MEASUREMENT OF HETEROGENEITY IN COMPUTATIONAL PSYCHIATRY

by

Abraham Nunes

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

at

Dalhousie University
Halifax, Nova Scotia
April 2020

For Sarah, Isabel, Gabriel, John Isaac, and Rosa

Table of Contents

List of Tables

List of Figures

Abstract

Psychiatric researchers are interested in heterogeneity because mental illnesses can have many (overlapping) causes and consequences. This limits the power and generalizability of basic and clinical research findings. This thesis first demonstrates that existing heterogeneity indices are inadequate for psychiatric research applications since they rely on (A) valid categorical grouping of subjects/observations, and (B) assumption that distance on the space of observable data is semantically relevant. To address these problems, we subsequently introduce the *representational Rényi heterogeneity* (RRH) framework and compare it to standard heterogeneity indices. Comparisons using simple systems (beta mixture distributions) and complex models of real world data (specifically a convolutional variational autoencoder trained on natural images) attest to the interpretability and flexibility of RRH.

For application, we introduce the largest-ever machine learning (ML) based study of lithium response prediction for patients with bipolar disorder, using only features collected from patient interviews. In a sample of 1266 patients pooled across 7 international sites, lithium response can be predicted with an area under the receiver operating characteristic curve (AUC) of 0.80 (95% CI [0.78,0.82]), but with limited generalizability due to substantial heterogeneity in classification signals between sites. We therefore used RRH to derive *exemplar scoring*, which identifies regions of a feature space that are most reliably classified with high accuracy. We consequently identify "canonical" clinical profiles of lithium responders and non-responders, and show that subjects with high exemplar scores are genetically distinct (AUC 0.88, IQR [0.83, 0.98]) compared to those with low exemplar scores (AUC 0.66 [0.61, 0.80]; $p<0.001$).

An ancillary contribution of this thesis is a counterexample to the common statistical dogma against dichotomization of continuous variables. To this end, we show that dichotomization of asymmetrically reliable variables can retain greater information and statistical power than their raw (continuous) forms.

In sum, this thesis develops a method for heterogeneity measurement that is interpretable, flexible, and useful, particularly for psychiatric research. As a result, we have also made significant advancements toward the understanding and management of bipolar disorder.

Acknowledgements

I must first acknowledge my wife, Dr. Sarah Nunes, whose leadership and diligent management of our family life has facilitated my pursuit of advanced clinical and academic training. She has given me four children and an ongoing example of fortitude, prudence, kindness, wisdom, and faith.

I am grateful for the many opportunities and guidance offered by my supervisors, Drs. Thomas Trappenberg and Martin Alda. The critical appraisal of my committee members, Drs. Dirk Arnold and Timothy Bardouille, have improved this thesis and my thinking more broadly. Dr. Tomas Hajek offered me the extraordinary opportunity to collaborate with the bipolar working group of the ENIGMA consortium (Enhancing NeuroImaging Genetics through Meta-Analysis). I am thankful for the excellent feedback received on manuscripts—often on evenings, weekends, and holidays—from our extensive list of collaborators (names are listed in manuscript citations for respective chapters).

The Dalhousie University Department of Psychiatry and Clinician Investigator Program gave me a fully funded three years to pursue this academic training despite the significant shortage of clinical psychiatrists in Nova Scotia at this time. Particular thanks are owed to Drs. Nick Delva, Michael Teehan, Mark Bosma, Sherry James, Jason Berman, and Michael Bezuhly for their advocacy and support. Drs. Sameh Hassan, Ahmed Saleh, and Jacob Cookey selflessly advocated for me to have this unparalleled opportunity.

I am grateful for financial support from the Killam Trusts, the Nova Scotia Health Research Foundation, the Nova Scotia Health Authority Research Fund, Genome Canada, the Canadian Institutes of Health Research, the Dalhousie University Department of Psychiatry, and the Nova Scotia Department of Health and Wellness. I am also thankful for the many clinical work opportunities made available to me by Dr. Hugh Maguire.

Finally, I owe an infinite debt to my parents, grandparents, and great-grandparents, who endured hardships I have never known, and offered me wisdom so I may never have to.

Chapter 1

Introduction

1.1 What is Heterogeneity?

Heterogeneity (or equivalently *diversity*) is a statistical concept that corresponds roughly to the number of distinct configurations in which a system may be found. When one samples from a *homo*geneous system, all observations in that sample will be identical. That is, a homogeneous system is in a state of *perfect conformity*. Conversely, a heterogeneous system yields samples in which observations have many differences between them. A heterogeneous system is therefore said to diverge from the ground state of perfect conformity.[1]

1.2 Why is Heterogeneity Important?

Understanding heterogeneity is important for the study of many natural phenomena, albeit for different reasons. Ecologists and conservationists must understand heterogeneity since ecological biodiversity improves ecosystems' productivity and robustness to perturbations (such as those related to invasive species and human factors) [2]. Economists and sociologists are interested in heterogeneity in the form of income inequality and industrial concentration (i.e. the degree of heterogeneity in resource ownership or profit generation), which can influence societal functioning, productivity, and political processes [3, 4]. Statistical physicists, network scientists, and many other disciplines, are also interested in heterogeneity in relation to understanding systems' complexity more generally [5].

The application domain of primary interest for the present thesis is medical science, particularly that subdomain related to psychiatric disorders. Here, heterogeneity is of concern mainly because it limits our ability to discover diagnostic tests and targeted therapies for various conditions.

Example 1. Consider a randomized trial of a new drug, denoted D_A, compared to drug D_B, for

[1] This definition is inspired by Eliazar's definition of inequality [1].

Table 1.1: Hypothetical schizophrenia (SCZ) drug trial results.

	N (%) Readmitted within 1 year	
	$\mathbf{D_A}$ (N=500)	$\mathbf{D_B}$ (N=500)
SCZ (N=1000)	250 (50%)	250 (50%)

the treatment of schizophrenia (SCZ). The primary outcome is the probability of hospital admission within 5 years. The investigators randomize 1000 SCZ patients into groups receiving either D_A or D_B, obtain the results in Table 1.1, and conclude that drugs A and B are similarly effective. However, if SCZ is a heterogeneous diagnosis—for instance, containing two hypothetical subtypes SCZ-1 and SCZ-2—then the results in Table 1.2 are plausible.

Table 1.2: Hypothetical schizophrenia (SCZ) drug trial results with subtyping.

	N (%) Readmitted within 1 year	
	$\mathbf{D_A}$ (N=500)	$\mathbf{D_B}$ (N=500)
SCZ-1 (N=500)	0	250 (100%)
SCZ-2 (N=500)	250 (100%)	0
SCZ (Pooled) (N=1000)	250 (50%)	250 (50%)

Tables 1.1 and 1.2 demonstrate the famous *Yule-Simpson effect*, more commonly known as *Simpson's Paradox* [6–9], where the relationship between two variables (here drug treatment and readmission) is eliminated or reversed upon conditioning for a third (here disease subtype). In our hypothetical case, both drugs A and B have extraordinarily high effect sizes, but these are obscured when heterogeneity of SCZ is ignored. Thus, heterogeneity of clinical conditions may attenuate effect sizes and limit progress toward improving diagnosis and treatment of various medical conditions.

Example 1 should not mislead the reader into believing that heterogeneity in medical applications is present only under conditions where the Yule-Simpson effect exists. Indeed, heterogeneity in the clinical group may exist, yet not be sufficient to observe the Yule-Simpson effect.

Example 2. Let us randomize 1000 hypothetical patients diagnosed with "psychosis" to treatment with drugs C (D_C) or D (D_D). Let us assume that patients in this sample have psychosis caused either by SCZ or the manic phase of bipolar disorder. The results in Table 1.3 show the absence of a Yule-Simpson effect, despite the fact that the pooled sample is heterogeneous.

Table 1.3: Hypothetical psychosis drug trial results.

	N (%) Readmitted within 1 year	
	$\mathrm{D_C}$ (N=500)	$\mathrm{D_D}$ (N=500)
Psychosis (SCZ) (N=500)	200 (80%)	25 (10%)
Psychosis (Mania) (N=500)	200 (80%)	25 (10%)
Psychosis (N=1000)	400 (80%)	50 (10%)

Heterogeneity is therefore an important statistical concept in medical sciences, primarily due to its impact on the estimation of diagnostic and therapeutic effect sizes [10]. However, Examples 1 and 2 demonstrate the fact that heterogeneity in clinical groups may exist regardless of effect size distributions.[2] Indeed, whereas heterogeneity directly impacted effect size estimates in Example 1, it had no impact (despite being present) in Example 2.

1.3 Why Should we Measure Heterogeneity?

Having established the importance of heterogeneity across various fields, we now turn to the question of why we should be interested in measuring it. Firstly, the fact that some collections are more diverse than others is a phenomenon observable through ordinary life experience. Since intuition ostensibly treats heterogeneity in a fashion similar to other measurable quantities (such as height, weight, and volume) then it stands to reason that heterogeneity, too, should have a quantitative measure.

Another reason to seek a measure of heterogeneity is that it may be an important property of a system in its own right. We have already noted that the robustness of an ecosystem to invasive species is related to its biodiversity [2]. The degree of cell-type heterogeneity in tumours may affect a cancer's resistance to chemotherapy [11]. Heterogeneity is also a defining feature of multiple sclerosis, which requires lesions in the brain's white matter to be "disseminated across space and time" (i.e. lesions are found at different locations and there are multiple attacks) [12]. Patients with and without bipolar disorder may also have different levels of heterogeneity in their mood states across time [13–15]. In such cases, among others [16], heterogeneity may be a useful feature variable for statistical and other modeling purposes. However, unlocking this potential requires development of quantitative

[2] Deriving a scenario in which one observes a Yule-Simpson effect *without* the presence of heterogeneity in the clinical group is left as an exercise for the reader. *Hint*: the clinical sample is not really heterogeneous in this case.

measures.

1.4 What Constitutes a Heterogeneity Measure?

Chapters 2 and 3 address the question of what constitutes a heterogeneity measure, with special consideration of limitations and constraints imposed by application to psychiatric research problems. Specifically, Chapter 2 summarizes the different statistical properties of heterogeneity implied by psychiatric research studies to date. These include *deviance* (roughly the degree to which observations from a system differ qualitatively) and *multimodality* (roughly the amount of "clustering" or "discrete states" present in a system). To be useful in psychiatric research applications, a heterogeneity measure must capture these properties. To this end, Chapter 2 also provides a cursory introduction to the *Rényi heterogeneity* indices (also known as the *Hill numbers* [17] or *Hannah-Kay indices* [18]), which capture both deviance and multimodality by measuring heterogeneity as the size of a system's event space (in units known as *numbers equivalent* [19, 20]). Since the Rényi heterogeneity requires precise definition of the system's event space, we argue that it has a corollary benefit of reducing the general vagueness with which psychiatric researchers have tended to use the term "heterogeneity."

Chapter 3 provides a more detailed and technical review of the meaning and measurement of heterogeneity by first defining the components of a *system*: an event space equipped with an abundance measure and a distance function. Heterogeneity measures are then divided into those applicable to event spaces with (A) categorical or (B) non-categorical topology. Categorical systems are those in which pairwise distances between all configurations are equal, rendering the event space permutation invariant. Conversely, non-categorical systems are those whose event spaces are *not* permutation invariant, such as when elements have ordinal or continuous distances.

1.5 Why Representational Rényi Heterogeneity?

Chapter 3 compiles an axiomatic basis for measures of heterogeneity, and proves that most are satisfied by the Rényi heterogeneity first broached in Chapter 2. Specifically, we emphasize the importance of a condition known as *the replication principle* [21–23], which arguably provides the strongest support in favour of the Rényi family as the standard indices for heterogeneity measurement. The replication principle states that if we aggregate K equally heterogeneous systems with non-overlapping event spaces, the pooled heterogeneity should be K-fold higher than that of each subsystem. Interestingly, common heterogeneity indices based on entropy [24–26] and variance [27, 28] violate this property (Chapter 3); consequently, these indices can be considered true measures of heterogeneity to the degree that a sphere's radius can be considered a measure of its volume [22].

