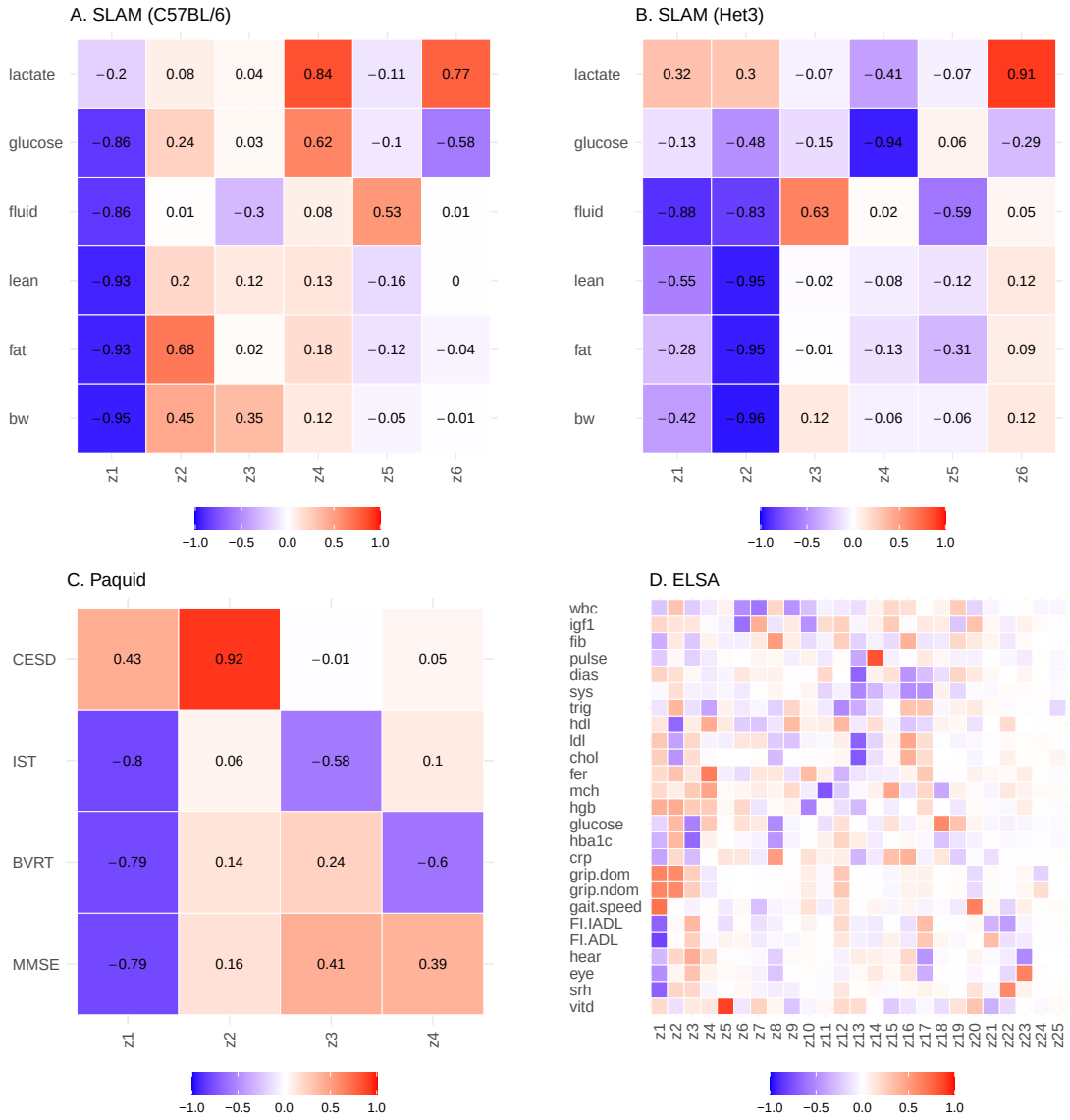

Figure C.11: Natural variable correlates — biomarkers (predictors). **A.** C57BL/6 mice (SLAM). **B.** Het3 mice (SLAM). **C.** Paquid (human, dementia). **D.** ELSA (human). This helps to describe what information is in each natural variable, z, and therefore what each natural variable is capable of controlling. The sign of each z is arbitrary due to idiosyncrasies of the eigendecomposition.

