LABORATORY HEALTH MEASURES AND OPTIMAL STRUCTURES FOR AGING

by

Garrett Stubbings

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
August 2021

Dedicated to any poor soul working on network optimization problems.

Table of Contents

Abstract

Aging in biological systems is complex. There is agreement that aging is a gradual accumulation of dysregulation and damage in the organism. However, there is little consensus on how to effectively measure this health state at any given time, let alone how to model the propagation or accumulation of poor health. The frailty index (FI), which measures the fraction of measured health aspects in a damaged state, is an effective measure of general health in an organism. The advantages of the FI are that it is largely insensitive to which health aspects are included and does not impose any structure on the accumulation or propagation of damage. In this work I aim to effectively include health aspects not typically included in the FI such as blood test results and other measurements not naturally dichotomized into a healthy or unhealthy state. The goal of this exercise being to extend the applicability of the FI to lower and more detailed levels of health. Following the practical work on measuring health on lower scales, I investigate how the levels of health interact in a network model of human aging. Through an optimization framework I investigate how the levels of health are arranged for maximizing lifespan and healthspan.

Acknowledgements

I'd like to thank everyone who has helped me along the way.

Chapter 1

Introduction

1.1 Health and aging

1.1.1 Aging as Deteriorating Health

The aging process, and eventual mortality, is a fundamental problems spanning all aspects of human thought. We associate health with the aging process. However, the relationship between health and aging must be quantified to study it properly. A promising approach is to assume that "healthy" aging is the state where the fewest aspects of health have deteriorated. Individuals who age gracefully have very little deterioration over a large span of time, continuing to be physically active and young in appearance longer than their peers. Operationalizing this intuitive understanding of aging has generally been approached by selecting a set of aspects of health, and monitoring them for deterioration or degradation over time [1, 2, 3]. This approach defines aging as a function of the health state, typically trending downward over time. This approach matches up well with intuition when done correctly because an individual aging gracefully can "look" younger than their chronological age by an appropriate measure, which can approximately match a casual assessment.

1.1.2 The Many Levels of Health

The main difficulty in applying a deterioration of health model of aging is establishing which aspects of health are important to measure. At the highest level, functional impairments and visible deterioration of the health state are most clearly associated with aging [2, 1]. However, these high-level impairments may be symptoms of deterioration of lower-level health attributes, things which must be measured using lab assays [4, 5]. Recently, there has been a push towards investigating the lowest levels of health by measuring health on the molecular level [6, 7, 3]. Furthermore, there is the question of how genetic and other hereditary factors affect the aging process

[8, 9]. The question of which level of health is most suitable to making predictions and planning interventions is a difficult one. It may be necessary to measure many levels of health, and to relate them to each other and to mortality.

The many levels of health are certainly associated in some manner [10, 11]. Thinking about them as levels of health implies a hierarchy of some combination of importance and precedence. However, quantitatively describing these relationships (or even determining their importance) has proven to be a challenging task [12]. Although the lower-level health aspects may precede the higher levels in flagging degradation, they are not as clearly linked to decline in health [10]. Whereas, the high-level deficits are intuitively linked to a decline in health.

1.1.3 Accruement of Acute Ailments - Summary Health?

At the very least there is consensus that the many aspects of health tend towards dysregulation and the aggregation of this dysregulation represents some aspect of the aging process [13, 14, 15]. In theory, proper aggregation of dysregulation across all levels of health would represent the aging process to the fullest extent in this framework. However, the question would remain of how to properly represent this information; can it be effectively compressed into one, or a small number, of interpretable health metrics?

Integrating any set of health aspects into a comprehensive health measurement presents a breadth of issues both theoretically and practically. Broadly, the approach has been to take an assay of measurements (often related in some way) and to use a linear combination and some transformation to predict negative outcomes. Common approaches include simply counting up the number of symptoms or precursors, linear combinations of measurements regressed against a scalar metric such as age, and logistic regression of measurements against categorical outcomes [16, 1, 17, 2].

Many approaches are founded on an assumption that there is a mechanism that drives aging (or accompanies it) and that the summary health metric they develop appropriately represents the contribution of that mechanism. For example, allostatic load suggests that the accumulation of stress drives dysregulation and always contains measurements relevant to neuroendocrine and immune responses [17]. More recently the focus has shifted to a less principled selection of health measurements with the

reasoning that a general health metric requires a breadth of included variables. Metrics which aim to include this breadth of measurements include the Frailty Index (FI) [1, 18] which counts the number of impairments across an arbitrarily broad set of measurements and the various biological clock approaches which select from an immense number of measurements using statistical methods [7, 6, 16].

1.1.4 Negative health outcomes

The aging process appears very linear at a glance since it is so intuitively tied to time. Famously, the Gompertz law of human mortality states that the mortality rate increases exponential with increasing age [19]. However, there are a variety of pathways to negative health outcomes including death [20, 21]. The question of where to aim composite metrics of health reflects much of the discussion on what level of health to measure. Is it best to predict outcomes or symptoms that are precursors to larger health events, or is it better to prioritize the risk of the larger outcome itself? The many different paths towards any individual negative health outcome suggests that predicting precursors may not be the best approach. However, the level at which interventions can be successful may be well before the large scale outcomes occur.

Overall, there is an impression that ultimately any adverse health outcome is a step towards mortality [22]. In this sense it seems like the obvious approach is to develop a composite measure which is an effective predictor of mortality. Although an accurate estimation of mortality risk or lifespan prediction would inform decision making, it is not useful from an intervention standpoint. Negative health outcomes and symptoms of dysregulation are treatable acutely. I believe the information useful for a clinician would be a metric describing the health state as it pertains to acute risk which is treatable. For example, using measurements to predict which pathway presents the greatest overall health risk and can be effectively treated.

1.1.5 Multidimensional Risk - Multidimensional Metrics?

The difficulty posed by having many distinct negative health outcomes (or pathways to negative health outcomes) is one that is not commonly addressed in composite health metrics. In part this is because age by itself does such a good job predicting risk within populations - seen later in this thesis. Intuitively, if one could assemble a

health metric which is effectively biological age (controlled for health state somehow) then the problem would be solved. However, biological ages (or equivalent) are more often than not tied to a narrow aspect of aging (or at least not the entire spectrum) [3]. Typically the relationship with health and aging is determined by the set of measurements used in the health measure, for instance which tissue was sampled for the analysis. However, even in the case of a broad spectrum clock which samples a wide range of health aspects tied to aging, is representing health as a scalar the most useful approach?

Given that the many different epigenetic clocks and other composite health metrics are not strongly correlated [23], would it not make sense to play towards each of their strengths? If each measurement has a particular advantage in predicting a certain class of outcomes, why flatten them into a worse metric of a more general outcome? A higher dimensional approach would better leverage the differences between approaches and assays. However, interpreting a high dimensional health metric could be difficult. A set of probabilities for experiencing distinct sets of negative health outcomes would be practical in a decision making setting but would be difficult from a population health or communication perspective. Alternatively there could be some mapping from probabilities to biological age metrics, for instance having a heart that looks 56 and lungs that look 34 would be more communicable. However, additional steps in generating metrics can obscure what is being measured.

1.1.6 Arguments for Interpretatibility

One aspect of composite metrics of health that can easily be overlooked with the proliferation of machine learning black-box approaches is the interpretability of the metric itself [24]. Without sufficient interpretability, it is unlikely that they provide any use either from a clinical or public health perspective. While a high-dimensional output such as a an array of probabilities of negative health outcomes could be beneficial from a decision making perspective, it does not clearly indicate the health state of an individual.

1.2 The Frailty Index Approach

The frailty index (FI) is one approach to composite measures of health which avoids much of the complication seen in other measures [1, 18].The FI is defined as the fraction of measured health aspects in an unhealthy state:

$$FI = \sum \frac{d_i}{N}. \tag{1.1}$$

Candidate health variables for the FI must only satisfy a handful of conditions [18]; prevalence of the health deficit - the unhealthy state of the health aspect - must increase with age after controlling for biases such as survivorship and selection, deficits must be related to health, and prevalence of the deficit must not saturate. The FI effectively predicts negative health outcomes including mortality with a wide range of included elements [25, 26, 4].

1.2.1 FI Against the field

Another advantage of the FI is that it is interpretable at all scales. For instance, communicating that an individual has about 30% of health attributes in an unhealthy state applies equally well to functional impairments as to blood work. These different scales of health or types of measurements are not necessarily equivalent and the behaviour of the resulting FIs will reflect these differences, but the interpretation remains the same.

This leads to the benefits of the FI from a modelling perspective: the FI is both manifestation and propagation agnostic. It does not matter what deficits are included or how they interact - so long as the accumulation of the damage represents a deterioration in health. The only limitation the FI puts on models is that health and aging is represented by a set of health attributes that can be in either a healthy or unhealthy state, with some type of damage propagation and relationship with mortality. The difference between FIs at different levels of health implies that some aspects of health behave differently than others in aggregate, but not specifically.

1.2.2 Pushing on FI / deficits in lower scales

The ability of the FI to capture all levels of health has not been rigorously tested empirically. Although the FI has been successfully extended to include laboratory-level

health deficits (FI-Lab) [27, 4, 28] there are concerns about the details of the implementation [11]. Measurements at the lower scales are typically continuous valued and therefore do not fit neatly into the dichotomous or categorical health deficit structure of the FI. Previous work used methods from acute care to transform these continuous measures into dichotomous health attributes using standard reference ranges [27, 28]. However, this approach raises some concerns; does this method match the definition of a valid deficit [18], can this method be used for measurements without precedent in acute care, is dichotomization of these deficits reasonable from a statistical perspective since they are not naturally binary? In chapter 2 of this thesis I address these issues and present possible solutions.

1.3 Health and aging: High Dimensional and Complex

The many pathways to a range of negative health outcomes and the large overlap between individual measurements with a variety of outcomes suggests that the problem of aging is inherently multidimensional and complex. The many interactions of health aspects leading to a large variety of possible health outcomes suggests a network structure of interactions. In general this means that not only does any particular health aspect affect the aging process, but that the contribution of many health aspects towards negative outcomes is more than the sum of their individual contributions. The idea of organism level health being a network of interconnected components is an idea which has been leveraged both qualitatively and quantitatively [29, 12].

1.3.0.1 Qualitative and Model Networks

Networks have proven to be a useful tool for developing an intuition about aging. One such qualitative model is the work done on the reliability theory of aging [15]. In this model health aspects are represented by nodes in a sequential network - much like an electrical circuit. Health attributes are combined in both series and parallel so there can be multiple pathways through the network to maintain successful function. Health aspects damage randomly and mortality occurs when there is no longer a complete path through the network. Modelling health in this way successfully recovered Gompertz law of human mortality [19]. Furthermore, this model provoked thought about how different aspects of human health might interact and how failure in biological systems compared to failure in mechanical systems. The reliability theory of aging is a successful model because it provided new insight in how to think about aging. However, the model is not directly useful when studying health and aging, since it does not capture the health of individuals over time.

1.3.0.2 Machine Learning Style Networks

On the quantitative side there are models which aim to correctly model the dynamics of health aspects over time, including up until mortality. In general these models assume that health aspects follow some trajectory over time, vary stochastically based on outside factors, and are influenced by the other health aspects in the system.

The goal of these approaches being to characterize these dynamics and interactions statistically using empirical data. Earlier work on this style approach assumed that deviations from "normal" aging behaviour drove the interactions between health aspects and mortality rates [30]. In that model the network is used to parametrize the strengths of the interactions between health aspects, and the relationships between health aspects and mortality rate have a convenient closed form. Despite the nice form of the model the authors were not able to deploy their model against a wide range of health aspects. However, more recent work has been done which leverages techniques from machine learning to tackle a similar problem [12]. The approach follows the same general principles but does not assume a closed form for much of the dynamics, instead using arbitrary non-linear functions. However, this model does use a closed form for the interactions between health aspects which shows the interaction strengths as a network of interactions. This combination of a high prediction quality machine learning approach with a network framework yields a model which is both useful from an empirical and theoretical perspective. The main issues concerning machine learning approaches have to do with the quality and quantity of available data. To effectively capture human health it would require a vast amount of data spanning the various levels of health. Current approaches are limited in their application by a small number of health variables, which cannot be effectively used to map out a network structure that spans many levels of health.

1.3.1 Generic "FI style" Networks

There is one modelling approach which combines the simplicity and general applicability of the FI with the potential of networks to capture complex behaviour. Dubbed the generic network model (GNM) this approach applies the detail-agnostic approach of the FI to a functional network model [31, 29, 5]. Both the health aspects and the interactions between aspects are identical across all aspects in the network. The health aspects included in the model do not map directly onto real-world health deficits, instead representing some general health deficits. The health and mortality phenomenology in the model are driven by how health aspects are connected and some specific network structures lead to phenomenology which parallels the levels of health discussion above [5].

1.3.2 Optimal Networks + Competing influences

The network structure which best captures human health data is a scale free network [5]. These networks are characterized by a large span of different health aspects, with relatively few aspects connected to a large number of lower-level health aspects [32]. However, it is not obvious why the scale free networks best model human health - do they represent some optimal structuring of health attributes for longevity. Alternatively, do they maximize the healthspan of the organism, where healthspan is vaguely defined as the period of highest function or good health [33, 34]. Supposing that the mapping of network structure to the levels of health discussion holds in general, what would an optimal network structure look like? For instance, would the higher level elements connect to overlapping sets of lower level aspects? Does a large variety in the levels of health help in terms of lifespan or healthspan? Are there competing factors which promote wide-spanning interactions between health aspects? We aim to answer some of these questions by optimizing the network structure for lifespan and healthspan.

1.4 Thesis Organization

1.4.1 Preview of Chapter 2

Chapter 2 of this thesis is concerned with pushing the limits of the FI in the case of continuous health deficits. The goal of this chapter is to verify and improve FI created using lab-based biomarkers using generic techniques. Contained in this chapter are two papers. The first of which is a paper published in the journal BioGerontology [11]. This publication builds upon work done in the summer preceding this masters work. However, the first 3 months of this masters work was spent refining and expanding both the analysis and the content of the paper. All analysis in the paper was done by me with the exception of some python functions which I adapted from work done by Spencer Farrell (rewriting standard functions to match my data formatting). The paper is primarily written by me with guidance from all coauthors, particularly Dr. Andrew Rutenberg. Editing contributions were made by all coauthors either directly or through implemented feedback. The second inclusion is a paper currently undergoing reviewer feedback changes for publication in Mechanisms of Aging and

Development. All analysis in this paper was performed by me. The paper was originally written by me with guidance from coauthors and editing and revision by Dr. Andrew Rutenberg. Code for analysis can be found on my github [35].

1.4.2 Preview of Chapter 3

Chapter 3 of this thesis contains the ongoing work on the GNM to determine how scale-free networks and the previous results of the GNM fits into an optimal health space in the model. Work on this project has been ongoing throughout the thesis. However, the previous approaches to the problem - which are not mentioned here - suffered from a host of problems both methodologically and computationally. The results presented in chapter 3 of this work are from our most recent approach to the optimization problem, which is looking promising. The results presented in this work largely match those found using other methods, but the current approach is better suited to tackling larger networks with more robust approaches. That being said, the results presented in this work summarize our findings so far, but will need to be expanded upon to fully answer the questions raised. All analysis in this chapter is performed by myself. The project has proceeded under the supervision and guidance of Dr. Andrew Rutenberg. All written work in this chapter is my own.

Chapter 2

FI-Lab

2.1 FI-Lab: Broad Context

Health and aging are multidimensional problems; there is a large variety of health aspects to measure and a large number of potential health outcomes. While we are generally well connected with measuring adverse health outcomes, measuring the wide variety of health aspects is challenging. Health measurements are often classified into a hierarchical "levels of health" structure, where the highest level represents function, and descending down the levels are standard laboratory assays, then molecular and genetic measurements further along [10, 36]. The different levels of health pose unique challenges from a measurement perspective and are often summarized using different techniques and approaches [2, 17, 1, 3]. The separation of levels of health and measurement techniques poses a problem when trying to model the organism as a whole. However, one health metric - the FI - is promising for application across multiple levels of health due to the broad inclusion of health aspects [18]. In this chapter we aim to rigorously adapt the FI to include lower levels of health measurements.

The first step down in the levels of health discussion is to go from functional and other clinically available deficits (visible aging) to lab-based measurements such as blood-work and other biomarker assays. Canonically these types of measurements are referred to as lab measures or biomarkers and consist of any set of continuous measurements that can be obtained through some sort of routine assay a clinician could request. There are a handful of biomarkers which have gained popularity in aging literature and occasionally popular science [37, 38, 39, 40]. Pushing to lower levels of health with biomarkers not in the commonly assigned assays has become popular in the aging field but is a level beyond what is considered in clinical practice [41]. We begin with the common biomarkers seen in bloodwork because they have been successfully used as health signals previously.

In the acute care setting the established approach is to flag when these biomarkers

are outside of established reference ranges [42]. These reference ranges have been directly used in the context of the FI adapted to biomarker measurements (FI-Lab) to dichotomize (or binarize) the labs into health deficits [27, 28]. However, it is not obvious that this approach is best suited for the FI-Lab. In the definition of the FI [18] deficits are described as being anything health-related which increases in prevalence with age (barring selection bias). Does it make sense to include deficits in FI-Lab whose prevalence and aging trends are artifacts of another application of the same measurements? Furthermore, with the additional information provided by having continuous measurements, does it make sense to define a deficit using a falsely dichotomized version of the deficit? In the following work two approaches to pre-processing biomarker data for inclusion in the FI-Lab are presented which argue against using clinical thresholds (Section 2.2). Indeed, we can build on this approach to develop FI-Lab measures that do not require dichotomization at all (Section 2.3). Together these approaches provide the tools for building FI measures from emerging sources of data where there are no preexisting clinical guidelines for treatment.

2.2 FI-Lab with Generic Binary Deficits

In this work we investigate the value of directly applying the Searle [18] definition of a deficit onto biomarker measurements to dichotomize them into health deficits. Biomarkers relevant to aging trend in some direction with age. We assume the direction of that trend is the primary risk direction for that measurement. Dichotomizing the measurement at some threshold such that measurements beyond the threshold in the risk direction are considered health deficits ensures that the deficit will increase in prevalence with age.

2.2.1 The Paper

2.2.1.1 Credits to authors + adaptations

Below is a manuscript primarily authored by myself with the guidance of Dr. Andrew Rutenberg [11]. Dr. Rutenberg, Dr. Arnold Mitnitski, Dr. Kenneth Rockwood, and Spencer Farrell all contributed significantly to the manuscript with either direct editing or feedback to the manuscript or during meetings. Some of the code used

for analysis in this work was adapted from work originally done by Spencer Farrell, the majority of the code was written by me. The manuscript is included in full with the exception of the abstract. The formatting of the figures, equations, and citations have been adapted for this document.

2.2.2 Introduction

Poor health is often associated with aging, a decrease in functional capacity, and an increased susceptibility to illness and injury. While chronological age is a convenient proxy for aging, it cannot capture individual variability of health at a given age. The frailty index (FI) is a well-tested way of incorporating large and varied aspects of health and function that can be easily used to differentiate between individuals of the same age. Defined as the fraction of selected health attributes that are in an unhealthy state (called deficits), the FI has been shown to be a robust measure of individual health over the aging process [1, 18]. The FI is observed to increase with age and the distribution of FI on a population level broadens with increasing age, describing the heterogeneity of aging [43]. The FI is predictive of mortality and of other adverse health outcomes [26, 25].

The health attributes considered in the FI are typically clinically observable or self-reported, such as disabilities in activities of daily living or physical or cognitive impairments [18]. Alternatively, standard laboratory measurements such as blood and urine biomarkers [28, 4, 27] as well as biomarkers of cellular senescence and oxidative stress [4] can be used to create a laboratory-test based FI known as FI-Lab. Cutpoints are used to binarize the quantitative biomarker measurements into deficits so that they can be naturally included in an FI. Normal reference ranges based on diagnostic or therapeutic utility [42] are commonly used as cutpoints.

Since the FI is an aggregate measure and is not used for the diagnosis or treatment of specific conditions, standard cutpoints are not necessarily best suited to its role of predicting risks. Furthermore, standard cutpoints are often not available for emerging biomarker measurements such as in epigenetic, proteomic, metabolomic, or other high-throughput "omics" approaches. Alternative "data-based" methods obtain cutpoints from the available data under consideration. Both normal reference ranges [28, 27] and data-based methods [4] can and have been used to create an effective

FI-Lab.

One data-based method of biomarker binarization is to select cutpoints to maximize some predictive aspect of the post-binarized biomarker. For example, cutpoints can be selected to maximize the difference between survival curves of people that are on either side of the cutpoint for each biomarker [4]. Equivalently, other predictive measures such as receiver operator characteristics (ROC) performance or mutual information [29] could be used with respect to a particular outcome such as mortality within 5-years to generate "optimal" cutpoints. While attractive in principle, such individual biomarker optimization approaches run the risk of creating FI that are overly specific to the study cohort and not generally applicable for other cohorts.

Another popular data-based method for binarizing continuous-valued data is to select cutpoints based on the quantile of the population. This approach is used in both the Fried frailty phenotype [2] (with quintiles) and in the exploration of the allostatic load theory of physiological disregulation [44, 17] (with quartiles). Here, a risk direction is chosen for each biomarker, e.g. by how the biomarker changes with age, and the cutpoint is selected for each biomarker by the quantile of that biomarker – i.e. the fraction of the population that has values of the biomarker above the cutpoint. This approach should be less susceptible to overfitting, since the quantile is chosen globally for all biomarkers rather than individually for each biomarker. Nevertheless, it raises the question of how to choose the best quantile and of how sensitively the results depend upon the quantile chosen. Investigation of allostatic load [44] found that deciles and quartiles behaved similarly, implying that the quantile approach may be robust with respect to choice of quantile. Nevertheless, no systematic investigation of the quantile approach in the context of the FI has been done before.

A systematic investigation of data-based approaches for the binarization of continuous-valued biomarkers used in the evaluation of the FI can explore the questions of overfitting due to optimization raised above. At the same time, we can examine the robustness (or insensitivity) of the FI as a predictive measure of health outcomes or mortality and the robustness of the FI maximum seen in observational studies of aging [18, 4], with respect to the details of any binarization approach. Robust and validated data-based approaches to binarization will facilitate the future development of FI for high-throughput 'omics data and for more model organisms of aging.

Here, we examine the effectiveness of data-based binarization schemes for building the FI from biomarker data. We use both the NHANES and CSHA data sets to check whether cohort effects are large; we find that they are not. We examine overfitting effects with cross-validation, and find that they are present when optimal cutpoints are chosen for each biomarker – but that they are small when global cutpoints are chosen for all biomarkers. We compare the predictive performance of data-based schemes against earlier published results, and find that the data-based schemes have comparable or slightly higher predictive value than the established FI with respect to predicting mortality and clinical deficits. Overall, we find that a generic quantile data-based binarization approach performs well.

A key characteristic of the FI is the relatively insensitivity [18, 4] to the particular choice of deficits. We show that this also holds for choosing cutpoints for FI-Lab, and we find that a broad range of cutpoints exist where the quantile binarized FI-Lab is effective. This demonstrates both the universality of the FI and the generality of our method of choosing cutpoints. Nevertheless, we identify the best range of quantiles to use and we find that they overlap with the quintiles used in the Fried frailty phenotype [2]. Furthermore, many aspects of the FI calculated at these quantiles such as maximum, minimum and overall distribution of FI in the population overlap with results from previous FI-Lab studies.

2.2.3 Methods

Data, evaluation, and cross-validation

The data used in this study are from the National Health and Nutrition Examination Study (NHANES) [45] and the Canadian Study of Health and Aging (CSHA) [46]. The NHANES data set consists of the 8881 individuals from the NHANES study with data for at least 11 of the 16 available biomarkers. This sample has an age range of 20 to 85. The data used from the CSHA study has 973 individuals aged 65+ for which data is available for at least 16 of the 22 biomarkers. Age distributions for these data sets are shown in supplemental Fig. A.1, which highlights the smaller and older cohort of the CSHA study.

Both of these data sets have previously been used to construct FI-Lab. Blodgett *et. al* [28] considered the NHANES data set, while Howlett *et. al* [27] considered

the CSHA. We will compare our results with both of these in this paper. Since a much larger sample size and a much larger range of ages are available, we focus on the NHANES data set. However, major results will be also validated in the CSHA data set. Both studies' FI-Lab consist of many shared deficits and cutpoints, so the differences in FI between the data sets are likely due to cohort effects. These two data sets have very different cohorts, so by applying our methods to both we test the generalizability of our approach.

The NHANES and CSHA cohorts differ in more than just age. In supplemental Fig. A.2 we show the distribution of FI-Lab for the CSHA cohort (white bars) together with a resampled NHANES cohort with the same (65-85 years) age distribution as the CSHA (blue bars). We see that the NHANES cohort has a significantly lower FI-Lab at the same age, i.e. it represents somewhat healthier individuals. This could be due to a large portion of the CSHA population being comprised of institutionalized individuals [27].

The purpose of binarizing data is to construct an FI. The FI is intended to be an inclusive and general indicator of individual health; it has been shown to correlate well with mortality [47] but also with institutionalization [48], postoperative complications [49], dementia [50], recovery time in hospital [51], and other adverse health outcomes [52]. Accordingly, we compare our newly constructed FI with the existing FI-Lab in their ability to predict 5 year mortality as well as by their ability to predict clinical outcomes from laboratory data. To evaluate prediction, we use the standard area under the curve (AUC) of the ROC curve. We obtain similar results using mutual information [29], as illustrated in supplemental Fig. A.3. We also check that our new FI behave similarly to the previously published FI-Lab, with respect the clinical FI, with respect to their distributions, and with respect the maximal observed FI in the population.

Each new FI is tested using cross-validation. Cutpoints are generated using a random half of the population, then those cutpoints are applied to the other half and the resulting FI are evaluated. This is repeated 100 times. Cross-validation allows us to characterize any over-fitting of cutpoints.

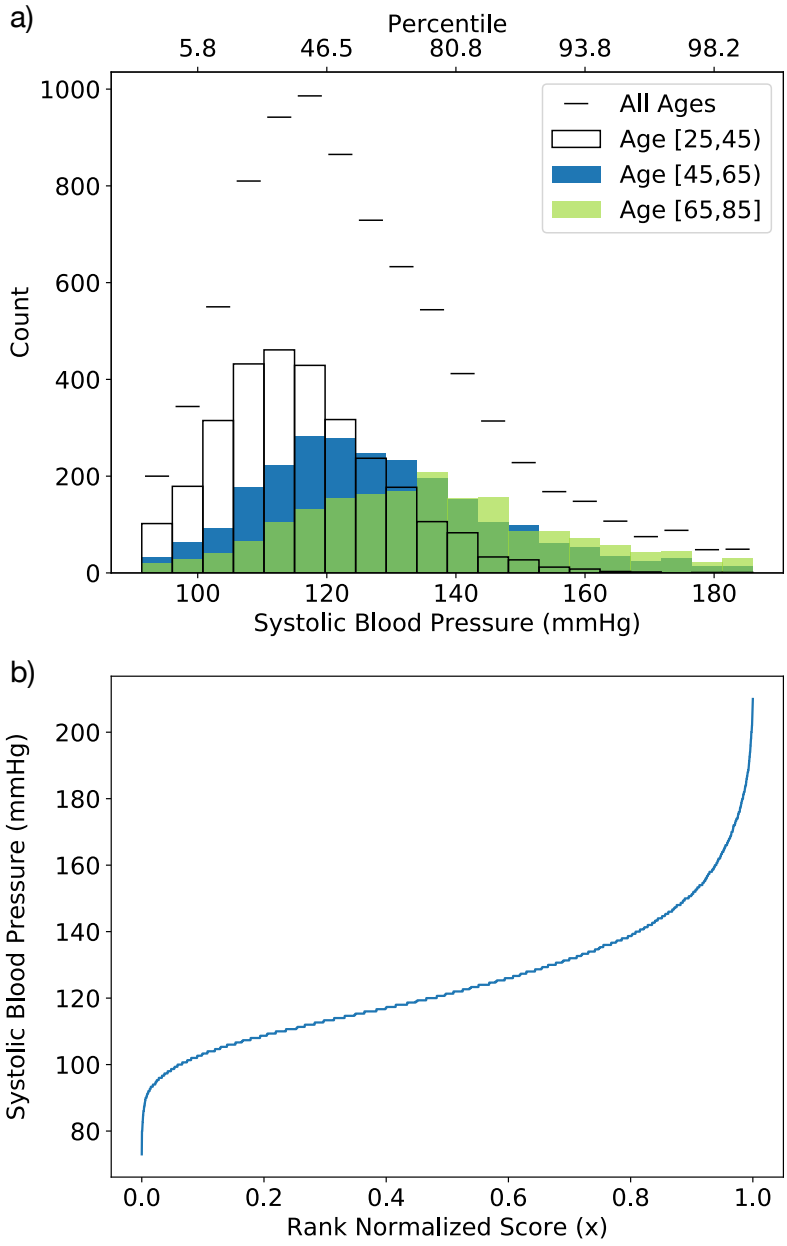

Figure 2.1: a) The distributions of systolic blood pressure measurements in the NHANES cohort [45]. Short horizontal lines indicate the whole population distribution, while unfilled, orange, and blue bars show the youngest [25, 45), middle [45, 65), and oldest [65, 85] age groups respectively (in years). The trend during aging is an upward shift of blood pressure. b) The rank normalized score x vs the corresponding systolic blood pressures. The median corresponds to $x = 0.5$. For this and other measures, the nonlinear mapping between x and corresponding value is always monotonic – but is either increasing or decreasing depending on the direction of risk.

Quantile-based cutpoints

We transform the biomarker data to a dimensionless form using quantiles. For each individual subject, each biomarker i is transformed to the proportion x_i of the population that has "less risky" values. This is illustrated in Fig. 2.1 for systolic blood pressure. If an individual has a value of 140 mmHg, which places them at the upper quintile of risk for blood pressure, their systolic blood pressure score is transformed to $x = 0.8$, corresponding to having a higher systolic blood pressure than 80% of the population.

Quantiles are implemented on a population scale by performing a rank normalization of the data, where each biomarker is sorted in ascending risk, then the ranks (position in the sorted list) are divided by the number of individuals. The rank normalized values x_i are given by

$$x_i = \frac{\text{Rank of biomarker } i \text{ in the population}}{\text{Number of individuals in the population}}, \tag{2.1}$$

and so $x_i \in [0, 1]$.

Implementing binarization is straightforward using these quantiles. We apply a global cutpoint (GCP) as a threshold value of the rank normalized values, X_{GCP}, applied identically across all biomarkers. We build the resulting FI as the average over these binarized deficits,

$$FI_{GCP} = \sum_{i=1}^{N} \frac{d_i}{N}, \qquad d_i = \begin{cases} 1 & \text{if } x_i > X_{GCP} \\ 0 & \text{otherwise} \end{cases}, \tag{2.2}$$

where N is the number of measured biomarkers. For each biomarker, a deficit $d_i = 1$ is assigned when x_i is above the threshold in the direction of risk.

Direction of risk

We determine a direction of risk for each biomarker, before applying quantile-based cutpoints. We then binarize with respect to the at-risk direction, as discussed above. We do not assert that biomarkers only have one direction of risk, but we do find that most biomarkers have one direction that is most often explored by the population, and so we assume this is the dominant direction of risk during aging. This is illustrated in the supplemental Fig. A.4.

We prefer a mortality-free approach to determining direction of risk to reduce potential over-fitting. We simply use the aging trends of the biomarkers to determine the risk direction. The relation between age and each biomarker is determined by the sign of Spearman's rank correlation. A positive value indicates the risk direction is towards large values of the biomarker, a negative value indicates risk towards small values. This method is effective at determining risk directions if the population has a reasonably large distribution of ages. Aging trends effectively classify risk direction in both the CSHA (ages 65-104 years) and NHANES (ages 20-85) data sets. However, we restrict the age range for calculating risk directions to ages 35+ to calculate relations based on normal aging behaviour.

Another method of determining risk direction is to use mortality data, or some other adverse health outcome. For each biomarker ROC curves can be generated with respect to the binary outcome (e.g. 5 year mortality) and an AUC can be calculated. An AUC above 0.5 indicates the primary risk direction is towards high values, an AUC below 0.5 indicates risk towards the low end. Equivalently, one could do a logistic regression of the biomarker against an adverse outcome and use the sign of the beta value (positive beta would indicate risk towards high values). This type of approach ensures that the risk directions generate the best FI for predicting that outcome, but they are potentially over-fit to that outcome. We find that risk directions from mortality data are predominantly the same as the aging trend directions. The predictive AUC of the resulting FI with respect to 5 year mortality is also essentially the same as with aging trends, as shown in supplemental Fig. A.5.

We have also considered a simple approach for two-sided cutpoints. For simplicity, we consider symmetric cutpoints with both $x_i > X_{GCP}$ and $x_i < 1 - X_{GCP}$ assigned as deficits with $d_i = 1$. The predictive AUC of the resulting FI is significantly worse than the one-sided approach, as indicated by the supplemental Fig. A.6. Accordingly, we focus on one-sided cutpoints in this paper.

Optimally predictive binarization

In addition to quantile binarization, we also compare with two different FI created with cutpoints selected for optimal prediction with respect to mortality. For both methods we treat the population uniformly; we do not stratify or control for possible

cohort effects such as age or sex. Additional details are provided in the supplemental information.

The first, $FI_{logrank}$, based on the separation of survival curves, has been used to create an FI-Lab [4]. For each biomarker, the cutpoint is found that maximizes the significance of separation between survival curves of individuals with and without the deficit by minimizing the p-value from a logrank test [53].

The second method for generating optimal cutpoints is based on information theory. FI_{info} uses cutpoints selected for the highest possible mutual information with respect to mortality at 5 years. In a manner similar to $FI_{logrank}$ every possible cutpoint is tested for every biomarker and the cutpoints which maximize the mutual information with respect to mortality are selected.

2.2.4 Results

To evaluate the various data-based approaches to binarization, we have calculated the AUC with respect to five-year mortality for both the NHANES and CSHA cohorts. The results are shown in Fig. 2.2. The performance of all measures was qualitatively similar for both the NHANES and CSHA data sets, though due to the smaller cohort the CSHA data showed greater variability in cross validation.

FI_{GCP}, assembled from quantile based global cutpoints, performed well. For all tested values of X_{GCP} the cross-validated and full dataset results agree, indicating minimal overfitting. For the extreme values of X_{GCP} equal to 0 (where all biomarkers are at risk) or 1 (where none are), there is no predictive value of FI_{GCP} and the AUC is equal to 0.5 – as expected. Between these extremes, we see a broad maximum of the AUC. Indeed, for global cutpoints between $0.5 - 0.9$ the FI_{GCP} slightly outperforms the published FI-Lab for both the NHANES and CSHA datasets.