Chapter 3 also highlights important limitations of existing heterogeneity measures as they apply to non-categorical systems. Specifically, we note that many such indices do not carry units of numbers equivalent, and are thus inconsistent with our more general view of heterogeneity as the size of a system's event space (i.e. the number of system configurations). Beyond some idiosyncratic limitations, the existing non-categorical heterogeneity indices measured in numbers equivalent are shown to carry important problems: they require the raw event space to be discretized into bins whose pairwise distances are measured using a standard distance function, such as those of the Minkowski family. These limitations virtually preclude their application to psychiatric research problems.

In Chapter 4, we propose an approach for measuring heterogeneity based on representation learning. This method maps observable data onto a latent space upon which (A) geometry of relevant semantic features are better captured and (B) the ordinary Rényi heterogeneity formula, or a parametric variation thereof, can be tractably applied. This obviates the need to both discretize the observable space and define a closed form distance metric upon it. Consequently, our *representational Rényi heterogeneity* method resolves the limitations of non-categorical numbers equivalent heterogeneity indices outlined in Chapter 3, and is thereby applicable to psychiatric research problems.

1.6 The Utility of Representational Rényi Heterogeneity for Computational Psychiatry[3]

The psychiatric research community has organized several large scale consortia such as ENIGMA (Enhancing Neuro Imaging Genetics through Meta Analysis [29]), ConLiGen (the International Consortium on Lithium Genetics; [30]), and the Psychiatric Genomics Consortium [31], whose data-pooling efforts can yield sample sizes that enable application

[3] Some material from this section was published in Nunes A. Two common questions about machine learning methods in psychiatric applications. *Bipolar Disorders*. 2019; In Press

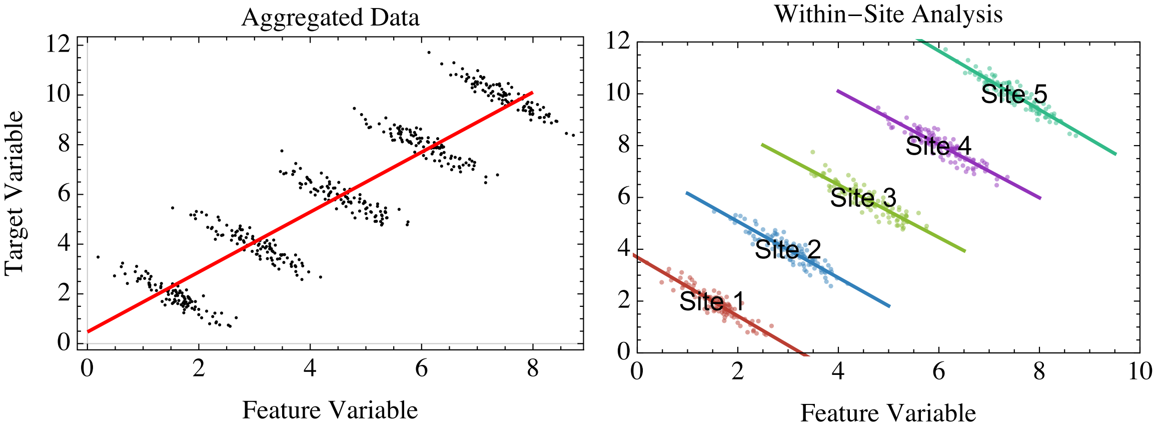


Figure 1.1: Demonstration of one instance of the Yule-Simpson effect in a multi-center analysis setting. When data collected from each of five sites is pooled, a positive association is observed (left plot). If this aggregate model were to be applied by individual sites, it would yield incorrect predictions, since the true relationship—when one controls for stratification—is reversed (rightmost plot).

of machine learning (ML) methods. Whereas classical statistical testing evaluates models' explanatory power, model criticism under the ML paradigm (at least in the supervised domain) is concerned only with predictive power. This reliance on generalization performance frees us from many of the asymptotic assumptions of orthodox statistical tests, thereby allowing us to model nonlinear relationships between features and some relevant target variable(s) [32].

Unfortunately, pooling samples collected at many international sites has also introduced new sources of heterogeneity into the pooled datasets. For instance, it is possible—and some might say nearly inevitable—that samples across sites are not independent and identically distributed (iid). This carries the potential to yield a Yule-Simpson effect (Figure 1.1). Indeed, it has most often been speculated that predictive performance is better when models are tested on within-site data, since they are presumably more homogeneous than the aggregated sets [33]. However, as we showed in Examples 1 and 2, the effects of sample heterogeneity on a given group-level effect are not straightforward. Indeed, we have observed multiple scenarios in which ML analysis of pooled—and nevertheless heterogeneous—datasets resulted in improved classification performances [34, 35]. That being said, the existence of between-site heterogeneity in multi-site datasets threatens the generalizability of prediction models trained on these datasets.

In Chapters 5 and 6, we introduce one of our central applied research problems: learning to predict the effectiveness of lithium treatment in patients with bipolar disorder (BD), based only on features that can be collected through clinical interviews [35]. Using a simple ML classifier, we show that lithium response can be predicted with an area under the receiver operating characteristic curve (ROC-AUC) of 0.80 (95% confidence interval, CI, [0.78-0.82]). However, the features relevant for classification varied systematically across constituent sites' datasets. Therefore, the model developed in Chapter 5 continues to have a significant generalization risk.

Chapter 7 develops a method to filter out heterogeneity introduced by multi-site sample pooling in the lithium response prediction dataset studied in Chapter 5. Our method, which we call *exemplar scoring*, is derived from an application of the representational Rényi heterogeneity to this multi-site prediction problem. Using the exemplar scoring approach—and by extension, the representational Rényi heterogeneity—we were able to rank subjects according to the reliability with which their lithium responsiveness could be predicted based on their clinical profiles. The best "exemplars" in the lithium response and non-response classes showed highly consistent clinical profiles. These empirically derived clinical profiles were similar to those found in past research [36–38]. Further evidence of validity is provided by the fact that the best exemplars of lithium response and non-response could be strongly discriminated using genomic data (ROC-AUC 0.88 [0.83, 0.98]), and that the most informative genetic features agreed with existing knowledge concerning the biological bases of lithium response in bipolar disorder [39–41]. These results, made possible by derivation of the exemplar scoring method from representational Rényi heterogeneity, are the current state-of-the art in genomic classification in computational psychiatry.

1.7 Summary

In sum, this thesis develops an approach to measure heterogeneity that retains the strong axiomatic properties and interpretability of the Rényi heterogeneity (Chapters 2 and 3), yet extends it to arbitrary data types while solving several limitations of existing state-of-the-art comparator indices (Chapter 4). Our approach is demonstrably useful for measurement of heterogeneity in computational psychiatric applications. Specifically, our measure enabled development of an approach to identify clinical phenotypes that are reliably predictive of lithium responsiveness in bipolar disorder (Chapters 5-7).

Chapter 2

Defining and Localizing Heterogeneity in Psychiatric Science[1]

Abstract. In this editorial, we note that while heterogeneity is often discussed at a relatively "high level" in the psychiatric literature, two implicit definitions emerge. Specifically, heterogeneity is generally viewed as one or both of (A) deviance (i.e. more or less qualitative differences between elements of a set) or (B) multimodality (i.e. multiple "clusters" or components in a mixture distribution). Thus, any measure that would be useful for psychiatric research applications must capture both deviance and multimodality. We briefly introduce the *Rényi heterogeneity* family of indices, which capture these properties. Moreover, we argue that the Rényi heterogeneity family's units, known as *numbers equivalent*, help improve the precision of heterogeneity as a concept in psychiatric research. More specifically, the units of numbers equivalent require that we concretely specify the feature space upon which heterogeneity is being measured.

2.1 Introduction

Despite advancements in research methods and the growth of large international data-sharing initiatives [29], our understanding of the biological underpinnings of psychiatric disorders remains limited. An often cited reason for this stagnation is the presence of "heterogeneity," whether intrinsic to the condition or an artifact of clinical assessment, sampling, experimental protocol, or otherwise. However, for a concept of such longstanding importance to psychiatric research, we have no consistent framework within which to study heterogeneity itself.

In this editorial, we argue that heterogeneity must be understood and communicated in two ways. First, we must have a sense of what heterogeneity is as a mathematical and statistical concept. In this respect, we highlight that heterogeneity is generally viewed as a combination of deviance (the degree of differences between elements in a set) and multimodality (the number of clusters in a set or modes in a mixture distribution), both of which can be expressed in a common and easily interpretable set of units known as the

[1] Published as Nunes A, Trappenberg T, and Alda M. We Need an Operational Framework for Heterogeneity in Psychiatric Research. *Journal of Psychiatry & Neuroscience*. 2020;45(1):3-6

numbers equivalent or *effective numbers* [20, 42]. Second, we must understand that the conceptual relevance of heterogeneity is linked to "where" (in terms of levels of analysis) it is expressed. That is, heterogeneity gains substantial conceptual power only when discussed with specific reference to the space of features being deemed "heterogeneous." Here, too, we argue that the units of numbers equivalent can clarify the level at which heterogeneity is being discussed. A central emphasis of this argument, overall, is that understanding heterogeneity requires us to separate our understanding of it as a quantity from the conditions and features that we deem to be "heterogeneous," and the causes thereof.

2.2 What is Heterogeneity?

This section provides a brief overview of the different perspectives with which heterogeneity has been viewed in psychiatric research: deviance and multimodality. We then unify these components under a single set of units, known in ecology, economics, and political science as "effective numbers" or "numbers equivalent" [4, 20, 21, 42].

2.2.1 Deviance

Deviance refers to the degree to which elements in a set or sample differ from one another along one or more characteristics. This is most commonly measured using variance and standard deviation [43], although model-based approaches are increasingly popular in the psychiatric literature [44]. There are many other deviance-based heterogeneity indices [45], but their use in the psychiatric literature remains limited at present for reasons we explore further in forthcoming work [46].

Perhaps the most familiar measure of heterogeneity in the sciences is simply the variance. Particularly notable are those versions employed in meta-analysis, including the variance of between-study effects in mixed-effects meta-analysis [27], and the i^2 statistic (which involves a decomposition of variance into within- and between-study components) [28]. Logarithmic ratios of variance (and coefficient of variation) and parametric models of variance have been used in the neuroimaging literature to compare structural brain heterogeneity of schizophrenic patients against controls [16, 47], Taking variance as one's heterogeneity index assumes that (squared) Euclidean distance of observations from their sample mean is the proper measure of variability in that given system. Unfortunately, this assumption may be overly simplistic in complex real-world data [48].

Recently, researchers in psychiatric neuroimaging have developed an increasingly popular method, known as *normative modeling*, for characterizing heterogeneity in clinical cohorts [44]. This approach begins by using a probabilistic model to learn a distribution of "normal" variation of some clinical or biological feature given some relevant covariate(s) such as age or neuropsychological function. Using extreme value statistics, one then evaluates the degree to which individual subjects in some cohort deviate from their predicted normative distribution, assuming that psychiatric disorders will tend to cause stronger deviations from normative ranges over relevant variables. However, although this method can be useful for characterizing sources of heterogeneity, it does not truly measure the *amount* of heterogeneity in a system.

2.2.2 Multimodality

Multimodality refers to different categories, strata, or distributions being represented within a given set or sample. In the psychiatric literature, the multimodality view of heterogeneity is implied in studies of symptom combination diversity [49], microbial biodiversity [50], and diversity of prescribing habits [51, 52], to name a few. However, it is the large number of clustering and latent class analyses that signify our field's tendency to view heterogeneity as reflective of multimodality in our data.