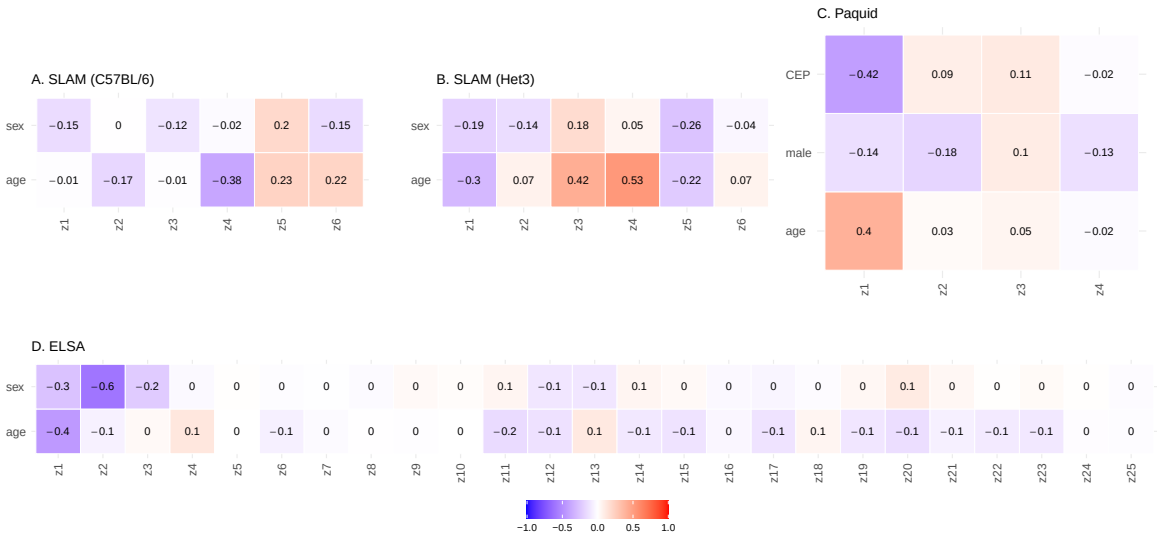

Figure C.12: Natural variable correlates — covariates. **A.** C57BL/6 mice (SLAM). **B.** Het3 mice (SLAM). **C.** Paquid (human, dementia). **D.** ELSA (human). This provides further information about what information each natural variable, z, contains. We expect the strongly drifting variables to exhibit correlations with age, though the sign of each z is arbitrary. Male is a binary sex indicator (1: male, 0: female); sex is the converse (0: male, 1: female). CEP is educational attainment level (1: attained primary, 0: did not).