The binarization approaches to maximize the mortality prediction for the full datasets gave comparable AUC values, as indicated by the columns to the right in Fig. 2.2. However, cross-validation of FI_{info} and $FI_{logrank}$ showed significantly lower AUC when compared to the full dataset calculation. The decreased performance in cross validation indicates that these cutpoints have poor out of sample performance and do not represent a generalizable risk threshold. While the cross-validated FI_{info}, using maximum information cutpoints, appears to perform as well as FI_{GCP} – the

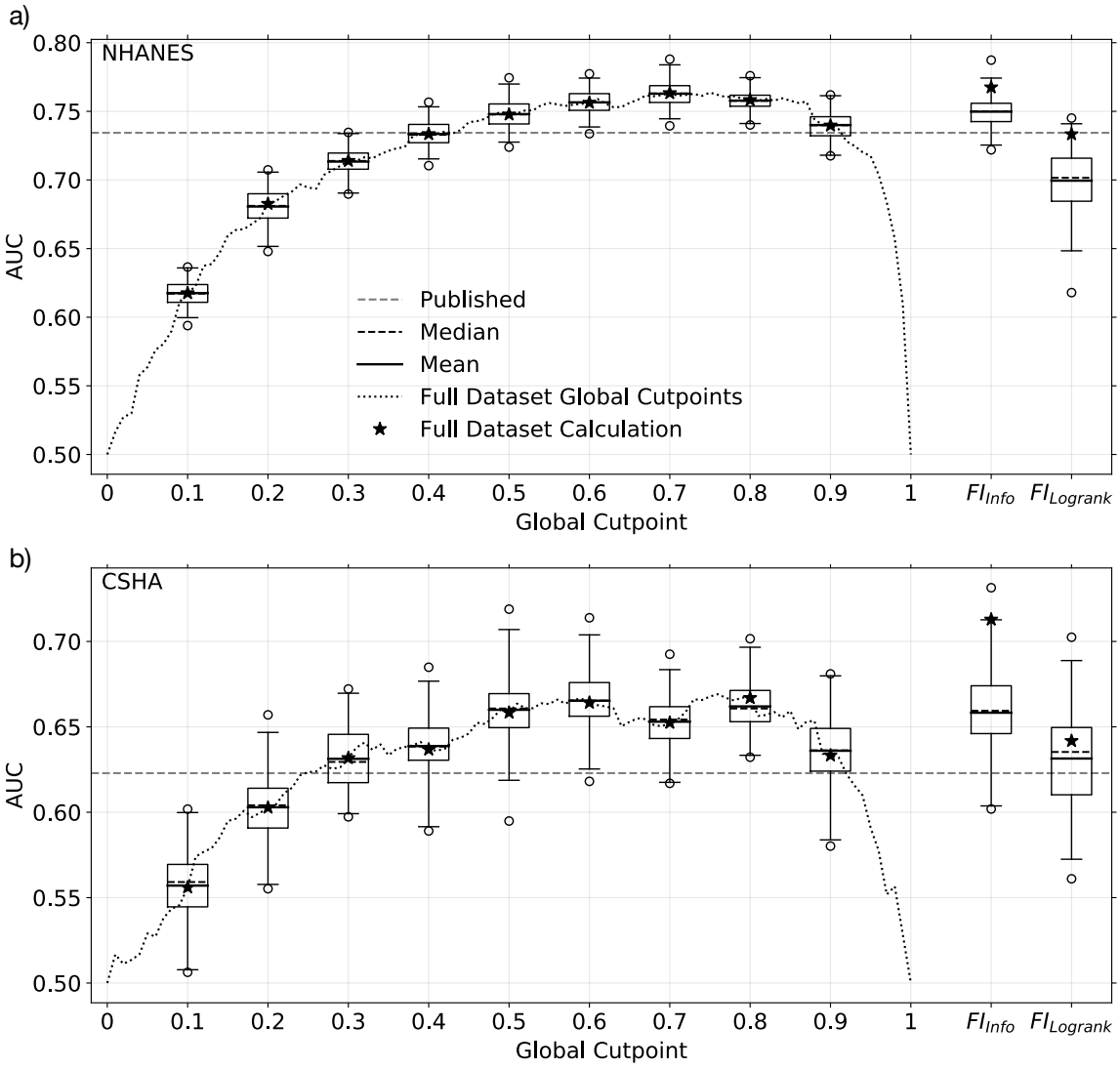

Figure 2.2: Cross-validated AUC of different data-based FI with respect to 5 year mortality for the a) NHANES and b) CSHA cohorts. The dotted line indicates the AUC of the quantile-based FI_{GCP} vs the global cutpoint X_{GCP}. Box and whisker plots display the data from cross-validation: the boxes represent the upper and lower quartiles, the whiskers go to the 99^{th} and 1^{st} percentiles, and the circles are remaining outliers. The short dashed line within each box is the median, the solid line the mean, and the star is the AUC for the full data set without cross validation. The horizontal grey dashed line shows the AUC of the published FI [28, 27]. The rightmost two columns, as indicated, show the AUC for FI_{info} constructed from maximum information cutpoints and $FI_{logrank}$ constructed from logrank minimum p-value cutpoints.

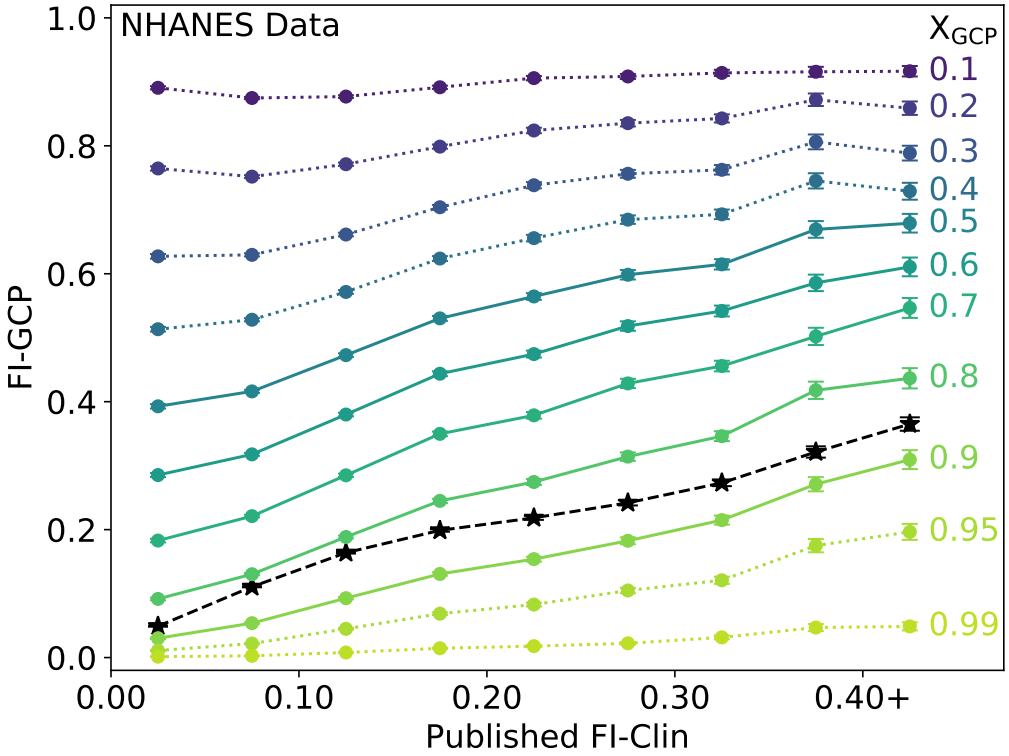

Figure 2.3: Average FI_{GCP} vs published FI-Clin for the NHANES dataset, for a variety of global cutpoints X_{GCP} as indicated by the coloured numbers at the right of each coloured line. The coloured markers indicate the middle of the bins used for averaging. The black dashed lines with stars show the published FI-Lab [28]. The FI_{GCP} lines are dotted when their AUC from Fig. 2.2 is below the published value, while they are solid when it is above.

cross-validated $FI_{logrank}$ does not. Since using individual cutpoints optimized for each biomarker to predict mortality leads to an FI that is prone to overfitting, we believe our quantile cutpoints method will apply more generally to more datasets (especially those with smaller cohorts), and so we focus on quantile cutpoints for the remainder of this paper.

We were surprised that the quantile-based cutpoints performed similarly well for both the NHANES and CSHA datasets, despite their significantly different age, health, and cohort sizes. Since quantile-based cutpoints are extracted from the cohorts being characterized, we wanted to investigate cohort effects more directly. Since the NHANES dataset has a large population and a large range of ages, we obtained

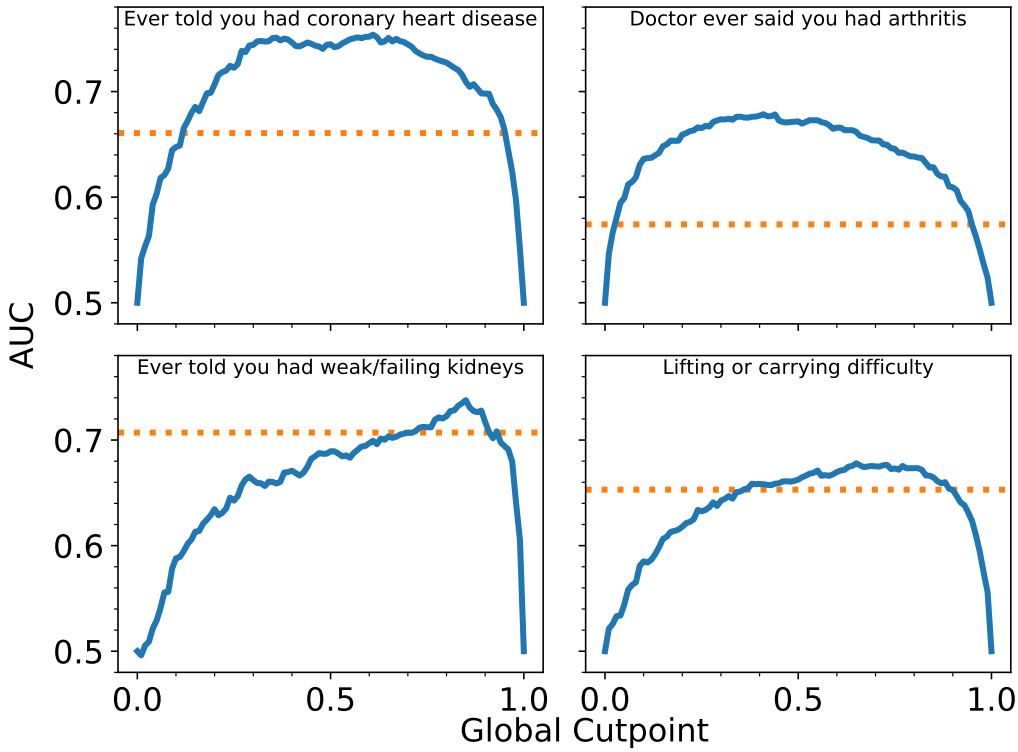

Figure 2.4: The blue lines indicate AUC of FI_{GCP} vs the global cutpoint X_{GCP} for four clinically observable deficits in the NHANES study. The horizontal dashed orange lines indicate the AUC from published FI-Lab [28]. FI_{GCP} performs at least as well as FI-Lab, although the range of cutpoints which are most effective varies. Similar plots for all clinical deficits are shown in supplemental Fig. A.8 for NHANES and Fig. A.9 for CSHA.

quantile-based cutpoints from sub-cohorts of NHANES for young (25-45), middle (45-65), or old (65-85) age groups. Remarkably, the cutpoints obtained from any one sub-cohort worked reasonably well applied to any other cohort. However, the range of X_{GCP} for best prediction decreased and shifted closer to 1 as shown in supplemental Fig. A.7. This supports our observation that cohort effects are not large with quantile-based cutpoints.

A crucial test of FI-Lab behavior is how well it corresponds to an established FI-Clin. The coloured lines in Fig. 2.3 shows average FI_{GCP} values binned by their corresponding FI-Clin values. For intermediate values of X_{GCP}, we see that FI_{GCP} is monotonically increasing with FI-Clin. Indeed, the published FI-Lab appears to

correspond to X_{GCP} values between 0.8 and 0.9 – where the FI_{GCP} also performs well with respect to both AUC and cohort effects. Conversely, for much larger or smaller values of X_{GCP}, where the AUC is significantly worse than for the published FI-Lab, we see that the FI_{GCP} is not strongly dependent on FI-Clin or even becomes non-monotonic.

We can test the versatility of the FI by its ability to predict outcomes other than mortality. In Fig. 2.4 we evaluate the prediction of four binary clinical deficits, where the blue lines indicate the AUC for FI_{GCP} vs the global cutpoint X_{GCP}. The corresponding AUC of the published FI-Lab [28] is indicated by the horizontal orange lines. We see that FI_{GCP} is as good as FI-Lab for a range of cutpoints – approximately where mortality prediction also performs best. (All clinical deficits are tested in supplemental Fig. A.8 for NHANES and Fig. A.9 for CSHA.).

We illustrate the distribution of FI_{GCP} in supplemental Fig. A.10 for $X_{GCP} = 0.85$ and for $X_{GCP} = 0.4$. Both perform as well as FI-Lab in terms of predicting mortality (see Fig. 2.2). However they have very different distributions when using the same NHANES population. While $X_{GCP} = 0.85$ has a similar distribution as the published FI-Lab, $X_{GCP} = 0.4$ leads to significantly higher FI values. While this is not unexpected, since the extreme value of $X_{GCP} = 0.0$ would lead to all FI being equal to 1, it does lead us to systematically examine the upper and lower limits of FI. In Fig. 2.5 we show the upper (light blue) and lower (dark blue) 1% of the FI_{GCP} distributions in the NHANES dataset vs X_{GCP}. We see that as X_{GCP} increases both the maximum and the minimum FI_{GCP} decrease. For $X_{GCP}0.7$ the minimum is zero. For $X_{GCP} = 0.85$ the range of maximal FI approximately corresponds to the range observed for the published FI-Lab (indicated in red, and labeled "Blodgett"). We also show that the 1^{st} and 99^{th} percentiles of FI_{GCP} in the CSHA dataset (black dashed lines) are similar to those of the NHANES dataset, despite the large differences in, e.g., the age distribution between these cohorts.

2.2.5 Discussion

For a large range of global cutpoints we have shown FI_{GCP} to predict mortality and adverse clinical outcomes as well or better than FI-Lab created using established clinical risk thresholds. This result was replicated in the NHANES and CSHA data

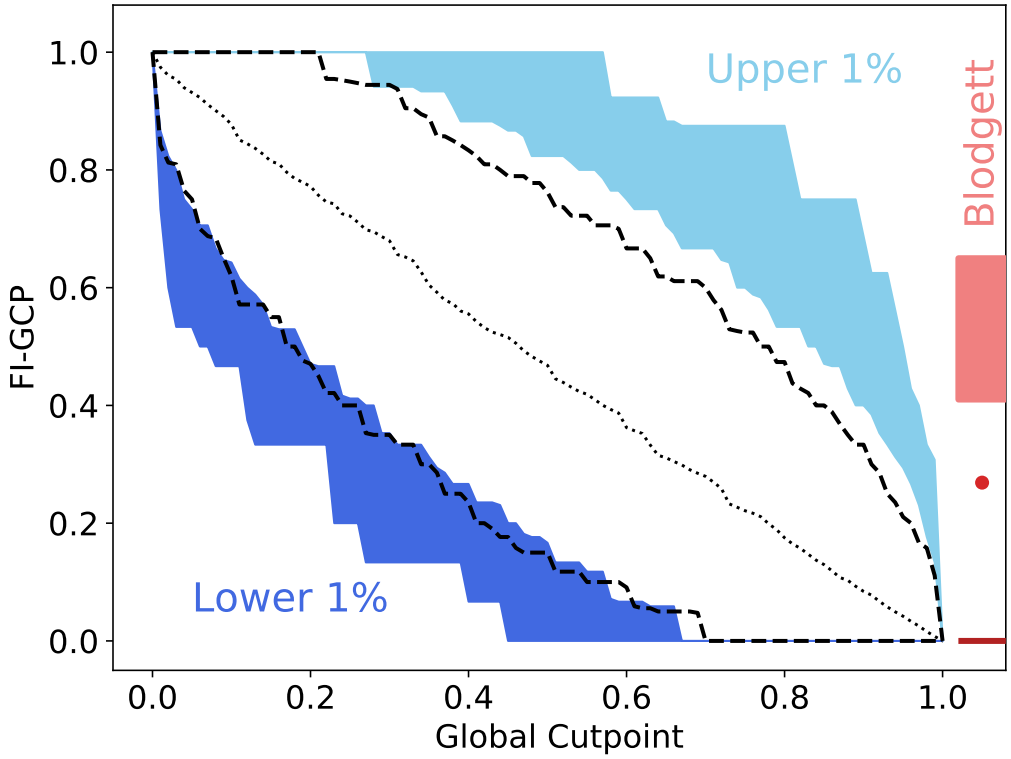

Figure 2.5: The upper 1% (light blue) and lower 1% (dark blue) of FI_{GCP} vs the global cutpoint X_{GCP} for the NHANES dataset. The dashed black lines show the 1^{st} and 99^{th} percentiles of FI_{GCP} in the CSHA dataset. The dotted diagonal black line shows the average FI_{GCP} in the NHANES dataset. The ranges and average for the published FI-Lab are indicated in red [28].

sets. Furthermore, FI_{GCP} was as informative in cross-validation, where cutpoints were calculated in one cohort and tested on another. Indeed, even applying cutpoints calculated in one age group to a cohort 20 to 40 years older remained effective. These results show FI_{GCP} is an effective method for generating an FI from biomarkers without prior knowledge of cutpoints, at least for cohorts of thousands of individuals or more.

The FI created using optimal cutpoints for each biomarker, $FI_{logrank}$ and FI_{info}, although highly informative, did not fare as well in cross-validation. Using these methods in one cohort did not yield an FI which was equivalently predictive in another cohort. Both the logrank and maximum-information based cutpoints strongly depend on the mortality of the particular cohort used and, as a result, do not represent general

risk thresholds. We suggest that cross-validation of cutpoints should always be done to ensure general applicability. Specifically, we caution against determining optimal cutpoints with respect to the outcome that the resulting FI will be tested on without extensive out of sample validation.

Notwithstanding this, there is room for improvement in the optimal cutpoint approaches. Considering the whole population uniformly ignores cohort differences which can be crucial when relating measurements of health to adverse outcomes. Accounting for these differences when calculating optimal cutpoints could increase the out of sample performance of these methods.

An important question to address when implementing FI_{GCP} is which global cutpoint is appropriate. We suggest that the cutpoint be selected such that the FI has good predictive value with respect to both health outcomes and mortality. However, in both the NHANES and CSHA data-sets there is a large range of cutpoints which are similarly predictive across many of these measures. Close study of Fig. 2.2 indicates that X_{GCP} of 0.6 or 0.7 would build FI_{GCP} that best predicts mortality, though this range of optimal X_{GCP} may depend on the cohort. Indeed, when we consider which X_{GCP} best predicts clinical deficits, the ranges of optimal cutpoints vary significantly (see supplemental Figs. A.8 and A.9, particularly). It appears that there is no one "best" global cutpoint for general prediction of health outcomes, or that applies equally well across cohorts.

Another criterion for picking the global cutpoint is the interpretability of the FI within and across studies. Within the range of cutpoints which are highly predictive there are large differences in the distributions of FI_{GCP}. Changing how the FI is constructed changes how individual values of the FI are assessed. For example, an FI of 0.2 has very different meaning depending on how biomarkers are binarized (see supplemental Fig. A.10).

In the context of current FI studies, an appropriate global cutpoint appears to be $X_{GCP} = 0.85$. The resulting $FI_{0.85}$ is highly predictive of both mortality and many of the clinical outcomes. Furthermore, the maximum, minimum, and mean of $FI_{0.85}$ are similar to the previously published medical threshold FI-Lab. As a result, individual values of $FI_{0.85}$ can be more easily interpreted between studies.

FI_{GCP} also provides a framework for investigating many aspects of the FI. Indeed,

we find that some common characteristics of the FI are not generally applicable. One of the results of changing X_{GCP} is the systematic change of the extremely high (or low) FI observed in a population, as shown in Fig. 2.5. Variations of the maximum FI has been observed in FI-Clin [18, 43, 54, 55, 56], FI-Lab [28, 27], between SHARE and SAGE multi-nation studies [57], in FI assembled from electronic health records [58] or primary care data [59], and were found to be necessary in network models of the FI [29]. We have shown that any explicit choice of binarization changes the observed FI_{GCP} limits. Indeed, any evaluation of binarized deficits – whether biomarker or clinical – should have similar effects. Because of the broad AUC maximum with respect to X_{GCP} we have shown that such variations of the FI-max do not imply that the quality of predictions of mortality or adverse health should be adversely affected. While cohort effects contribute to observed differences of FI-max between studies, we suggest that binarization effects may dominate. In Fig. 2.5, the difference between the upper and lower 1% of FI_{GCP} between the NHANES and CSHA cohorts is less than when X_{GCP} is changed by only 0.1.

How might we compare FI that use different binarization approaches within the same cohort? Perhaps we shouldn't: since the ability of FI to predict various clinical outcomes sometimes improves and sometimes degrades as X_{GCP} is changed, we can't expect one FI to behave exactly like another. However, qualitative comparisons may be possible with reference to extremal values of FI such as shown in Fig. 2.5. For quantile cutpoints, we also have a formal relationship between the global cutpoint and the population average of the FI that should facilitate such qualitative comparisons:

$$\langle FI_{GCP} \rangle = 1 - X_{GCP}. \tag{2.3}$$

This follows since it is precisely the fraction $1 - X_{GCP}$ of the biomarkers which are labelled at risk, across all biomarkers. This relationship is shown as a dotted black line in Fig. 2.5 and appears to hold approximately for the NHANES dataset. This remains to be better explored in future work.

Cohort effects become evident when the same cutpoint approaches are used between studies. While FI_{GCP} behaved qualitatively similarly in the NHANES and CSHA cohorts, it exhibits quantitative differences (see e.g. Fig. 2.2) that indicate cohort effects. FI_{GCP} is convenient for exploring cohort effects since it allows a complete separation of the cohorts at the level of biomarker binarization. For example, in

previous work on FI-Lab in the CSHA and NHANES studies [28, 4] some cutpoints were sex specific (blood pressure, creatinine, blood urea, and hemoglobin) and some were not. Using FI_{GCP} we could treat all biomarkers in a generic sex specific manner by first separating the population by sex then calculating the rank normalized scores. This approach does not require previous knowledge of the cohort dependence of the biomarkers, and should be useful in future studies of general cohort dependence of the FI – including sex differences. Brief analysis using sex-specific cutpoints show only marginal improvements in predictive value of FI_{GCP} over the favourable range of X_{GCP} (Supp. Fig. A.11). However, qualitative differences in the sex specific FI_{GCP} were also observed which could provide a new tool for analyzing sex differences in FI. Detailed analyses of cohort effects are beyond the scope of this paper but FI_{GCP} is well suited for these questions.

More generally, we have shown that FI created using population-based approaches can effectively treat biomarkers without prior medical knowledge. The same data-based approaches could also be useful in approaching FI for metabolomics, proteomics, and other omics-style applications. There is no Henry's clinician's handbook [42] to select omics cutpoints from, and the large number of measurements in an omics dataset necessitates an automated method for treating potential deficits. An FI based on omics data (FI-omics) would provide insight into how frailty manifests itself on the most fundamental levels, and a quantile approach should facilitate FI-omics.

Similarly, a general method of creating the FI from biomarker measurements opens the door to many more animal model applications. Previous work has been done to create an FI-Lab in laboratory mice [60]. Since there is no clinical guide for treating mice, cutpoints were selected in reference to measurements in young mice. This requirement of having a healthy cohort to use as a benchmark is incompatible with studies where there is no clearly defined healthy group available. A generic approach which can be applied to any set of biomarker measurements allows the FI to be used more generally, and should then facilitate comparisons of health and aging between organismal models and human studies.

In this study we have created and explored an effective quantile-based method of creating FI from biomarker data. We demonstrated that our methods performs

as well as or better than established methods which use diagnostic thresholds. Furthermore, we show that they are more robust to cohort effects than methods based on optimal prediction. These methods are applicable to any set of continuous valued data where information on the age or mortality of the population is available, and they do not require previous knowledge of how each measurement relates to health. We believe that our global cutpoint approach will be a powerful tool for examining cohort differences since cutpoints can be calculated for that cohort without prior knowledge of the biomarkers. We found that the main limitations in our approach are based on choosing an appropriate global cutpoint for a given study. Accordingly, we have raised the question of how to compare individuals across studies which have used different approaches for creating an FI. Nevertheless, we show that there is overlap between FI created using global cutpoints around $X_{GCP} = 0.85$ and other methods for creating FI-Lab.

2.3 FI-Lab without Binarizing Deficits

The FI-GCP approach to integrating bimoarkers into FI-Lab was successful. However, there are some questions raised by the methods used in the approach.

Despite it being clear that there is a wide range of viable cutpoints, the fact that a cutpoint must be chosen to begin looking at the FI may pose a problem. Furthermore, despite there not being large differences when mixing and matching cutpoints between cohorts, there are measurable differences. This raises some questions about which group to use when selecting cutpoints and how to effectively communicate what those cutpoints are. In the following section we develop a method which eliminates arbitrary decisions about cutpoints and helps to clarify the effects of cohorts on the resulting FI.

Another question raised is why, given that the population quantile measurement is naturally constrained between 0 and 1, is there any need to dichotomize in the first place. By definition the risk quantiles used to binarize the measurements increase with age. So, although it's not exactly the prevalence of the deficit that increases in the population with age, the average quantile score certainly does. Since the quantile scores would fit into the FI framework directly, is there any point in dichotomizing them? We investigate in the following work.

Additionally, given that we are forcing deficits to increase in prevalence (or average value) with age, is the resulting measure simply a proxy for age? In the following work we also consider the impacts of implicitly including age and explicitly excluding age in FI-Lab measures.

2.3.1 The Paper

2.3.1.1 Credits to authors + adaptations

Below is a manuscript primarily authored by myself with the guidance of Dr. Andrew Rutenberg. Dr. Rutenberg, Dr. Arnold Mitnitski, and Dr. Kenneth Rockwood all contributed significantly to the manuscript with either direct editing or feedback to the manuscript or during meetings. All code used in the analysis was written by me. The manuscript is included in full with the exception of the abstract.

2.3.1.2 Introduction

Population health declines with advancing age, but health trajectories vary considerably between individuals [61]. There are many distinct measures used to assess aspects of health on both the individual and population level. These range from molecular details of epigenetic methylation, to laboratory blood and metabolite tests, to clinical assessment measures in the comprehensive geriatric assessment, to self-assessed functional measures such as in the activities of daily living (ADL) or independent ADL (IADL). In principle, tens of thousands of distinct measurements are accessible for any individual. Nevertheless, any one measurement varies both intrinsically and due to measurement quality control [42]. Furthermore, any one measurement paints an incomplete picture of individual health. To obtain a fuller picture, summary measures of health can be assembled from many disparate measurements.

Summary measures of health combine many aspects of individual health into one. They include frailty [1, 2, 62], prognostic measures [63], Allostatic Load [17], epigenetic clocks [6, 3], and biological age [23, 64]. These metrics span the range from the tissue level of biological age, to the standard laboratory evaluations of the Allostatic Load, to the functional level of the FI or the frailty phenotype. Functional-level summary measures are strongly associated with a wide array of adverse health outcomes

[65].

While many summary measures of health overlap in how they are constructed or how they perform, they are generally not identical [62, 63, 64, 3, 23]. This reflects multidimensional aging – including organismal scales ranging from cellular, to tissue, to functional [36, 16, 66]. To assess multidimensional health more completely, we need to continue to both develop new summary measures of health and to improve existing ones. For example, controlling for both age and sex is important in assessing and comparing individual health. How to conveniently and effectively do this for a given summary measure of health is a persistent challenge.

Here, we focus on the frailty index (FI) [1] because it is simply constructed and can be effectively adapted to a broad variety of health aspects. The FI has been defined as the proportion of measured health aspects which are considered to be in the unhealthy state. Candidate health variables considered in the FI include anything health-related that increases in prevalence with age [18]. These have typically been high-level health deficits such as impairments in acts of daily living, self-rated health, and other clinically observable deficits in a FI-Clin [67]. However, biomarkers such as the results of blood tests can also be used to create an effective FI-Lab [28, 4, 27, 11]. Both FI-Clin and FI-Lab are strongly associated with adverse health outcomes including mortality.

One challenge in calculating FI-Lab is how to properly incorporate measurements which are not already dichotomized. Typically, measurements are dichotomized based on normal reference ranges such as those found in clinician's handbooks – such as [42] [28, 4, 27]. However, diagnostic thresholds – intended to guide treatment – may not be appropriate for a summary measure of health [11]. Furthermore, many biomarkers do not have associated diagnostic thresholds. With larger omics-style biomarker assays becoming more prevalent this absence will become increasingly pressing [68].

There are also significant intuitive and empirical issues with dichotomization (or "binarization") of continuous variables [69, 70, 71, 72, 73, 74]. These are well understood for predictive measures since there are quantifiable losses in statistical power when imposing dichotomy on a continuous variable [69]. Individual dichotomized variables are sensitive to small variations around the cutpoint. Consider an individual measurement with a value close to the dichotomization threshold. Any small

variation of that measurement could result in a switch from absence to presence of deficit – the maximum penalty for a minimal variation. Frailty indices reduce these issues by averaging a large number of variables [75, 4], but the scale of these effects have not been systematically explored within the FI literature.

To assemble an FI-Lab without dichotomization, we first need to pre-process health measurements in order to be able to combine them into a single measure. The common approach of using Z-scores (or standard scores), which shift measurements by their mean and then rescale by their standard deviation, does not naturally fit into the 0 (maximal health) to 1 (maximal unhealth) range of FI scores. However, ranking individuals by age-related health risk with respect to a reference population is an effective way of pre-processing an arbitrary set of biomarker measurements that naturally leads to a 0 to 1 range [11]. Rank normalization is often used in e.g. pre-processing of gene expression data [76], and is illustrated in Fig. 2.6. We found that by imposing a single global quantile cutpoint (GCP) on all of the individual rank-normalized scores, the resulting FI-GCP outperformed pre-existing FI-lab with the same data and was effective for a broad range of GCP [11]. We show in the Methods how this quantile approach can be adapted to assemble an FI without dichotomization. Nevertheless, while quantile approaches avoid artificially grouping individual measurements, preserve aging trends, and treat all biomarkers similarly, they do require an explicit reference population.

Any health assessment is implicitly with respect to one or more reference populations. For example, dichotomization of any one variable creates two reference populations – a healthy one and an unhealthy one. Repeating this for many health variables creates many small reference populations that have identical dichotomized scores across many health measures. In contrast, quantile approaches can share a reference population across many health measures. With a small number of reference populations we can more easily treat them as independent, or controllable, ingredients of our summary measure of health. In particular, we can use reference populations to control for age and sex effects.

By construction, deficits included in the FI increase in prevalence with age [18]. As a result, there is often a significant correlation between included biomarkers and age. For dichotomized biomarkers, this raises the question of how to age-control thresholds.

Doing this by prognosis raises the question of whether multiple age-related outcomes may lead to distinct thresholds. However age-control is done, or not done, it will affect the resulting FI-lab. We can think of age as a confounding variable in terms of assessing aging health. How much we can learn about individual health independently of an individual's age? This is broad question that has also been raised in the context of biological age [77], and other summary measures of health.

Issues of dichotomization are compounded when state variables, such as sex, are considered. Summary measures of health should reflect differences in health between the sexes. In many studies men are measured as "healthier" despite having greater prevalence of negative outcomes [78]. Selecting clinical-level health deficits based on sex-dependent prevalence affects sex-dependent mortality prediction of composite measures [79]. When biomarker measurements are used the prevalence of each deficit can be tuned by the dichotomization threshold. However, using sex-dependent diagnostic thresholds still results in large sex differences in the FI [27]. Furthermore, new biomarkers do not yet have known diagnostic relevance or sex-dependent relevance. A transparent approach may be best: treat sexes as independent populations and use identical methods for calculating FI for each sex.

While a broad reference population with natural demographics is used in e.g. the frailty phenotype [2] (gender, height, BMI) or allostatic load [17] (non-stratified), we find that three smaller reference populations are particularly useful. One is a population of older adults (80-85 year-olds). This group is more prone to adverse health and outcomes than younger adults, but is still very well represented in population studies since they are slightly below the average human lifespan in Canada. This reference population is useful since it leads to a FI that is most similar to existing FI-lab measures in appearance. The second reference population we explore is a set of age-matched populations for each individual. We use this to critically examine the explicit and implicit role of age in the FI, particularly with respect to its association with adverse health outcomes. The third type of reference population is to use sex-specific reference cohorts in combination with either of the others. This allows investigation of sex differences in the FI with a non-parametric approach.

In this work, we show that the quantiles of age-related risk lead to a predictive and interpretable FI-lab, which we call the quantile frailty index (QFI). The QFI predicts

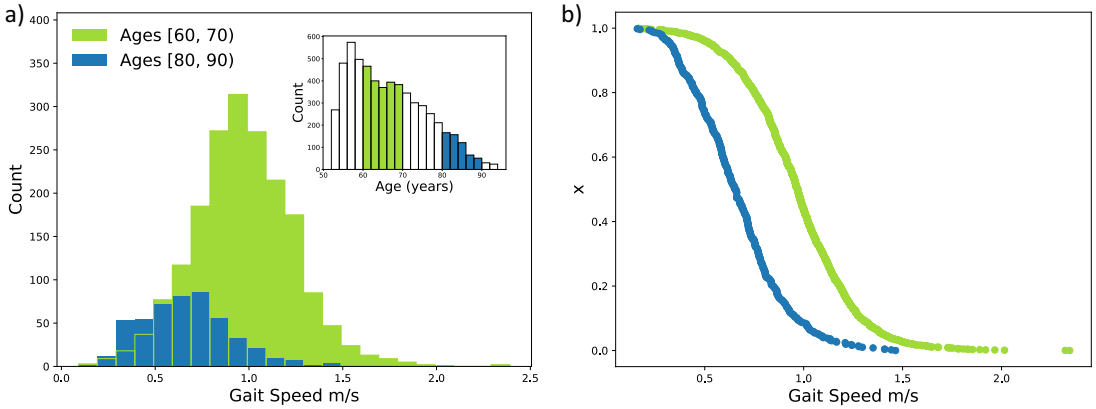

Figure 2.6: Risk quantile calculation example with the ELSA dataset. We show a) the distribution of gait speeds for example reference populations of 60-70 year-olds (green) and 80-90 year-olds (blue), with b) the associated risk quantile x *vs.* gait speed. The inset in b) shows the age distribution. Gait speed decreases with age, so the highest risk quantiles are associated with low gait speeds.