The nature of clinical psychiatric nosology as a set of symptom checklists has prompted many authors to combinatorially enumerate the number of possible symptom groupings for different conditions. In these studies, each symptom combination is a categorical "mode" in the set of all presentations for a given condition. For example, the number of symptom combinations for major depressive disorder in the Diagnostic and Statistical Manual of Mental Disorders (5th edition) [53] can be shown to equal 227 [54, 55], whereas generalized anxiety disorder (GAD) and borderline personality disorder (BPD) can be shown to have upper bounds of 42 and 256 combinations, respectively [49]. Under this perspective, a condition's heterogeneity is related to the size of the space of all possible clinical presentations. In real world practice, however, there is significant inequality in the distribution of symptom combination incidence. Consider that if each of the 42 presentations of GAD were equally likely, but 99.999% of BPD patients fulfilled all nine criteria, then BPD would be effectively less heterogeneous than GAD, despite having a larger absolute "space of presentations."

To address the insensitivity of simple combinatorial enumeration to inequality in the probability of different events, several indices view heterogeneity as a combination of both (A) size of the event space and (B) the level of inequality in those events' probabilities. As we will see, these indices do not directly measure heterogeneity, but rather properties that are correlated with heterogeneity. For instance, one may measure the degree of uncertainty in the process of sampling from a population (this index is the Shannon entropy) [25]. Indeed, the contents of samples from a more heterogeneous system should be more uncertain. Heterogeneous sets should also be associated with a lower probability of sampling identical pairs, and a greater expected absolute difference (with respect to some normalized feature variables). These two properties of heterogeneity are captured by the famous Gini index [24]. Both the Shannon and Gini indices, or variations thereof, have been used to quantify diversity in psychiatric symptom presentations [49] and gut microbial flora [50] in psychiatric disorders, as well as heterogeneity of psychotropic prescribing patterns [51, 55]. However, these indices can be difficult to interpret and synthesize because they do not measure heterogeneity directly, but rather common secondary properties of heterogeneous sets [22].

Perhaps the most common approach for characterizing heterogeneity in the psychiatric literature has been to count the number of latent clusters or factors inferred from data under some unsupervised learning model. A comprehensive review of these studies is beyond our scope, but many of these studies have been reviewed elsewhere [56, 57]. The central point to appreciate in our context is that these studies all implicitly prioritize multimodality over deviance as the *sine qua non* of heterogeneity. When one measures heterogeneity by latent cluster counting, he is not interested in the absolute amount of deviation between observations, but rather only in the *aggregation* of samples into effectively homogeneous groups. Once the individuals are aggregated into defined clusters, they are treated as now belonging to categorical groups between which deviance is maximal and symmetrical, and within which deviance is absent, since observations are now treated categorically.

Unfortunately, cluster counting approaches have several problems. Perhaps the most significant is related to cluster validity, reproducibility, and the appropriateness of one clustering approach compared to another [56]. Second, since latent classes are viewed as categorical, these methods ignore any within- and between-cluster heterogeneity after the classes have been inferred; for instance, there would be no accounting for the fact that apples are more similar to pears than they are to asphalt. Finally, and perhaps most straightforward, is that the absolute number of clusters will encounter similar problems as the combinatorial symptom enumeration methods discussed above, wherein inequality in cluster sizes is not accounted for the reported "amount" of heterogeneity.

2.2.3 The Effective Numbers (or Numbers Equivalent)

Deviance and multimodality are distinct insofar as they evince one's assumptions about the "smoothness" of differences between observations in a sample. In situations where the phenomenon of interest is thought to be a spectrum, then heterogeneity is typically formalized and communicated in terms that emphasize relative "distances" between subjects. However, when the phenomenon of interest is thought to have an internal stratification, the multimodality view is dominant. The normative modeling paradigm takes a combined perspective where extreme value testing can be used to identify "deviant modes." Yet, these perspectives all manifest in the same practical conclusion: heterogeneous systems all generate a larger number of unique observations.

If a system generates a larger number of observations, then it must have an effectively larger event space. This will be the case regardless of whether one is considering heterogeneity as deviance or multimodality. The word "effective" here is critical, because it accounts for the fact that some systems with large potential event spaces may be "effectively" small if most of the sampling probability is attached to only a few events (as in our example comparing GAD and BPD in Section 2.2.2).

Through an index known as Rényi heterogeneity (synonymous with the Hill numbers [17] in ecology or the Hannah-Kay [18] indices in economics), we can in fact measure a system's heterogeneity in units of numbers equivalent. For a system X with a probability mass function $\mathbf{p} = (p_k)_{k=1,2,\dots,K}$ over discrete event space $\mathcal{X} = \{1, 2, \dots, K\}$, the Rényi heterogeneity is defined as

$$\Pi_q(\mathbf{p}) = \left(\sum_{k=1}^{K} p_k^q \right)^{\frac{1}{1-q}} \tag{2.1}$$

where $q \geq 0$ is a parameter governing sensitivity to rare events. As a simple illustration, consider a bipolar patient who spends 90% of his time depressed, 8% of his time manic and

2% of his time euthymic. Plugging the distribution into Equation 2.1 gives

$$\Pi_q\left(\{0.9, 0.08, 0.02\}\right) = \left(0.9^q + 0.08^q + 0.02^q\right)^{\frac{1}{1-q}}. \qquad (2.2)$$

At $q = 0$ the relative probabilities are ignored and we obtain the patient's *effective number of total mood states* ($\Pi_0 = 3$). At the limit of $q \to 1$ we obtain the patient's *effective number of typical mood states* ($\Pi_1 = 1.5$), and at $q = 2$ we obtain the patient's *effective number of common mood states* ($\Pi_2 = 1.2$). Given a set of mood-state labels for the same subject within two or more time "windows," significance of differences in mood-state heterogeneity may be computed most flexibly by comparison of bootstrap estimated confidence intervals of the Rényi heterogeneity in those two states. In the specific case of affective time-series data, for example, such statistical procedures may enable a more precise quantification of the "evolution" of heterogeneity of mood states within and between individuals.

Note that as we increased q, the measure becomes progressively less sensitive to the presence of the less common states. When we cannot be assured that our sample covers the whole event space—that is, when a system of interest is thought to have a large event space populated mainly by many very rare events (such as the set of species in a gut microbiome)—the value of q is generally set higher (typically $q = 2$). We recommend a default setting of $q = 1$, which proportionally weights common and rare classes, and corresponds to the commonly used perplexity measure.

Although the resolution parameter q introduces some nuances that are beyond our current scope, the central feature of this measure can be nonetheless observed. Specifically, its results are always reported in terms of the size of the event space. This has three benefits. First, it is easily understood since it relies only on the intuitive concepts of counts and sizes. Second, it respects a scaling law known as *the replication principle* [22] which means that doubling the effective number of observations will result in a doubling of the Rényi heterogeneity. Conversely, other indices such as the Shannon entropy, Gini index, and variance will respond idiosyncratically to changes in the size of the event space, and none will respect this doubling property.

The final benefit of measuring heterogeneity in terms of event space size is that it forces us to clearly specify the characteristics and event space of the system whose heterogeneity is being measured. For instance, we defined the affective event space as {Depressed, Manic, Euthymic$\}$ in the toy example above. Readers who astutely identified that the heterogeneity value reported could be invalidated by the overly simplistic three-category "affective event space," have (perhaps implicitly) exploited this very benefit of Rényi heterogeneity. Under our formulation, it is insufficient to simply refer to a disorder as "clinically heterogeneous," "genetically heterogeneous," or worse still, "heterogeneous" in a more general sense. Where one given disorder may be thought of as heterogeneous because of a large effective number of presentations, another may be considered heterogeneous by virtue of a large effective number of causal genetic variants. To this end, we bring further attention to the "localization" of heterogeneity measurement at different levels of analysis.

2.3 Where is the Heterogeneity?

If we are to report heterogeneity in terms of an effective number, we must clearly answer the question: "effective numbers of what?" This question is nontrivial, since the heterogeneity of psychiatric (and other) disorders may differ in degree and relevance across levels of analysis (e.g. genetic, structural, physiological, symptomatic, or otherwise). For instance, syphilis is counted among one of medicine's "great imitators" chiefly because of its large number of clinical presentations. However, it is etiologically homogeneous, with all cases caused by the spirochete, *Treponema pallidum*. Therefore, conditions such as syphilis may be understood as entailing a sort of "distal expansion" of heterogeneity, with the point of expansion beginning at the infection.

In relation to syphilis, other conditions such as, for instance, amyotrophic lateral sclerosis (ALS), might be thought to entail a "contraction" in heterogeneity across levels of analysis (i.e across genetic $\rightarrow$ molecular $\rightarrow$ cellular $\rightarrow \cdots \rightarrow$ clinical levels). Historically, this condition has been sufficiently homogeneous from clinical and electrophysiological perspectives to be distinct from other motor neuron diseases, but it has substantial underlying genetic diversity. There are at least 20 autosomal dominant genetic causes alone of familial ALS, the most prominent of which may be those involving the superoxide dismutase gene (*SOD1*): itself a family of at least six mutations (A4V missense, I113T, A4T, H46R, A89V, G93C) [58].

The metaphor of a condition such as ALS representing a "contraction" of heterogeneity from etiology to clinical presentation may seem clear only in relation to a clear "expansion" associated with syphilis infection. However, identifying relative differences in heterogeneity across levels of analysis is not straightforward, since one may always identify novel but insignificant variations in genetic makeup, biological structure, or clinical presentation. This problem provides still further motivation for emphasizing the feature space upon which heterogeneity is being reported, because comparing effective numbers of 10 and 20 genetic variants is certainly more meaningful than comparing an effective number of 10 genetic variants to 5 clinical phenotypes.

An additional point at which heterogeneity measurement may be relevant is with respect to factors outside of the patient entirely, instead being associated with diagnostic instruments, clinical practices, treatment protocols, and research methods. Quantifying heterogeneity at these levels is an important step toward better isolating and measuring heterogeneity of psychiatric disorders proper.

2.4 Concluding Remarks

Heterogeneity is the degree to which a system diverges from a state of perfect internal conformity. Many psychiatric studies have attempted to describe heterogeneity of clinical cohorts by either quantifying some form of deviance or multimodality in their data. However, we have yet to develop a consistent operational framework within which to measure and communicate heterogeneity. Developing such a framework will first require (A) adopting a common set of easily understandable units for heterogeneity measures, and (B) clarifying the different levels of analysis at which heterogeneity manifests.

Adopting measures with units of numbers equivalent is an important first step to advance the precision with which we can study heterogeneity in psychiatric research. These measures are well developed and accepted particularly for in ecological applications [22], but we must further evaluate their strengths and limitations for psychiatric research applications. One particular limitation that must be confronted is the fact that numbers equivalent heterogeneity measures currently require the system's event space to have a categorical component. As it stands, this will be problematic in scenarios where the categorical groupings of patients are either (A) unreliable or (B) of questionable validity.

The formulation of Rényi heterogeneity makes it clear that the "causes" of heterogeneity will depend on the system whose heterogeneity is being measured. For instance, the effective number of clinical presentations of major depressive disorder will depend on one's diagnostic criteria. Alternatively, the effective number of neurostructural phenotypes in bipolar disorder may be influenced by pharmacological treatments and diversity thereof. The Rényi measure will fortunately admit a statistical procedure for identifying causes or correlates of heterogeneity. In the latter example, if one can model a probability distribution over the space of structural brain images (which can be done using standard unsupervised learning methods), then the effects of medication use on neurostructural heterogeneity can be isolated by exploiting a decomposition of the Rényi heterogeneity (originally proved by the ecologist Lou Jost [59]), whose technical details we expand upon in Chapters 3 and 4. Such a procedure for isolating heterogeneity caused by exogenous factors may better enable us to characterize the heterogeneity intrinsic to the primary system of interest; the ability to precisely quantify and decompose heterogeneity using the Rényi measure is a step in this direction.