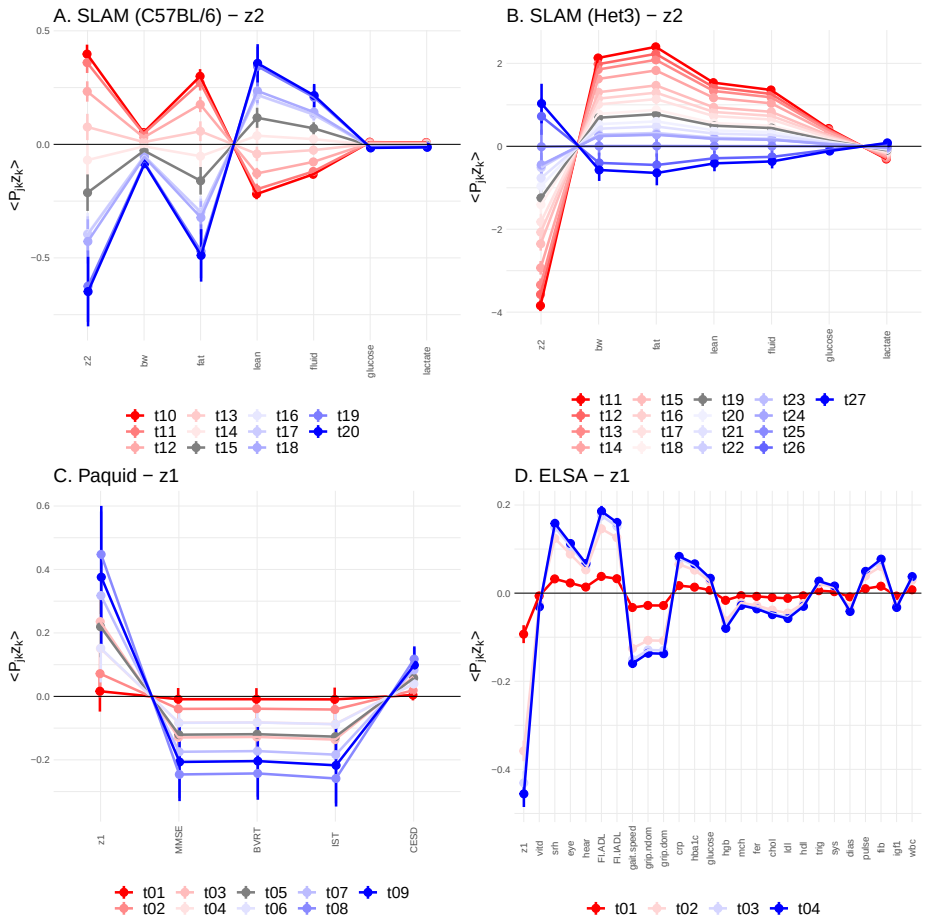

Figure C.13: Natural variable drift drives biomarker drift. **A.** C57BL/6 mice (SLAM). **B.** Het3 mice (SLAM). **C.** Paquid (human, dementia). **D.** ELSA (human). We consider the drift of the primary risk natural variables: z_1 for ELSA and Paquid and z_2 for SLAM. We observe a continuous drift in the natural variables. We also plot the drift of the biomarkers which is directly caused by each z via $\boldsymbol{P}$. In this manner, a few natural variables can drive drift across several biomarkers. Since $\boldsymbol{P}$ is orthogonal (length-preserving) the drift of each natural variable must be diluted across biomarkers (at most a single biomarker can drift at the same rate). See also the correlation matrices, Figures C.11 and C.12. For the SLAM datasets we've included only timepoints where the average age was over 80 weeks.

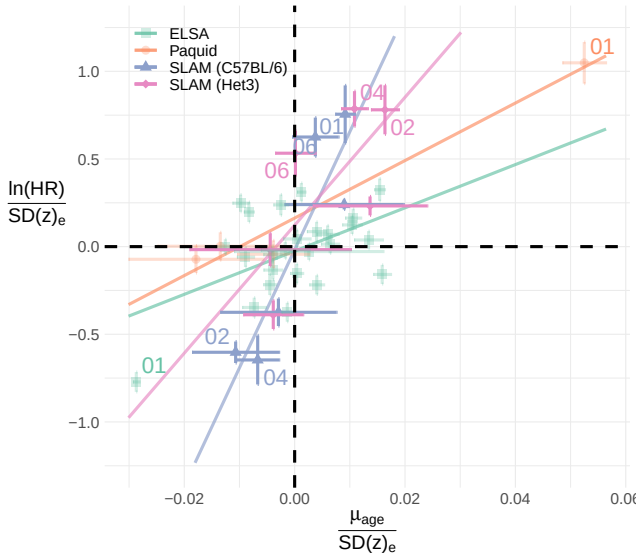

Figure C.14: Allostasis drifts towards the risk direction. We fit a Cox model for each natural variable including age and sex as covariates. The Cox coefficient — i.e. log–hazard ratio (HR) per unit increase — correlates with the steady-state drift rate, μ_{age}. The dominant risk direction for each dataset has been labelled by eigenvalue rank (e.g. z_1 is 01). The equilibrium standard deviation provides a native scale for each variable.