5 year mortality significantly better than previous dichotomized methods of creating FI-Lab. The QFI is strongly correlated with the number of accumulated diagnoses, with number of new diagnoses accumulated at a one year follow-up, and is strongly associated with independent FI-Clin for the same individuals. Furthermore, we show that changing the reference population does not significantly affect prediction, but does affect the observed distribution of the QFI. Using different reference populations, we investigate the role of age and sex in the predictive quality of the QFI.

2.3.1.3 Methods

2.3.1.3.1 Quantile Frailty Index (QFI)

Consider N biomarkers that are assessed for every individual in a reference population, so that the ith biomarker (y_i) has a distribution $P(y_i)$. We take the risk quantile x_i as the position of the corresponding biomarker value y_i in its cumulative distribution. For biomarkers which increase with age (e.g. c-reactive protein), we take the quantile to increase with increases of the biomarker:

$$x_i = \int_0^{y_i} P(y_i')dy_i'. \tag{2.4}$$

In the case of biomarkers that decrease with age (e.g. gait speed), we take the quantile to increase with *decreases* of the biomarker:

$$x_i = \int_{y_i}^{\infty} P(y_i')dy_i'. \tag{2.5}$$

In both cases, the quantile $x_i \in [0, 1]$ and increases in the same "risk" direction as the biomarker increases with age on average [11]. Obtaining the quantile is equivalent to performing a rank normalization of the score with respect to the population. Because many biomarkers have limited measurement precision there are frequent ties in biomarker scores. We use the minimum rank of tied scores; other methods of tie-breaking lead to similar results.

Our definition of the risk quantile means that x_i corresponds to the proportion of the population that has lower health-risk associated to that biomarker. So, x_i is equivalent to the "fraction unhealthier than" for a given biomarker with respect to a reference population. For example, Fig. 2.6 shows that having a gait speed of 1 m/s is slower than about 50% of the 60-70 year-olds, so $x_i = 0.5$ is the fraction of 60-70 year-olds that an individual with a gait speed of 1.0 m/s is unhealthier than.

We then average the N non-dichotomized risk quantile measures for every biomarker to obtain an individual frailty index, the QFI:

$$QFI = \sum_{i=1}^{N} \frac{x_i}{N} = \langle x_i \rangle, \tag{2.6}$$

where the angle-brackets indicate an average. We have $QFI \in [0, 1]$.

We can quantify the advantage of having a continuous score by also examining m discrete risk categories such as dichotomization ($m = 2$) [11], tertiles ($m = 3$), quartiles ($m = 4$), or general m. For m risk categories, our risk scores would then be

$$d_i^{(m)} = floor(x_i * m)/(m - 1), \tag{2.7}$$

where the $floor(z)$ function returns the greatest integer less than or equal to z. So for dichotomization ($m = 2$) scores of $x_i \in [0, 0.5)$ would give $d_i^{(2)} = 0$ while $x_i \in [0.5, 1)$ would give $d_i^{(2)} = 1$. We can then construct discrete $QFI_m = \sum_{i=1}^{N} d_i^{(m)}/N$. As $m \to \infty$ we obtain $d_i \to x_i$.

The QFI can be calculated with respect to an arbitrary reference populations. We examine two age-related reference populations. The first is a fixed-age reference,

which was defined as all individuals from a particular study (NHANES, CSHA, or ELSA) that were within a fixed range of ages – 80-85 year olds unless otherwise stated. We will use this 80-85 year old reference population as the default reference for the QFI – unless otherwise mentioned this is the the reference population used.

A second reference population was age-matched. Here we used the same fixed-range bins for both the reference population and the individuals (so, e.g., the quantiles of 50-55 year olds were determined with respect to 50-55 year olds). For all reference populations, and unless otherwise stated, our results are for non-overlapping 5-year ranges of ages and the age of the population for plotting purposes was taken to be the middle of the range.

We also consider sex-matched reference populations. Some measurements (e.g. grip strength) vary substantially between the sexes and comparing individuals only within their group is desirable. We can combine sex and age to make very specific reference populations, for instance comparing all women in the study to a subset of 80-85 year old women, or women of similar age.

2.3.1.3.2 Assessment

We evaluate predictive performance of the FI using the area under the receiver operating characteristics curve (AUC) [80]. We re-sample random halves of the population 100 times to estimate errors, while FI-GCP measures are cross validated as described in [11]. We present distributions using box and whisker plots with whiskers extending to the 99^{th} percentiles. We exclude bins with less than 20 individuals. All analysis is available on GitHub [35]; logistic regression is done using the statsmodels Python package [81].

2.3.1.3.3 Data

We have explored the QFI with cross-sectional data from the National Health and Nutrition Examination Study (NHANES) [45] and the Canadian Study of Health and Aging (CSHA) [46]. The NHANES data set consists of the 8881 individuals from the NHANES study with data for at least 11 of the 16 available biomarkers. This sample has an age range of 20 to 85 years. The data used from the CSHA study has 973 individuals aged 65+ for which data is available for at least 16 of the 22 biomarkers. We use the same NHANES and CSHA data examined previously with other frailty

indices [11, 27, 28]. We also use data from the ELSA study [82], described in detail in Supplemental information. We focus on data from the second and fourth waves of the ELSA study, and examine predictive value on available data in subsequent waves.

2.3.1.3.4 Replication

All figures are replicated in waves 2 and 4 of the ELSA dataset, as well as in the NHANES and CSHA datasets when applicable. The exceptions being the figures where diagnosis data is used, since it is not available in the NHANES and CSHA datasets. Preferentially, we show data from the NHANES study and wave 2 of the ELSA data due to their larger sample sizes and number of mortality events. Replicated figures are available in the Supplemental material (Figs. S1-S15).

2.3.1.4 Results and Discussion

2.3.1.4.1 Advantages of not Dichotomizing

We compared the effects of transforming biomarker measurements into categorical variables with m categories and found that using more categories gives better prediction in the NHANES data-set (Fig. 2.7). There are notable improvements from using 2 categories (equivalent to binarizing at the median) to using 5 risk categories. Using more risk categories with fewer variables can sometimes even result in better prediction than dichotomization with more variables. Here, using 5 or more risk categories on the 5 highest predicting biomarkers outperforms 2 risk categories used for all 17 biomarkers. In the NHANES data-set, quintiles perform as well as any finer grouping, which suggests that variation within 20% of the population does not have a significant effect on outcomes. However, using the QFI – with as many risk categories as is possible with the available data – does not negatively affect prediction and is more convenient than restricting everything to quintiles. In the other data-sets the plateau of prediction vs number of risk categories occurs in different places, but the QFI never under-performs a coarser grouping of risk. For the remainder of the paper we will use the QFI.

In Fig. 2.8 we show that the QFI performs modestly better than using a global cutpoint to binarize biomarkers based on risk quantile [11], and is also better than

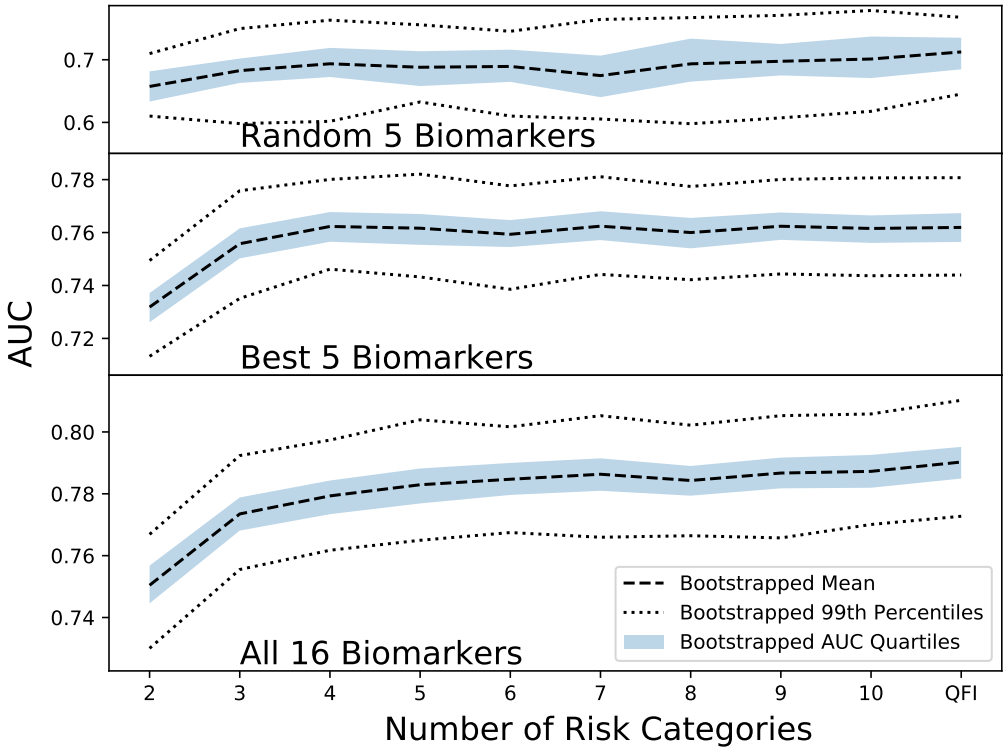

Figure 2.7: The relationship between number of risk categories and the predictive value with respect to mortality within 5 years in the NHANES dataset. 2 risk categories is equivalent to dichotomization at the median, 3 risk categories equivalent to risk tertiles, and so forth. The upper plot shows the effect using 5 randomly selected biomarkers, the middle plot shows the best 5 biomarkers selected by AUC with respect to 5 year mortality, and the bottom plot shows the results using all available biomarkers. We resample the data using half the population size 400 times for each point, with the random 5 biomarkers also being re-selected 20 times. The dotted lines show the upper and lower 1^{st} percentiles of AUC, the shaded blue region shows the upper and lower quartile range of AUC, and the dashed line shows the average AUC. Note that AUC ranges improve from the top to the bottom plots.

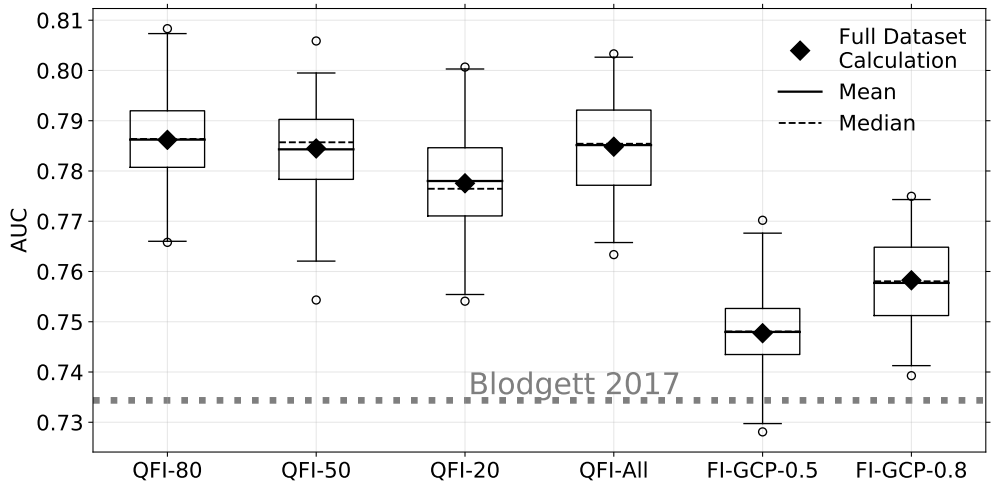

Figure 2.8: The predictive value of various FI-Lab with respect to 5 year mortality in the NHANES study. From left to right we show the QFI using the 80-85 year-old reference population, the QFI with a 50-55 year-old reference, QFI with a 20-25 year-old reference, QFI using the whole NHANES population, and FI-GCP with the cutpoint at 0.5 or at 0.8 [11]. Box and whisker plots display the data from resampling and cross-validation: the boxes represent the upper and lower quartiles, the whiskers go to the 99^{th} and 1^{st} percentiles, and the circles are remaining outliers. The short dashed line within each box is the median, the solid line the mean, and the diamond is the AUC for the full data set without cross validation or resampling. The horizontal grey dashed line shows the AUC of the published FI-Lab using the same data [28].

binarizing biomarkers using diagnostic thresholds [28]. Interestingly, using the full data-set as a reference population performs the same as using an 80-85 year-old reference population. We find that the choice of fixed-age reference population does not significantly affect the prediction quality of the QFI.

2.3.1.4.2 QFI is interpretable

The detailed characteristics of the QFI are similar to other types of FI. We show the relationship between the QFI and the FI-Clin in wave 2 of the ELSA data in Fig. 2.10a. The relationship between QFI and FI-Clin is close to linear at larger values. Consistent with this, the aging trend of the QFI is very similar to that of FI-Clin – as shown in Fig. 2.10b. Nevertheless, the average QFI is significantly larger than FI-Clin at all ages. Very few individuals exhibit a QFI below 0.3.

These detailed characteristics of the QFI are dependent on the choice of reference

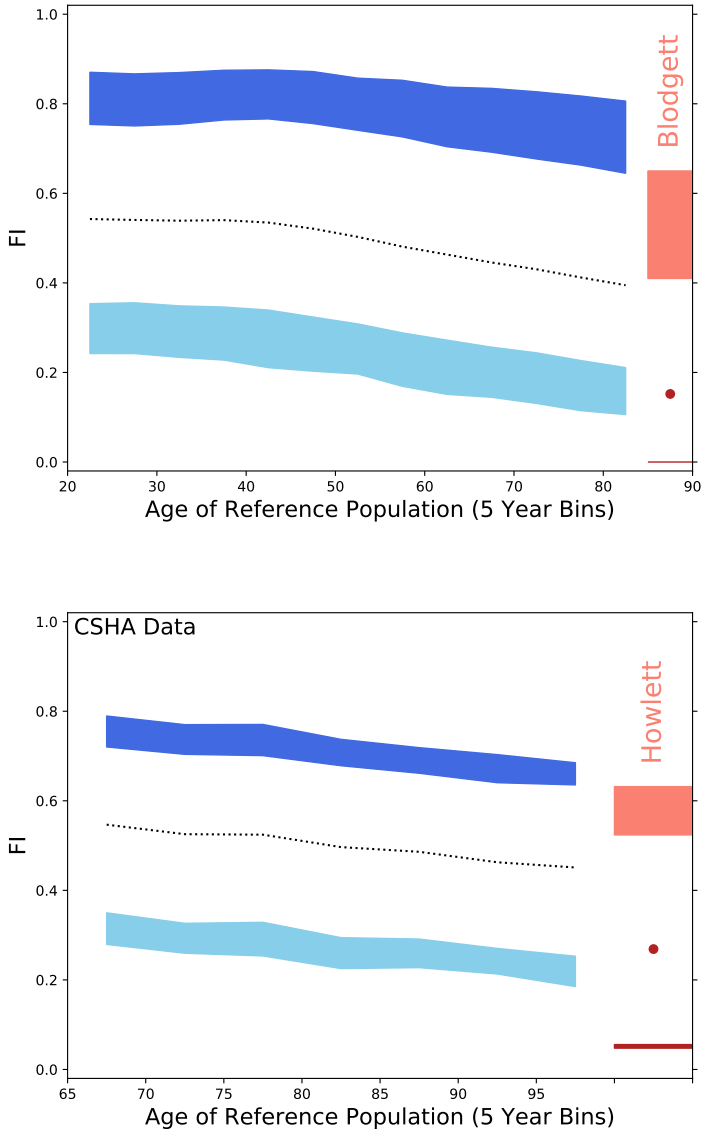

Figure 2.9: The effects of changing the reference population on the distribution of QFI scores in the NHANES (top) and CSHA (bottom) studies. The upper 1% (light blue) and lower 1% (dark blue) of the QFI distributions as the age of the reference cohort changes in 5 year bins. On the right side of the plots, the red blocks show the upper 1% (light red), lower 1% (dark red), and average (red point) for the respective published FI-Lab using diagnostic thresholds [28, 27]. We require each bin to have at least 20 individuals, removing only the 100-105 year-old bin in the CSHA study that had 2 individuals.

population, though we have seen that predictive performance is not. We know that if most of the study is younger than 80 years old that selecting an 80 year old reference will make the bulk of the study appear relatively healthy. The effect of switching the reference population to an unhealthier group is an overall lowering of QFI scores in the population. As seen in Fig. 2.9, selecting an older cohort as the reference population leads to a general downwards shift in the distribution, and a slight positive skew. Using an older reference makes the distribution of the QFI look much more like a typical FI. However, to achieve a QFI of 0 an individual would have to be the healthiest individual compared to the population across every single biomarker measurement. Intrinsic variability and measurement noise make this unlikely even if there is someone in perfect health.

Although the QFI looks more like a standard FI with older reference populations, we do not think that aiming to look exactly like a standard FI is necessary. The QFI has a natural interpretation as being the average relative health with respect to the reference population.

We have also used ELSA data to test the association of the QFI with the various non-mortality outcomes available. For simplicity we use a reference population ages 80-85 in all cases where the QFI is calculated. The ELSA dataset has a list of reported diagnoses recorded at every wave (see supplemental information for details). We use the wave of first reported diagnosis to relate the QFI to these diagnoses in a number of ways. Firstly we look at the total number of accumulated diagnoses as it relates to the QFI in Fig. 2.10c. This figure shows a strong relationship between the QFI and the total number of diagnoses before this wave of the ELSA. Fig. 2.10d shows the proportion of individuals with one or more new diagnoses in the wave directly following the QFI assessment (1 year later). A higher QFI is associated with an increased probability of new diagnoses in the coming year on average. The difference in expected number of diagnoses almost doubles from a QFI of 0.3 to a QFI of 0.6.

2.3.1.4.3 Role of age within the QFI

Since we can define health with respect to a specific age group we can also remove the confounding effects of age from the QFI. We do this by calculating the QFI with respect to a group of individuals of the same age. In this age-paired QFI we group

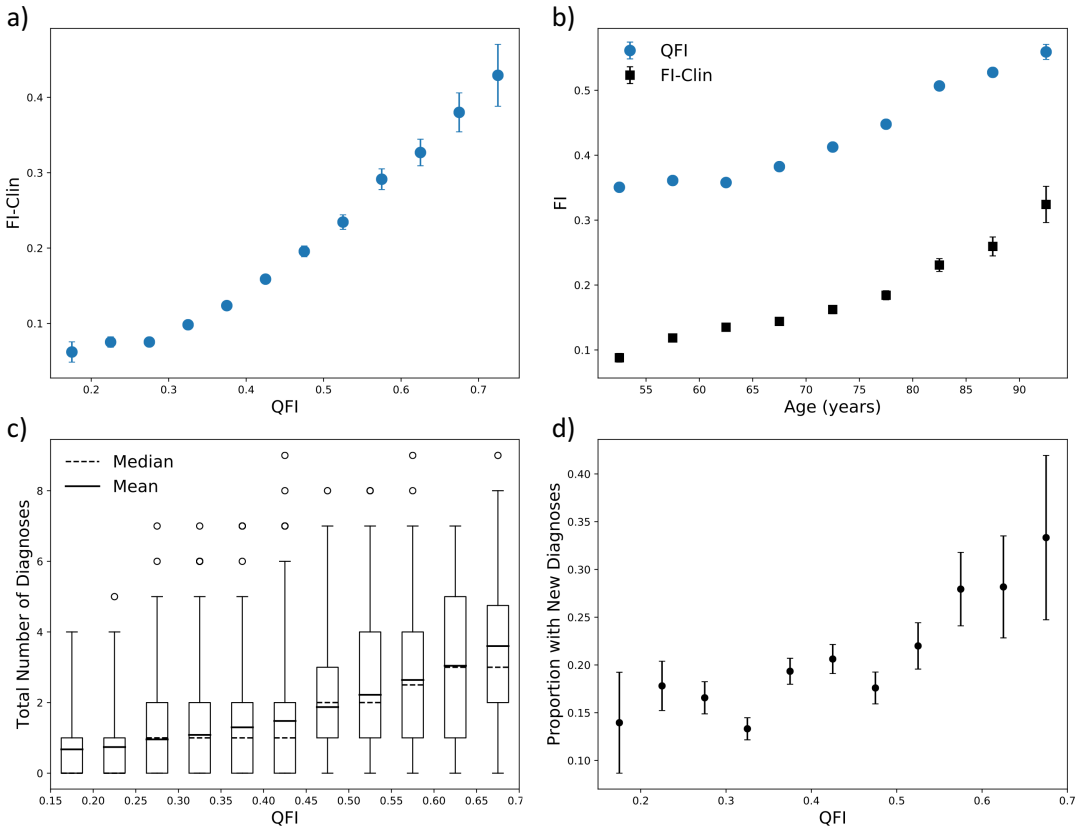

Figure 2.10: QFI on wave 2 of the ELSA dataset, using a reference 80-85 year-old population. a) The average FI-Clin binned by QFI for wave 2 of the ELSA dataset. b) The average QFI (blue points) and FI-Clin (black squares, described in the Supplemental Information) binned by age. c) The relationship between the QFI and the total number of existing or previous diagnoses. The boxes represent the upper and lower quartiles, the whiskers go to the 99^{th} and 1^{st} percentiles, and the circles are remaining outliers. The short dashed line within each box is the median and the solid line is the mean. d) The fraction of the population with 1 or more new diagnoses in the year following the QFI evaluation.

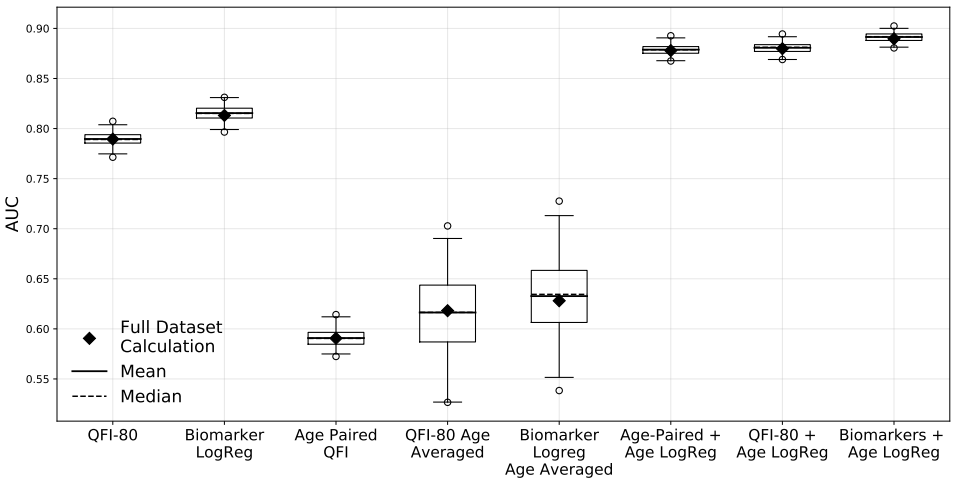

Figure 2.11: Comparing the predictive value for the different methods of calculating the QFI against 5 year mortality in the NHANES dataset. From left to right we first have two "raw" points: the QFI with an 80-85 year-old reference population (QFI-80) and a logistic regression model using all of the biomarkers regressed against 5 year mortality. Then three age-controlled points: the age-paired QFI, QFI-80 with AUC averaged across performance within 5 year age bins, and a logistic regression of the biomarkers against 5 year mortality with AUC averaged across performance within 5 year age bins. Then three age-supplemented points: the age-paired QFI combined with age in a logistic regression, the QFI-80 combined with age in a logistic regression, and the raw biomarkers included with age in a logistic regression. Box and whiskers are from resampling with half the population. The boxes represent the upper and lower quartiles, the whiskers go to the 99^{th} and 1^{st} percentiles, and the circles are remaining outliers. The short dashed line within each box is the median, the solid line the mean, and the diamond is the AUC for the full data set without resampling. We use logistic regression to control for age in the prediction rather than for testing a logistic model, so performance is evaluated on the same individuals as the model is fit on.

individuals into 5 year age bins and calculate the QFI using those binned individuals as the reference population.

We compare the predictive value of the age-paired QFI to the QFI with an 80 year old reference population in Fig. 2.11. We find that the QFI-80 substantially outperforms the age-paired QFI. However, we find that if we calculate the AUC as the average AUC across a set of age-binned QFI measurements (see Supplemental Figs.B.11B.12) the QFI-80 performs similarly to the age-paired QFI. Furthermore, we find that if we add age back in to prediction using a logistic regression of both QFI and age – then both the QFI-80 and age-paired QFI perform similarly. As a benchmark for mortality prediction with biomarker data we have included the results of logistic regression on the raw biomarker measurements. We find that raw regression of the biomarker measurements performs better than the raw QFI-80. However, the biomarker measurements perform as badly as other FI measures when age controlled and only slightly better than other FI measures when combined with age.

2.3.1.5 Sex-Specific Reference Populations

We also consider sex-specific reference populations, where quantile scores for each sex are calculated with respect to a reference cohort of only that sex (age restricted or not). In the ELSA data there is a large sex difference in the adjusted QFI scores due to the presence of grip strength measurements. Fig. 2.12a-b shows the difference in quantile scores for dominant hand grip strength and the resulting shift in QFI. Fig. 2.12c shows that controlling for sex in the reference population has the effect of narrowing the difference between male and female across the age range for ages below 90 years. Fig. 2.12d shows that the AUC for 5-year mortality predictions improve when controlled for sex. The age-paired QFI also saw improved prediction when matching sex (see Fig.B.10a for comparison).

2.3.1.6 Discussion and Summary

We have shown that the dichotomization of continuous biomarker data into binary health deficits negatively affects the predictive value of the resulting FI-Lab (Fig. 3). Using categorical variables for deficit scores increases predictive value of the resulting

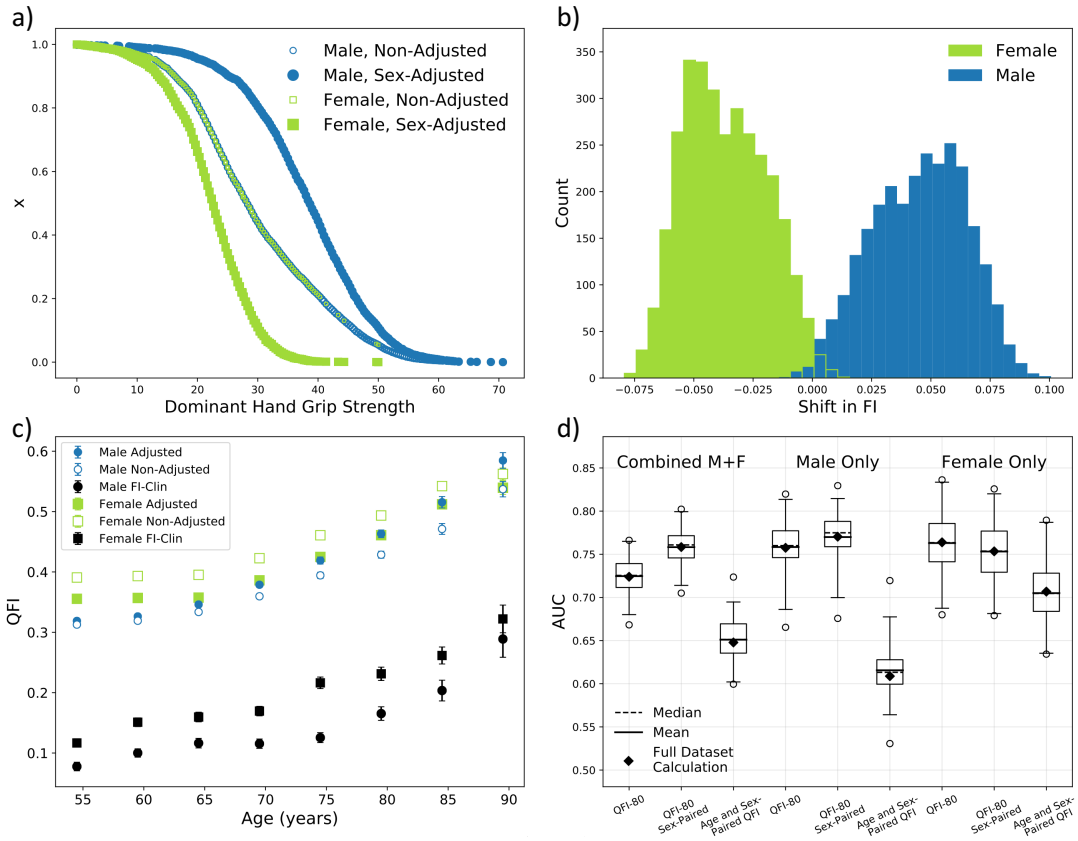

Figure 2.12: The effects of using sex-specific reference populations on QFI-80 in wave 2 of the ELSA dataset. All plots show female in green and male in blue. a) The risk quantiles for dominant hand grip strength with (filled points) and without (no-fill) using a sex-specific reference population. The non-adjusted male and female scores overlap since they are using the same reference population. b) The difference between sex-adjusted QFI-80 and non-adjusted QFI-80 using all 80-85 year-olds as a reference population. Sex-adjusted QFI-80 uses only 80-85-year-olds of the respective sex as the reference population. c) The average QFI for the sex-adjusted QFI (filled) and non-adjusted (no fill) binned by age in 5 year bins, the black points show the associated FI-Clin. d) The AUC of various QFI with respect to mortality at 5 year follow up. We compare (from left to right) the QFI-80, sex-paired QFI-80, and age-and-sex-paired QFI for the combined, male, and female populations (from left to right). Box and whiskers are from resampling with half the population. The boxes represent the upper and lower quartiles, the whiskers go to the 99^{th} and 1^{st} percentiles, and the circles are remaining outliers. The short dashed line within each box is the median, the solid line the mean, and the diamond is the AUC for the full data set without resampling.

quantile frailty index (QFI) when compared to dichotomization: increasing the number of risk categories improves the predictive value of the FI (Fig. 2). These results are replicated in the CSHA, NHANES and ELSA datasets.

The QFI allows us to easily explore the average relative risk with respect to the reference population across many biomarkers. For example, an individual with a QFI score of 0.3 is healthier than 70% of the reference population. Selecting an appropriate reference population can enhance the interpretability of the QFI: a QFI score of 0.6 with respect to a reference population of 80 year olds means that the individual is in worse health than 60% of 80 year olds. By using a common reference population, the relative health of individuals in different populations or subpopulations can be assessed. This could be useful when study populations are heterogeneous. For example, with mixtures of community dwelling and institutionalized individuals.

We have critically addressed the implicit inclusion of age in the QFI through the age-correlation of included health attributes. By using age-paired reference populations, or by considering predictive value of narrow ranges of age, we see that the predictive value of the QFI is strongly degraded (Fig. 2.11). Conversely, by combining the QFI with age explicitly within a logistic regression we see that predictive power is greatly enhanced. Including age explicitly in this way leads to approximately the same predictive power whether we use an age-paired QFI, a reference population on 80-year olds, or raw biomarker values. This indicates that the mortality-associated age-independent health information contained in the biomarkers comprising the QFI is retained in the QFI.

In clinical practice, any summary health measure for an individual will be available together with age – so both should be used for prognosis. A single summary health measure may also be desirable. By including age explicitly in assessing predictive power of the QFI, we can assess how much a single summary measure of health could be improved by constructing it with more age-associated components – either implicitly or explicitly. For the QFI this requires a fixed-age reference population.

If we prefer a summary measure of health that excludes age, then we need to show that the age-averaged predictive quality agrees with the overall quality. For the QFI, we can construct this age-excluded measure using age-paired reference populations.

If we want to compare the predictive power of two summary measures of health

through, e.g., the AUC of an ROC, it is clear that any differences in the implicit inclusion of age will dominate the comparison. Age should be either explicitly added or explicitly controlled for in such a comparison – ideally both.

A similar discussion of age-dominated composite measures exists in the biological age literature [3]. When chronological age was controlled for, early epigenetic clocks lost many of their significant associations with health outcomes [83]. Later epigenetic clocks addressed this issue by including biomarkers associated with adverse health outcomes independently of age [16].

The QFI can also straight-forwardly and non-parametrically control for sex. Using sex-dependent reference populations ensures that the male and female individuals are treated the same. The results are intriguing. In Fig. 7c we found a crossing of male and female QFI as a function of age, so that males have a higher average QFI at later ages (above 85 years). This does not exhibit a mortality-morbidity paradox, since male mortality is somewhat higher than female at higher ages. Accordingly, we see slightly improved AUC for the sex-adjusted QFI. While there are real biological differences between male and female aging populations [78], our finding raises the intriguing possibility that the mortality-morbidity paradox could be significantly reduced with proper control populations. This is worth further study with different aging populations using agnostic approaches to controlling for sex, such as the QFI.

Individual health is high-dimensional. There are a vast number of individual characteristics of good or poor health. In contrast, populations are often described by only a few characteristics such as just age and sex. Nevertheless, it is important to condition individual health on comparable populations. For binarized health variables this can be done with population-dependent cutpoints [28, 27, 42]. For the QFI, we can explicitly choose the reference population. In this paper we have explored the role of age and sex, but any demographic differences in health and aging can be addressed with our approach.

We have mostly used a fixed-age reference population of 80-85 year olds. This leads to a natural interpretability of the resulting QFI. By varying the fixed-age reference, we see in Fig. 5 that commonly reported maxima and minima of the FI (0.7 and 0.0, respectively) [18, 4] appear to be recovered as we approach a supercentenarian reference. While that would gives an appealing interpretation of the QFI as your

health quantile with respect to the "very old", we do not yet have a large enough sample of the very old to explore that limit.

We have developed a summary health measure, the quantile frailty index (QFI), from continuous biomarker or health measurements without any dichotomization. The QFI is both predictive of mortality, and interpretable as a frailty index. Different reference populations can be easily used to construct the QFI. We have investigated the role of age in the QFI, and demonstrate that the QFI effectively includes the non-age related aspects of considered biomarkers. The QFI can control for other important population state variables with appropriate reference populations.