The greater precision afforded by developing rigorous measures of heterogeneity will undoubtedly require us to speak of conditions' heterogeneity in terms of more specific levels of analysis. This will likely bring about another challenge: measuring only that heterogeneity which is "relevant" to the phenomenon in question. For example, two brain images of the same person may deviate from each other based on scanner noise, yet the semantic content of those images—which may be known a priori or identifiable only by unsupervised feature learning models such as autoencoders—is homogeneous. The specificity enforced by reporting heterogeneity as "the effective number of X" could serve as such a filter, since presumably one must justify why the heterogeneity of X is sufficiently important to measure its numbers equivalent. However, answers to these questions await the results of these measures' real world applications to psychiatric research problems.

Chapter 3

The Meaning and Measure of Heterogeneity[1]

Abstract. Heterogeneity is an important concept in psychiatric research and science more broadly. It negatively impacts effect size estimates under case-control paradigms, and it exposes important flaws in our existing categorical nosology. Yet, our field has no precise definition of heterogeneity proper. We tend to quantify heterogeneity by measuring associated correlates such as entropy or variance: practices which are akin to accepting the radius of a sphere as a measure of its volume. Under a definition of heterogeneity as the degree to which a system deviates from perfect conformity, this paper argues that its proper measure roughly corresponds to the size of a system's event/sample space, and has units known as numbers equivalent. We arrive at this conclusion through focused review of more than 100 years of (re)discoveries of indices by ecologists, economists, statistical physicists, and others. In parallel, we review psychiatric approaches for quantifying heterogeneity, including but not limited to studies of symptom heterogeneity, microbiome biodiversity, cluster-counting, and time-series analyses. We argue that using numbers equivalent heterogeneity measures could improve the interpretability and synthesis of psychiatric research on heterogeneity. However, significant limitations must be overcome for these measures—largely developed for economic and ecological research—to be useful in modern translational psychiatric science.

3.1 Introduction

Psychiatric discussions of heterogeneity are largely motivated by limitations of the case-control paradigm: ignorance of (A) inter-individual differences within groups, and (B) the fact that some group differences may be larger than others. These assumptions may compromise effect size estimation [10], thereby impeding progress in understanding psychopathology and its treatment. Chapter 1 provided several examples of this phenomenon using the concept of the Yule-Simpson effect.

More broadly, the psychiatric literature has discussed heterogeneity in terms of meta-analysis, the combinatorial enumerations of symptom profiles (i.e. the "number of ways" disorder X can present) [49, 54, 55, 60, 61], cluster analyses of data from psychiatric populations [62, 63], dimensional models [64], concentration or inequality measures [51, 52],

[1] Nunes A, Trappenberg T, and Alda M. The Meaning and Measure of Heterogeneity. *Submitted manuscript*.

time-series complexity [65], and more recently in terms of "normative models" [44, 66]. These approaches evince a problem that has been noted in other fields: amidst a jungle heterogeneity indices, we have neither a unified definition nor clear measure for this concept [42]. If we are to seriously tackle the problem of heterogeneity in psychiatry, we believe it is necessary to have a consistent, easily interpretable, and problem-agnostic framework for its definition and measurement.

In this paper, we define heterogeneity as *the degree to which a system diverges from a state of perfect conformity* (inspired by Eliazar's definition of inequality [1]) and undertake a focused review of more than 100 years of research concerning its measurement. Measures developed in ecology, economics, statistical physics, and more are reviewed along with some of their known psychiatric research applications. We broadly, though somewhat artificially, split these measures into those that operate on categorical or non-categorical data. Importantly, we highlight that generalizable and well-behaved heterogeneity measures share a set of units known in ecology and economics as *the numbers equivalent* [17, 19–23], which allow these measures to roughly capture the "size" of a system's sample/state space (one can also think about this as the number of states that a random variable can take). However, since these measures have largely been developed outside of psychiatric research, we identify several problems to be overcome before they can be widely applicable in modern translational psychiatric science.

3.2 A Definition of Heterogeneity and Measurement in Categorical Systems

A system's heterogeneity is the *degree to which it diverges from a state of perfect conformity*. A "system" has three components (Figure 3.1A): (A) *a set*, "event space," or "sample space" $\mathcal{X}$ of distinct potential observations which one can also think of as "elements," "partitions," "groups," or "categories," (B) *a measure of distance* $d(x_i, x_j)$ between any two potential elements x_i and x_j in $\mathcal{X}$, and (C) *a measure of abundance* of each element in $\mathcal{X}$. If the abundance function sums to 1 over the entire set $\mathcal{X}$, then the abundance measure is a probability distribution.

In this section, we consider only categorical systems since they are an excellent starting point for developing intuition about the measurement of heterogeneity. Categorical systems are effectively defined by the following distance function (the discrete metric):

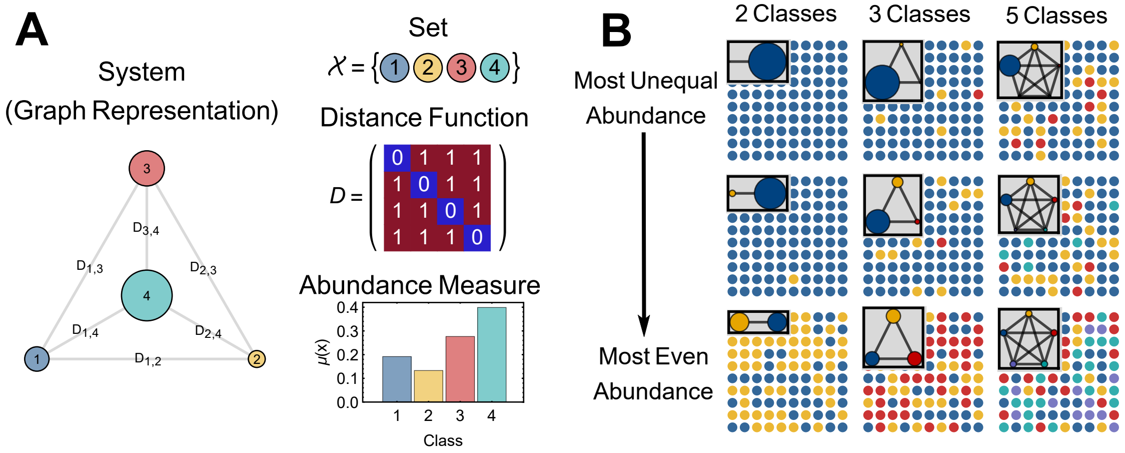


Figure 3.1: Illustration of system components and influence on heterogeneity of samples. **Panel A** depicts a categorical system comprised of a set four categories (equivalently "elements" or "partitions") connected by undirected edges whose lengths are proportional to the distance between categories. In this case, the distances between categories are all equal (symmetric), and the within-category distance is 0 (as evident in the depicted distance matrix). These properties define the set as categorical. The size of the nodes represents their relative abundance, which is also shown in the corresponding bar chart. **Panel B** demonstrates samples from nine categorical systems with varying number of categories (2, 3, and 4) and varying levels of inequality in the abundance distribution. Systems in the upper row have the highest level of inequality in abundance, whereas the systems shown in the bottom row have perfectly even abundance distributions. Together, these plots demonstrate that heterogeneity increases with both (A) increases in the number of categories and (B) more evenly distributed abundance across categories.

$$D_{ij} = d\left(x_i, x_j\right) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \tag{3.1}$$

Like the case-control assumptions, this function states that (A) there are no inter-individual differences within a group or category, and (B) all categories are maximally different, thus meaning no two categories are more similar than any other two. This is one of the central problems we seek to address by better understanding heterogeneity in computational psychiatry. In Section 3.3 we will consider approaches that do not rely on Equation 3.1, and further describe limitations precluding their application to computational psychiatric research. Then, in Chapter 4, we will provide a framework to address those problems. However, for the time being, it is important to begin with an understanding of heterogeneity measurement in categorical systems, since this is the foundation of existing approaches.

It must be emphasized that in this section we acknowledge that most data are not strictly categorical. For example, in classification problems, we attempt to predict some categorical labels based on some other features of arbitrary type, and in clustering we often assign categorical labels to clusters on a non-categorical space. We will deal with these problems in Section 3.3. In this section, however, we are only considering data that consist of a set of mutually exclusive categories, without reference to any other properties those categories may be associated with in reality.

3.2.1 Measuring Heterogeneity by Partition Counting

A system in *a state of perfect conformity* is one whose event space $\mathcal{X}$ effectively has only one element. All observations from this system will be identical. All else being equal, systems that deviate further from perfect conformity will thus have larger event spaces (Figure 3.1B). We therefore require that heterogeneity measures be strictly positive unless the system consists of an empty event space.[2]

Axiom 1 (Non-negativity)**.** For all vectors $\mathbf{y}$ sampled from an n-dimensional space of abundance distributions $\mathcal{Y} \subseteq \mathbb{R}^n_{\geq 0}$, the heterogeneity measure $h(\mathbf{y})$ is strictly positive.

[2] Axiom 2 is trivial in practice, but we include it here simply for completeness.

Axiom 2 (Null empty set)**.** The heterogeneity measure $h : \mathcal{Y} \to \mathbb{R}_{\geq 0}$ equals 0 iff $\mathcal{Y} = \emptyset$

Clearly, a simple count of the number of elements or partitions in $\mathcal{X}$ will satisfy Axioms 1 and 2. Partition counting methods are canonical examples of measures that fulfill the axiom of *monotonicity to set size* (Axiom 3), which states that heterogeneity must increase if a system's event space grows in size.

Axiom 3 (Extensivity or Monotonicity to Set Size)**.** Given a family of distributions $\mathbf{y}(n) = (y_i)_{i=1}^{n}$ with a constant level of inequality for all $n \in \mathbb{N}_+$, the heterogeneity measure h must satisfy

$$h(\mathbf{y}(n+\delta)) > h(\mathbf{y}(n)) \; \forall \delta \in \mathbb{N}_+ \tag{3.2}$$

Such partition counting methods work on the assumption that the size or "cardinality" of $\mathcal{X}$—the number of distinct partitions or elements it contains—measures that system's heterogeneity.

Partition counting methods are often used to quantify a disorder's clinical heterogeneity by the number of criteria-satisfying symptom combinations [49, 54, 55, 60, 61, 67, 68]. Here, one assumes that the "system" is the disorder in question. For each diagnosis, the set $\mathcal{X} = \{1, 2, \dots, n_c^*\}$ consists of a total of n_c^* (the asterisk denotes that this is the "true" value, which may or may not be known) categorically unique symptom combinations or "presentations." Estimating n_c^* amounts to estimating the system's heterogeneity. The next few sections will describe several approaches for this estimation problem.

Combinatorially Estimating an Upper Bound for $\mathbf{n_c^*}$

Many studies estimate an upper bound for n_c^* using combinatorial methods. In these cases, one is not obtaining n_c^* from empirical data (such as by counting the number of distinct observations in a dataset); rather, one directly calculates the total number of unique configurations that may be realized by that categorical system. Hence, this is an *upper bound* on n_c^* since empirical data could not exceed the computed value. For example, a diagnosis of generalized anxiety disorder (GAD) under the Diagnostic and Statistical Manual of Mental Disorders (5th ed.) [69], requires three or more of six symptoms. If we denote the total number of available symptoms as N and the number of required symptoms as K, the number of unique symptom combinations is

$$S\left(N,K\right)=\sum_{k=K}^{N}\frac{N!}{k!\left(N-k\right)!} \tag{3.3}$$

One calculates that GAD has at most $S\left(6,3\right)=42$ unique presentations. Similarly, one can verify that for borderline personality disorder $S(9,5)=256$, for catatonia $S(12,3)=4017$. For major depressive disorder (MDD), which has mandatory symptoms of either low mood or loss of interest, one can show that there are 227 symptom combinations.