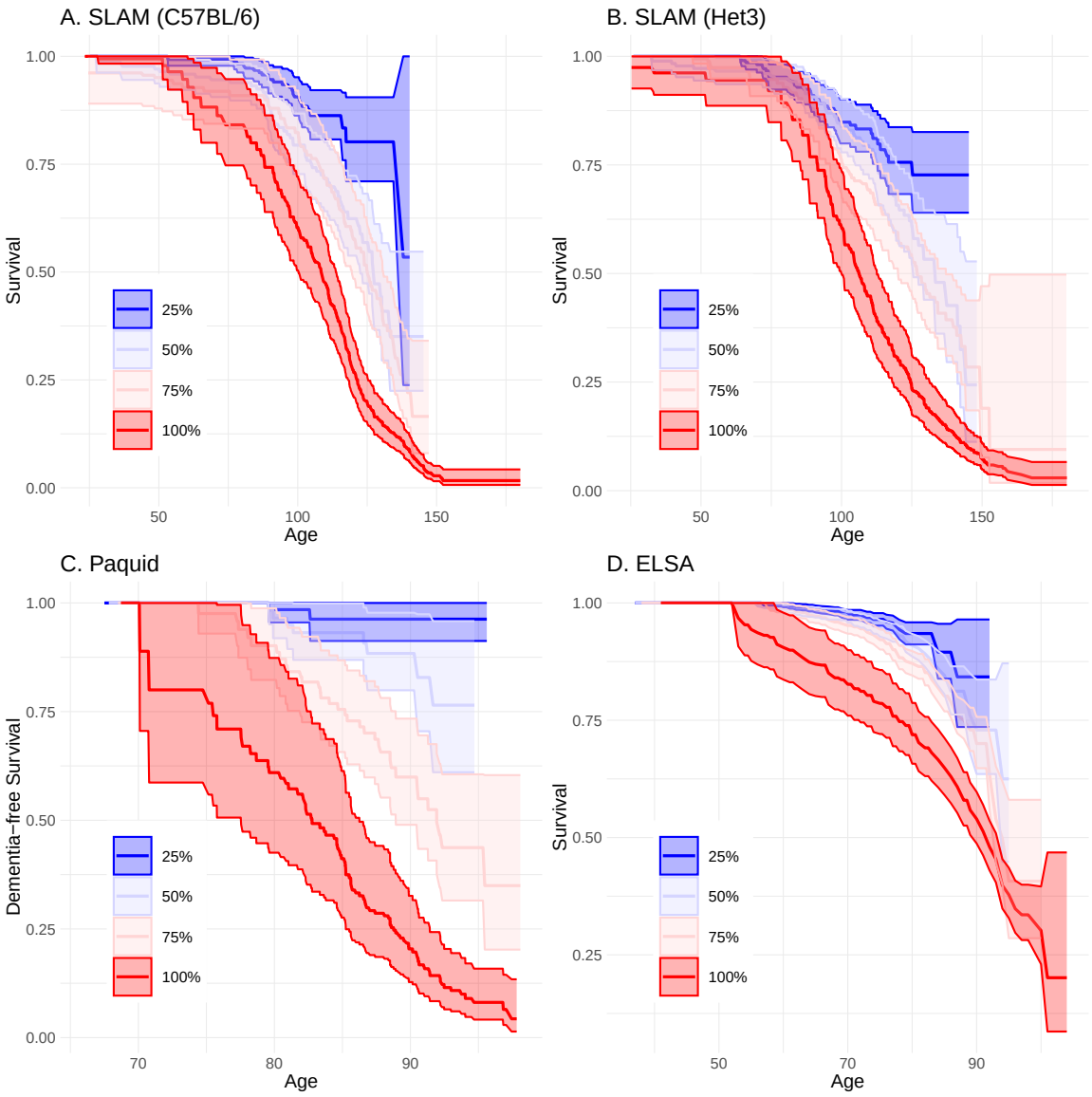

Figure C.15: Composite health measure performance. **A.** C57BL/6 mice (SLAM). **B.** Het3 mice (SLAM). **C.** Paquid (human, dementia). **D.** ELSA (human). A simple estimator of health is $\vec{\mu}_{age}^{T}\vec{z}$. This leverages mallostasis to infer individual health. Large separation between quartiles (colours) indicates a strong predictor of adverse outcome. Fill is 95% confidence interval.