2.3.2 Two-Sided Risk

In the work done on FI-GCP we saw that a generic 2-sided approach to risk was less effective than a single sided approach for predicting mortality. Using a continuous approach to including biomarkers in the FI allows a variety of methods for including 2-sided approaches to risk. However, throughout the work developing the QFI we were unable to find a method which significantly out-performed our 1-sided approach. Approaches like quadratic ($d_i \sim b(x_i - a)^2$) or gaussian-well ($d_i \sim 1 - e^{-b(x_i-a)^2}$) transformations of the risk quantiles did not offer substantial improvements to prediction. Furthermore, using these types of 2-sided approaches introduced a set of parameters to the problem, which is undesirable given the simple nature of the FI. Using transformations of the risk quantiles also results in a loss of interpretability for the QFI, since the score is no longer just the average relative standing of health in the population. It is possible that there is a clever approach to including 2-sided risk in an approach similar to the QFI, but we did not find one in this work. However, given the modest improvements to prediction when using more involved biomarker transformations, it seems unlikely that any gains in prediction would be worth the loss in interpretability.

2.3.3 Results of QFI Paper in Broader Context

Creating an FI-Lab measure without dichotomizing the included biomarkers is certainly an improvement in the quality of the metric. However, is it really an FI? Despite having the appearance of an FI, the QFI has an independent interpretation. The largest distinction being that deficits in the QFI are determined by relative

health, even amongst the healthiest individuals in the population. This is in contrast to typical FI where the accumulation of damage is intended to be absolute [18]. The implications from a practical perspective are minimal; it is an effective and interpretable way to measure FI-Lab. However, it poses some difficulties theoretically and conceptually.

Theoretically the issue is as follows: in a model system how does one encode the progression of a biomarker measurement through the deficit scores. The obvious solution is to change from accumulation of binary deficits to incrementing categorical deficits. However, there will always be a discordance with empirical measurements of the QFI since it is nearly impossible to have a QFI of 0. Consider blood pressure as an example, all healthy measurements of blood pressure have a deficit score greater than 0 since individuals with acute risk will be represented in the extremes.

There are two approaches to reconciling the empirical lack of healthy QFI scores with the general FI approach. The first is to redesign the QFI to effectively include a healthy state ($d_i = 0$). Consider something like an elbow shaped deficit score where measurements before a certain quantile are assigned a deficit score of 0 and measurements above have some increasing score. However, this is likely to just reopen the parametrization can of worms with the selection of elbow positions etc.

Alternatively, one could take the QFI style of health deficits at face value and reconsider the FI-Lab conceptually. The QFI-based reinterpretation of FI-Lab would have biomarker scores incur some sort of health cost on the system even in the healthy range. The current model of the FI is that damage is accumulated stochastically over time originating from an ultimately healthy baseline state as measured with those deficits. However, the mechanisms for driving this accumulation of damage have not been rigorously explored. One plausible mechanism could be a health cost incurred by biomarker-level deficits even in relatively healthy states.

Chapter 3

Network Optimization

3.1 Introduction

The interactions between the many aspects of health involved in the aging process can be effectively represented using networks. In the case of aging - and many other complex problems - the network framework is convenient for understanding health aspects and their interactions. However, developing a network model of aging has some substantial obstacles; which elements of health should be included and how should they interact? Supposing that these problems could be solved, there is the additional task of interpreting the resulting network.

One approach to modelling human health as a network is the generic network model (GNM) [31, 29, 5]. This approach leverages the broadly defined health aspects of the frailty index (FI) [1, 18] to effectively model human health on a population level. This model does not directly tackle the issues of which health aspects are included, instead using generic deficits which map indirectly onto observable health attributes. Furthermore, the GNM uses generic interactions between the health attributes, so all phenomenology in the model is determined by the network structure. Although the GNM does not map directly on to individual health - it could not be used for clinical intervention - analogies between the network structure and common health aspects considered in the study of aging has provided valuable insight.

The GNM uses a scale free network structure to capture the phenomenology of human health. Scale-free networks describe a large family of networks that have been observed in a variety of systems both biological and man-made [84, 32]. It was shown that the scale-free network structure was critical for capturing certain aspects of human health and aging phenomenology [5], but it is not clear why this structure would emerge. In this work we use an optimization procedure to determine whether the scale-free network structure is an optimal structure for prolonging lifespan and healthspan, or whether there are competing influences on the network structure.

Optimizing network structures is a problem of increasing relevance; many components of modern infrastructure such as transportation is best described using networks, so anything from databases to air traffic is effectively implemented - or modelled - using networks [85, 86]. As a result, optimization of the structure of these networks could mean large performance increases and potentially profits or savings. These network oriented problems range from purely mathematical - many np-complete problems involve graphs [87]- to practical [88, 89]. For instance, how does optimal air traffic look when maintenance costs is weighed against travel costs [89]? The optimal solutions have structures which depend on the incentive structures. In organic systems it is typically assumed that any observed network is optimal due to it persisting through any selection processes over time [90, 91]. However, it is not always clear what aspect of the system is being optimized in a given system. Furthermore, it is possible that there are factors competing against whatever metrics are driving the network to a given structure.

In this work we assume that the primary driver of the network structure in the GNM is longevity. However, any competing influences are hard to describe since there is no direct mapping of human health aspects on to the health aspects in the GNM. Therefore, we use a maximization of entropy on the network structure as a competing factor which drives a diversity in the types of nodes and the connections between those nodes. We argue that since no distinct influence on the network structure can be imposed, that a maximum entropy approach will represent a generic set of exterior factors. Network entropy has previously been shown to promote random network structures [92, 93, 94], with some definitions being extended to characterizing diffusion processes on networks [95]. We use a definition of network entropy which promotes a diversity of connections in the network.

3.1.1 Network Structure Notation and Metrics

3.1.1.1 Useful Jargon

The networks considered in this work are fully connected, unweighted, undirected, simply-connected graphs. This means that all nodes in the network are can be reached by all other nodes via edge-traversal, all edges in the graph have identical weight, the interactions between two nodes proceeds identically in both direction, there can be

at most one connection between nodes, and nodes cannot connect to themselves.

One way to describe network structures is to consider network motifs, also commonly referred to as sub-graphs. A sub-graph is typically a small section of a network where the "neighbourhood" is well defined [96, 97, 98]. In the case of a social network, an example of a sub-graph would be a particular friend-group. Individuals within the friend group share many interactions within the group but do not share many connections outside of that group. Here, the friend-group is a sub-group of a social structure and is represented by a sub-graph element in a social network. Note that a sub-graph must be connected to the rest of the network, otherwise it is just an isolated graph and should be considered separately.

3.1.1.2 Mathematical Representations

The number of neighbours a node has is referred to as the degree of the node and is denoted by k. In the case of unweighted, undirected graphs the degree is often the easiest way to classify nodes in the network. If all nodes and all edges behave identically, then the number of edges is the only thing distinguishing the behaviour of one node from another. A useful framework for describing networks is the d-K framework [99]. The general idea is to assume that the structure of the network is defined by the connection between nodes based only on their degrees.

The 0-th order (d-0) description of a network is to only fix the average degree

$$\langle k \rangle = \sum_i \frac{k_i}{N}. \tag{3.1}$$

The average degree describes the density of connections between nodes in the network. In this work, the average degree is used to constrain the optimization, so the density of connections in the network cannot change throughout optimization.

d-1 graphs are constrained by the distribution of degrees in the network $P(k)$. $P(k)$ (or P_k) is the probability that a randomly chosen node in the network has degree k. The average degree the network can be retrieved from $P(k)$ description following

$$\langle k \rangle = \sum_i k_i P(k_i) = \sum_k k P_k. \tag{3.2}$$

Practically, a network will never exactly follow a prescribed degree distribution due to finite size constraints. Often the more practical measure is the number of nodes of a

given degree $D(k)$ (or D_k). Typically this D_k is sampled from the degree distribution or is a product of the graph generating methodology. In the latter case the degree distribution is calculated using

$$P(k) = \frac{D(k)}{N}. \tag{3.3}$$

The degree distribution defines much of the characteristics of a graph. For instance, scale-free networks are defined by their powerlaw degree distributions $P(k) \sim k^{\alpha}$.

The most detailed description of graphs relevant to this work is the d-2 series of graphs. d-2 graphs are determined by the probability of an edge being present between nodes of degree k and nodes of degree k'. The relevant metric for describing these graphs is the joint degree distribution (or degree correlation matrix) $P(k', k)$. $P(k', k)$ (or $P_{k',k}$) is the probability that a randomly selected edge in the network is connected on one side to a node of degree k and other to a node of degree k'. In the case of undirected graphs $P(k', k)$ is symmetric. The degree distribution can be calculated using

$$P(k) = \frac{\langle k \rangle}{k} \sum_{k'} P(k', k). \tag{3.4}$$

Similarly to the degree distribution, the joint degree distribution (unless measured) is never exactly represented by a finite size network. The useful quantity in this case is the joint degree matrix $J(k', k)$ which describes the number of edges in the network connected on one side to a node of degree k and other to a node of degree k'. Which can be used to calculate the degree correlations using

$$P(k', k) = \frac{J(k', k)}{N\langle k \rangle}. \tag{3.5}$$

The joint degree distribution is useful for characterizing which types of connections are prevalent in the network. For instance, are most of the connections in the network between nodes of similar degree? Networks with nodes which connect to nodes of similar degree are called assortative networks [100]. If nodes connect to nodes of very different degrees the network is dissassortative. The joint degree distribution can be used to tune and measure the assortativity of networks.

Higher-order structural features such as those describing how 3 or more nodes are connected (e.g. in a line vs in a triangle) and how that depends on degree are not investigated in this work.

3.2 Generic Network Model

In this work we follow the implementation of the GNM described in Farrell et al. (2018) exactly, unless otherwise noted. The GNM is an unweighted, undirected, simply connected, complete network with $N = 10^4$ nodes. Each of the nodes represents some health attribute in the system which can be in either a healthy ($d_i = 0$) or unhealthy ($d_i = 1$) state. All nodes begin simulations in the healthy state and damage randomly with a base rate of $\Gamma_o = 0.00183 year^{-1}$. The damage rates of a node increases when it's neighbouring neighbours damage. The damage rate of the nth node is

$$\Gamma_n^+ = \Gamma_o e^{\gamma^+ f_n}, \tag{3.6}$$

where f_n is the FI calculated over the nodes neighbours

$$f_n = \sum_{\langle n \rangle} \frac{d_i}{k_n}. \tag{3.7}$$

Here the sum over $\langle n \rangle$ denotes the nodes neighbouring the nth node, of which there are k_n. γ^+ is set to 7.5 in the GNM.

Mortality in the GNM was implemented using the damage state of the top 2 most connected nodes in the network, where damage of both would constitute mortality. However, in this work we use a proportional hazards model of mortality [101]. The mortality rate in the network increases proportionally to the FI measured across the whole network:

$$\Gamma_M = \Gamma_d e^{\gamma_m \sum \frac{d_i}{N}}. \tag{3.8}$$

This mortality is not massively different from the two node mortality model proposed in the earlier work; mortality is effectively implemented using just one mortality node which is fully connected to the network, albeit with slightly different parametrization. The phenomenology of the proportional hazards mortality is similar to the two node mortality model with minor tweaks to parameters. Here, $\Gamma_d = 0.01$, $\Gamma_m = 8$, and a scale free exponent of $\alpha = 2.35$ (in stead of the original $\alpha = 2.27$) recover the mortality rate and FI curves as a function of age from the original model. The rationale for changing the mortality rule will be explained in later sections.

The network structure originally used to model human aging data is a scale free network constructed with the Barabasi-Albert preferential attachment method [32].

The scale free exponent used is $\alpha = 2.27$, the average degree of the network is $\langle k \rangle = 4$, and the minimum degree is 2. Other common network structures investigated in previous work include random graphs [102], small world networks [103], and assortativity-tuned scale free networks. All of these alternate network structures failed to properly capture the phenomenology of human health and aging.

An important detail of a human aging captured by the GNM is the mapping between types of nodes and the levels of health discussed in the aging literature [36]. It was shown that the higher degree nodes mapped on to the highest level health deficits such as functional impairments, while lower degree nodes behaved more along the lines of lab-based deficits [5]. This behaviour was only observed for dissassortative scale free networks. The other network structures investigated would have high value for high degree nodes but would not have high information content in the lower degree nodes.

3.3 Optimization

3.3.1 Approaches

3.3.1.1 Driving by Degree Distribution

In this work we assume that the degree distribution is the largest factor contributing to the performance of the network. It is likely that higher order factors such as the degree correlations play a role, but previous work suggests that degree distributions must be broad for correlations to have a significant impact [5]. Also, having tried both degree correlation and degree distribution optimizations, the same or better results can be achieved with degree distribution tuning in a small fraction of the time; optimizing in degree correlation space adds a dimension to the optimization space. Furthermore, imposing degree correlations directly on a network is computationally expensive.

3.3.2 Steering by Assortativity

To tune the degree correlations in our networks we use a parametric approach. We select between assortative, dissassortative, and random assignments of edges between degrees k and k'. The proportion of which type of assignment is used tunes the

assortativity using only 3 parameters. This approach does not allow direct changes to the degree correlations, but changing the proportions of which assignment is used has intuitive results.

3.3.2.1 Non-Parametric Approach

This approach allows the degree distribution to vary arbitrarily while maintaining average degree and normalization. We use an evolutionary algorithm - making random changes and keeping successful changes [104]. The main challenge with this approach is that the degree distributions relevant to this application are long-tailed, so the potential optimization space is large. Furthermore, the differences in performance of nodes of large degree do not vary significantly, so much of the long tail of the distribution is largely redundant. For computational convenience we down-sample the available degrees in this approach. It is possible that this reduction in possible degrees affects the results of our optimization. For instance, maximum entropy solutions will be skewed by not including all degrees. However, the entropy measures are primarily used to promote a diversity of connections between types of nodes. So, if a set of nodes in a given range of degrees has similar behaviour, the set can be effectively reduced to one representative degree. Results thus far do not contradict this assumption, but further computation will be needed for verification.

3.3.2.2 Variational Approach

In this approach we investigate the neighbourhood of networks similar to those that best represent human health. Here we tune the assortativity of scale-free networks used previously in the GNM work to investigate the effects on life and healthspan.

3.4 Methods

3.4.1 Measuring Healthspan

Measuring the lifespan of simulated individuals in the GNM is straightforward. However, expressing healthspan in an interpretable metric is somewhat more difficult. Since health is represented entirely by the FI in the GNM, we integrate the health as measured by the FI across the lifespan of the individual to represent health span.

Specifically, we use a measure called the quality adjusted life years (QALY) that integrates the remaining health until death age (t_d)

$$QALY = \int_0^{t_d} (1 - FI(t))dt. \tag{3.9}$$

However, this measure can be dominated by extended longevity, even in the case of extended periods of poor health. A complimentary measure which controls for extended periods of poor health is the healthy aging index (HA)

$$HA = \frac{1}{t_d} \int_0^{t_d} (1 - FI(t))dt. \tag{3.10}$$

The HA measures the proportion of life spent healthy, as opposed to the QALY which has units of healthy years.

3.4.2 Measuring Network Entropy

We use the shannon entropy [105] measured over the degree distribution

$$S(P_k) = -\sum_k P(k) log(P(k)) \tag{3.11}$$

and over the degree correlations

$$S = -\sum_{k,k'} P(k', k) log(P(k', k)). \tag{3.12}$$

We use the entropy measured over the degree correlations unless otherwise mentioned.

3.4.3 Merit Functions

We combine the health performance and entropy of our generated networks to define the evolutionary merit as

$$M = \lambda S + (1 - \lambda)\langle t_d \rangle. \tag{3.13}$$

Here the lagrange multiplier λ determines how heavily the optimization is weighted towards a maximum entropy solution. We have deliberately set the range of the lagrange multipliers to $\lambda \in [0, 1]$. However, entropy measures are typically order 1 and lifespan is typically order 100 so the main behaviour is observed around $\lambda \sim 0.9$. Here we show death age t_d as the health metric, average HA or average QALY could also be used.

3.4.4 Generating Networks

The network generation pipeline is as follows:

- Degree counts $D(k)$ are sampled from a degree distribution $P(k)$ or generated using a graph generation algorithm.
- Edges are assigned between nodes of degree k and k' through the connection matrix $J(k', k)$, depending on the assortativity parametrization.
- The network is generated from the connection matrix following Gjoka et al. [106].

The important details of our methods lie in how the edges are assigned in the connection matrix. Fundamentally, the connection matrix consists of taking weighted random selections between purely assortative, dissassortative, and random connections.

An assortative connection will select the degree k' which is closest to degree k, and assign an edge if there is one available. A dissassortative connection selects the degree k' which is furthest from degree k. Again, since k is sorted descending k' begins at the minimum degree and increases when the constraints are met. A random edge connection selects from the set of degrees k' with probabilities weighted by the number of remaining possible edges between nodes of degree k and nodes of degree k'. Things which influence the number of available connections are the number of outgoing edges from nodes of a given degree $E(k) = kD(k)$, the number of edges already assigned to nodes of that degree, and the number of unique pairs of nodes - since there can only exist one edge between any two nodes.

These three methods have associated probabilities which must sum to 1. An illustration of this approach is shown in figure 3.1. When the degree distribution does not have a long tail the resulting degree correlations are very well behaved. In the case of fully assortative connections it is mostly a linear correlation between k and k'. In terms of common metrics, the degree correlation coefficient $r \in [-1, 1]$ for the fully assortative case approaches 1. The degree correlation coefficient for the fully dissassortative case does not saturate as effectively, typically not reaching below $r \sim -0.95$. Increasing the probability of random connections adds the desired

uncorrelated blob effect, with the fully random limit resembling a slightly skewed gaussian. The degree correlation coefficient of these graphs is close to 0.

3.4.5 Optimization Procedure

The optimization proceeds using an evolutionary algorithm [104]. A degree distribution is initialized. This degree distribution is sampled using the random hubs method described in [107]. A network is generated using these sampled degree as described above and network characteristics such as joint degree distributions are measured. The network is input into the GNM to measure health and longevity statistics. The GNM performance is used alongside entropy measured over the joint degree distribution of the network to determine the merit score as described above. The merit is compared with the merit of the previous best performing network, and is selected using simulated annealing procedures [108]. The best performing networks degree distribution is then modified randomly and the procedure repeats.

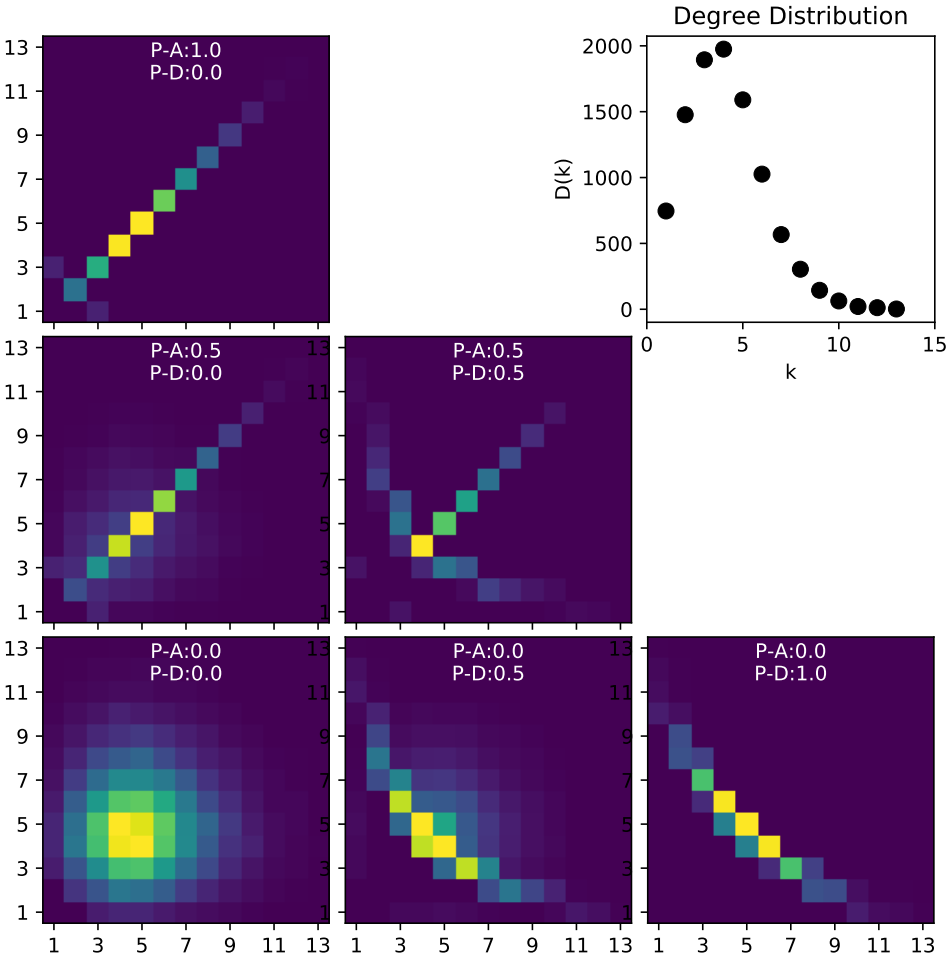

Figure 3.1: Example degree correlations for networks generated using our assortativity tuning algorithm. Here we use a degree sequence generated using an erdos-renyi random graph generator. Network size for all graphs is 10^4 nodes, probability of generating an edge is set to 4/N. All graphs are fully connected.

3.5 Results

3.5.1 Non-Parametric Optimization

3.5.1.1 Verification of Approach

To verify that our non-parametric manipulation of the degree distribution can find optimal solutions, we first optimized the entropy of the degree distribution. Maximum entropy solutions to degree distributions with constrained average degree can be solved analytically and have exponential form [109]. In figure 3.2 we show that our non-parametric degree distribution optimization can replicate this result. Note that this optimization was not performed by generating networks and measuring the distribution, so the degree distribution does not suffer from sampling or rounding errors.

3.5.1.2 Death Age Optimization

Optimization of the network structure for death age leads to an extremely dissassortative network structure. The degree correlations are characterized by connections between the lowest degree nodes and highest degree nodes. This dissassortative limit is observed for a variety of allowed degrees; in figure 3.3 the degree 1 nodes are all connected to the degree 30 nodes which are the highest in this optimization. In figure 3.4 the minimum degree is increased to $k = 2$ and the maximum is increased to $k = 500$, the same extreme dissassortativity observed. There is one critical difference when changing minimum degree; when degree 1 nodes are allowed there is a characteristic highly assortative block in the degree correlations, seen in the degree 15 interconnectedness in figure 3.3.

3.5.1.3 Network Motifs and Sub-Graphs

Across all of the approaches used in this work there have been a handful of network motifs which have consistently emerged as optimal death age networks. The first of which we refer to as the star motif, seen in figure 3.5a. In this substructure of size n there are $n-1$ nodes of degree 1 connected to 1 node of degree $n-1$. To understand why this structure emerges consider the following. The damage rates of all of the

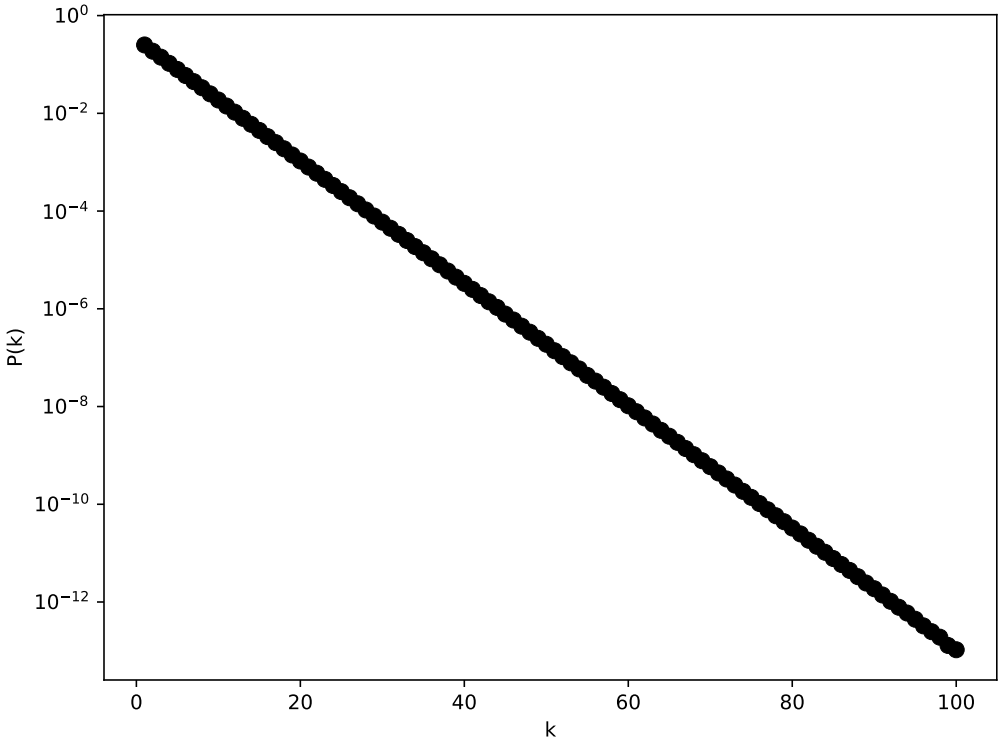

Figure 3.2: The optimal degree distribution for maximizing entropy measured over the degree distribution. Here average degree is constrained to 4. The analytic result of an exponential degree distribution is retrieved using our optimization procedure.

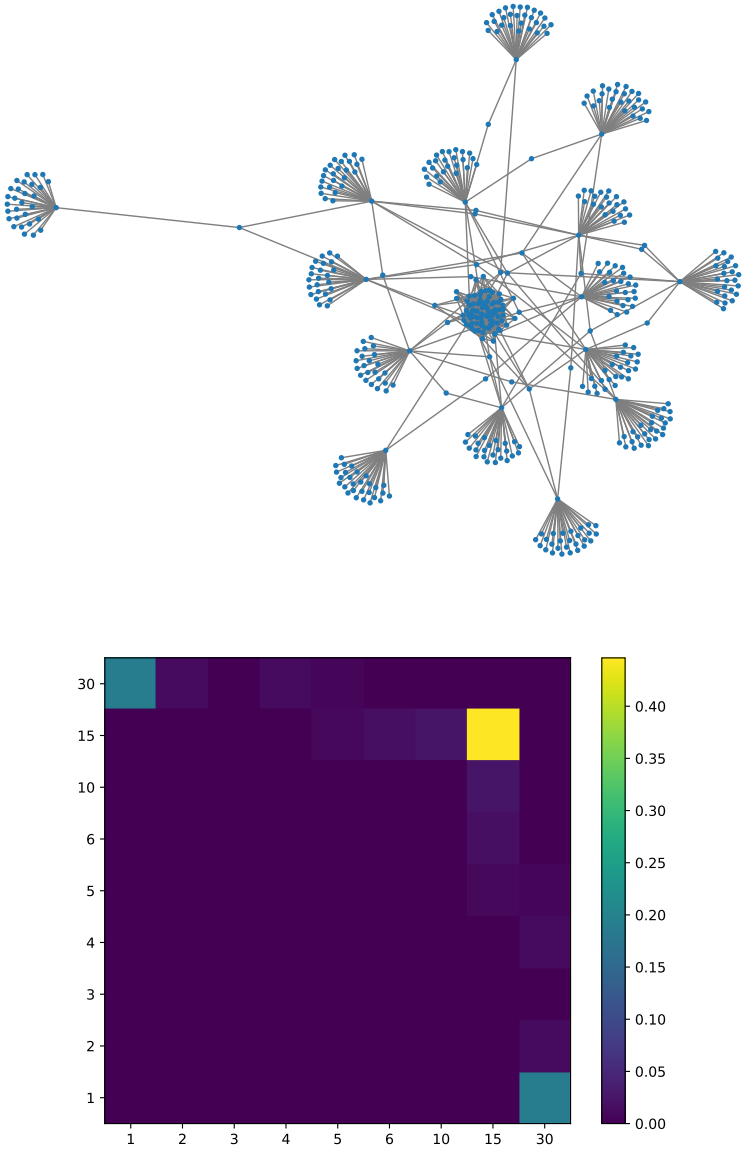

Figure 3.3: An example of a death age optimization using a network size of $N = 500$, an average degree of $\langle k \rangle = 4$, and a minimum degree of 2. The top plot shows a visualization of the network using a spring layout. Note that the many hubs connected to degree 1 nodes seem to be isolated from one-another and from the densely connected core. The bottom plot shows the joint degree distribution for the network. Here the network consists mostly of only 2 types of edges, edges between degrees 1 and 30, and edges between nodes of degree 15.

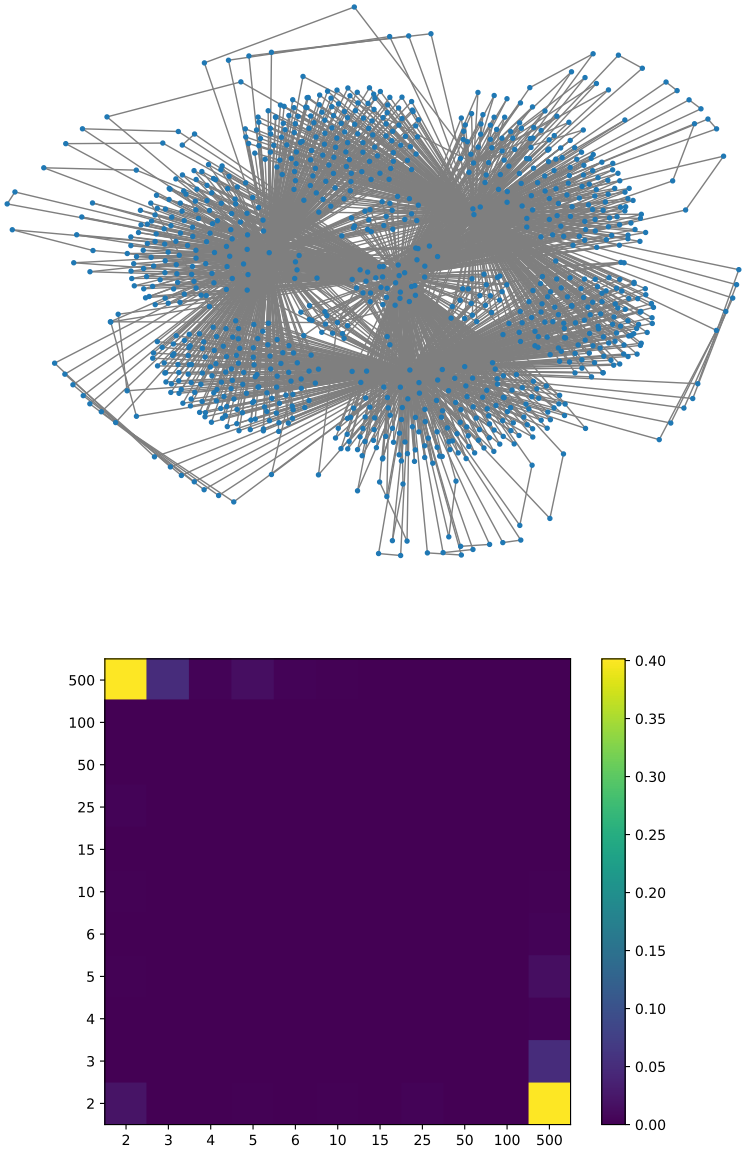

Figure 3.4: The results of a death age optimization on a network constrained to $N = 1000$ nodes, average degree $\langle k \rangle = 4$, and minimum degree 2. The upper figure shows the network structure arranged using a spring layout. There are 4 central hub nodes of degree $k = 500$ predominantly attached to degree 2 nodes in the parents motif. The lower plot shows the degree correlations of the optimal death age network. The correlation matrix is extremely sparse with almost all connections being between nodes of degree 2 and nodes of degree 500.

peripheral nodes is determined by the local frailty of their neighbours, in this case this is exactly the damage state of the hub node. So the peripheral nodes damage at the base damage rate until the hub node damages, after which they damage in short order. The hub node's damage rate is based on the FI of all of the peripheral nodes, so any peripheral node damaging only increases the local FI by $1/(n-1)$, which is vanishingly small. Furthermore, since the damage rates of the peripheral nodes will remain at the base rate, these increments in the damage rate are infrequent. Other than the individual who had their hub node damage randomly early on, this star sub-graph is the most efficient way to prolong lifespan and healthspan, seen in figure 3.6.

The other motif that emerges in optimizations with minimum degree 1 is a fully-connected sub-graph with all n nodes connected to all $n-1$ other nodes (figure 3.5b). Referred to here as a ball due to the resemblance to a rubber band ball, this motif is a set of nodes of high degree with most edges being between nodes within the set. Any individual node damaging only increments the local frailty of it's neighbours by $1/(n-1)$, but there many shared neighbours. The ball motif is not particularly long-lived, as seen in figure 3.6, so why does the ball show up alongside the star? The ball motif has average degree of $n-1$, the highest possible average degree in without allowing multiple connections or self-connections. This high average degree means that adding a ball to the network is the most node-efficient way to raise the average degree, which is necessary since the average degree of the star motif is $2(1-\frac{1}{n})$. So the optimal network structure for longevity is a star sub-graph to promote longevity and a ball sub-graph to satisfy the average degree constraint with minimal negative health contribution.

When the minimum degree is constrained to be greater than one the motifs which appear change significantly. The networks used to model human health have minimum degree of 2, disallowing the star motif. However, there is an analogous structure which simply increases the degree of the peripheral nodes by connecting them to another high degree node. We refer to this structure as the parents motif, shown in figure 3.5d. This parents motif is common in the networks used to model human health data due to the dissassortativity of those networks. In figure 3.6 we show that it is excellent at prolonging lifespan and healthspan.

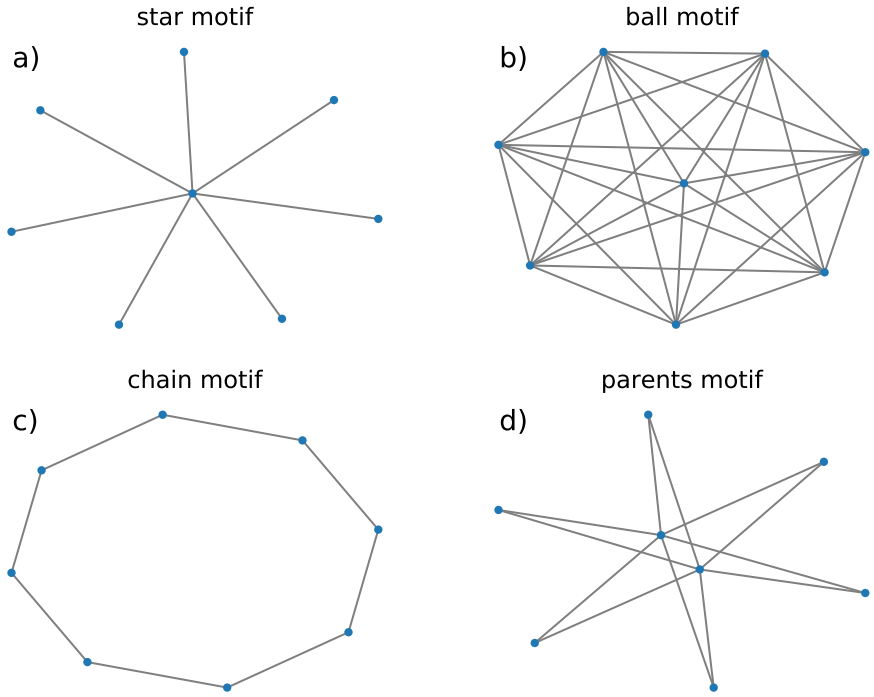

Figure 3.5: Visualizations of the motifs which appear most often in networks during optimization. a) The star motif, most efficient for extending lifespan. b) The ball motif, most efficient at increasing average degree. c) Chain motif, the low degree assortative connection limit. d) The parents motif, the dissassortative limit for minimum degree 2.