Estimating $\mathbf{n_c}$ Empirically from Data

Zimmerman et al. [54] found a total of 170 unique symptom combinations in a survey of 1500 MDD patients, suggesting that 25% of theoretical symptom combinations do not occur. Similarly, Park et al. [55] found 119 unique combinations in 853 subjects further highlighting that empirical estimates of n_c^* are important complements to combinatorial enumeration. Unfortunately, any sample short of a complete census will underestimate n_c^*, particularly if many of the categories in $\mathcal{X}$ are rare. Several approaches address this problem.

The simplest, but most biased (lower limit), estimator of n_c^* is the *observed richness* (also known as *species richness* to ecologists) [42, 70], which is the observed number of categories in the sample. We denote this quantity as $\Pi_0=n_c$ (the lack of asterisk denotes it is an estimate).

A less biased approach for estimating n_c^* is to compute a more appropriate lower bound [70, 71], using the *Chao* estimators. These indices, which are standard in ecology, use information about the frequency of rare categories to speculate on how many further rare categories may exist who have not yet been sampled. If we denote f_K as the number of categories observed only K times, then the corresponding Chao estimator is as follows [72]:

$$\text{Chao}_1\left(f\right)=\begin{cases}\Pi_0+\frac{f_1^2}{2f_2} & f_2>0\\ \Pi_0+\frac{1}{2}\left(f_1\left(f_1-1\right)\right) & f_2=0\end{cases} \tag{3.4}$$

If $f_2>0$ then the Chao1 estimator has a corresponding variance estimator that can be used to construct confidence intervals.

The observed richness values reported by Zimmerman et al. [54] and Park et al. [55] underestimate the true number of MDD presentations. After abstracting the presentation frequency tables from these papers (Figure 3.2A), we used the Chao estimator to recalculate

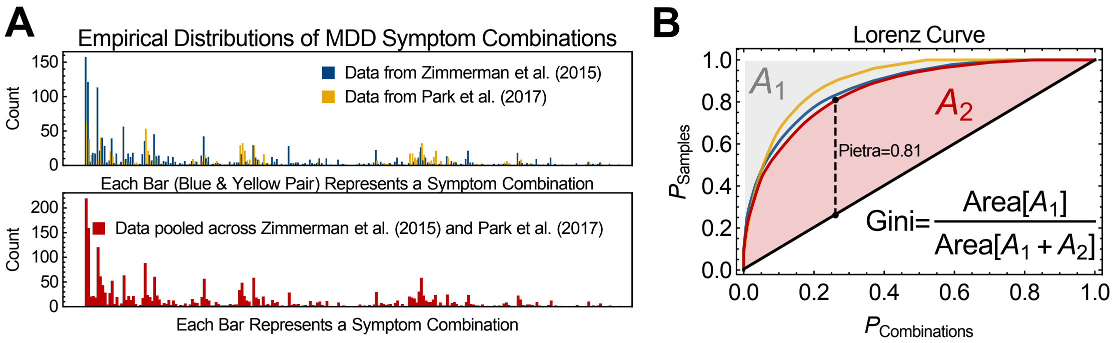


Figure 3.2: **Panel A**: Distribution of symptom presentations in patients with major depressive disorder as reported by Zimmerman et al. [54] and Park et al. [55] (data extracted from their published tables). **Panel B**: Lorenz curves for the empirical distributions shown in Panel A. Curve colours are matched between panels. In this case, the Lorenz curve demonstrates the proportion of symptom combinations ($P_{Combinations}$) that account for at least $P_{Samples}$ proportion of observed presentations in the datasets. The diagonal (black) line represents the line of perfect equality, which would occur only if all symptom combinations accounted for the same proportion of observed presentations. The closer a Lorenz curve is to the upper corner, the more inequality exists in the abundance distribution, which in this case would indicate greater homogeneity of symptom presentations. Geometric calculation of the Gini coefficient and Pietra indices is also demonstrated. The Gini index is the ratio of (A) the area between the Lorenz curve and the line of perfect equality to (B) the total area above the Lorenz curve. The Pietra index is the maximum distance from the Lorenz curve to the line of perfect equality, and represents the proportion of observations that would need to be transferred from the most common to the least common symptom combinations in order to reach the line of perfect equality.

lower bound estimates on the number of MDD symptom combinations. In the Zimmerman et al. [54] data, this was 189.8 (95% confidence interval, CI [189.3, 190.2]), compared to 144.1 (143.4, 144.9) for the Park et al. [55] data, and 200.6 (200.4, 200.9) in the pooled sample. Thus, the heterogeneity of symptom combinations in MDD may be larger than previously estimated using empirical data.

Observed richness and the Chao estimator have been used to quantify gut microbiomic heterogeneity in people with psychiatric disorders, finding no difference between healthy controls and males with attention deficit-hyperactivity disorder (ADHD) [50], but lower microbiome diversity in patients with MDD [73].

The Chao estimators are notably related to *capture-recapture* methods [70, 74], which estimate the size of difficult-to-sample population by examining overlap in repeated samples. Applications include estimation of the prevalence of alcohol-related disorders [75], opioid addiction [76], and other conditions [77–83]. Krebs [84] provides an accessible introduction to these approaches.

Limitations of Partition Counting Approaches

Partition counting methods ignore the abundance distribution's skewness. For example, imagine 99.999% of all patients showed a single presentation of MDD, with the remaining 0.001% spread across the other 226 symptom combinations. This system is *effectively* close to perfect conformity, yet partition counting methods would nonetheless overestimate a heterogeneity value of 227 presentations.

3.2.2 Measures Accounting for Inequality in Category Abundance

Consider a scenario in which 99.999% of all patients have the same presentation of MDD, with the remaining 0.001% evenly spread across the other 226 symptom combinations. In this section, we compute how far this system diverges from perfect conformity given the highly skewed abundance distribution. We restrict our search to those indices that satisfy several axioms, including symmetry (Axiom 4), which is applicable only to heterogeneity measures on categorical spaces.

Axiom 4 (Symmetry)**.** Given an abundance distribution $\mathbf{y} = (y_i)_{i=1}^n \in \mathcal{Y} \subseteq \mathbb{R}^n_{\geq 0}$ and a permutation function $\sigma : \mathbb{N}_+ \to \mathbb{N}_+$, the heterogeneity measure $h : \mathcal{Y} \to \mathbb{R}_{\geq 0}$ satisfies

$$h\left([y_1, y_2, \dots, y_i, \dots y_n]\right) = h\left(\left[y_{\sigma(1)}, y_{\sigma(2)}, \dots, y_{\sigma(i)}, \dots, y_{\sigma(n)}\right]\right) \tag{3.5}$$

We also require that our measure be continuous and differentiable (Axiom 5).

Axiom 5 (Continuity and Differentiability)**.** The heterogeneity measure $h : \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is continuous and differentiable $\forall \mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^n_{\geq 0}$.

Further, we require satisfaction of the *axiom of transfers* [85, 86]. That is, any transfer of abundance from a more abundant category to any less abundant category (thereby making the abundance distribution more even) must increase heterogeneity. This is sensible, since in the opposite scenario—progressively stacking all abundance onto a single category—would push the system toward perfect conformity.

Axiom 6 (Principle of Transfers)**.** Given an abundance vector $\mathbf{y} = (y_i)_{i=1}^{n}$, if we define a new vector $\mathbf{y}'$ by the following transfer of some small amount of abundance,

$$y'_k = \begin{cases} y_k - \epsilon & k = j \\ y_k + \epsilon & k = i \\ y_k & k \neq i \wedge k \neq j \end{cases} \tag{3.6}$$

where $y_j > y_i$ then heterogeneity must increase, with a maximal value attained iff $y_i = y_j$.

Assuming that abundance measures are normalized such that they represent probability distributions, the most common of these heterogeneity indices are entropies derived from the Tsallis family [26], most notably the *Shannon entropy* [25],

$$H\left[\mathbf{p}\right] = -\sum_{i=1}^{n_c} p_i \log p_i \tag{3.7}$$

which measures the average amount of uncertainty in the system. If the logarithm is taken with base 2, then Shannon entropy gives the average number of yes/no questions required to classify an observation from the system.

The *Gini-Simpson index* (GSI), perhaps more commonly known simply as the *Gini index* or *Gini coefficient* [24], is another historically important entropy:

$$\text{GSI}\left(\mathbf{p}\right) = 1 - \sum_{i=1}^{n_c} p_i^2 \tag{3.8}$$

Whereas Shannon entropy uses units of information, the GSI is the probability that two observations from our system (sampled with replacement) will belong to different categories.

The GSI is related to a concentration index commonly attributed to *Simpson* or *Herfindahl* [87, 88]:

$$\text{Simpson}\left(\mathbf{p}\right) = \sum_{i=1}^{n_c} p_i^2 = 1 - \text{GSI}\left(\mathbf{p}\right) \tag{3.9}$$

The Simpson index gives the probability that two samples from our system will belong to the same category. Psychiatric researchers have used this to measure the *homogeneity* of physicians' and health systems' prescription repertoires [51, 52].

Olbert et al. [49] used the GSI and a normalized version of the Shannon entropy to empirically quantify symptom heterogeneity in MDD and PTSD. Using data from $n_s = 84,103$ subjects with MDD in the National Comorbidity Survey Replication (NCS-R) [89], they found an observed richness of 137 unique symptom combinations. The probability of sampling two individuals with MDD whose symptom profiles were different (i.e. the GSI) was 0.96, suggesting a high degree of symptomatic diversity in MDD. However, their Shannon entropy index (with base 2) was 3.9 bits, meaning that approximately four yes/no questions could precisely identify a typical subject's specific symptom profile given only knowledge of their MDD diagnosis.

If one accepts that the GSI and Shannon entropy are both *measures* of heterogeneity, then the results obtained by Olbert et al. [49] are puzzling. On the one hand, the GSI suggests that most pairs of MDD patients will have different symptom profiles (GSI=96%). Conversely, the Shannon entropy amounted to 55% of its theoretical maximum (3.9 of 7.09 bits), suggesting less heterogeneity than the GSI, illustrating the problem of multiple meanings between entropic-based heterogeneity indices. Synthesizing the results from such indices with different meanings can be challenging, and thus we seek measures with conceptually standard units.

Another important problem with the entropy-based heterogeneity indices is that they do not satisfy the *axiom of replication* (also known as the *replication principle* in ecology) [21–23, 90]. The replication principle states that if we pool K completely unique independent systems with equal amount of heterogeneity, h, then the heterogeneity should measure $K \times h$. Jost [22] noted this is akin to merging two spheres, each with volume V; the resulting volume of the pooled sphere should be $2V$, which would not be the result if we treated the sphere's radius (a mere *index* of volume) as a measure.

Axiom 7 (The Replication Principle)**.** We are given $n_s \in \mathbb{N}_{\geq 2}$ systems, with respective distributions $\mathbf{y}_i \; \forall i \in \{1, 2, \dots, n_s\}$, whose domains of support are non-overlapping, but whose heterogeneities are equal:

$$h\left(\mathbf{y}_i\right) = h\left(\mathbf{y}_j\right) \; \forall (i, j) \in \{1, 2, \dots, n_s\} \tag{3.10}$$

Letting $\bar{\mathbf{y}}$ be the abundance distribution on the pooled n_s systems, the replication principle states that

$$h\left(\bar{\mathbf{y}}\right) = n_s h\left(\mathbf{y}_i\right) \; \forall i \in \{1, 2, \dots, n_s\} \tag{3.11}$$

Proposition 1. Entropic heterogeneity measures of the Tsallis family [26] fail to satisfy the replication principle.