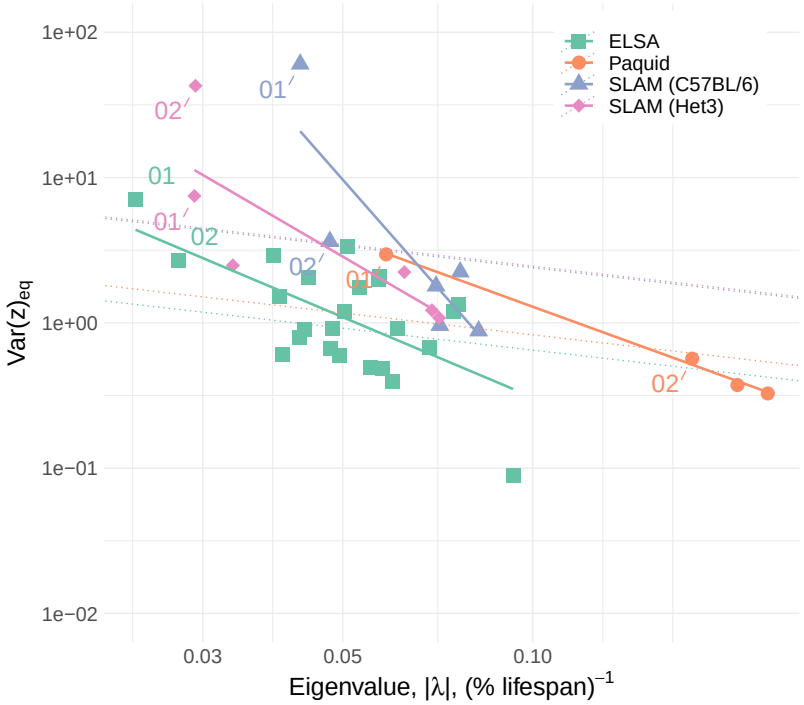

Figure C.16: Equilibrium dispersion is primarily determined by eigenvalue strength, $|\lambda|$ equation (C.48). Smaller eigenvalues are predicted to have larger equilibrium variances. The range of equilibrium variances spans 3 orders of magnitude. The largest variance will drive the observed variation in biomarkers in the steady-state e.g. rank 1 will become principal component 1 (equation (C.49)). Dotted lines illustrate what the equilibrium variance would be if each dimension had the same noise strength, σ^2. The fitted solid lines indicate that the noise makes the smaller eigenvalues even more dominant than expected.