One defect from the parent motif that appears in death age optimizations is having multiple degree 2 nodes form a loop connected to the high-degree hub (seen in figure 3.4). We refer to this motif as a chain of degree 2 nodes, shown in figure 3.5. This chain motif becomes very detrimental to life and healthspan if the chain becomes longer than a few nodes, as seen in figure 3.6.

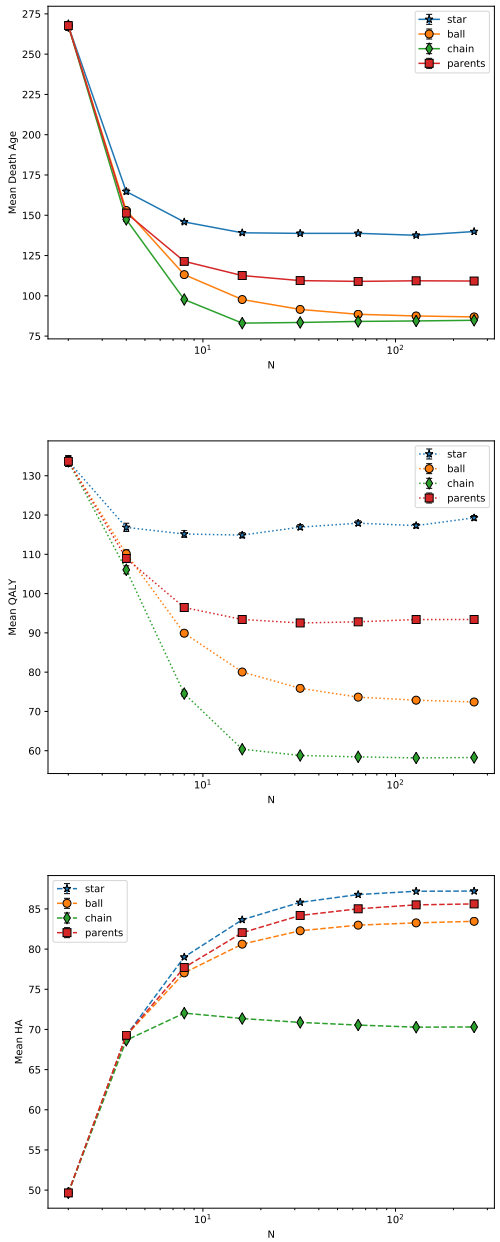

Figure 3.6: Evaluating the performance of the common network motifs in the GNM. In a) we show the average lifespan. The star and parents motifs - the dissassortative limits - have significantly better lifespans. However, the star graph with minimum degree 1 significantly outperforms the parent motif which has minimum degree 2. b) Shows the healthspan of the graph motifs measured by QALY. The healthspan results follow the lifespan results, except for a significant improvement in healthspan for the ball motif over the chain motif despite having similar lifespans. c) Shows the lifespan-adjusted healthspan as measured by the HA. Here the differences between the chain motif and the others is large. The star, ball, and parents motifs are ranked the same as the other health measures but are more similar in performance.

3.5.1.4 Hand-Built Optimal Death Age Networks

Using the results of our death age optimizations, we build "optimal" death age networks by hand using the observed motifs. These optimal networks consist of a tuneable number of star sub-graphs chained together with a ball sub-graph attached on one end. The size of the ball sub-graph is tuned to fulfill the desired average degree in the network. We show examples of these networks in figure 3.7 with 1 and 10 star sub-graphs. The average death age of these networks increases with increasing network size - as they become more heavily weighted towards the star sub-graphs - as shown in figure 3.8a. In the case of one star sub-graph the average death age appears to saturate around the same performance as a pure star sub-graph. Similarly, the health-spans for the one-hub optimal graph saturates to the performance of the pure star sub-graph. The optimal networks which consist of multiple star motifs chained together do not attain the same death age performance for network sizes up to 10^4 nodes. Furthermore, the healthspans of these networks are significantly lower than the 1-hub optimal network and show decreasing trends with increasing number of nodes. As seen in the motifs investigation, chaining nodes together is not an effective way of prolonging healhtspan.

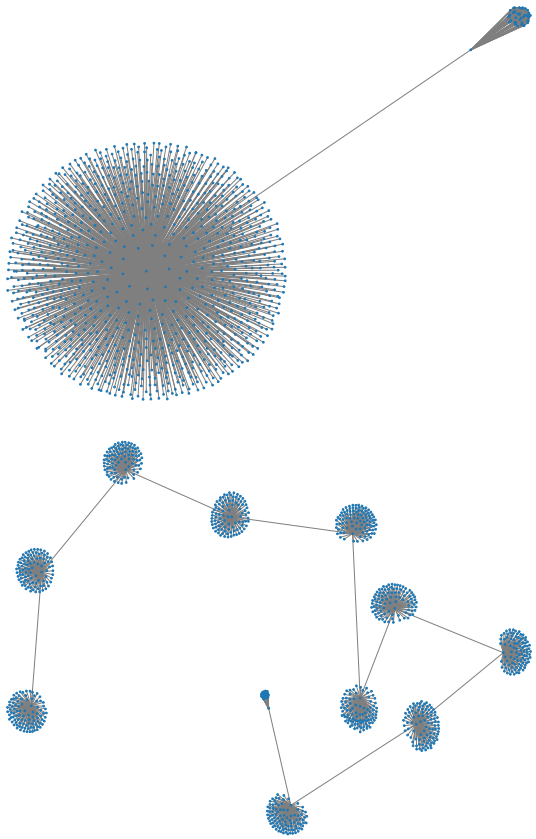

Figure 3.7: Visualizations of optimal death age networks using common network motifs. Both networks pictured have an average degree of 4 and use network sizes of 10^3 nodes. In the top plot we use 1 star motif, in the bottom plot we use 10 star sub-graphs chained together.

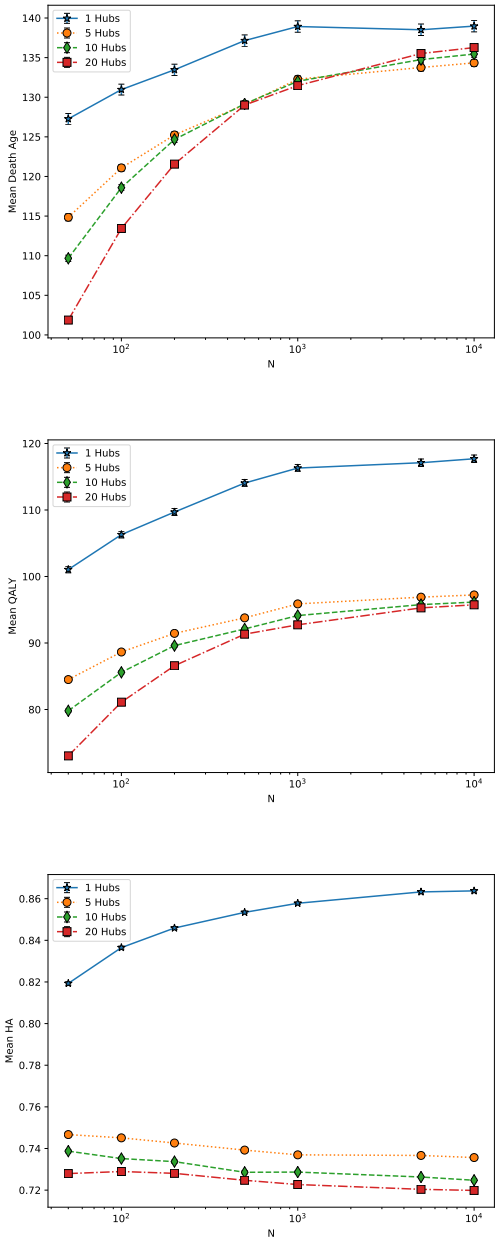

Figure 3.8: Evaluating the performance of the hand built optimal death age networks. In a) we show the average lifespan. At lower network sizes using fewer star motifs is best for extending longevity. b) Shows the healthspan of the same networks. The healthspan results follow the lifespan results, except we do not observe a crossing in performance for larger networks. c) Shows the lifespan-adjusted healthspan as measured by the HA. Here the differences between using only one star motif and the others is extremely large. Furthermore, using more than 1 star motif leads to a decrease in HA with increasing network size. It appears that chaining these star subgraphs together is highly detrimental to HA. In general the 1-hub optimal network has the largest error-bars (standard errors), indicating a much larger variance in health and lifespan compared to the other networks.

3.5.1.5 Adding Entropy to Death Age Optimization

Adding entropy to the merit function by increasing λ has the desired effect of broadening the distribution of connections in the network. We see in figure 3.9 that the extreme dissassortative connections still remain, but the range of high degree and low degree nodes that are included in those categories is expanded. Furthermore, we see that the low degree chain motif is expanded to more nodes of low degree and is more prevalent in the case of higher entropy weighting. There is some representation of higher degree nodes with assortative connections, but not to the degree of a ball motif forming.

The degree distributions of networks also become more broad with increasing λ. Figure 3.10 shows that increasing lambda decreases the "slope" of the distribution. The distributions are characterized by a very steep decrease at low degrees, more data of higher quality is needed to be confident in an assertion of functional form. Also of note is the decrease in prevalence of extremely high degree nodes. So, although it appears that the tail of the distribution is becoming thicker, the extremely high degree nodes are decreasing in number.

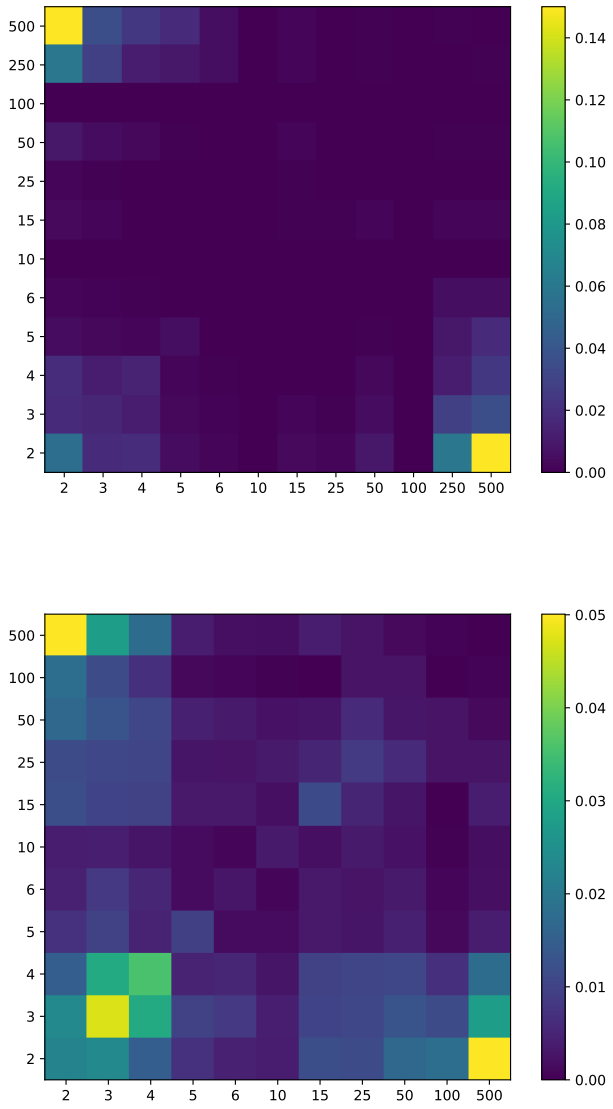

Figure 3.9: An example of the degree correlation matrix following death age optimization with entropy weighted at $\lambda = 0.8$ (top) and $\lambda = 0.9$ (bottom). We find that there is still the characteristic connections between the very low and very high degree nodes, but the range of high and low degree nodes has broadened. Furthermore, the number of connections between low degree nodes is increased and there is some prevalence of assortative connections between higher degree nodes.

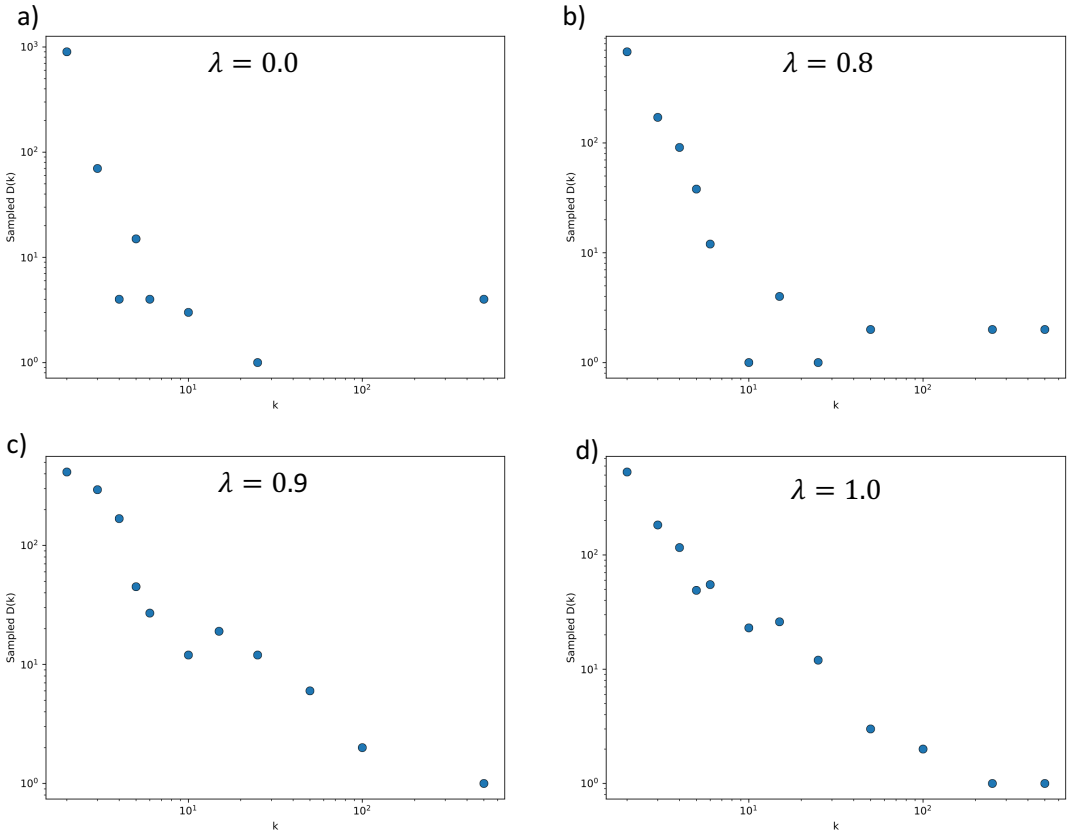

Figure 3.10: An example of the degree correlation matrix following death age optimization with entropy weighted at $\lambda = 0.9$. We find that there is still the characteristic connections between the very low and very high degree nodes, but the range of high and low degree nodes has broadened. Furthermore, the number of connections between low degree nodes is increased and there is some prevalence of assortative connections between higher degree nodes.

3.5.2 Assortativity and Performance of Scale-Free Networks

In figure 3.11 we show how our assortativity tuning approach affects the performance of the degree distribution used to capture human data. We find that the networks which maximize lifespan are the most assortatively tuned in our approach. However, these networks still have a degree assortativity coefficient below 0, indicating a slightly dissassortative network. Furthermore, we find that the degree assortativity coefficient does not effectively capture the variability in the networks when tuning assortativity using our approach; the proportion of dissassortative connections does not significantly change the coefficient, but does significantly effect the entropy and the health measures. The average QALY increases with the fraction of dissassortative connections in the network, as well as with decreasing number of assortative connections. Changing the assortativity parameters in this set of networks can gain 1.5 health adjusted years of life despite decreasing absolute lifespan by about the same amount. The average HA follows the QALY fairly closely.

3.5.2.1 Unexpected Results for Scale-Free Networks

The results of our parametric reconstruction of scale-free networks are not what we expected. Better performance in healthspan measures using dissassortative construction match our expectations, but the assortative structures out-living the dissassortative is shocking. In figure 3.12 we show visualizations of a scale-free degree distribution of exponent $\alpha = 2.27$ and network size $N = 500$ constructed into purely assortative and purely dissassortative networks. We find that the dissassortatively wired network has a high prevalence of parents sub-graph type structures, which we expect to indicate very high lifespan. Furthermore, in the assortative case the majority of nodes in the network are in a degree 2 chain sub-graph, indicating poor lifespan. Both of these observations point towards the dissassortative network out-living the assortative one, but that is not the case. It appears that there are structures in these networks which have significant impact on health which are more involved than a purely motif-motivated understanding of the problem.

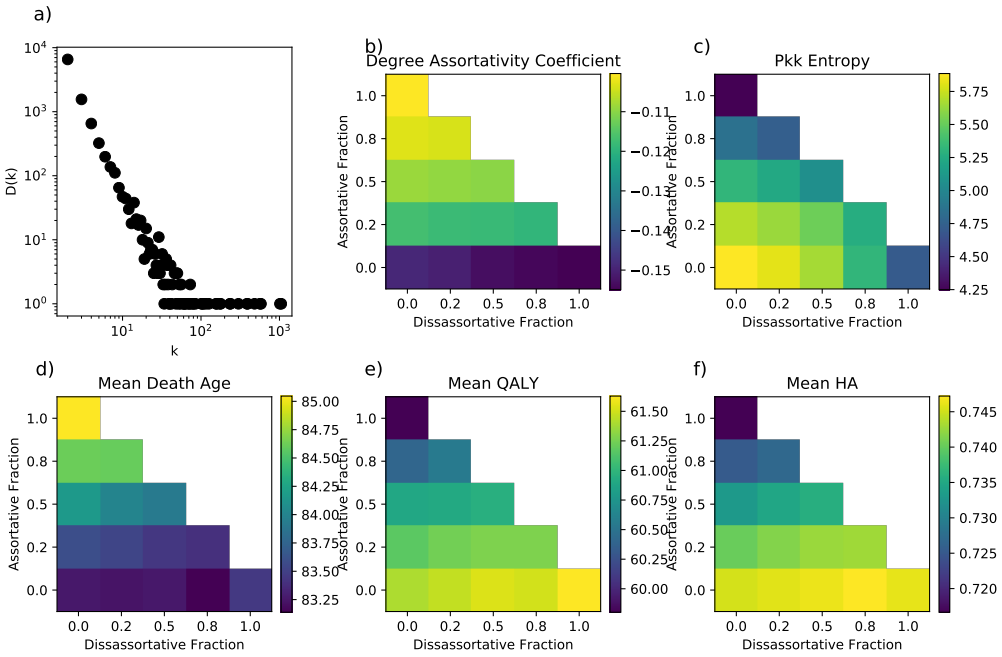

Figure 3.11: The network characteristics and simulated health data for scale free networks with 10^4 nodes. Degree distribution is generated with scale free exponent $\alpha = 2.35$ and average degree $\langle k \rangle = 4$ using the Barabasi-Albert preferential attachment method. The degree sequence is identical for all plots. a) Shows the sampled degree distribution $D(k)$. b) Shows the degree assortativity coefficient. c) Shows the entropy of the degree correlations. d) Shows the average death age in the GNM for the generated networks. e) Shows mean QALY from GNM simulations. Interestingly the QALY does not follow death age. f) Shows mean HA (QALY normalized by death age), which largely follows the QALY along the assortative axis, but differs slightly along the dissassortative axis. Simulations are using 10^4 individuals, and only 1 instantiation of the network.

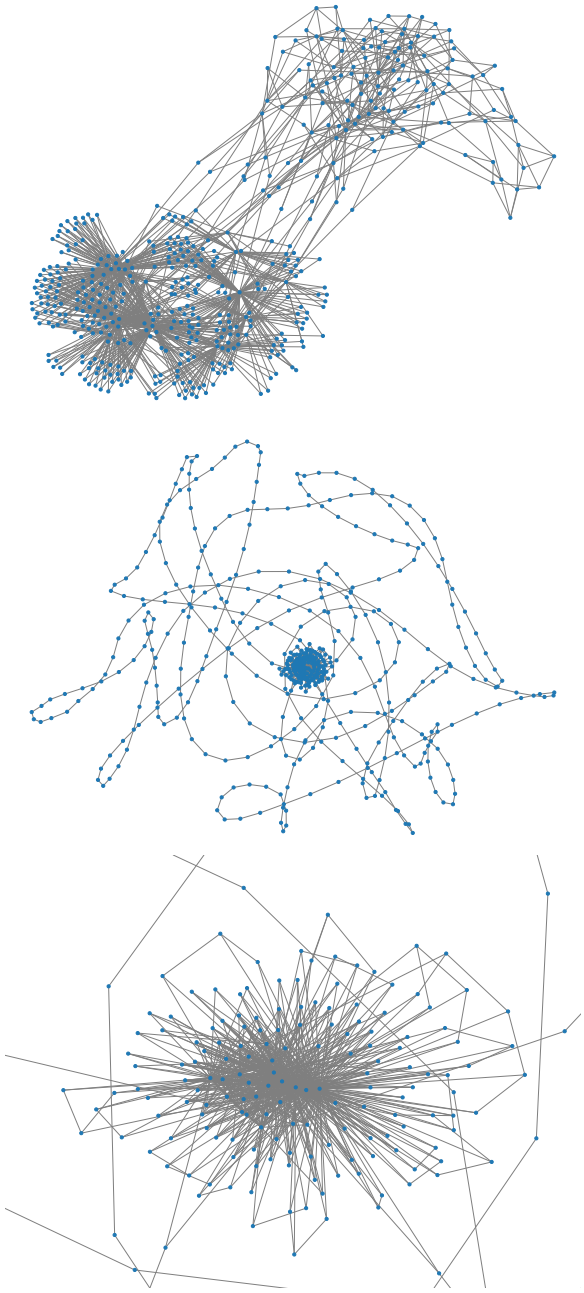

Figure 3.12: Examples of scale free networks with 500 nodes and scale free exponent $\alpha = 2.27$. The top figure shows the network configured with purely dissassortative attachments. The middle plot shows the same degree sequence connected using purely assortative moves. The bottom plot is the same network zoomed in on the central hub-node. The dissassortative network has lower average lifespan, but higher healthspan measured by QALY and HA than the assortative network.

3.6 Discussion

3.6.1 Non-Parametric Optimization: Discovering the Network Motifs of Longevity

Direct optimization of the degree distribution has yielded some interesting results. A handful of network motifs have emerged which dominate the large lifespan and healthspan limits. Furthermore, these motifs are fairly "cheap" from an optimization perspective; they can be combined with inefficient aspects of the network and still perform very well. Finding network motifs at the extreme ends of optimizations is not without precedent [89, 94], but it is not clear whether the motifs solutions are reasonable structures to represent aging of biological systems. Furthermore, networks composed only of simple motifs lack a reasonable scales of health interpretation, since the nodes in the motifs have clear purposes.

Adding entropy into the merit function was successful in promoting a diversity of nodes and connections between nodes. Adding progressively more entropy had the effect of broadening both the degree distribution and the aspects of the degree correlations seen in optimal death age networks.

3.6.2 Quality of Optimization Results

The results presented in this work are representative but not definitive. Optimizing death age non-parametrically has good results because there was a clear trend towards certain sub-graphs as the optimization quality increased. Furthermore, extending those trends by hand lead to even greater performance, solidifying understanding. In the case of the mixed optimization with entropy, our results thus far show the desired mixing between the GNM performance driven network motifs and the broadening effects of entropy. That being said, the quality of the optimizations presented here are not high enough to present definitive results on aspects such as the degree distribution or degree correlations. Compelling results are within reach using the techniques described in this chapter and will be further explored in the coming months when developing this work for publication.

3.6.3 Scale-Free Results: Pushing for a Variational Approach

The unexpected results shown for the longevity of scale-free networks is exciting. These results point towards higher-level structural effects of the network on longevity in the GNM, which is beyond our current understanding. To better explore these results it is necessary to fully develop a variational approach to the optimization. The variational approach should be less likely to fall into the simple network motif structures observed in the non-parametric optimization. Forms of the degree distribution which vary smoothly may be able to clear up which aspects of the network structure are contributing positively to lifespan, without the expected network motifs. Furthermore, a variational approach will allow rigorous inclusion of long-tailed distributions without the associated computational overhead. Future work will develop a parametric degree distribution that is some combination of powerlaw, gaussian, and exponential functions that should be able to effectively probe the health-space. Certainly such an approach would be able to recover the scale-free results seen here.

3.6.4 Reflection on 2-Node Mortality Condition

The emergence of the star motif as an optimal structure for prolonging the health of a node drove the switch from the 2-node mortality condition to the proportional hazards mortality condition. In the case of 2-node mortality the optimization could extend lifespan by prioritizing just the mortality nodes. Forming the star motif around the mortality nodes would massively improve lifespan while the rest of the network could be in any arbitrary state. That is, the mortality condition was not representative of the health of the network, just of the local health of 2 arbitrary nodes. Given these results, is it valid to use a 2-node mortality model in the GNM? In the case of scale free networks the 2-node mortality is not unreasonable; the mortality nodes in that case are of degree $k \sim 0.1N$, so they are affected by the health of much of the network. However, 2-node mortality may not be reasonable in other network structures with narrow degree distributions such as small world [103] and random [32] networks since the mortality nodes do not widely sample the health of the network.

The optimization of healthspan for a subset of identified nodes being very effective raises further questions about the network structure effects observed in Farrell et al. [5]. One of the main results from that work is that the damage of low degree nodes

can be informative of mortality. Specifically, the degree 2 nodes had increasing information content with increasing neighbour degree. However, a clearer effect would be to classify the degree 2 nodes by their neighbours. For instance, are the most informative degree 2 nodes attached to the mortality nodes? Without checking whether or not a given node is attached to a mortality node it is impossible to know that the observed effect is truly a network effect - driven by neighbour degree - as opposed to an artifact of the mortality condition. It would be interesting to see a re-analysis using proportional hazards mortality or the neighbour degree results broken down by whether or not the nodes are connected directly to the mortality nodes.

3.6.5 Degree 1 Nodes

Degree 1 nodes pose an interesting problem in the GNM. The optimization approaches show that attaching degree 1 to any node in the networks is the most effective way to prolong life and health span. Furthermore, for extreme longevity it is most effective to maximize the number of degree 1 nodes, even the cost of using a ball sub-graph to satisfy average degree. This star motif style of longevity optimized networks does not seem like a network which could reasonably represent a biological system. Following the scales of health argument this star motif represents a system which has one (or some small number) of functional health aspects which are connected solely to a set of base-level health aspects. This structure is not a model of health that represents the many interactions between health aspects in an organism since all nodes in the network are there to satisfy one function. The entire star sub-graph is effectively behaving as just one node with increased health impacts attached to its health state; if it damages it contributes massively to the mortality rate since the peripheral nodes damage soon after, but there are minimal down-stream health effects since the peripheral nodes cannot propagate the damage to elsewhere in the network.

How to deal with the effectiveness of degree 1 nodes is an interesting question. It is not obvious that degree 1 nodes cannot exist in the network structure as a viable health aspect. However, in the current model they are overpowering in a health and lifespan optimization. Our definition of network entropy is aimed at promoting a diversity of connections between health aspects through connections between nodes of different degrees. This approach is effective for nodes of degree greater than 1 because

they can connect different levels of health. Consider nodes of degree 2, a node which is connected to a high degree node and a low degree nodes has different function than one is connected to 2 nodes of high degree. This distinction can not be drawn for degree 1 nodes. Although they can be distinguished by the type of node they are connected to, the function of the degree 1 node will always be to simply buffer the health of whatever node it is attached to. In this case our degree correlation entropy does drive the degree 1 nodes to connect to a variety of different degree nodes, but this does not truly drive a difference in function.

3.6.6 Merit Functions and Evolutionary Fitness

Another approach to dealing with the degree 1 nodes is to develop a more refined approach to the merit used in the optimization. For instance, incentivizing multiple pathways through the network, similarly to the redundancy argument used by Gavrilov [15], would reduce the prevalence of degree 1 nodes since they are dead-ends. However, the question of what merit function is a difficult one in general and it could be that an evolutionarily appropriate fitness does not effectively deal with degree 1 nodes.

Effectively defining evolutionary fitness in a model which only effectively captures health is a difficult task. The organism must be in relatively good health to reproduce, but the parameters of the goodness of health are not clear. Evolutionary entropy is a model which argues that the uncertainty in the age of a mother is an effective measure of fitness [110]. Evolutionary entropy uses healthspan as the foremost indicator of evolutionary fitness, a mother must be sufficiently healthy to reproduce for a large span of time. However, it is placing particular emphasis on having great health for an extended period, and does not consider any extended period of good health beyond the period of sufficient health to reproduce. It is likely that effectively implementing this definition of health would require strict parametrization of sufficient health to reproduce, where the usual motifs for prolonging node health would likely surface. Previously in this network optimization approach we have implemented evolutionary entropy, but we could not effectively disentangle it from results maximizing QALY on a subset of nodes labelled for reproductive fitness. Furthermore, if fitness is described directly pertaining to health, it is likely that the same network motifs would reappear.

It is likely that without defining fitness mechanically in the network - essentially something layered on top of health - that fitness and healthspan will go hand in hand.

Chapter 4

Conclusion

4.1 Tying it all together

In this thesis we have explored health and aging through the context of the FI. We have determined that the FI can be effectively extended to include laboratory-level health deficits. These health aspects can be included using the generic definition of a deficit in the FI to dichotomize the naturally continuous measurements. We further improved the predictive quality of the FI-Lab by including the biomarkers without dichotomizing them. However, there are legitimate concerns about whether our QFI approach is truly an FI given the lack of a healthy state. Nonetheless, our work on FI-Lab has raised some important questions for the field of aging metrics such as the appropriate pre-processing of continuous data and being explicit about how cohort effects and confounding factors affect the resulting health metrics.

We have pushed a network model of health - which leverages the general applicability of the FI - to its limits of lifespan and healthspan. Our results show that at the extremes of life and healthspan the popular levels of health analogy breaks down in our model. We suggest that there must also be some exterior influences on the network structure - some element outside of maintaining good health - which is driving the interactions between health aspects. However, it is also clear that the health aspects are not arranged randomly, in some sort of maximum-entropy configuration. Further work is required on this problem to determine where in-between those two extremes the network structure which best represents human aging lies.

4.2 Opinions

The FI-Lab approaches laid out in this thesis are both effective as summary health metrics. However, there is the question of chasing predictive value in the FI. The

QFI approach certainly seems one step removed from the general FI approach; lacking a healthy state and having continuous deficits. However, as a health metric it is effective at retrieving the information content from the constituent elements and it is interpretable. That being said, I do not think it significantly impacts the understanding of the FI as a framework for understanding health and aging. Most measurements included in the FI in typical studies could be graduated to include more than dichotomous health states and some are already included with more than 2 states. This tension between the FI as a framework and the FI as a health metric is exacerbated when the health aspects are inherently continuous. Shoe-horning more information into an FI-like measurement would certainly be nice from a public health and decision making perspective. However, communicating the FI is much more effective when it is simply defined as the fraction of health aspects which are damaged.

From a modelling perspective I do not think that the QFI being a better predictor in an empirical setting is a significant issue. The modest gains in prediction on a population level are dwarfed by the gains made by more sophisticated predictive models using the same data. In that sense the QFI does not detract from the current models based on the FI - a generic model of health deficits is not ineffective at the laboratory level of health. Descending further towards molecular measurements may still pose a problem when applying an FI approach, but that remains to be seen.

In terms of the network optimization, I believe that this approach can get to a satisfying endpoint. Now, this is not the first time that I have believed that a specific optimization approach will be suitable to the problem, but this one has definite advantages. Spending a few clock-months of compute time optimizing ensembles of larger networks, even if it just solidifies the results presented in this work, will yield interesting results. If the non-parametric optimization flops, even under ideal computing circumstances, a variational approach will at least determine how tightly the GNM depends on the specific scale-free degree distribution. Furthermore, optimizing a network structure to fit the health data of a non-human organism is more attainable with highly tune-able networks like those developed in this work.

Bibliography

[1] A B Mitnitski, A J Mogilner, and K Rockwood. Accumulation of deficits as a proxy measure of aging. *The Scientific World*, 1:323–36, 2001.

[2] L P Fried, C M Tangen, J Walston, A B Newman, C Hirsch, J Gottdiener, T Seeman, R Tracy, W J Kop, G Burke, M A McBurnie, and Cardiovascular Health Study Collaborative Research Group. Frailty in older adults: evidence for a phenotype. *Journals of Gerontology Series A*, 56(3):M146–56, March 2001.

[3] Morgan E Levine. Assessment of epigenetic clocks as biomarkers of aging in basic and population research. *The Journals of Gerontology: Series A*, 75(3):463–465, 2020.

[4] Arnold Mitnitski, Joanna Collerton, Carmen Martin-Ruiz, Carol Jagger, Thomas von Zglinicki, Kenneth Rockwood, and Thomas B. L. Kirkwood. Age-related frailty and its association with biological markers of ageing. *BMC Medicine*, 13(1), 2015.

[5] Spencer G. Farrell, Arnold B. Mitnitski, Olga Theou, Kenneth Rockwood, and Andrew D. Rutenberg. Probing the network structure of health deficits in human aging. *Physical Review E*, 98(3), 2018.

[6] Steve Horvath. DNA methylation age of human tissues and cell types. *Genome Biology*, 14:R115, 2013.

[7] Gregory Hannum, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, Srini-Vas Sadda, Brandy Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, Rob Deconde, Menzies Chen, Indika Rajapakse, Stephen Friend, Trey Ideker, and Kang Zhang. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular Cell*, 49(2):359–367, 2013.

[8] J Graham Ruby, Kevin M Wright, Kristin A Rand, Amir Kermany, Keith Noto, Don Curtis, Neal Varner, Daniel Garrigan, Dmitri Slinkov, Ilya Dorfman, Julie M Granka, Jake Byrnes, Natalie Myres, and Catherine Ball. Estimates of the heritability of human longevity are substantially inflated due to assortative mating. *Genetics*, 210(3):1109–1124, 2018.