Proof. Assume the total number of partitions in a system composed of K subsystems is $n = \sum_{i=1}^{K} n_i$, where n_i is the number of partitions in the i'th sample. The probability distribution for system i is $\mathbf{p} = (p_{ij})_{j=1}^{n_i}$. Recall that the domains of support for $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_K$ are disjoint. The Tsallis entropy of a single system is

$$T_q(\mathbf{p}_i) = \frac{1 - \sum_{j=1}^{n_i} p_{ij}^q}{q-1} \tag{3.12}$$

and the Tsallis entropy of the pooled system, whose probability distribution is $\bar{\mathbf{p}}$ is as follows:

$$T_q(\bar{\mathbf{p}}) = \frac{1 - \sum_{k=1}^{n} \bar{p}_k^q}{q-1} \tag{3.13}$$

Since the domains for each of the K subsystems is disjoint, we have that

$$\bar{\mathbf{p}} = \frac{1}{K}\left(p_{11}, p_{12} \dots, p_{1n_1}, p_{21}, p_{22}, \dots, p_{2n_2}, \dots, p_{i1}, p_{i2}, \dots, p_{in_i}, \dots, p_{K1}, p_{K2}, \dots, p_{Kn_K}\right). \tag{3.14}$$

Substituting Equation 3.14 into Equation 3.13 yields the following expression

$$\begin{aligned} T_q(\bar{\mathbf{p}}) &= \frac{1 - \sum_{i=1}^{K} \sum_{j=1}^{n_i} K^{-q} p_{ij}^q}{q-1} \\ &= 1 - K^{-q} \sum_{i=1}^{K} \sum_{j=1}^{n_i} p_{ij}^q \\ &= \frac{1 - K^{1-q}\lambda_i}{q-1}, \end{aligned} \tag{3.15}$$

where the last step follows from the facts that $\sum_{j=1}^{n_i} p_{ij}^q = \sum_{j=1}^{n_k} p_{kj}^q \ \forall i, k$, and where $\lambda_i = \sum_{j=1}^{n_i} p_{ij}^q$. Under these assumptions, we also have that

$$T_q(\mathbf{p}_i) = \frac{1 - \lambda_i}{q-1}. \tag{3.16}$$

The replication principle asserts that $T_q(\bar{\mathbf{p}}) = K T_q(\mathbf{p}_i)$, which we henceforth prove to be false.

$$T_q(\bar{\mathbf{p}}) = KT_q(\mathbf{p}_i) \tag{3.17}$$

$$1 - \lambda_i K^{1-q} = K - K\lambda_i \tag{3.18}$$

$$\lambda_i = \frac{K-1}{K-K^{1-q}}. \tag{3.19}$$

The last inequality holds only at $q = 1$, which is irrelevant since $T_q(\mathbf{p})$ is undefined at that point (we will prove this in the limiting case of $q \to 1$, below). Showing that the above equality is not generally true is done by counter example. Substituting $\lambda_i \to \sum_{j=1}^{n_i-2} p_{ij}^q + (p_{in_i} - \epsilon)^q + \left(p_{i(n_i-1)} + \epsilon\right)^q$ and differentiating with respect to ϵ, we obtain our result:

$$\frac{\partial}{\partial\epsilon}\left(\sum_{j=1}^{n_i-2} p_{ij}^q + \left(p_{i(n_i-1)} + \epsilon\right)^q + (p_{in_i} - \epsilon)^q\right) = \frac{\partial}{\partial\epsilon}\left(\frac{K-1}{K-K^{1-q}}\right) \tag{3.20}$$

$$\frac{1}{K-K^{1-q}} - \frac{(K-1)\left(1-(1-q)K^{-q}\right)}{\left(K-K^{1-q}\right)^2} = 0 \tag{3.21}$$

$$qK^{1-q} - K^{1-q} - qK^{-q} + K^{-q} + K - 1 = K - K^{1-q} \tag{3.22}$$

$$qK^{1-q} - 1 = qK^{-q} - K^{-q} \tag{3.23}$$

$$K^{-q} = 0 \tag{3.24}$$

Since $K \geq 1$ and $q \geq 0$, the final equality is false. Thus, for $q \neq 1$, the Tsallis family of entropies does not satisfy the replication principle.

We now show the $q \to 1$ case. We use L'Hôpital's rule to obtain $T_1(\mathbf{p})$:

$$T_1(\mathbf{p}_i) = -\sum_{j=1}^{n_i} p_{ij}\log(p_{ij}), \tag{3.25}$$

and similarly for $T_1(\bar{\mathbf{p}})$,

$$T_1(\bar{\mathbf{p}}) = -\sum_{i=1}^{K}\sum_{j=1}^{n_i} \frac{p_{ij}\log\left(\frac{p_{ij}}{K}\right)}{K}. \tag{3.26}$$

The replication principle states that

$$KT_1\left(\bar{\mathbf{p}}\right) = K^2 T_1\left(\mathbf{p}_i\right) \tag{3.27}$$

$$-\sum_{i=1}^{M}\sum_{j=1}^{n_i} p_{i,j}\log\left(\frac{p_{ij}}{K}\right) = -K^2\sum_{j=1}^{n_i} p_{ij}\log\left(p_{ij}\right), \tag{3.28}$$

$$\tag{3.29}$$

and since $\sum_{i=1}^{K}\sum_{j=1}^{n_i} p_{ij} = K$, with $\sum_{j=1}^{n_i} p_{ij}\log p_{ij} = \lambda_i$, we have

$$-K\lambda_i - K\log K = -K^2\lambda_i. \tag{3.30}$$

Solving for λ yields

$$\lambda_i = \frac{\log K}{K-1}, \tag{3.31}$$

which is false since λ_i is a property of a single subsystem, independent of the number of pooled subsystems K.

□

3.2.3 Numbers Equivalent Measures of Heterogeneity

The *Rényi heterogeneity* family of indices—which are the exponential of the Rényi entropy [91]—satisfies the replication principle, and its units are the same units as partition counting methods: the (effective) number of distinct elements in an event space,

$$\Pi_q\left(\mathbf{p}\right) = \left(\sum_{i=1}^{n_c} p_i^q\right)^{\frac{1}{1-q}}. \tag{3.32}$$

This family is also known as the *Hill numbers* in ecology [17], and the *Hannah-Kay* indices in economics [18]. The parameter $q \geq 0$ serves as an "importance" attributed to more abundant categories. When $q = 0$, the abundances are ignored, and we recover the observed richness:

$$\Pi_0\left(\mathbf{p}\right) = \sum_{i=1}^{n_c} p_i^0 = n_c \tag{3.33}$$

Taking the limit as $q \to 1$ yields the exponential of the Shannon entropy, which is the *perplexity* [43] or *the effective number of typical categories* in the system:

$$\Pi_1\left(\mathbf{p}\right) = e^{-\sum_{i=1}^{n_c} p_i \log\ p_i}. \tag{3.34}$$

An alternative derivation of Equation 3.34 makes the connection to typical set size more clear. Consider sampling n_s observations from a system with n_c classes. The observed frequencies over classes is denoted by $\mathbf{p} = (p_i)_{i=1,2,\dots,n_c}$. The effective number of typical samples of size n_s is

$$N = \frac{n_s!}{\prod_{i=1}^{n_c} p_i!} \tag{3.35}$$

Taking logs of both sides, we have

$$\log N = \log n_s! - \sum_{i=1}^{n_c} \log p_i!, \tag{3.36}$$

which by Stirling's approximation, $\log x! = x \log x - x$, gives

$$\log N = -\sum_{i=1}^{n_c} p_i \log \frac{p_i}{n_s}, \tag{3.37}$$

which reduces to Equation 3.7 when $n_s = 1$. Its exponential is the effective number of typical categories in the system, and is equal to Equation 3.34.

At $q = 2$, we have the *inverse* Simpson concentration [42],

$$\Pi_2\left(\mathbf{p}\right) = \frac{1}{\sum_{i=1}^{n_c} p_i^2} \tag{3.38}$$

which is *the effective number of common categories* in the system, known to political scientists as *the effective number of parties* [4]. This measure has been used to estimate the effective number of common bacterial species in the microbiome of patients with MDD [73].

The units of Rényi heterogeneity are known as *numbers equivalent* [19–21] by ecologists and economists. Numbers equivalent can be intuitively understood as follows: for any system A with a given abundance distribution, we can find a "hypothetical" categorical system B whose abundance distribution is perfectly even, and whose heterogeneity is equal to that of A. The number of partitions in this "equivalent" system B serves to measure the heterogeneity of A. Numbers equivalent allow us to account for inequality in the abundance distribution while retaining the units of set size.

It is trivial to show that the Rényi heterogeneity satisfies Axioms 1 (non-negativity) and 4 (symmetry), as well as the axiom of invariance to scaling of the abundance distribution (Axiom 8).

Axiom 8 (Scale-invariance)**.** Given an abundance vector $\mathbf{y} = (y_i)_{i=1}^n$ and a positive scalar $k \in \mathbb{R}_+$, $h(k\mathbf{y}) = h(\mathbf{y})$.

Axiom 8 holds because probabilities are normalized: that is, when $p_i = y_i/(\sum_{j=1}^n y_j)$. Here, we prove that Rényi heterogeneity satisfies Axioms 3 and 6.

Proposition 2. The Rényi heterogeneity is monotonic to set size, and thereby obeys Axiom 3.

Proof. For $\mathbf{p} = (p_i)_{i=1}^n$, recall that $\Pi_0(\mathbf{p}) = n$. One can show that the derivative of the Rényi entropy with respect to q is proportional to the negative Kullback-Leibler divergence $\sum_{i=1}^n z_i \log \frac{z_i}{p_i}$, where $z_i = p_i^q / \sum_{i=1}^n p_i^q$:

$$\frac{\partial}{\partial q}\Pi_q(\mathbf{p}) = \frac{1}{(q-1)^2}\sum_{i=1}^n z_i \log \frac{z_i}{p_i}, \tag{3.39}$$

which means that $\Pi_q(\mathbf{p})$ is non-increasing with respect to q, and thus

$$\frac{\Pi_q(\mathbf{p})}{\Pi_0(\mathbf{p})} \leq 1. \tag{3.40}$$

Now define a hypothetical family of distributions $\mathbf{p}(n) = (p_i)_{i=1}^n$ with a constant level of evenness, $\Pi_q(\mathbf{p}(n))/\Pi_0(\mathbf{p}(n))$. Thus

$$\frac{\Pi_q(\mathbf{p}(n))}{\Pi_0(\mathbf{p}(n))} = \frac{\Pi_q(\mathbf{p}(n+1))}{\Pi_0(\mathbf{p}(n+1))} \tag{3.41}$$

$$\frac{\Pi_q(\mathbf{p}(n))}{n} = \frac{\Pi_q(\mathbf{p}(n+1))}{n+1} \tag{3.42}$$

$$\frac{n+1}{n} = \frac{\Pi_q(\mathbf{p}(n+1))}{\Pi_q(\mathbf{p}(n))} \tag{3.43}$$

□

Proposition 3. The Rényi heterogeneity obeys Axiom 6 (the principle of transfers).

Proof. For some small ϵ transfer from p_n to p_{n-1}, where prior to transfer $p_n > p_{n-1}$, the Rényi heterogeneity is

$$\Pi_q(\mathbf{p}') = \left((p_{n-1} + \epsilon)^q + (p_n - \epsilon)^q + \sum_{i=1}^{n-2} p_i^q \right)^{\frac{1}{1-q}} . \tag{3.44}$$

Solving for $\frac{\partial}{\partial \epsilon}\Pi_q(\mathbf{p}') = 0$ gives

$$\epsilon^* = \frac{1}{2}\left(p_n - p_{n-1}\right), \tag{3.45}$$

which is a transfer that would set $p_n = p_{n-1}$. Recalling that $(p_n + p_{n-1}) > 0$ and $q > 0$, with simple algebra one can show that $\Pi_q(\mathbf{p}')$ has constant negative curvature at ϵ^* and therefore that no further transfers can increase heterogeneity beyond establishment of equality. □

Rényi heterogeneity satisfies the replication principle (Axiom 7).