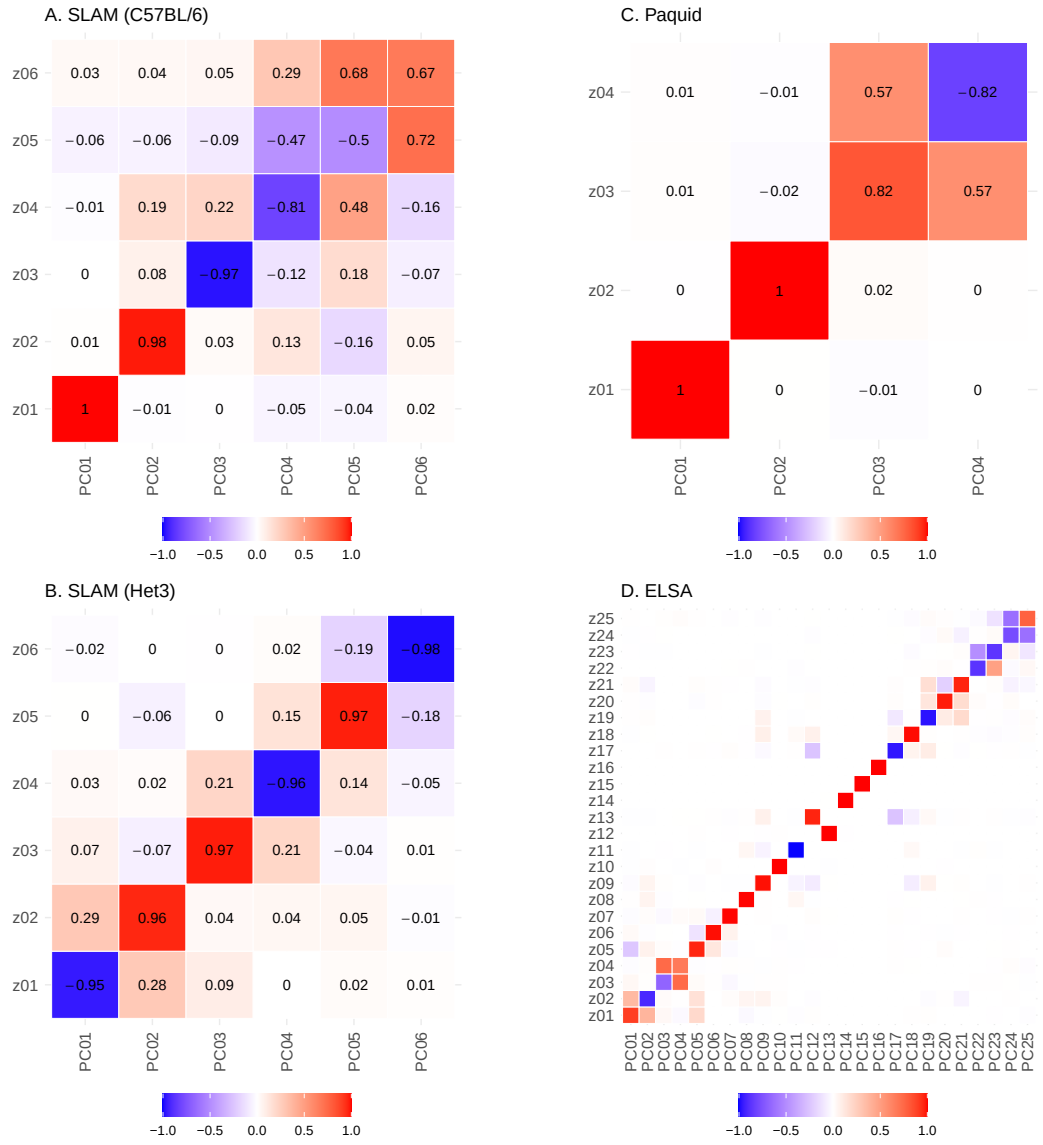

Figure C.17: Principal components are very similar to the natural variables. **A.** C57BL/6 mice (SLAM). **B.** Het3 mice (SLAM). **C.** Paquid (human, dementia). **D.** ELSA (human). Shown are the dot products between the principal component rotation and $\boldsymbol{P}$. The dot product assesses similarity between the transformations ranging from 1: identical, 0: orthogonal, and -1: identical with opposing sign. Identical transformations will generate identical natural variables. If the transformations are identical then all values on the diagonal should be ± 1 (sign is arbitrary[146]). We see that the dot products are often close to ± 1, indicating that the transformations are very close, although they do not perfectly coincide.

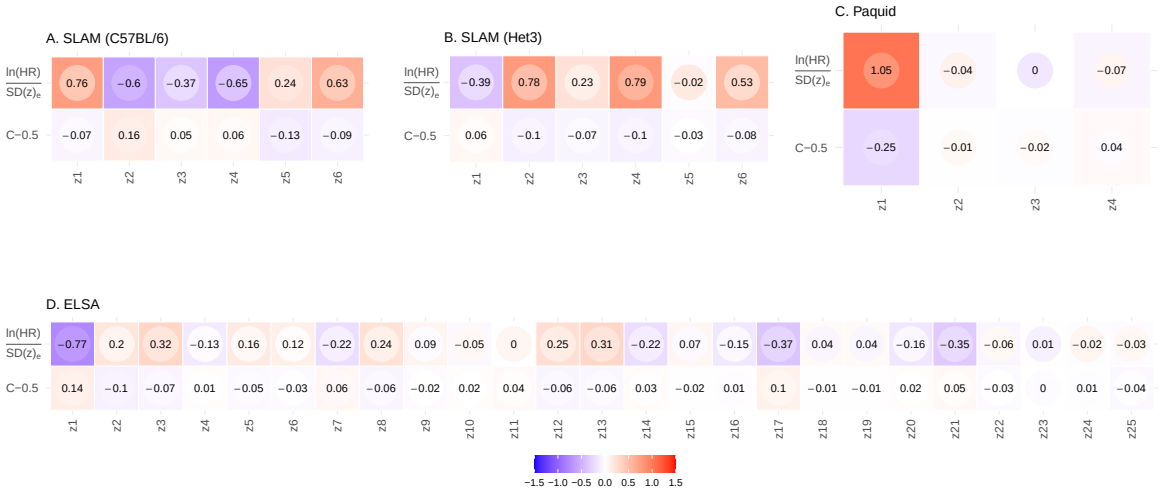

Figure C.18: Survival summary. **A.** C57BL/6 mice (SLAM). **B.** Het3 mice (SLAM). **C.** Paquid (human, dementia). **D.** ELSA (human). For each dataset, the top row corresponds to the Cox coefficient standardized by the equilibrium dispersion ($\ln(HR)/SD(z)_e$) while the bottom is the C-index centered to 0 ($C - 0.5$). A Cox coefficient greater than 0 indicates that higher values are at increase risk and vice versa. A centered C-index greater than 0 indicates that higher values are at reduced risk and vice versa (opposite of the Cox coefficient). The Cox model is conditioned on age and sex (the same as Figure C.14); the C-index is unconditioned. We see that in humans, the first dimension is the dominant determinant in risk of death (ELSA) or dementia (Paquid). It is less clear in mice, where allostatic drift is a better way to identify important survival dimensions (Figure C.14 or Figure 6.4A). Inner colour indicates the limit of 95% confidence interval (CI) closest to zero (non-significant are red on blue or blue on red).