[9] Anne Maria Herskind, Matthew McGue, Niels V. Holm, Thorkild I. A. Sørensen, Bent Harvald, and James W. Vaupel. The heritability of human longevity: A population-based study of 2872 danish twin pairs born 1870–1900. *Human Genetics*, 97(3):319–323, 1996.

[10] Ksenia S. Kudryashova, Ksenia Burka, Anton Y. Kulaga, Nataliya S. Vorobyeva, and Brian K. Kennedy. Aging biomarkers: From functional tests to multi-omics approaches. *PROTEOMICS*, 20(5-6):1900408, 2020.

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

[12] Spencer Farrell, Arnold Mitnitski, Kenneth Rockwood, and Andrew Rutenberg. Interpretable machine learning for high-dimensional trajectories of aging health, 2021.

[13] Brian T. Weinert and Poala S. Timiras. Invited review: Theories of aging. *Journal of Applied Physiology*, 95(4):1706–1716, 2003.

[14] L. P. Fried, Q.-L. Xue, A. R. Cappola, L. Ferrucci, P. Chaves, R. Varadhan, J. M. Guralnik, S. X. Leng, R. D. Semba, J. D. Walston, C. S. Blaum, and K. Bandeen-Roche. Nonlinear multisystem physiological dysregulation associated with frailty in older women: Implications for etiology and treatment. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 64A(10):1049–1057, 2009.

[15] LEONID A. GAVRILOV and NATALIA S. GAVRILOVA. The reliability theory of aging and longevity. *Journal of Theoretical Biology*, 213(4):527–545, 2001.

[16] Morgan E. Levine, Ake T. Lu, Austin Quach, Brian H. Chen, Themistocles L. Assimes, Stefania Bandinelli, Lifang Hou, Andrea A. Baccarelli, James D. Stewart, Yun Li, Eric A. Whitsel, James G Wilson, Alex P Reiner, Abraham Aviv, Kurt Lohman, Yongmei Liu, Luigi Ferrucci, and Steve Horvath. An epigenetic biomarker of aging for lifespan and healthspan. *Aging*, 10(4):573–591, 2018.

[17] Robert-Paul Juster, Bruce S. McEwen, and Sonia J. Lupien. Allostatic load biomarkers of chronic stress and impact on health and cognition. *Neuroscience & Biobehavioral Reviews*, 35(1):2–16, 2010.

[18] Samuel D Searle, Arnold Mitnitski, Evelyne A Gahbauer, Thomas M Gill, and Kenneth Rockwood. A standard procedure for creating a frailty index. *BMC Geriatrics*, 8(1), 2008.

[19] 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'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1666):20140379, 2015.

[20] Casey Crump, Marilyn A Winkleby, Kristina Sundquist, and Jan Sundquist. Comorbidities and mortality in persons with schizophrenia: a swedish national cohort study. *American Journal of Psychiatry*, 170(3):324–333, 2013.

[21] Miguel Divo, Claudia Cote, Juan P de Torres, Ciro Casanova, Jose M Marin, Victor Pinto-Plata, Javier Zulueta, Carlos Cabrera, Jorge Zagaceta, Gary Hunninghake, et al. Comorbidities and risk of mortality in patients with chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 186(2):155–161, 2012.

[22] Mary Charlson, Ted P. Szatrowski, Janey Peterson, and Jeffrey Gold. Validation of a combined comorbidity index. *Journal of Clinical Epidemiology*, 47(11):1245–1251, 1994.

[23] 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*, 9:132, February 2020.

[24] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[25] S J Evans, M Sayers, A Mitnitski, and K Rockwood. The risk of adverse outcomes in hospitalized older patients in relation to a frailty index based on a comprehensive geriatric assessment. *Age and Ageing*, 43(1):127–132, 2014.

[26] K Rockwood, X Song, C MacKnight, H Bergman, D B Hogan, I McDowell, and A Mitnitski. A global clinical measure of fitness and frailty in elderly people. *Canadian Medical Association Journal*, 173(5):489–495, 2005.

[27] Susan E Howlett, Michael Rockwood, Arnold Mitnitski, and Kenneth Rockwood. Standard laboratory tests to identify older adults at increased risk of death. *BMC Medicine*, 12(1), 2014.

[28] 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, 2017.

[29] Spencer G Farrell, Arnold B Mitnitski, Kenneth Rockwood, and Andrew D Rutenberg. Network model of human aging: Frailty limits and information measures. *Physical Review E*, 94(5):052409, November 2016.

[30] A.I. Yashin, K.G. Arbeev, I. Akushevich, A. Kulminski, S.V. Ukraintseva, E. Stallard, and K.C. Land. The quadratic hazard model for analyzing longitudinal data on aging, health, and the life span. *Physics of Life Reviews*, 9(2):177–188, June 2012.

[31] Swadhin Taneja, Arnold B. Mitnitski, Kenneth Rockwood, and Andrew D. Rutenberg. Dynamical network model for age-related health deficits and mortality. *Phys. Rev. E*, 93:022309, 2016.

[32] Albert-László Barabási and Eric Bonabeau. Scale-free networks. *Scientific american*, 288(5):60–69, 2003.

[33] Eileen M Crimmins. Lifespan and healthspan: past, present, and promise. *The Gerontologist*, 55(6):901–911, 2015.

[34] Alejandro Martin-Montalvo, Evi M Mercken, Sarah J Mitchell, Hector H Palacios, Patricia L Mote, Morten Scheibye-Knudsen, Ana P Gomes, Theresa M Ward, Robin K Minor, Marie-José Blouin, et al. Metformin improves healthspan and lifespan in mice. *Nature communications*, 4(1):1–9, 2013.

[35] Garrett Stubbings. Quantile frailty index. https://github.com/GarrettStubbings/QuantileFrailtyIndex, 2021.

[36] Luigi Ferrucci, Morgan E Levine, Pei-Lun Kuo, and Eleanor M Simonsick. Time and the metrics of aging. *Circulation Research*, 123(7):740–744, September 2018.

[37] Xian Xia, Weiyang Chen, Joseph McDermott, and Jing-Dong Jackie Han. Molecular and phenotypic biomarkers of aging. *F1000Research*, 6:860, 2017.

[38] Lida Katsimpardi, Nadia K Litterman, Pamela A Schein, Christine M Miller, Francesco S Loffredo, Gregory R Wojtkiewicz, John W Chen, Richard T Lee, Amy J Wagers, and Lee L Rubin. Vascular and neurogenic rejuvenation of the aging mouse brain by young systemic factors. *Science*, 344(6184):630–634, 2014.

[39] Hillary Klonoff-Cohen, Elizabeth L Barrett-Connor, and Sharon L Edelstein. Albumin levels as a predictor of mortality in the healthy elderly. *Journal of clinical epidemiology*, 45(3):207–212, 1992.

[40] Thomas E Johnson. Recent results: biomarkers of aging. *Experimental gerontology*, 41(12):1243–1246, 2006.

[41] Catharine Sturgeon, Robert Hill, Glen L. Hortin, and Douglas Thompson. Taking a new biomarker into routine use - a perspective from the routine clinical biochemistry laboratory. *PROTEOMICS - Clinical Applications*, 4(12):892–903, 2010.

[42] Richard McPherson. *Henry's clinical diagnosis and management by laboratory methods*. Elsevier, St. Louis, Mo, 2017.

[43] D. Gu, M. E. Dupre, J. Sautter, Haiyan Zhu, Yuzhi Liu, and Zeng Yi. Frailty and mortality among chinese at advanced ages. *Journal of Gerontology: Social Sci*, 64B(2):279–289, 2009.

[44] C Seplaki, N Goldman, D Glei, and M Weinstein. A comparative analysis of measurement approaches for physiological dysregulation in an older population. *Experimental Gerontology*, 40(5):438–449, May 2005.

[45] Centers for Disease Control and Prevention National Center for Health Statistics. National health and nutrition examination survey data, Updated 2014.

[46] Canadian Study of Health and Aging Working Group. Canadian study of health and aging: study methods and prevalence of dementia. *Canadian Medical Association Journal*, 150(6):899, 1994.

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

[48] Kenneth Rockwood, Arnold Mitnitski, Xiaowei Song, Bertil Steen, and Ingmar Skoog. Long-term risks of death and institutionalization of elderly people in relation to deficit accumulation at age 70. *Journal of the American Geriatrics Society*, 54(6):975–979, June 2006.

[49] Vic Velanovich, Heath Antoine, Andrew Swartz, David Peters, and Ilan Rubinfeld. Accumulating deficits model of frailty and postoperative mortality and morbidity: its application to a national database. *The Journal of surgical research*, 183(1):104–110, July 2013.

[50] Xiaowei Song, Arnold Mitnitski, and Kenneth Rockwood. Age-related deficit accumulation and the risk of late-life dementia. *Alzheimer's research & therapy*, 6(5-8):54, 2014.

[51] Oliver L Hatheway, Arnold Mitnitski, and Kenneth Rockwood. Frailty affects the initial treatment response and time to recovery of mobility in acutely ill older adults admitted to hospital. *Age and Ageing*, pages 1–6, January 2017.

[52] 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 and Ageing*, 45(4):463–468, April 2016.

[53] N. Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50(3):163–170, 1966.

[54] Stephanie Bennett, Xiaowei Song, Arnold Mitnitski, and Kenneth Rockwood. A limit to frailty in very old, community-dwelling people: a secondary analysis of the Chinese longitudinal health and longevity study. *Age and Ageing*, 42(3):372–377, May 2013.

[55] Ruth E Hubbard, Nancye M Peel, Mayukh Samanta, Leonard C Gray, Brant E Fries, Arnold Mitnitski, and Kenneth Rockwood. Derivation of a frailty index from the interRAI acute care instrument. *BMC geriatrics*, 15(1):27, 2015.

[56] Joshua J Armstrong, Arnold Mitnitski, Lenore J Launer, Lon R White, and Kenneth Rockwood. Frailty in the Honolulu-Asia Aging Study: deficit accumulation in a male cohort followed to 90% mortality. *Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 70(1):125–131, January 2015.

[57] Kenneth Harttgen, Paul Kowal, Holger Strulik, Somnath Chatterji, and Sebastian Vollmer. Patterns of frailty in older adults: comparing results from higher and lower income countries using the Survey of Health, Ageing and Retirement in Europe (SHARE) and the Study on Global AGEing and Adult Health (SAGE). *PLoS ONE*, 8(10):e75847, 2013.

[58] 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 and Ageing*, 45:353 – 360, 2016.

[59] Irene Drubbel, Niek J de Wit, Nienke Bleijenberg, René J C Eijkemans, Marieke J Schuurmans, and Mattijs E Numans. Prediction of adverse health outcomes in older people using a frailty index based on routine primary care data. *Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 68(3):301–308, March 2013.

[60] Alice E Kane, Kaitlyn M Keller, Stefan Heinze-Milne, Scott A Grandy, and Susan E Howlett. A murine frailty index based on clinical and laboratory measurements: Links between frailty and pro-inflammatory cytokines differ in an sex-specific manner. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 74(3):275–282, 2019.

[61] Emily J Nicklett. Socioeconomic status and race/ethnicity independently predict health decline among older diabetics. *BMC Public Health*, 11(1), 2011.

[62] Gloria A Aguayo, Michel T Vaillant, Anne-Françoise Donneau, Anna Schritz, Saverio Stranges, Laurent Malisoux, Anna Chioti, Michèle Guillaume, Majon Muller, and Daniel R Witte. Comparative analysis of the association between 35 frailty scores and cardiovascular events, cancer, and total mortality in an elderly general population in England: An observational study. *PLoS Medicine*, 15(3):e1002543, March 2018.

[63] Sandra M Shi, Ellen P McCarthy, Susan L Mitchell, and Dae Hyun Kim. Predicting mortality and adverse outcomes: Comparing the frailty index to general prognostic indices. *Journal of General Internal Medicine*, 60(10):1–7, February 2020.

[64] Daniel W Belsky, Terrie E Moffitt, Alan A Cohen, David L Corcoran, Morgan E Levine, Joseph A Prinz, Jonathan Schaefer, Karen Sugden, Benjamin Williams, Richie Poulton, and Avshalom Caspi. Eleven telomere, epigenetic clock, and biomarker-composite quantifications of biological aging: Do they measure the same thing? *American Journal of Epidemiology*, 187(6):1220–1230, June 2018.

[65] Alberto Zucchelli, Davide L. Vetrano, Giulia Grande, Amaia Calderón-Larrañaga, Laura Fratiglioni, Alessandra Marengoni, and Debora Rizzuto. Comparing the prognostic value of geriatric health indicators: A population-based study. *BMC Medicine*, 17(1), 2019.

[66] S Michal Jazwinski and Sangkyu Kim. Examination of the dimensions of biological age. *Frontiers in Genetics*, 10:263, 2019.

[67] K Rockwood, A Mogilner, and A Mitnitski. Changes with age in the distribution of a frailty index. *Mechanisms of Ageing and Development*, 125(7):517–519, 2004.

[68] Konrad J. Karczewski and Michael P. Snyder. Integrative omics for health and disease. *Nature Reviews Genetics*, 19(5):299–310, 2018.

[69] Jacob Cohen. The cost of dichotomization. *Applied Psychological Measurement*, 7(3):249–253, 1983.

[70] Altman, Douglas G and Royston, Patrick. The cost of dichotomising continuous variables. *BMJ (Clinical research ed.)*, 332(7549):1080, May 2006.

[71] Patrick Royston, Douglas G Altman, and Willi Sauerbrei. Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine*, 25(1):127–141, January 2006.

[72] Valerii Fedorov, Frank Mannino, and Rongmei Zhang. Consequences of dichotomization. *Pharmaceutical Statistics*, 8(1):50–61, January 2009.

[73] O Naggara, J Raymond, F Guilbert, D Roy, A Weill, and D G Altman. Analysis by categorizing or dichotomizing continuous variables is inadvisable: An example from the natural history of unruptured aneurysms. *American Journal of Neuroradiology*, 32(3):437–440, March 2011.

[74] Neal V Dawson and Robert Weiss. Dichotomizing continuous variables in statistical analysis: A practice to avoid. *Medical Decision Making*, 32(2):225–226, March 2012.

[75] 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 Geriatrics*, 14(1):25, 2014.

[76] A. Tsodikov, A. Szabo, and D. Jones. Adjustments and measures of differential expression for microarray data. *Bioinformatics*, 18(2):251–260, 2002.

[77] Arnold Mitnitski, Susan E Howlett, and Kenneth Rockwood. Heterogeneity of human aging and its assessment. *Journals of Gerontology Series A*, 72(7):877–884, July 2017.

[78] EH Gordon and RE Hubbard. Physiological basis for sex differences in frailty. *Current Opinion in Physiology*, 6:10–15, 2018.

[79] Alexander M. Kulminski, Irina V. Culminskaya, Svetlana V. Ukraintseva, Konstantin G. Arbeev, Kenneth C. Land, and Anatoli I. Yashin. Sex-specific health deterioration and mortality: The morbidity–mortality paradox over age and time. *Experimental Gerontology*, 43(12):1052–1057, 2008.

[80] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[81] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference*, 2010.

[82] Z. Oldfield, N. Rogers, A. Phelps, M. Blake, A. Steptoe, A. Oskala, M. Marmot, S. Clemens, J. Nazroo, and J. Banks. English Longitudinal Study of Ageing: Waves 0-9, 1998-2019, 2020.

[83] Joanne Ryan, Jo Wrigglesworth, Jun Loong, Peter D Fransquet, and Robyn L Woods. A systematic review and meta-analysis of environmental, lifestyle, and health factors associated with DNA methylation age. *The Journals of Gerontology: Series A*, 75(3):481–494, 2019.

[84] Albert-László Barabási. Scale-free networks: a decade and beyond. *science*, 325(5939):412–413, 2009.

[85] Renzo Angles and Claudio Gutierrez. Survey of graph database models. *ACM Computing Surveys (CSUR)*, 40(1):1–39, 2008.

[86] Stavros A. Zenios. Network based models for air-traffic control. *European Journal of Operational Research*, 50(2):166–178, 1991.

[87] Michael R. Garey and David S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman Co., USA, 1990.

[88] Marián Boguá, Romualdo Pastor-Satorras, and Alessandro Vespignani. Epidemic spreading in complex networks with degree correlations. In *Statistical Mechanics of Complex Networks*, pages 127–147. Springer Berlin Heidelberg, 2003.

[89] Michael T Gastner and Mark EJ Newman. The spatial structure of networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 49(2):247–252, 2006.

[90] Yuri I. Wolf, Georgy Karev, and Eugene V. Koonin. Scale-free networks in biology: new insights into the fundamentals of evolution? *BioEssays*, 24(2):105–109, 2002.

[91] Vittoria Colizza, Jayanth R Banavar, Amos Maritan, and Andrea Rinaldo. Network structures from selection principles. *Physical review letters*, 92(19):198701, 2004.

[92] RJ Mondragón. Estimating degree–degree correlation and network cores from the connectivity of high–degree nodes in complex networks. *Scientific reports*, 10(1):1–24, 2020.

[93] Li Ji, Wang Bing-Hong, Wang Wen-Xu, and Zhou Tao. Network entropy based on topology configuration and its computation to random networks. *Chinese Physics Letters*, 25(11):4177, 2008.

[94] Ramon Ferrer i Cancho and Ricard V. Solé. Optimization in complex networks. In *Statistical Mechanics of Complex Networks*, pages 114–126. Springer Berlin Heidelberg, 2003.

[95] Jesús Gómez-Gardeñes and Vito Latora. Entropy rate of diffusion processes on complex networks. *Physical Review E*, 78(6), 2008.

[96] R. Milo. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[97] Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007.

[98] Shalev Itzkovitz and Uri Alon. Subgraphs and network motifs in geometric networks. *Physical Review E*, 71(2), 2005.

[99] Priya Mahadevan, Dmitri Krioukov, Kevin Fall, and Amin Vahdat. Systematic topology analysis and generation using degree correlations. *ACM SIGCOMM Computer Communication Review*, 36(4):135–146, 2006.

[100] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.

[101] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[102] P. Erdös and A. Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290, 1959.

[103] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[104] Thomas Bäck and Hans-Paul Schwefel. An overview of evolutionary algorithms for parameter optimization. *Evolutionary computation*, 1(1):1–23, 1993.

[105] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[106] Minas Gjoka, Bálint Tillman, and Athina Markopoulou. Construction of simple graphs with a target joint degree matrix and beyond. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 1553–1561. IEEE, 2015.

[107] Maria Letizia Bertotti and Giovanni Modanese. The configuration model for barabasi-albert networks. *Applied Network Science*, 4(1):1–13, 2019.

[108] Peter JM Van Laarhoven and Emile HL Aarts. Simulated annealing. In *Simulated annealing: Theory and applications*, pages 7–15. Springer, 1987.

[109] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[110] Lloyd Demetrius, Stéphane Legendre, and Peter Harremöes. Evolutionary entropy: a predictor of body size, metabolic rate and maximal life span. *Bulletin of mathematical biology*, 71(4):800–818, 2009.

[111] Cameron Davidson-Pilon, Jonas Kalderstam, Paul Zivich, Ben Kuhn, Andrew Fiore-Gartland, AbdealiJK, Luis Moneda, , Gabriel, Daniel WIlson, Alex Parij, Kyle Stark, Steven Anton, Lilian Besson, , Jona, Harsh Gadgil, Dave Golland, Sean Hussey, Ravin Kumar, Javad Noorbakhsh, Andreas Klintberg, Dylan Albrecht, Dhuynh, Dmitry Medvinsky, Denis Zgonjanin, Daniel S. Katz, Daniel Chen, Christopher Ahern, Chris Fournier, , Arturo, and André F. Rendeiro. Camdavidsonpilon/lifelines: v0.22.8, 2019.

Appendix A

FI-GCP Supplemental

A.1 Optimal cutpoints

Optimal cutpoints are calculated by testing binarization at every possible cutpoint in the data for every biomarker with respect to mortality at 5 years.

Mutual information can be calculated following [29]. The information entropy with respect to binary mortality $M \in \{0, 1\}$ is calculated as

$$S(M) = -m \ln(m) - (1 - m) \ln(1 - m), \tag{A.1}$$

where m is the proportion of the population dead at 5 years. The information entropy conditional on the presence of a deficit is the average of the entropy conditioned on each state of the deficit:

$$S(M|D) = p(d = 1)S(M|d = 1) + p(d = 0)S(M|d = 0), \tag{A.2}$$

where $p(d)$ is the proportion of the population with $(d = 1)$ or without $(d = 0)$ the deficit. The mutual information gained by knowing the status of a given deficit is then

$$I = S(M) - S(M|D). \tag{A.3}$$

For the logrank cutpoints we use a Python implementation of the logrank test [53] from the survival analysis package Lifelines [111]. Since its use of χ^2 statistics for estimating p-values underestimates them systematically for small sample sizes, the logrank test has a bias to select cutpoints with extremely few individuals in one group so as to artificially decrease p-values. To compensate for this bias, we imposed a minimum group size of 20 individuals.

A.2 FI-GCP Supplemental Figures

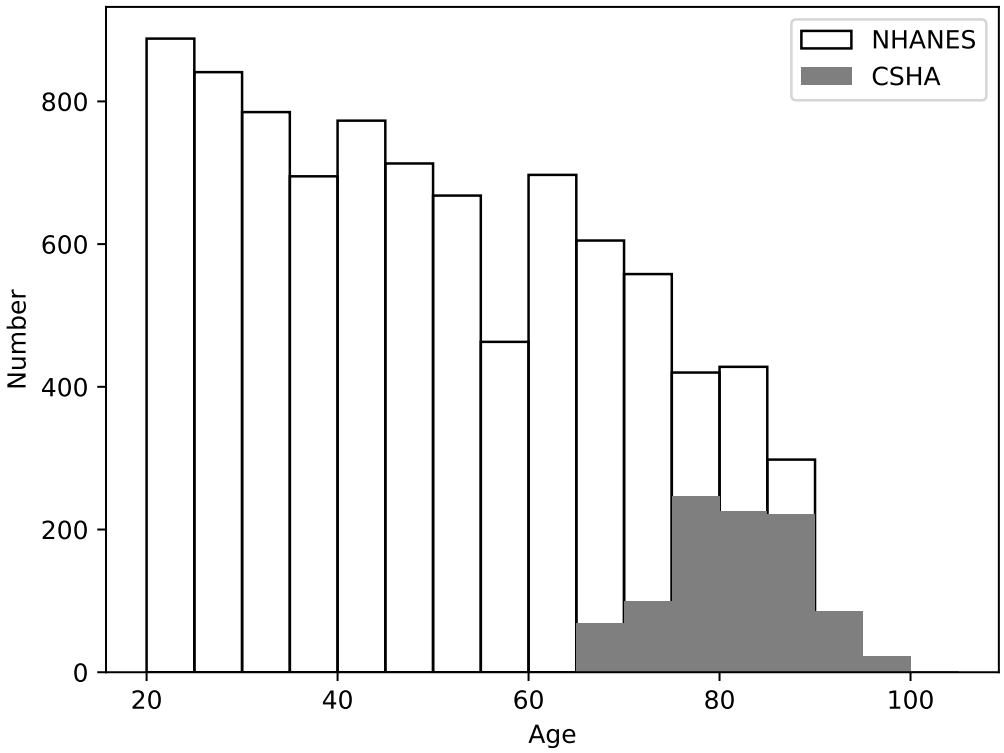

Figure A.1: Age distributions of the CSHA [46] (grey fill) and NHANES [45] (no fill) data sets. The NHANES data set has 8881 individuals with an age range of 20 to 85 years, and is considerably larger than the CSHA study with 973 individuals. The CSHA study was limited to older individuals with ages from 65-104.

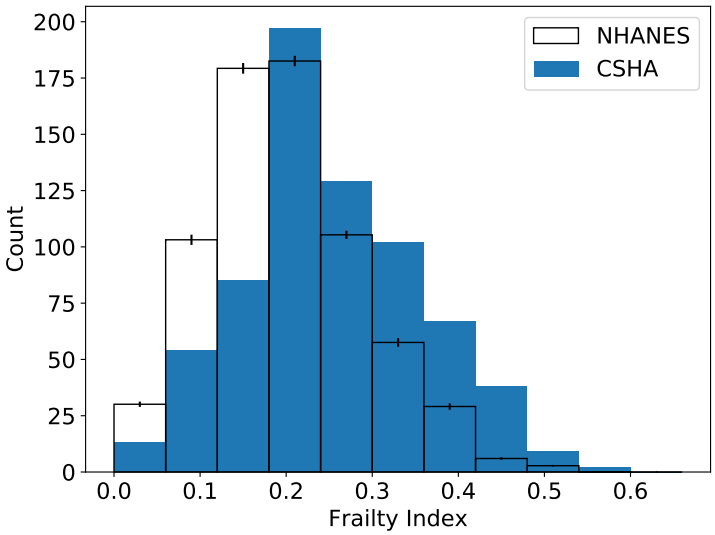

Figure A.2: FI distributions of individuals between the ages of 65 to 85. The NHANES data ([28]) has been randomly resampled to have the same age distribution as the CSHA data set within this age range ([27]).

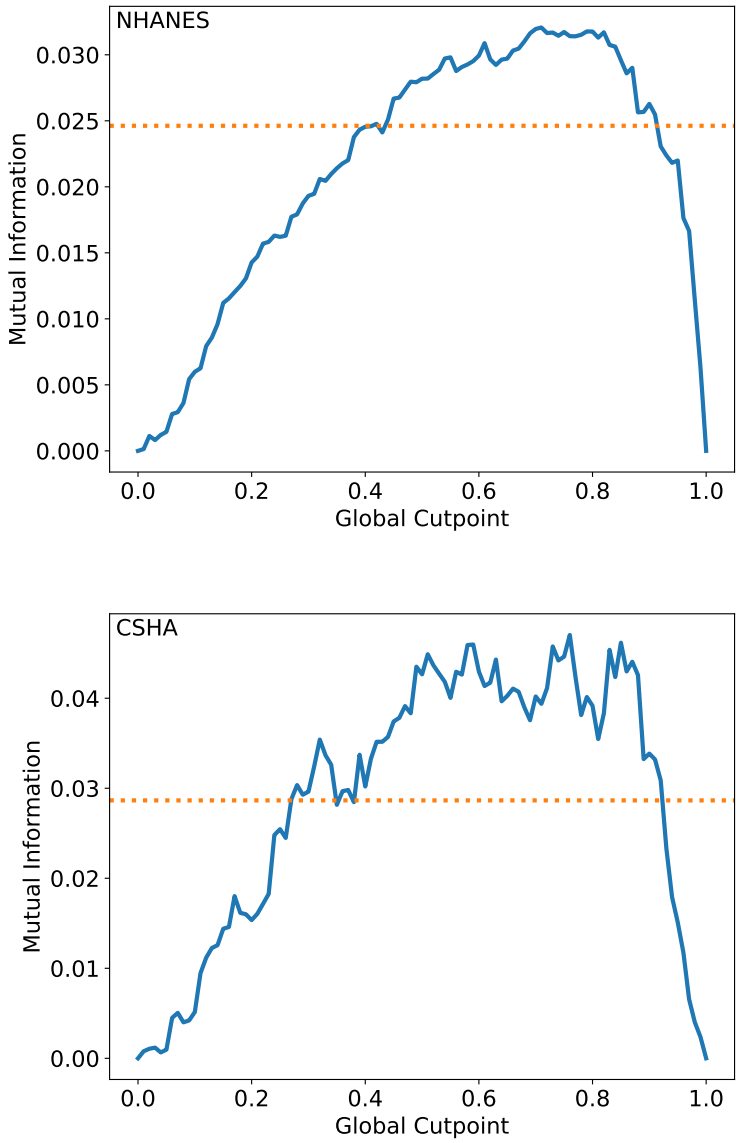

Figure A.3: Mutual information with respect to mortality at 5 years in the NHANES (top) and CSHA (bottom) datasets for FI_{GCP} (blue lines) vs the global cutpoint X_{GCP}. The orange dashed lines show the mutual information of the published FI-Lab [28, 27]. The behavior is qualitatively like that of AUC in Fig. 2.2.

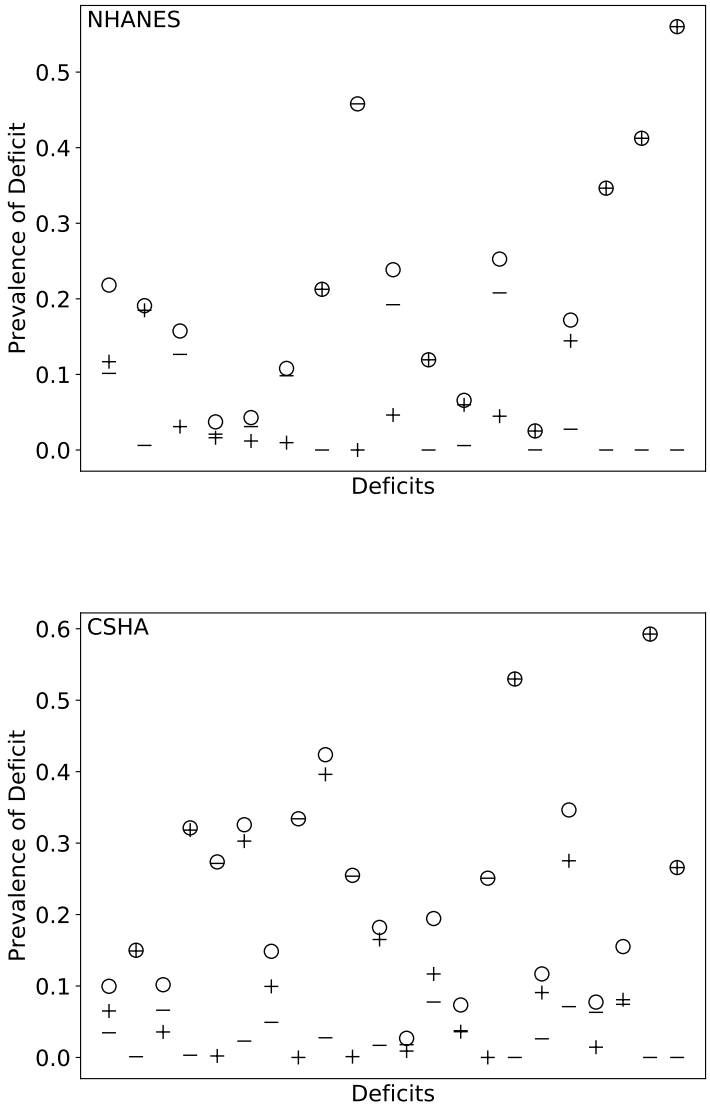

Figure A.4: Deficit prevalence for each deficit included in the published FI-Lab (circles) broken down into proportion at risk in high (+) and low (-) categories. Most of the deficits are predominantly on a single side of the risk direction.

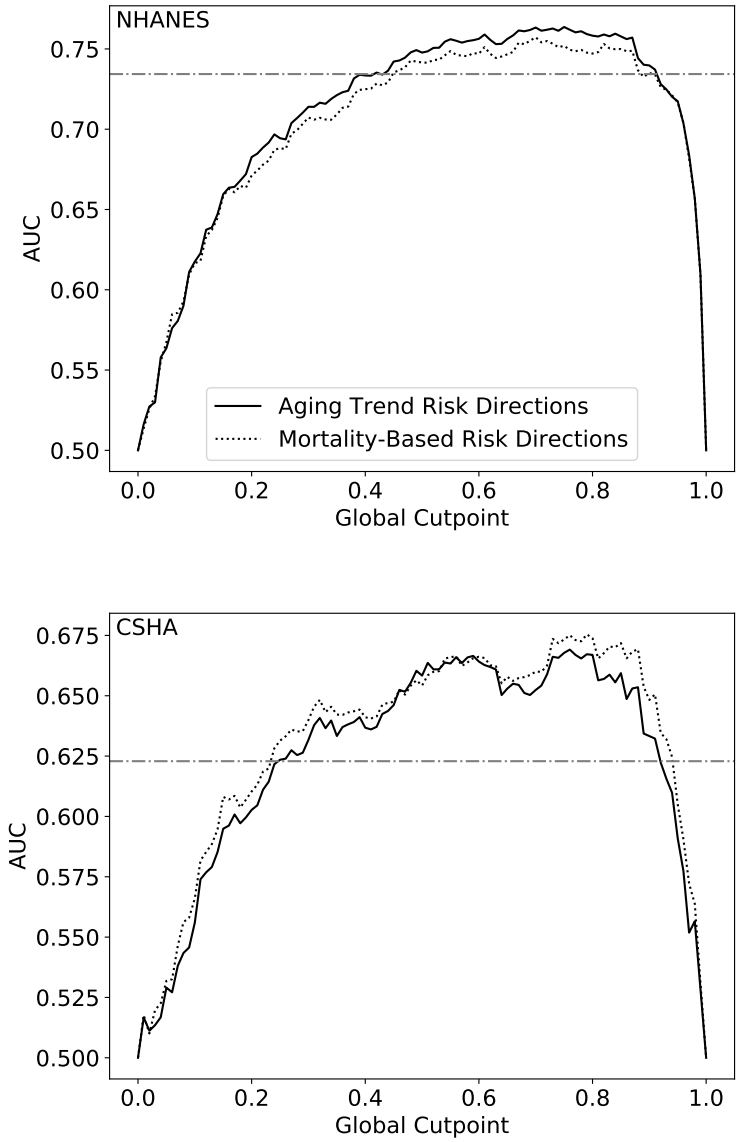

Figure A.5: Comparing the differences in predictive value of the FI using directions of primary risk calculated with respect to mortality (dotted line) and the aging trend method (solid line) for NHANES (top) and CSHA (bottom). Note, for NHANES the age conditions are calculated only in individuals age 35 or greater, while predictive value includes the whole population. The AUC of the published FI (horizontal dot-dashed line) is provided as a benchmark.