Proposition 4. Rényi heterogeneity obeys the replication principle (Axiom 7).

Proof. The Rényi heterogeneity for a single distribution $\mathbf{p}_i = (p_{ij})_{j=1,2,\dots,n_i}$, where $n_i \in \mathbb{N}_+$ is the size of the state space in system j, is

$$\Pi_q(\mathbf{p}_i) = \left(\sum_{j=1}^{n_i} p_{ij}^q \right)^{\frac{1}{1-q}} \tag{3.46}$$

and for the aggregation of K subsystems is

$$\Pi_q(\bar{\mathbf{p}}_i) = \left(\sum_{i=1}^{K} \sum_{j=1}^{n_i} \left(\frac{p_{ij}}{K} \right)^q \right)^{\frac{1}{1-q}} . \tag{3.47}$$

The replication principle asserts that

$$\Pi_q(\bar{\mathbf{p}}_i) = K\Pi_q(\mathbf{p}_i), \tag{3.48}$$

which simple algebra shows to be true. Let $\lambda_i = \sum_{j=1}^{n_i} p_{ij}^q$ and recall that $\lambda_i = \lambda_k \ \forall (i,k) \in \{1, 2, \dots, K\}$. Then, $K\Pi_q(\mathbf{p}_i) = K\lambda_i^{\frac{1}{1-q}}$, and expanding the left hand side, we have

$$
\begin{aligned}
\Pi_q(\bar{\mathbf{p}}_i) &= \left(K^{-q} \sum_{i=1}^{K} \sum_{j=1}^{n_i} p_{ij}^q \right)^{\frac{1}{1-q}} \\
&= \left(K^{-q} \sum_{i=1}^{K} \lambda_i \right)^{\frac{1}{1-q}} \\
&= \left(K^{1-q} \lambda_i \right)^{\frac{1}{1-q}} \\
&= K \lambda_i^{\frac{1}{1-q}}.
\end{aligned}
\tag{3.49}
$$

□

Proof that the Rényi heterogeneity satisfies the *axiom of decomposability* can be found in Jost [59]. We also provide a more thorough treatment of decomposition in Chapter 4. Briefly, if a system is composed of K pooled groups, then the overall heterogeneity (known as γ-heterogeneity) must be decomposable into *within*- and *between*-group components ("α-heterogeneity" and "β-heterogeneity," respectively). Decomposition in categorical systems must satisfy some important criteria (detailed by Jost [59]), including the fact that within-group heterogeneity (α)—which can be interpreted as the average heterogeneity within the composition's subgroups (see Chapter 4)—must always be less than the pooled system heterogeneity (γ). This is sensible, since pooling categorical systems should never reduce heterogeneity. Heterogeneity decomposition is commonly employed in meta-analysis (via the i^2 statistic), albeit not using units of numbers equivalent.

3.2.4 Inequality Indices for Comparing Heterogeneity of Differently Sized Sets

It is sometimes useful to measure abundance inequality independently of the event space size (but see Jost [92] for counterpoints). For instance, let each individual in a population be a "partition" in our system, and the abundance measure his or her share of the total populations' wealth. If we collect such data from two populations of different sizes and compare their Rényi heterogeneity values, our results will be confounded by the population sizes; the larger population will tend to have a higher heterogeneity despite potentially having more wealth inequality. For this reason, isolated measures of inequality tend to be invariant to the size of the event space: a property known as *non-extensivity* or *the population principle* [93, 94]. This is essentially the opposite of Axiom 3, meaning that increases of the number of states in a system should not change the heterogeneity index. From a practical standpoint, this would mean that a measure of economic inequality should not change solely based on the size of a population.

Axiom 9 (Non-extensivity or The Population Principle)**.** Given a family of distributions $\mathbf{y}(n) = (y_i)_{i=1}^{n}$ with a constant level of inequality for all $n \in \mathbb{N}_+$, the heterogeneity measure h must satisfy

$$h(\mathbf{y}(n+\delta)) = h(\mathbf{y}(n)) \; \forall \delta \in \mathbb{N}_+ \tag{3.50}$$

There are two main approaches to compute these inequality measures: methods based on the Lorenz curve [95], and derivations based on normalization of the Rényi heterogeneity [92].

The most classically important characterization of inequality is based on the Lorenz curve [95], which represents the percentage of total abundance in a system belonging to the top $x\%$ of categories. For example, when examining the distribution of abundance across presentations of MDD [54, 55], the Lorenz curve (shown in Figure 3.2B) shows that 50% of all observed samples were attributable to only 7.1% of MDD symptom combinations in the pooled sample. Several summary indices can be computed from the Lorenz curve, such as the Gini coefficient (which we also discussed above) [24] or the *Pietra index* (also known as the *Robin Hood*, *Hoover*, or *Schutz* coefficient) [96]. Some other Lorenzian inequality indices are well reviewed elsewhere [94, 97].

The distribution and utilization of psychiatric resources has been quantified with Lorenz curves [98–100], although other questions have also been addressed [101–104]. However, (direct) Lorenzian inequality analysis is univariate, which limits applicability to modern translational psychiatric research.

An alternative to the Lorenzian approach is to define a measure of "evenness" (conceptually the opposite of inequality) by expressing Rényi heterogeneity relative to its theoretical maximum (the observed richness):

$$\widetilde{\Pi}_q\left(p\right) = \frac{\Pi_q\left(p\right)}{n_c} \tag{3.51}$$

This is based on the more general concept of a *diversity profile* discussed in detail elsewhere [59]. The range of Equation 12 is the (0, 1] interval, and it can be used to derive many well-known inequality indices such as *Heip's index* [105], *Pielou's J* [106], and economists' lauded *Generalized Entropy Index* (GEI) [93, 94], which is itself generalizes several important indices [86, 107, 108]. This approach has not clearly been used for inequality measurement in psychiatry.

Limitations of Categorical Heterogeneity Measures

The main problem with categorical heterogeneity measures are the assumptions of categorical data. First, the categories to which one's data belong must be (A) known a priori and (B) scientifically valid. In some cases this will be more problematic than in others. For example, defining species as categories (as ecologists do) is likely of greater validity than defining the categories as DSM-5 diagnoses. There is simply more certainty about the validity of the former than of the latter.

Second, one must assume that all members of the same category are identical in every way, and that all between-category differences are equal. These assumptions about the within- and between-category dissimilarity are surely violated in most psychiatric research applications. For example, the analyses of Zimmerman et al. [54] and Park et al. [55] (and our reanalysis thereof) did not account for the fact that different presentations will share symptoms in common. Clearly, the proper distance metric for these data is not the discrete metric (Equation 1), and they are thus not categorical data.

Despite these limitations, categorical heterogeneity measures—and particularly the Rényi heterogeneity family—have advantages related to interpretation. The "size" of a system's event space is an intuitive and principled measure of deviation from perfect conformity. In our MDD example, we spoke in terms of the easily understandable units of "number of symptom combinations" rather than of bits or probabilities. Rényi heterogeneity also respects the replication principle and can be decomposed into within- and between-group components. We now seek a measure that retains these useful properties of Rényi heterogeneity, while not being restricted to a categorical system.

3.3 Non-categorical Heterogeneity Indices

Non-categorical systems have elements that can vary in the degree to which they are similar to each other. These indices are subdivided into two broad groups: those that split the observations into categories defined *a priori*, and those that either (A) do not assume such a stratification at all or (B) attempt to learn it from the data.

3.3.1 Methods Requiring *a priori* Stratification

These methods first split observations from a system into one of n_c pre-defined categories (e.g. diagnoses or species). However, (A) the within-category distance can exceed 0 (e.g. acknowledging that "tall" people still vary in height), and (B) the distance between pairs of categories can be asymmetrical (e.g. lobsters are "further" from elephants than they are from crabs).

The experimenter must choose a relevant distance measure, which will significantly impact the heterogeneity estimates. Returning to our re-analysis of the MDD symptom combination data [54, 55], we clarify that each of the 227 unique symptom combinations is a distinct category in the event space $\mathcal{X}$. However, we now specify the dissimilarity between symptom combinations x_i and x_j using the *Jaccard distance* [109]:

$$D_{ij} = 1 - \frac{\#\text{Symptoms occurring in both } x_i \text{ and } x_j}{\#\text{Symptoms occurring in either } x_i \text{ or } x_j} \tag{3.52}$$

which takes values between 0 (complete overlap of symptoms) and 1 (no symptoms in common). This results in a 227×227 matrix, $\mathbf{D}$, of distances between symptom combinations.

To quantify heterogeneity, D must be summarized into a single non-negative value. The most common approaches are related to Rao's Quadratic Entropy (RQE) [110],

$$Q\left(\mathbf{D}, \mathbf{p}\right) = \sum_{i=1}^{n_c} \sum_{j=1}^{n_c} D_{ij} p_i p_j \tag{3.53}$$

which is the average pairwise distance between categories in the system. For our present example, we have an RQE=0.35 for the Zimmerman et al. [54] data, RQE=0.38 for the Park et al. [55] data, and RQE =0.37 in the pooled sample. Note that the RQE of one of the subsets (Park et al. [55]) is greater than the pooled sample's heterogeneity, which is problematic, since pooling non-identical systems should monotonically *increase* the overall heterogeneity. By using a different distance metric (the Hamming distance), this problem disappears; we obtain RQE estimates of 2.89 (Zimmerman et al. [54]), 3.04 (Park et al. [55]), and 3.05 (pooled). How are we to compare these estimates which are on ostensibly different scales? Moreover, is one set of estimates "more correct" than the other?

To solve this problem, researchers have sought to develop RQE-based measures with units of numbers equivalent, since they do not appeal to the units of a given distance metric [90, 111–114]. An added bonus is that numbers equivalent measures will obey the replication principle [90, 113]. Unfortunately, current RQE-based numbers equivalent measures have some idiosyncratic limitations that virtually obviate their psychiatric research applicability. A more detailed exposition of these issues is given in Chapter 4, but for the present instance, we note the *functional Hill numbers* [111] become insensitive to distance between categories when they are equally abundant. Not surprisingly, we are unaware of any studies in the psychiatric literature that employ non-categorical heterogeneity indices with *a priori* stratification.

Two additional problems stand out with RQE-based and related approaches. First, these heterogeneity indices are entirely dependent on the imposed stratification. In cases where the imposed strata are unreliable or invalid (such as the case in which strata are DSM-5 psychiatric diagnoses), these aforementioned non-categorical heterogeneity indices will unlikely be useful.

There is also a problem with defining the distance metric *a priori*. The distance metric chosen determines which paths between points A and B in the data space are “allowed.” An appropriate distance metric should allow only realistic paths between these points (Figure 3.3). For example, the *straight-line* distance between Toronto and Tokyo is irrelevant to travelers, since that path cannot be traversed. In that vein, many real-world data are thought to be embedded on lower dimensional manifolds in the data space [115]. In such cases, the distance between points should be measured on paths along that manifold, which may be curved. Since the manifolds of support will vary between datasets, it is unlikely that predefined distance metrics (such as a global Euclidean distance) will accurately describe the dispersion of one’s data. To our knowledge, this problem remains unaddressed in the heterogeneity measurement literature.

3.3.2 Methods That do not Require *a priori* Stratification

There are three main approaches to quantify heterogeneity when no compelling *a priori* stratification exists: (A) treating heterogeneity as the “volume” of a space that completely encloses one’s data points, (B) clustering-based methods, and (C) dendrogram-based methods.