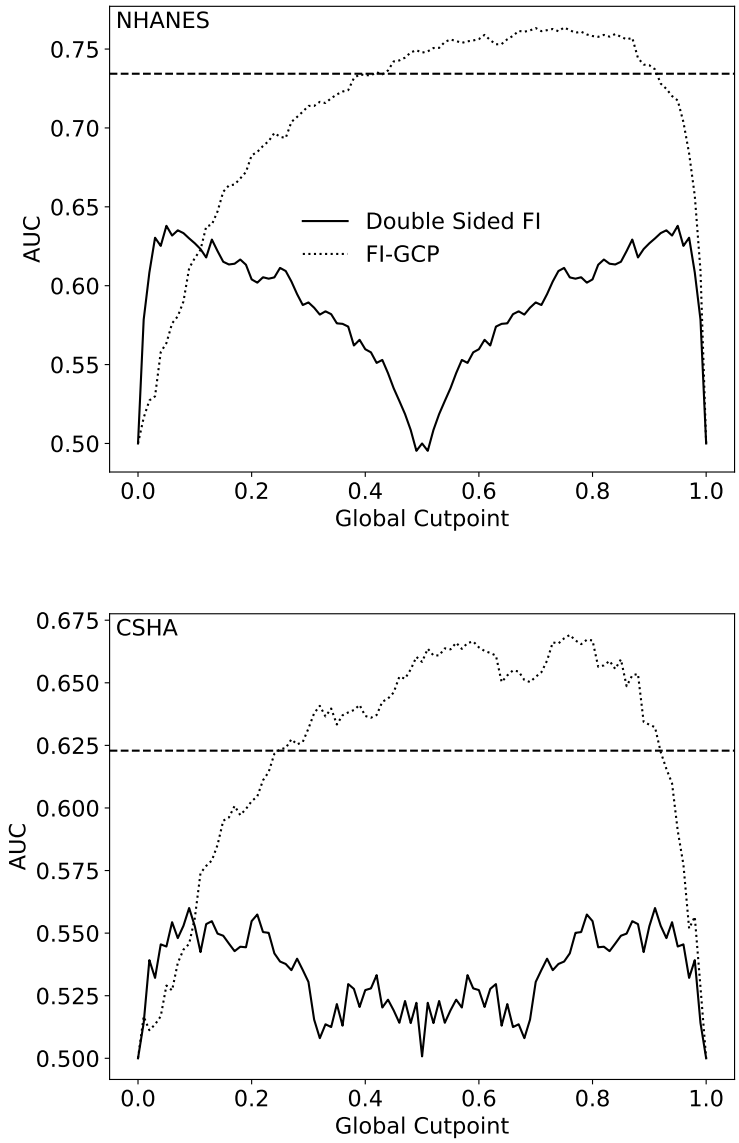

Figure A.6: The predictive value of a symmetric two-sided binarization approach (solid line), cutpoints move out from 0.5 symmetrically. The dotted line shows the AUC of FI_{GCP}, while the horizontal dashed line shows the AUC of the published FI-Lab.

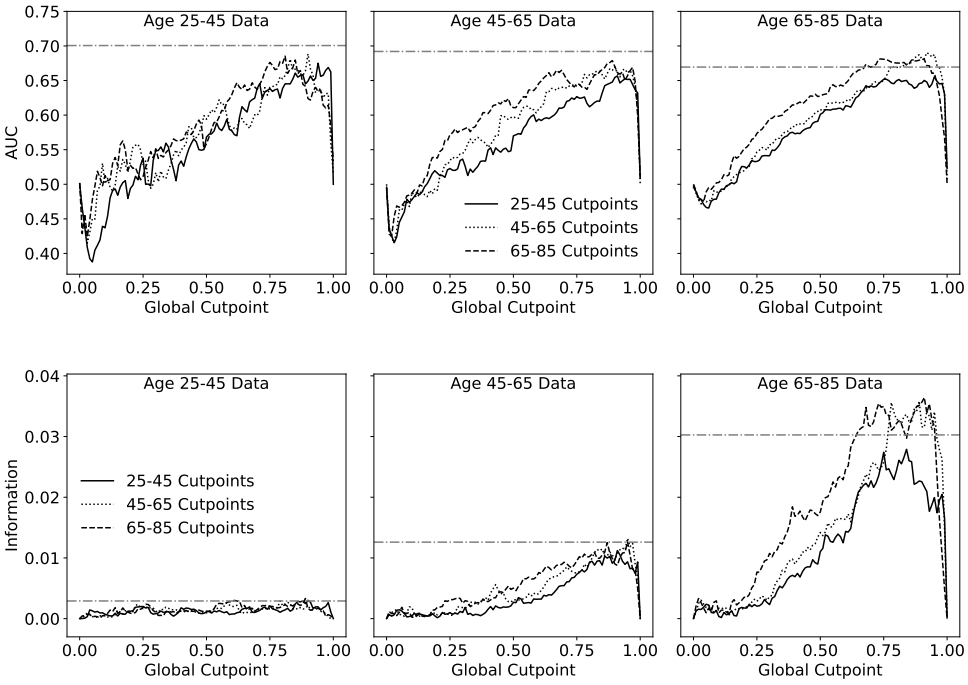

Figure A.7: Predictive value of FI_{GCP} when cutpoints are calculated in one age group and used in another. AUC with respect to 5 year mortality is shown on top. The bottom plot shows information with respect to mortality at 5 years. Note that the information captures the poor predictive value of any FI for the youngest age group, which has very few mortality events, while the AUC does not.

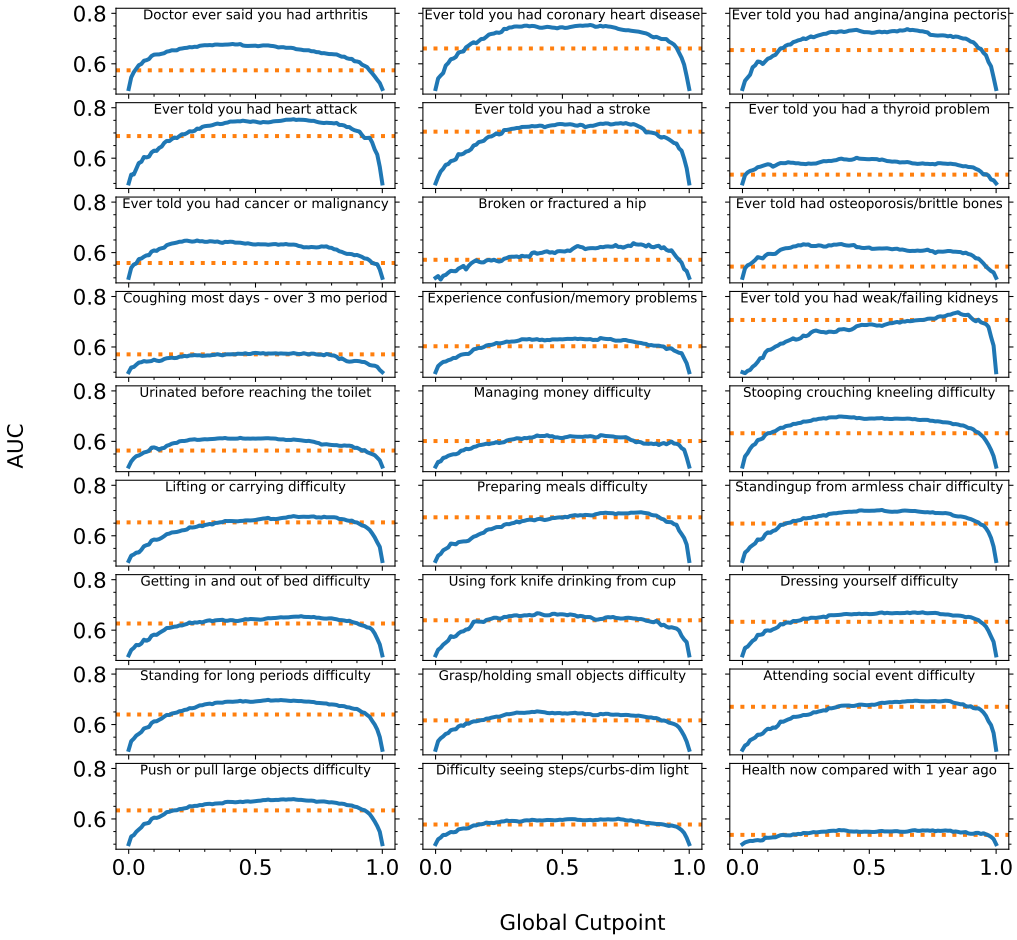

Figure A.8: Using FI_{GCP} to "predict" clinical deficits (solid blue lines) is at least as effective as using the published FI-Lab (horizontal dotted lines) for all deficits in the NHANES study.

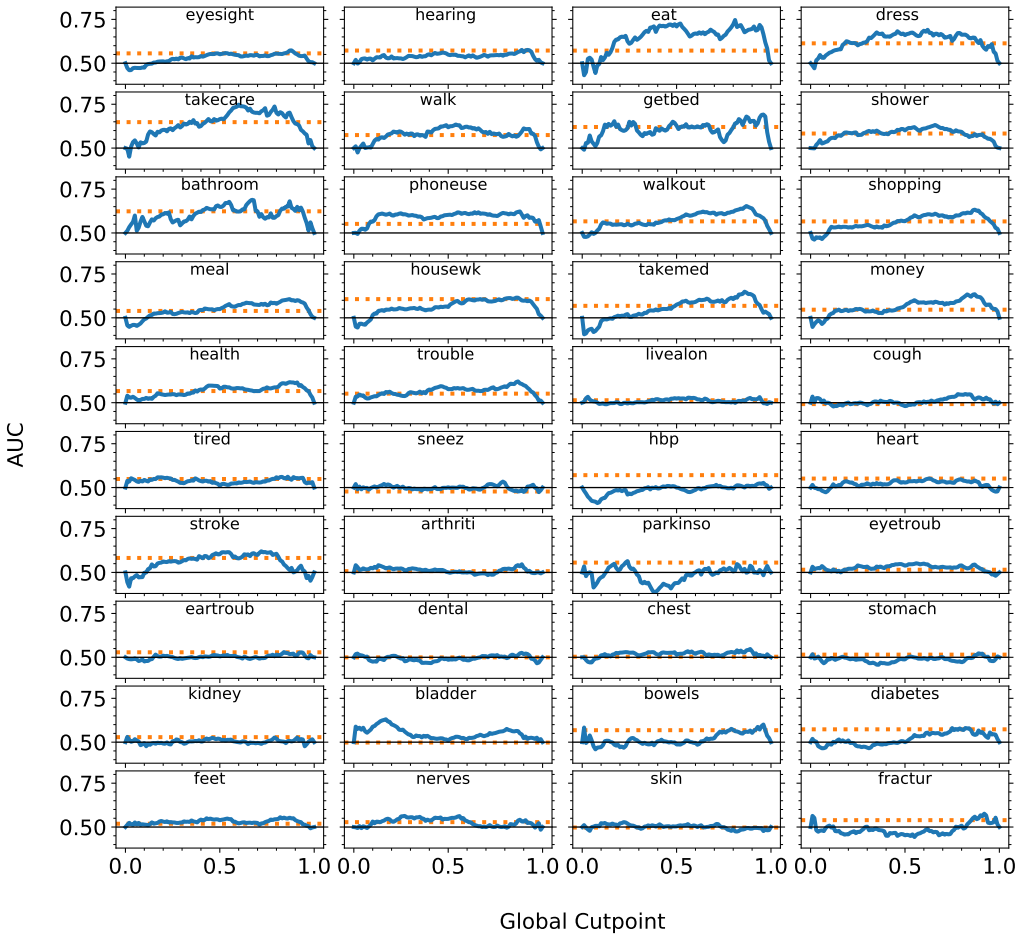

Figure A.9: Using FI_{GCP} to "predict" clinical deficits (solid blue lines) is at least as effective as using the published FI-Lab (horizontal dotted lines) for most deficits in the CSHA study. The horizontal black line shows the benchmark AUC of 0.5 for visual reference.

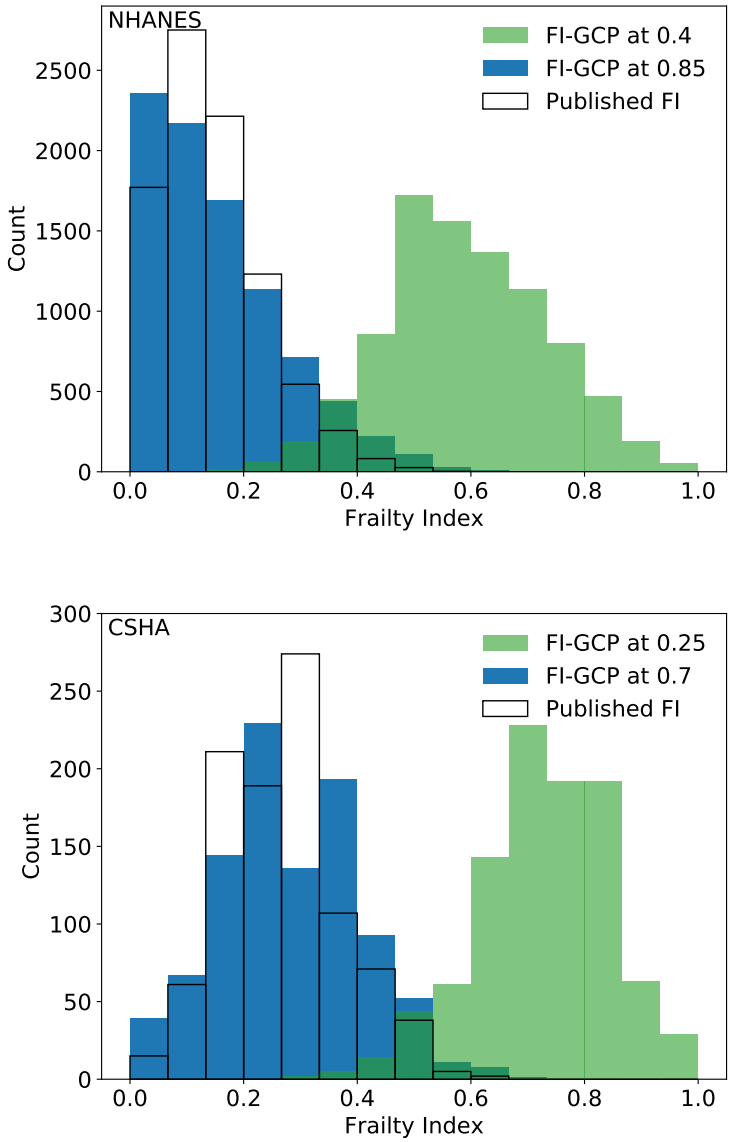

Figure A.10: FI distributions in the NHANES (top) and CSHA (bottom) using the published FI (no fill), FI_{GCP} with cutpoint at the minimum cutpoint with similar prediction to the published FI (0.4 and 0.25, for NHANES and CSHA respectively, in green), and FI_{GCP} with cutpoint where the distributions are most similar to the published FI (0.85 and 0.7, for NHANES and CSHA respectively, in blue).

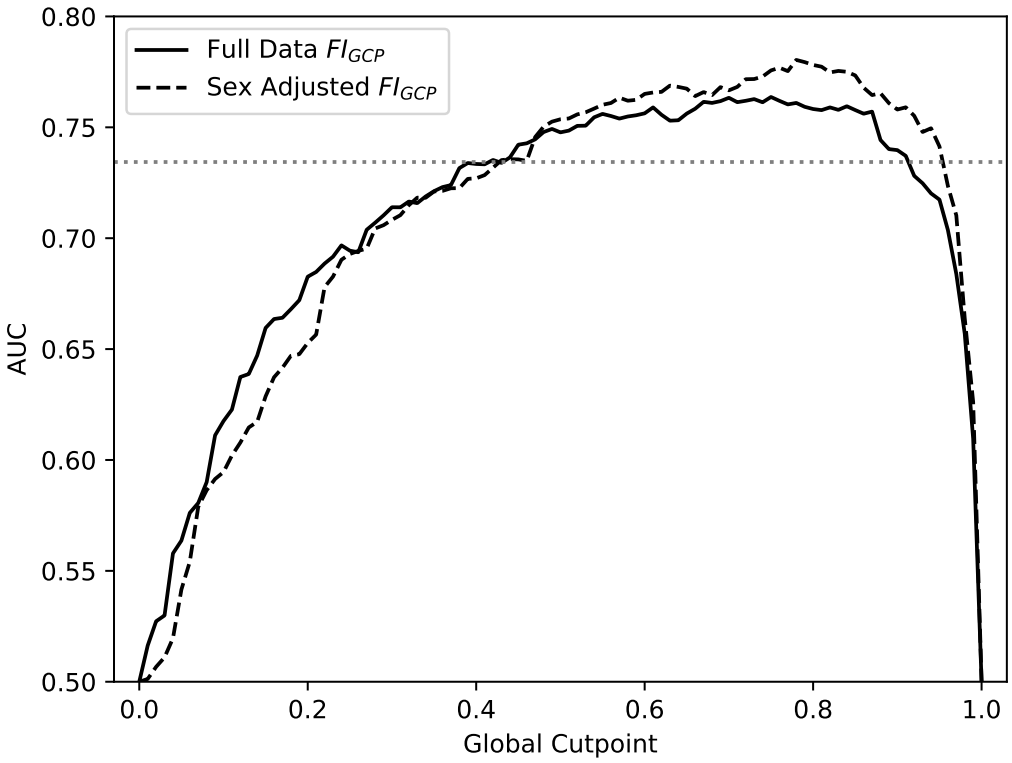

Figure A.11: Predictive value of FI_{GCP} with cutpoints calculated over the whole NHANES dataset (solid black line) compared to cutpoints calculated seperately for male and female (black dashed line). Horizontal grey dotted line shows the published FI-Lab performance.

Appendix B

QFI Supplemental

B.1 ELSA Data Description

In the ELSA dataset we use 15 biomarkers, requiring no less than 12 per individual (in wave 2, we use 14 requiring no less than 11): dominant hand grip strength, non-dominant hand grip strength, gait speed, diastolic blood pressure, systolic blood pressure, body mass index, c-reactive protein, triglycerides, hdl cholesterol, ldl cholesterol, insulin-like growth factor 1 (not reported in wave 2), fibrinogen, ferritin, hemoglobin, glucose.

For FI-Clin in the ELSA dataset we use 10 ADL (walking 100 yards, sitting for about 2 hours, getting up from a chair after sitting for long periods, climbing several flights of stairs without resting, climbing 1 flight of stairs without resting, stooping kneeling or crouching, reaching or extending arms above shoulder level, pulling/pushing large objects like a living room chair, lifting/carrying over 10 lbs like a heavy bag of groceries, picking up a 5p coin from a table) and 13 IADL deficits (dressing including putting on shoes and socks, walking across a room, bathing or showering, eating such as cutting up your food, getting in or out of bed, using the toilet including getting up or down, using a map to get around in a strange place, preparing a hot meal, shopping for groceries, making telephone calls, taking medications, doing work around the house or garden, managing money e.g. paying bills and keeping track of expenses).

The ELSA data available for diagnoses are: high blood pressure, angina, myocardial infraction, congestive heart failure, heart murmur, arrhythmia, diabetes, stroke, hedibonic lung disease, asthma, arthritis, osteoporosis cancer, Parkinson's disease, psychiatric disorder, Alzheimer's, dementia/memory impairment, glaucoma, diabetic retinopathy, macular degeneration, and cataracts. We exclude high blood pressure and high cholesterol as diagnoses since they are directly encoded in the available biomarkers. We focus on waves 2 and 4 since we have diagnosis data up to wave 5.

B.2 QFI Supplemental Figures

The demographics of the ELSA (Fig. B.1) and NHANES and CSHA datasets (Fig. B.2), together with the distribution of mortality. The demographics can be compared with wave 2 of the ELSA dataset (inset of Fig. B.1).

The 5 year mortality AUC vs quantile versions of the QFI are show in Fig. B.3 for ELSA data, and Fig. B.4 for CSHA data. These can be compared with the NHANES results in Fig. 2.7.

The 5 year mortality AUC performance for different QFI are shown in Fig. B.5 for CSHA data and Fig. B.6 for ELSA data. These can be compared with NHANES results in Fig. 2.8.

The comparison of QFI and FI-Clin for the 4th wave of ELSA data are shown in Fig. B.7. This can be compared with results from the 2nd wave of ELSA data in Fig. 2.10.

The effects of changing the age of the reference population is shown in Fig. B.8 for ELSA data. This can be compared with Fig. 2.9 for NHANES and CSHA data.

The effects of controlling for age is explored in Fig. B.9 for CSHA data and Fig. B.10 for ELSA data. These can be compared with Fig. 2.11 for NHANES data. The age-stratified data used to obtain the age-averaged points are shown in Fig. B.11 for NHANES and CSHA data, and Fig. B.12 for ELSA data.

Sex adjusted effects are explored in Fig. B.13 for NHANES data, Fig. B.14 for CHSA data, and Fig. 2.12 for wave-4 of ELSA data. These can be compared with Fig. 2.12 for wave-2 of ELSA data.

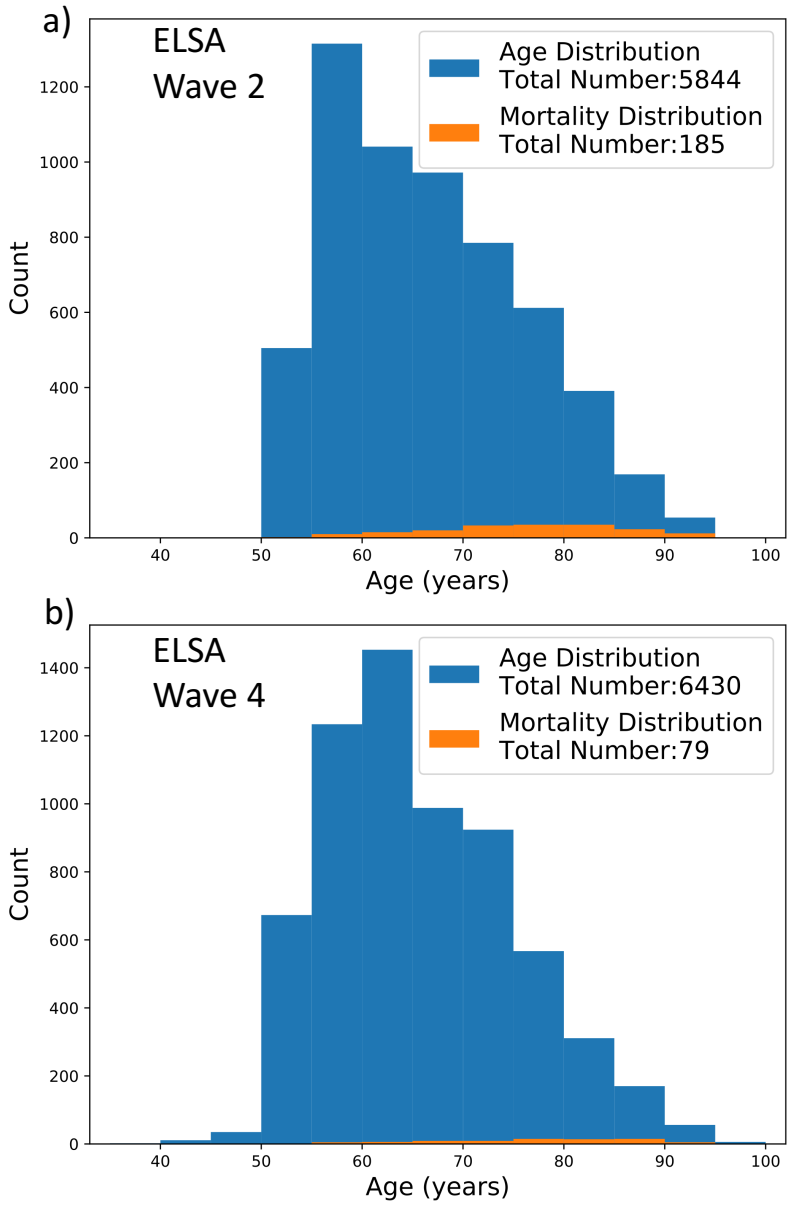

Figure B.1: The age distribution in waves 2 (a) and 4 (b) of the ELSA study (blue) with the associated distribution of mortality events within a 5 year followup (orange).

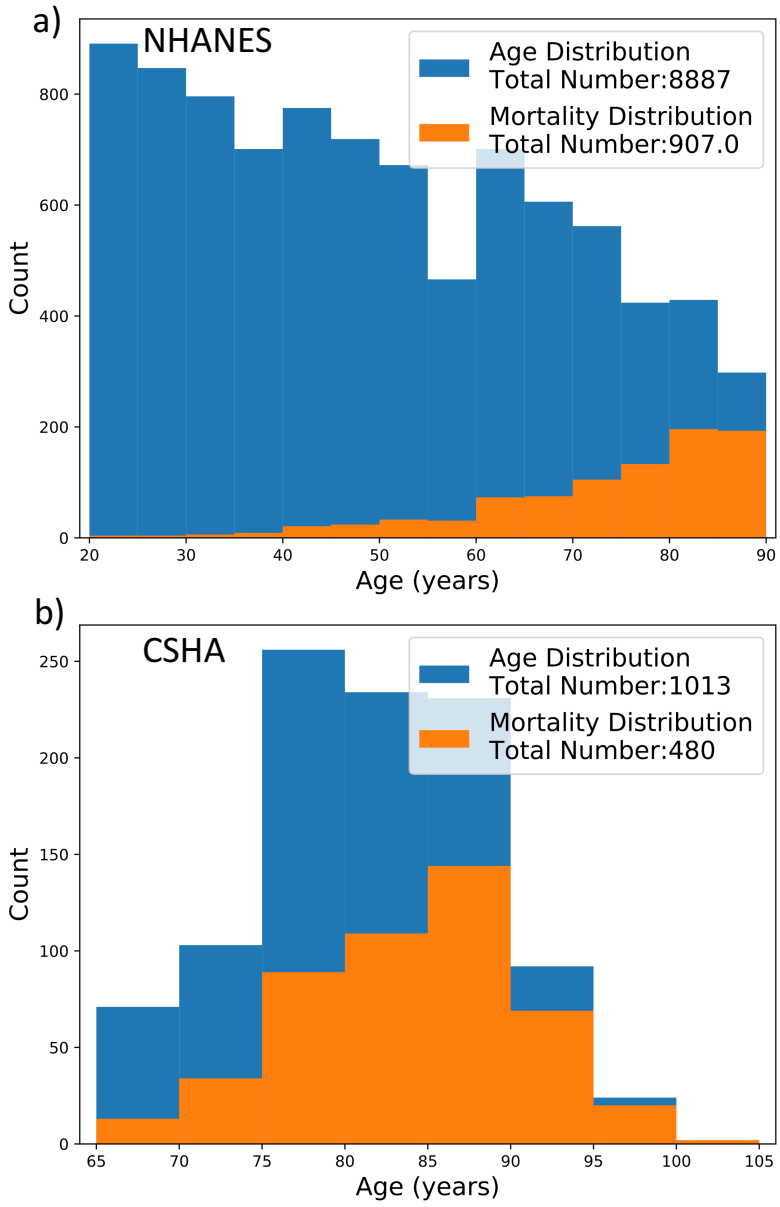

Figure B.2: The age distribution in the NHANES study (a) and CSHA study (b) with the associated distribution of mortality events within a 5 year followup (orange).

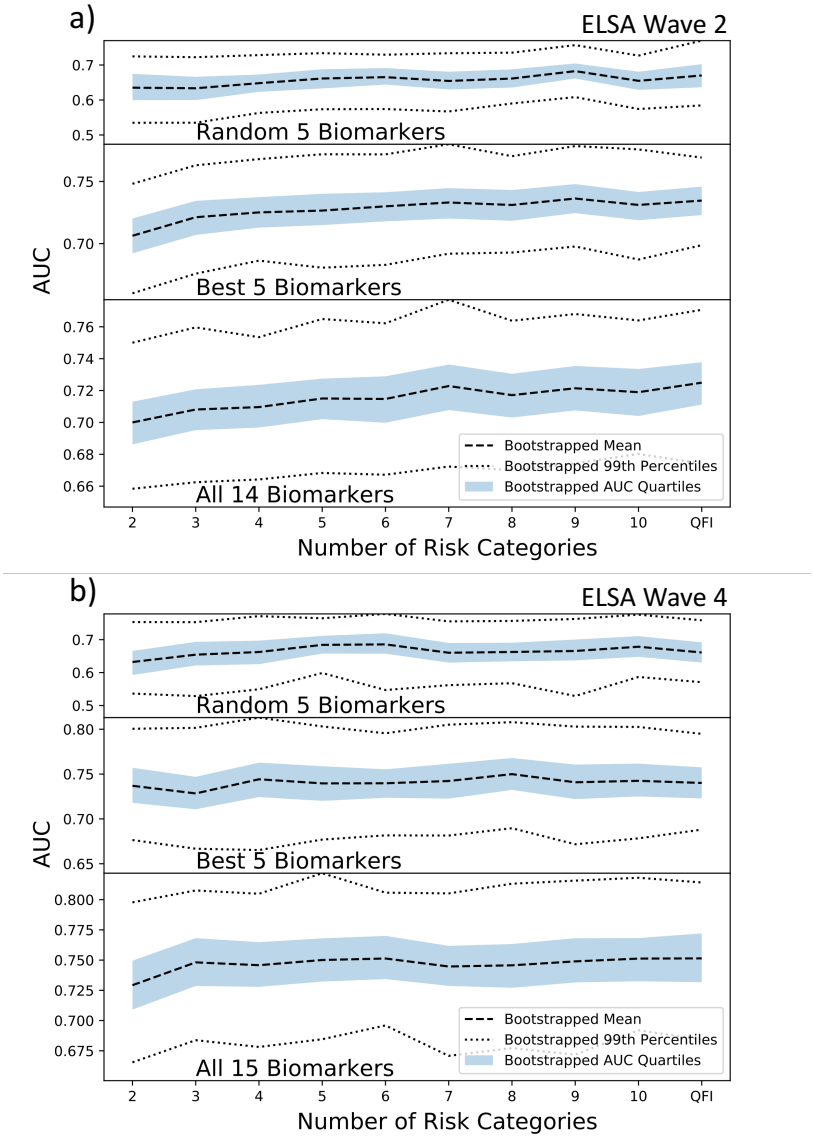

Figure B.3: The relationship between number of quantiles considered and the predictive value with respect to mortality within 5 years in the 2nd (a) and 4th (b) waves of the ELSA dataset. 2 risk categories is equivalent to dichotomization at the median, 3 risk categories equivalent to risk tertiles, and so forth. The upper plot shows the effect using 5 randomly selected biomarkers, the middle plot shows the best 5 biomarkers (selected by AUC with respect to 5 year mortality), and the lower showing the results using all available biomarkers. We resample the data using half the population size a total of 400 times for each point, with the random 5 biomarkers being re-selected 20 times. The dotted lines show the upper and lower 1^{st} percentiles of AUC, the shaded blue region shows the upper and lower quartile range of AUC, and the dashed line shows the average AUC.

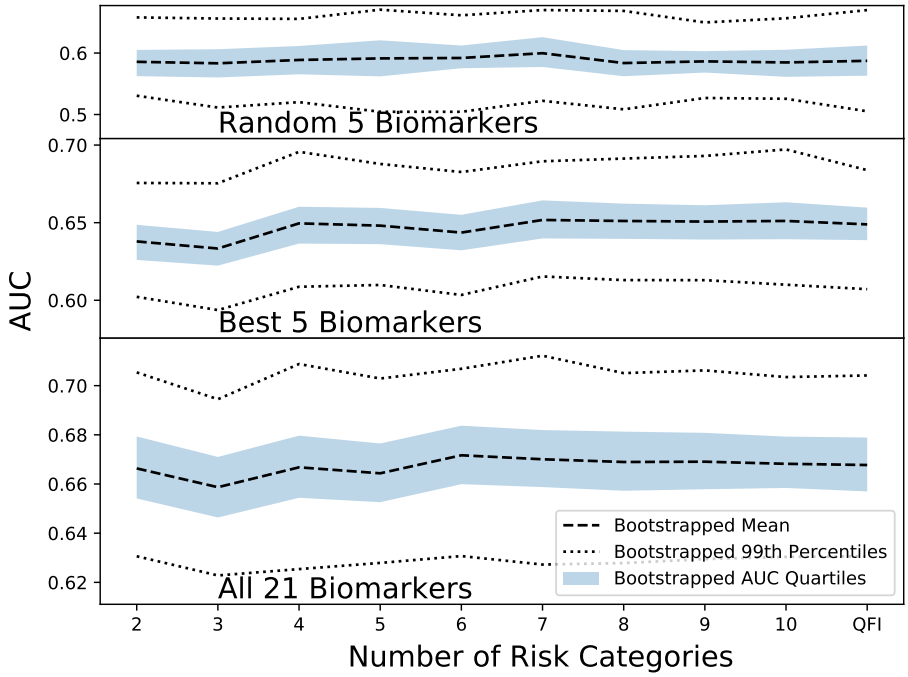

Figure B.4: The relationship between number of quantiles considered and the predictive value with respect to mortality within 5 years in the CSHA dataset. 2 risk categories is equivalent to dichotomization at the median, 3 risk categories equivalent to risk tertiles, and so forth. The upper plot shows the effect using 5 randomly selected biomarkers, the middle plot shows the best 5 biomarkers (selected by AUC with respect to 5 year mortality), and the lower showing the results using all available biomarkers. We resample the data using half the population size a total of 400 times for each point, with the random 5 biomarkers being re-selected 20 times. The dotted lines show the upper and lower 1^{st} percentiles of AUC, the shaded blue region shows the upper and lower quartile range of AUC, and the dashed line shows the average AUC.

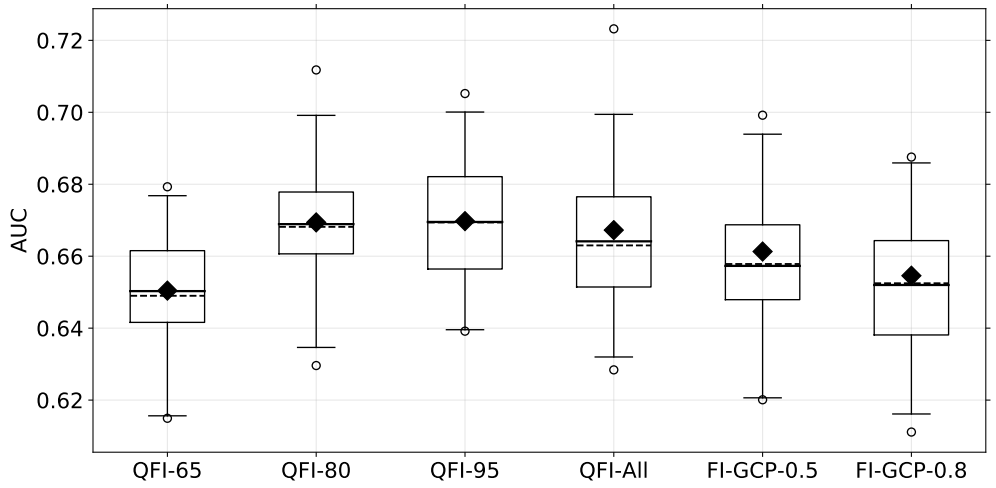

Figure B.5: The predictive value of various FI-Lab with respect to 5 year mortality in the CSHA study. From left to right we have the QFI using a 65-70-year-old reference population, the QFI with a 80-85-year-old reference, QFI with a 95-100-year-old reference, QFI using the whole study as a reference, FI-GCP with the cutpoint at 0.5, and FI-GCP with the cutpoint at 0.8. Box and whisker plots display the data from resampling and cross-validation: the boxes represent the upper and lower quartiles, the whiskers go to the 99^{th} and 1^{st} percentiles, and the circles are remaining outliers. The short dashed line within each box is the median, the solid line the mean, and the diamond is the AUC for the full data set without cross validation.

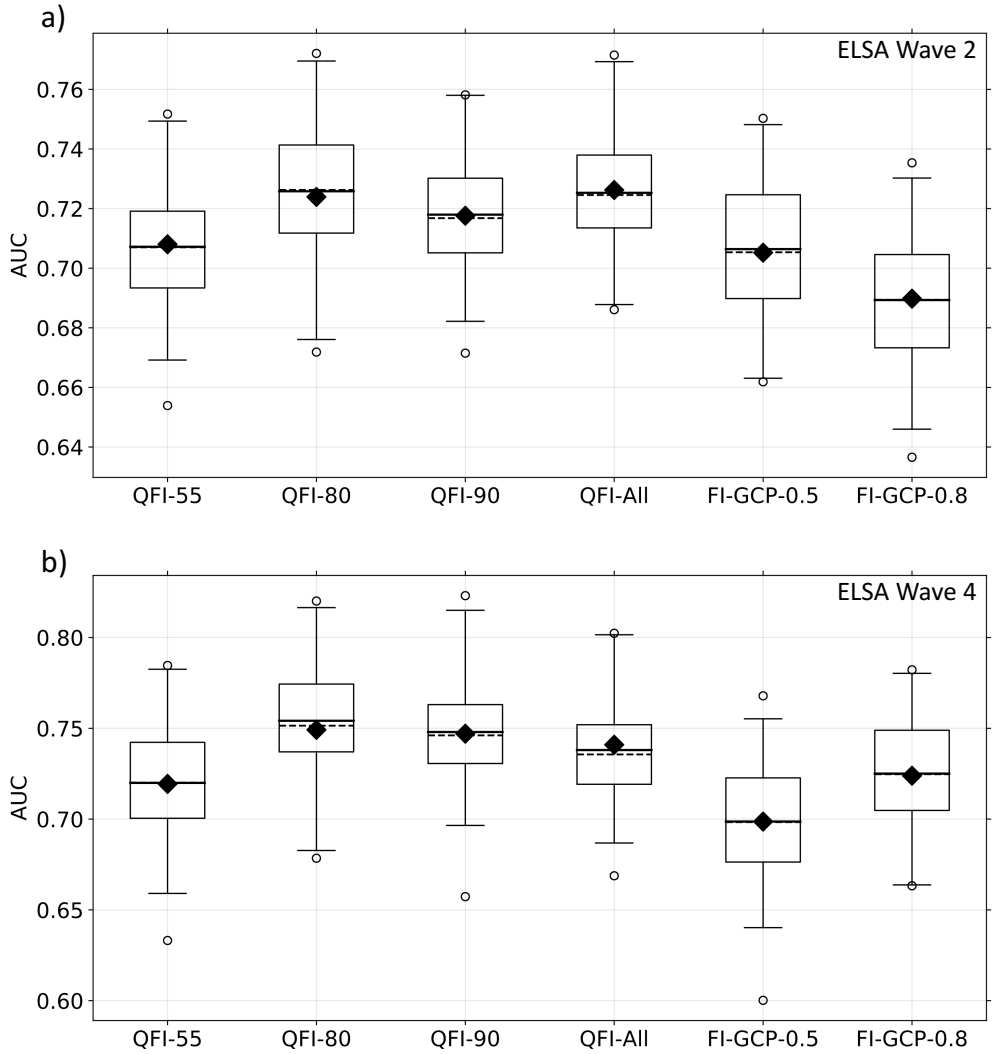

Figure B.6: The predictive value of various FI-Lab with respect to 5 year mortality in the 2nd (a) and 4th (b) waves of the ELSA study. From left to right we have the QFI using a 55-60-year-old reference population, the QFI with a 80-85-year-old reference, QFI with a 90-95-year-old reference, QFI using the whole study as a reference, FI-GCP with the cutpoint at 0.5, and FI-GCP with the cutpoint at 0.8. Box and whisker plots display the data from resampling and cross-validation: the boxes represent the upper and lower quartiles, the whiskers go to the 99^{th} and 1^{st} percentiles, and the circles are remaining outliers. The short dashed line within each box is the median, the solid line the mean, and the diamond is the AUC for the full data set without cross validation.

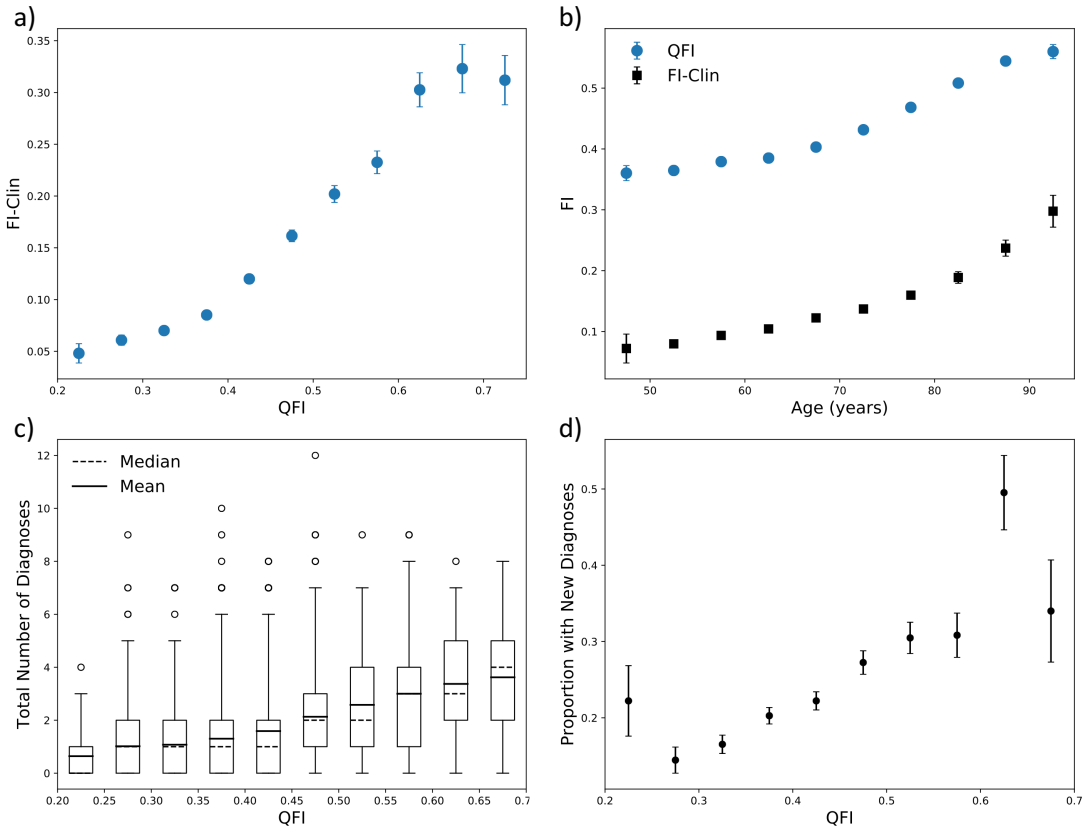

Figure B.7: a) The average FI-Clin binned by QFI. b) shows the average QFI and FI-Clin binned by age. c) The relationship between the QFI and the total number of existing or previous diagnoses in the fourth wave of the ELSA dataset. The boxes represent the upper and lower quartiles, the whiskers go to the 99^{th} and 1^{st} percentiles, and the circles are remaining outliers. The short dashed line within each box is the median and the solid line is the mean. d) shows the fraction of the population with 1 or more new diagnoses in the year following the QFI evaluation.

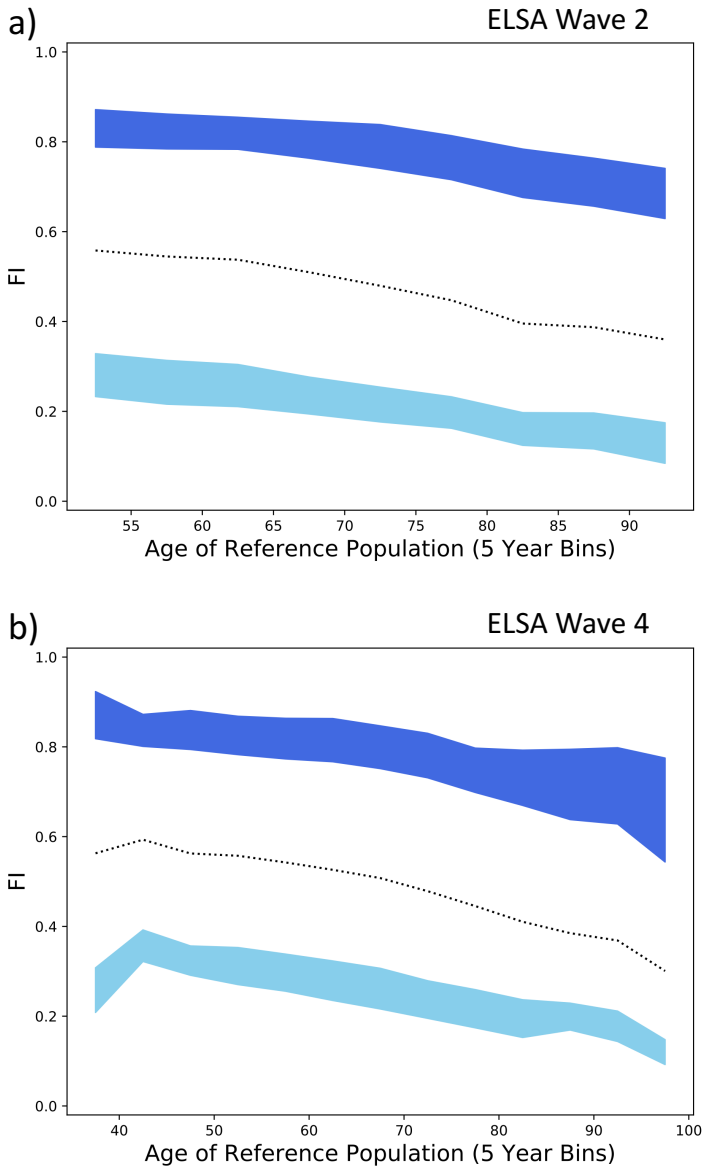

Figure B.8: The effects of changing the reference population on the distribution of QFI scores in waves 2 (a) and 4 (b) of the ELSA dataset. The upper 1% (light blue) and lower 1% (dark blue) of the QFI distributions as the age of the reference cohort changes (in 5 year bins).

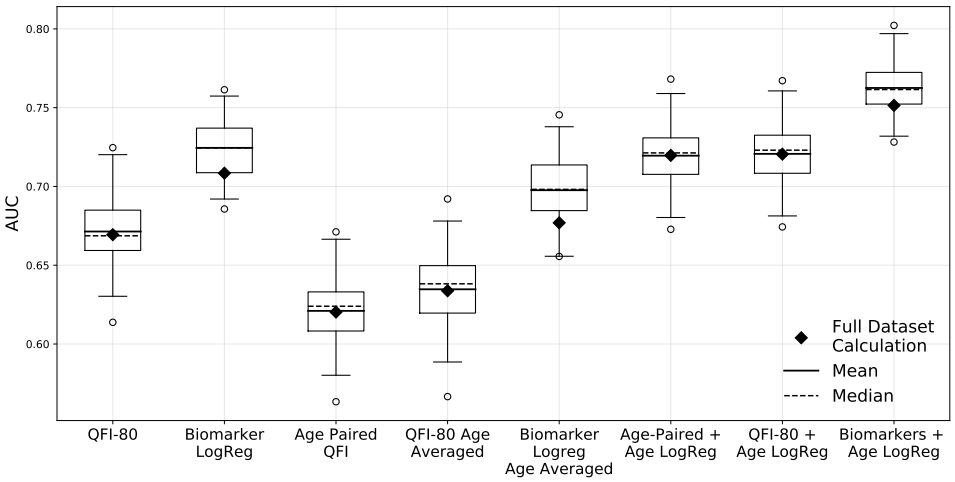

Figure B.9: Comparing the predictive value of the different methods of calculating the QFI against 5 year mortality in the CSHA dataset. From left to right we have the QFI with an 80-85-year-old reference population (QFI-80), a logistic regression model using all of the biomarkers regressed against 5 year mortality, the age-paired QFI, QFI-80 with AUC averaged across performance within 5 year age bins, a logistic regression of the biomarkers against 5 year mortality with AUC averaged across performance within 5 year age bins, the age-paired QFI combined with age in a logistic regression, the QFI-80 combined with age in a logistic regression, and the raw biomarkers included with age in a logistic regression. Box and whiskers are from resampling with half the population. The boxes represent the upper and lower quartiles, the whiskers go to the 99^{th} and 1^{st} percentiles, and the circles are remaining outliers. The short dashed line within each box is the median, the solid line the mean, and the diamond is the AUC for the full data set without resampling. Here we are using logistic regression to control for age in the prediction, not testing a logistic model, so performance is evaluated on the same individuals as the model is fit on.

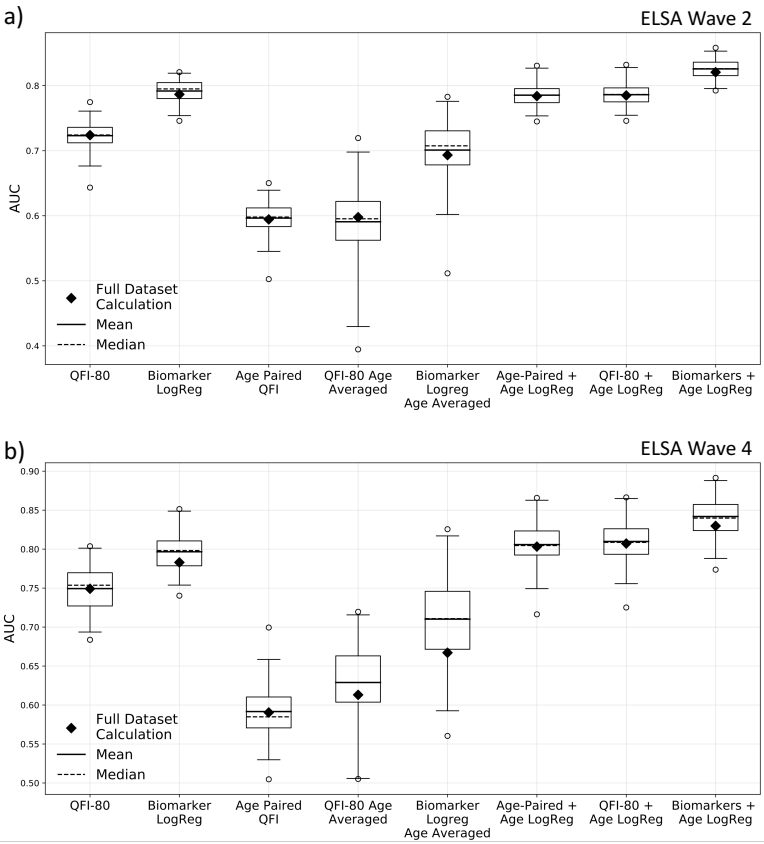

Figure B.10: Comparing the predictive value of the different methods of calculating the QFI against 5 year mortality in waves 2 (a) and 4 (b) of the ELSA dataset. From left to right we have the QFI with an 80-85-year-old reference population (QFI-80), a logistic regression model using all of the biomarkers regressed against 5 year mortality, the age-paired QFI, QFI-80 with AUC averaged across performance within 5 year age bins, a logistic regression of the biomarkers against 5 year mortality with AUC averaged across performance within 5 year age bins, the age-paired QFI combined with age in a logistic regression, the QFI-80 combined with age in a logistic regression, and the raw biomarkers included with age in a logistic regression. Box and whiskers are from resampling with half the population. The boxes represent the upper and lower quartiles, the whiskers go to the 99^{th} and 1^{st} percentiles, and the circles are remaining outliers. The short dashed line within each box is the median, the solid line the mean, and the diamond is the AUC for the full data set without resampling. Here we are using logistic regression to control for age in the prediction, not testing a logistic model, so performance is evaluated on the same individuals as the model is fit on.

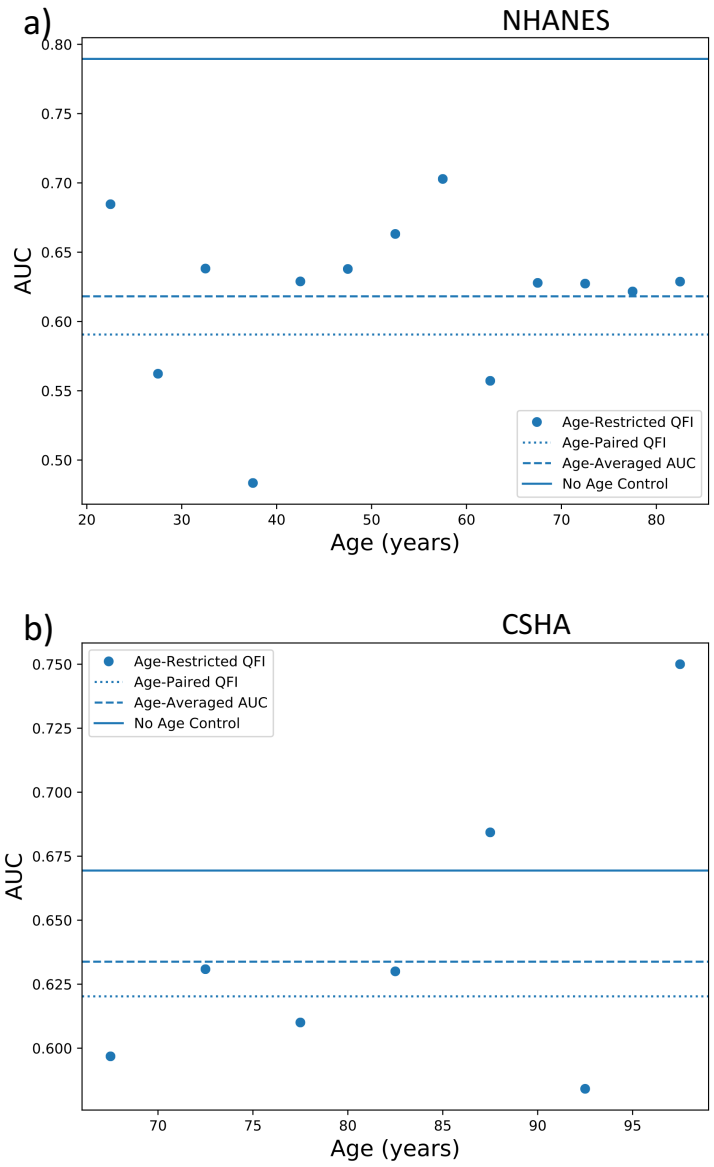

Figure B.11: Predictive value of the QFI restricted to 5 year age bins in the NHANEs (a) and CSHA (b) datasets. The points show the AUC with respect to 5 year mortality calculated only within the 5 year age bins. The dashed line shows the weighted average of these points. The dotted line shows the AUC using the age-paired QFI. The solid line shows the AUC using only the QFI without controlling for other factors.

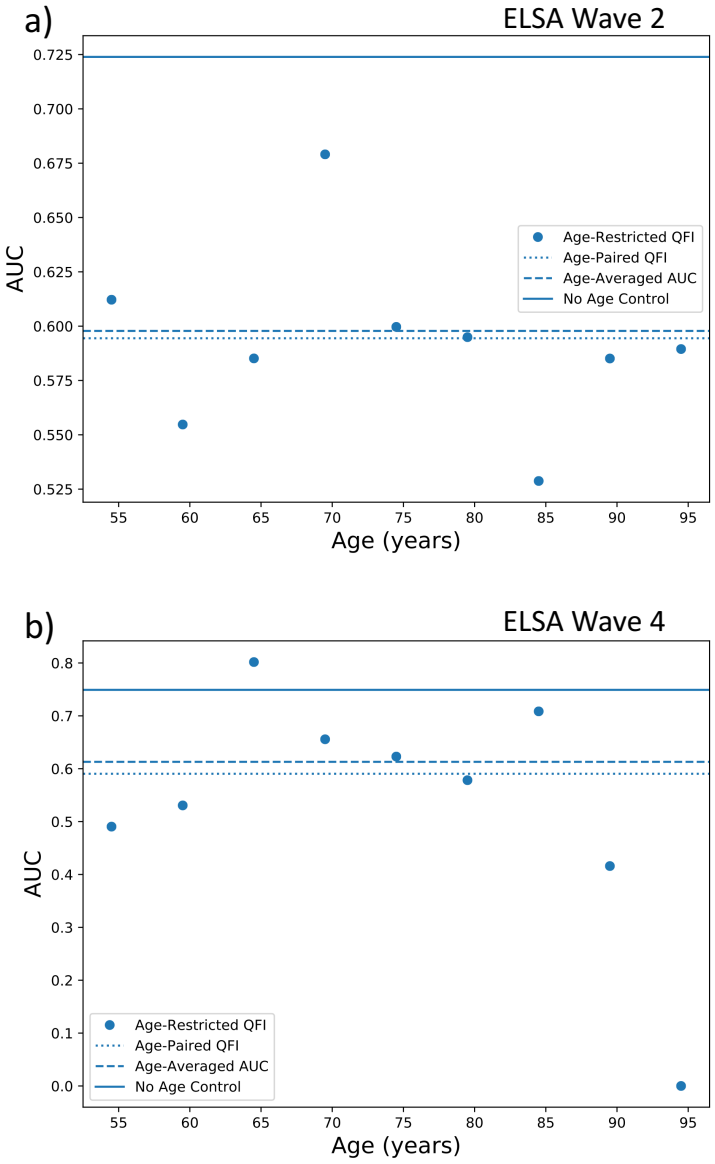

Figure B.12: Predictive value of the QFI restricted to 5 year age bins in waves 2 (a) and 4 (b) of the ELSA dataset. The points show the AUC with respect to 5 year mortality calculated only within the 5 year age bins. The dashed line shows the weighted average of these points. The dotted line shows the AUC using the age-paired QFI. The solid line shows the AUC using only the QFI without controlling for other factors.

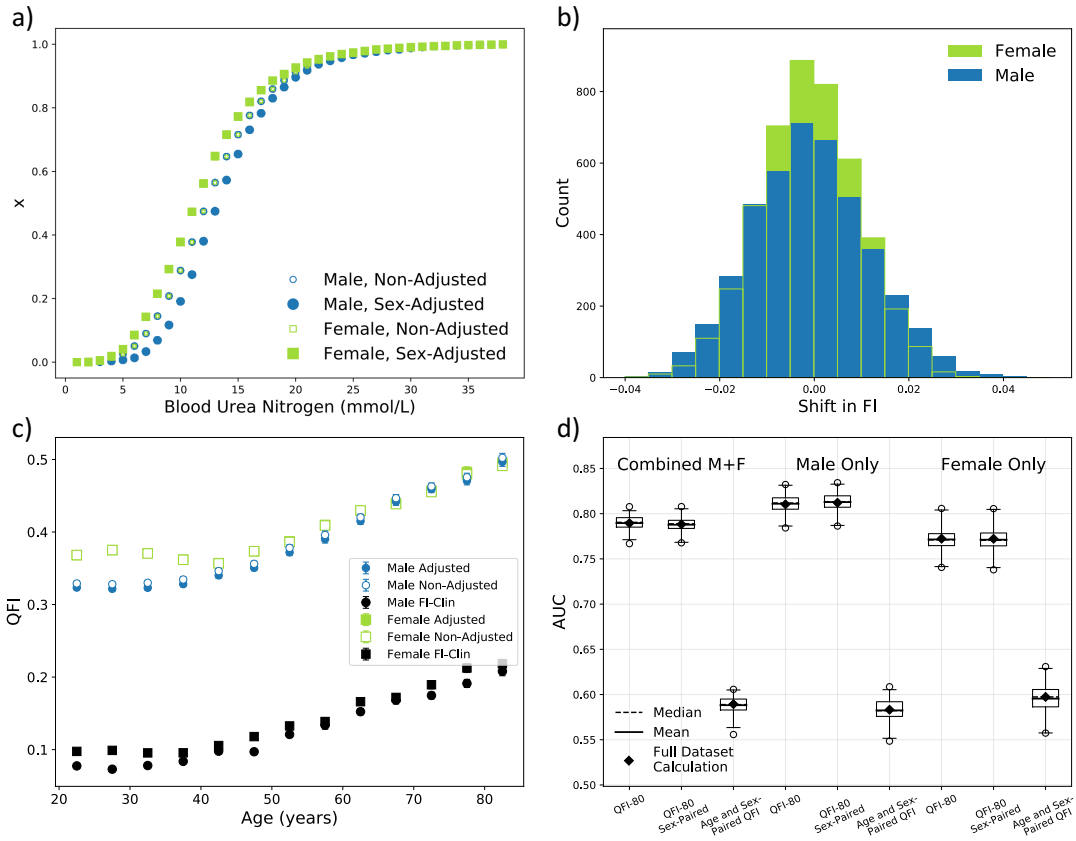

Figure B.13: The effects of using sex-specific reference populations on QFI-80 in the NHANES dataset. All plots show female in green and male in blue. a) The risk quantiles for blood urea nitrogen with (filled points) and without (no-fill) using a sex-specific reference population. Note that the non-adjusted male and female scores overlap since they are using the same reference population. b) The difference between sex-adjusted QFI-80 and non-adjusted QFI-80 using all 80-85 year-olds as a reference population. Sex-adjusted QFI-80 uses only 80-85-year-olds of the respective sex as the reference population. c) The average QFI for the sex-adjusted QFI (filled) and non-adjusted (no fill) binned by age in 5 year bins, the black points show the associated FI-Clin. d) The AUC of various QFI with respect to mortality at 5 year follow up. We compare (from left to right) the QFI-80, sex-paired QFI-80, and age-and-sex-paired QFI for the combined, male, and female populations (from left to right). Box and whiskers are from resampling with half the population. The boxes represent the upper and lower quartiles, the whiskers go to the 99^{th} and 1^{st} percentiles, and the circles are remaining outliers. The short dashed line within each box is the median, the solid line the mean, and the diamond is the AUC for the full data set without resampling.

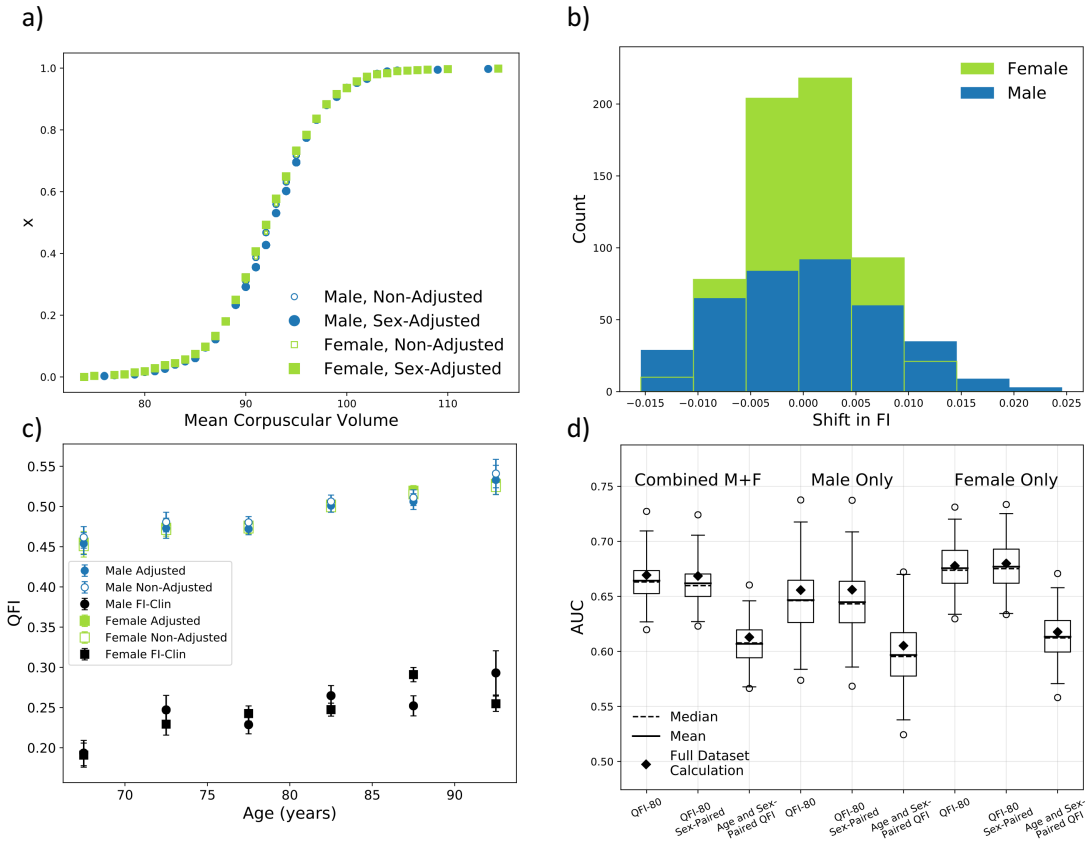

Figure B.14: The effects of using sex-specific reference populations on QFI-80 in the CSHA dataset. All plots show female in green and male in blue. a) The risk quantiles for mean corpuscular volume with (filled points) and without (no-fill) using a sex-specific reference population. Note that the non-adjusted male and female scores overlap since they are using the same reference population. b) The difference between sex-adjusted QFI-80 and non-adjusted QFI-80 using all 80-85 year-olds as a reference population. Sex-adjusted QFI-80 uses only 80-85-year-olds of the respective sex as the reference population. c) The average QFI for the sex-adjusted QFI (filled) and non-adjusted (no fill) binned by age in 5 year bins, the black points show the associated FI-Clin. d) The AUC of various QFI with respect to mortality at 5 year follow up. We compare (from left to right) the QFI-80, sex-paired QFI-80, and age-and-sex-paired QFI for the combined, male, and female populations (from left to right). Box and whiskers are from resampling with half the population. The boxes represent the upper and lower quartiles, the whiskers go to the 99^{th} and 1^{st} percentiles, and the circles are remaining outliers. The short dashed line within each box is the median, the solid line the mean, and the diamond is the AUC for the full data set without resampling.

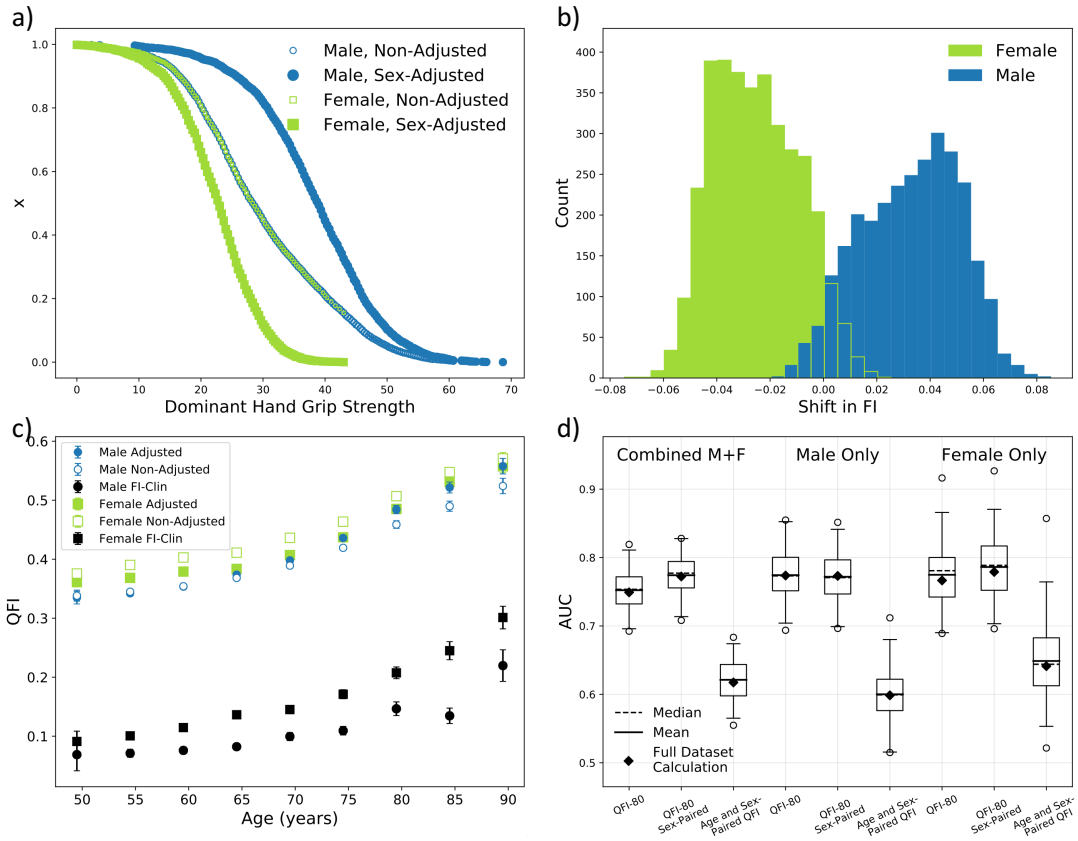

Figure B.15: The effects of using sex-specific reference populations on QFI-80 in wave 4 of the ELSA dataset. All plots show female in green and male in blue. a) The risk quantiles for dominant hand grip strength with (filled points) and without (no-fill) using a sex-specific reference population. Note that the non-adjusted male and female scores overlap since they are using the same reference population. b) The difference between sex-adjusted QFI-80 and non-adjusted QFI-80 using all 80-85 year-olds as a reference population. Sex-adjusted QFI-80 uses only 80-85-year-olds of the respective sex as the reference population. c) The average QFI for the sex-adjusted QFI (filled) and non-adjusted (no fill) binned by age in 5 year bins, the black points show the associated FI-Clin. d) The AUC of various QFI with respect to mortality at 5 year follow up. We compare (from left to right) the QFI-80, sex-paired QFI-80, and age-and-sex-paired QFI for the combined, male, and female populations (from left to right). Box and whiskers are from resampling with half the population. The boxes represent the upper and lower quartiles, the whiskers go to the 99^{th} and 1^{st} percentiles, and the circles are remaining outliers. The short dashed line within each box is the median, the solid line the mean, and the diamond is the AUC for the full data set without resampling.