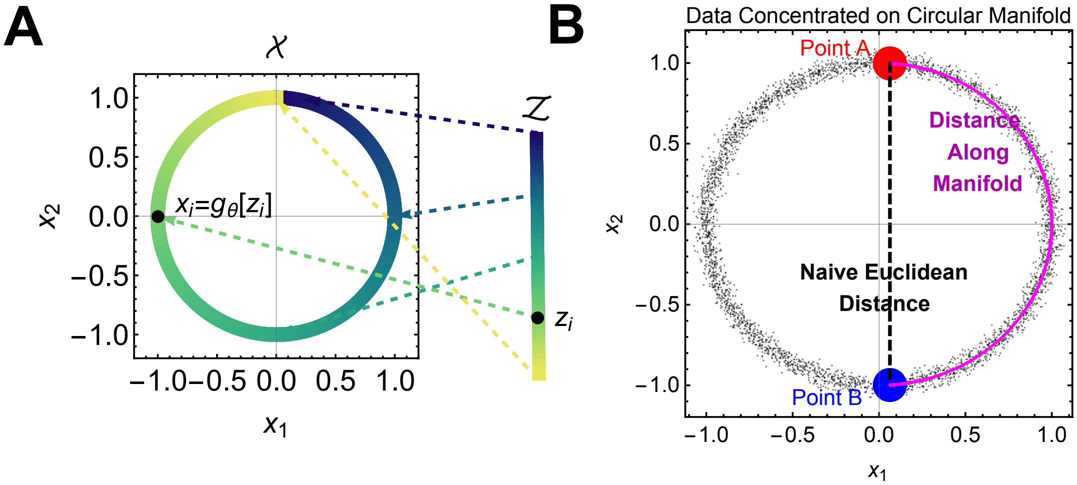


Figure 3.3: Demonstration of how data in an observable space $\mathcal{X}$ can be concentrated along a manifold (here just a curve). **Panel A** shows how the curve is simply an image of a latent space $\mathcal{Z}$ projected through a generator function $x_i = g_\theta(z_i)$. **Panel B** demonstrates noisy data along the circular curve illustrated in Panel A. Measurement of the Euclidean (straight-line) distance between Points A and B implies traversal across a region of $\mathcal{X}$ in which no data lie. The correct approach is instead to measure distance with respect to the data's manifold of support.

Heterogeneity as a Convex Hull Volume

Roughly speaking, the space enclosed by the smallest perimeter around all pairwise paths in one's data is a *convex hull*. The volume of this space is sometimes used as a heterogeneity index [116, 117], but if there are outliers or the data are not distributed uniformly within the convex hull, heterogeneity will be overestimated (Figures 3.4A-C). We know of no psychiatric study using convex hull volume to quantify heterogeneity.

Methods based on clustering and dendrogram construction

Psychiatric studies often characterize a heterogeneity as the number of latent categories in some data. For example, cluster analytic studies of MDD have reported discovery of between 1 and 5 strata (depending on the data), although these groups are qualitatively inconsistent [57]. Similarly, cluster analyses in schizophrenia [62, 118–127], attention deficit-hyperactivity disorder (ADHD) [128–133], autism [134–139] and other conditions [63, 140–142] have returned proposals for various stratifications, with heterogeneity implicitly "measured" by cluster counting. Further review of the psychiatric cluster analysis literature can be found elsewhere [56, 57]. However, we note that measurement of heterogeneity in cluster analyses by mere cluster counting will prove exquisitely sensitive to the method by which one determines the optimal number of clusters, itself a difficult problem that has been discussed extensively elsewhere [56].

There are three other prominent limitations of cluster counting. First, the cluster count does not capture inequalities in the distribution of subjects across clusters. Cluster counting is therefore a variation on observed richness. Second, the clusters themselves are consequently assumed to be internally homogeneous and maximally dissimilar from the other clusters (effectively re-instantiating the case-control paradigm). Finally, simply identifying the statistically optimal number of clusters in a dataset does not guarantee that those clusters are optimal in terms of their biological or scientific validity. To address this, many reports have validated their inferred clusters using external data [143–146]. Notwithstanding, there remain several open areas for improvement in measuring heterogeneity using cluster analysis, particularly with respect to (A) evaluation of whether a clustering approach (i.e. mapping some data onto a categorical space) is appropriate for some data in the first place, and (B) accounting for uncertainty in the number of clusters, which overlaps with our above discussion of partition counting methods.

An alternative approach involves measuring heterogeneity by first performing agglomerative clustering, and then computing the sum of all branch lengths in the resulting hierarchical tree (also known as a “dendrogram”) [147, 148]. This requires computing pairwise distances between observations in one’s dataset. Figure 3.4 demonstrates this approach on some synthetic data. It may be possible to compute an effective number from dendrogram-based analyses [149]. Whereas the convex hull approach defines heterogeneity by the most extreme points in a dataset, the dendrogram-based methods are sensitive to the *density* of sample space coverage. Unfortunately, this will create a problem if there are truly multiple groups in one’s data, since the dendrogram-based heterogeneity index *increases* if the groups’ feature distributions become more similar (Figure 3.4E). To our knowledge, there are no applications of dendrogram-based heterogeneity measures in the psychiatric literature, although gene co-expression studies (such as those employing *Weighted Gene Correlation Network Analysis*) are ostensibly immediate targets for these indices [150–152].

Normative modeling is a recent notable alternative for characterizing heterogeneity in clinical cohorts [66]. Briefly, this approach learns a model of normal variation, much like

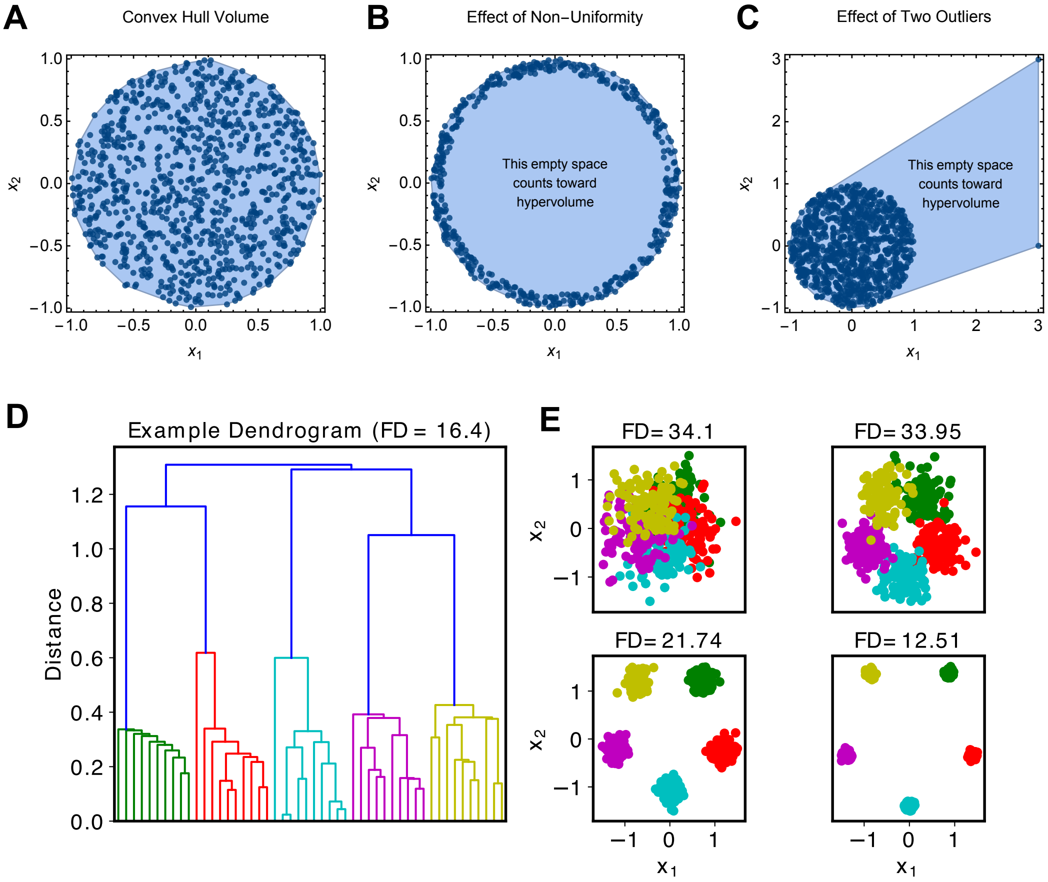


Figure 3.4: Illustration of convex hull and dendrogram-based heterogeneity indices for non-categorical systems. **Panel A** illustrates the basic concept of a convex hull on synthetic 2-dimensional data. The volume of the hull is taken as an index of heterogeneity. **Panel B** shows one problem with the convex hull method, which occurs when data lie along a lower dimensional surface (here just a curve). In this example, the data are all concentrated along the outer border of the hull, leaving the core unoccupied. However, the convex hull volume index will nonetheless count the empty space toward the heterogeneity value. **Panel C** illustrates the effect of outliers on convex hull volume. Since a convex hull is found by creating a "shell" around one's data, outlying points will expand this shell in ways that leave much of the convex hull empty (though still counting toward the heterogeneity value). **Panel D** shows the dendrogram computed using agglomerative clustering for a simple mixture of five 2-dimensional (2D) Gaussians. The Functional Diversity (FD) measure, shown in the title, is the sum of all branch lengths in this tree. **Panel E** shows a simple simulation with five 2D Gaussians (standardized to lie within the bounds [-1.5, 1.5] in both axes) that were progressively separated further. One can appreciate that the FD measure decreases as the distributions become more distinct. This is the opposite effect demonstrated by the convex hull volume, insofar as FD increases as the space becomes more densely populated with data points.

growth chart models in pediatrics, then evaluates the degree and uncertainty with which individual subjects deviate from this normative range. The assumption is that pathological states tend to deviate more extremely. Applications include (predominantly neuroimaging) studies of autism [153, 154], ADHD [66, 155, 156], schizophrenia and psychosis [157, 158], bipolar disorder [157], and neurocognitive disorders [159, 160]. Although normative modeling offers an important and novel non-categorical system within which to frame the heterogeneity of psychiatric disorders, no study employing this method has offered a *measurement* of heterogeneity. Thus, it would be of great interest to develop numbers equivalent measures applicable within the normative modeling framework.

3.3.3 A Note on Meta-Analytic Heterogeneity

Standard mixed-effects meta-analysis employs a parametric index of heterogeneity on non-categorical spaces [27]. A full discussion of this (likely familiar) topic is beyond our present scope, but an illustration of the mixed-effects formulation is provided in Figure 3.5. However, to motivate meta-analytic Rényi heterogeneity, it is important to demonstrate that the current meta-analytic heterogeneity—which is the variance of the Gaussian distribution over study effects—is an index that fails to satisfy the replication principle.

To illustrate failure of the variance to satisfy Axiom 7 (replication), we consider the example of pooling n uniform distributions on the respective intervals $\{[\gamma_0, \gamma_1), [\gamma_1, \gamma_2), \dots, [\gamma_{n-1}, \gamma_n]\}$. Since the replication principle requires that the pooled systems have equal heterogeneity, we specify that $\gamma_i - \gamma_{i-1} = \gamma_{j-1} - \gamma_j \; \forall (i,j) \in \{1, 2, \dots, n\}$ and $\gamma_{i-1} < \gamma_i \; \forall i$.

The probability density function (PDF) for the i^{th} uniform distribution is defined on the half open interval $[\gamma_{i-1}, \gamma_i)$ as follows:

$$U(x|\gamma_{i-1}, \gamma_i) = \begin{cases} \frac{1}{\gamma_i - \gamma_{i-1}} & \gamma_{i-1} \leq x < \gamma_i \\ 0 & \text{Otherwise} \end{cases} \tag{3.54}$$

The variance for the uniform distribution on the half-open interval is the same as that of the uniform distribution on the closed interval

$$\text{Var}(\gamma_{i-1}, \gamma_i) = \frac{1}{12} \left(\gamma_i - \gamma_{i-1}\right)^2, \tag{3.55}$$

and the variance of n pooled uniform distributions is thus

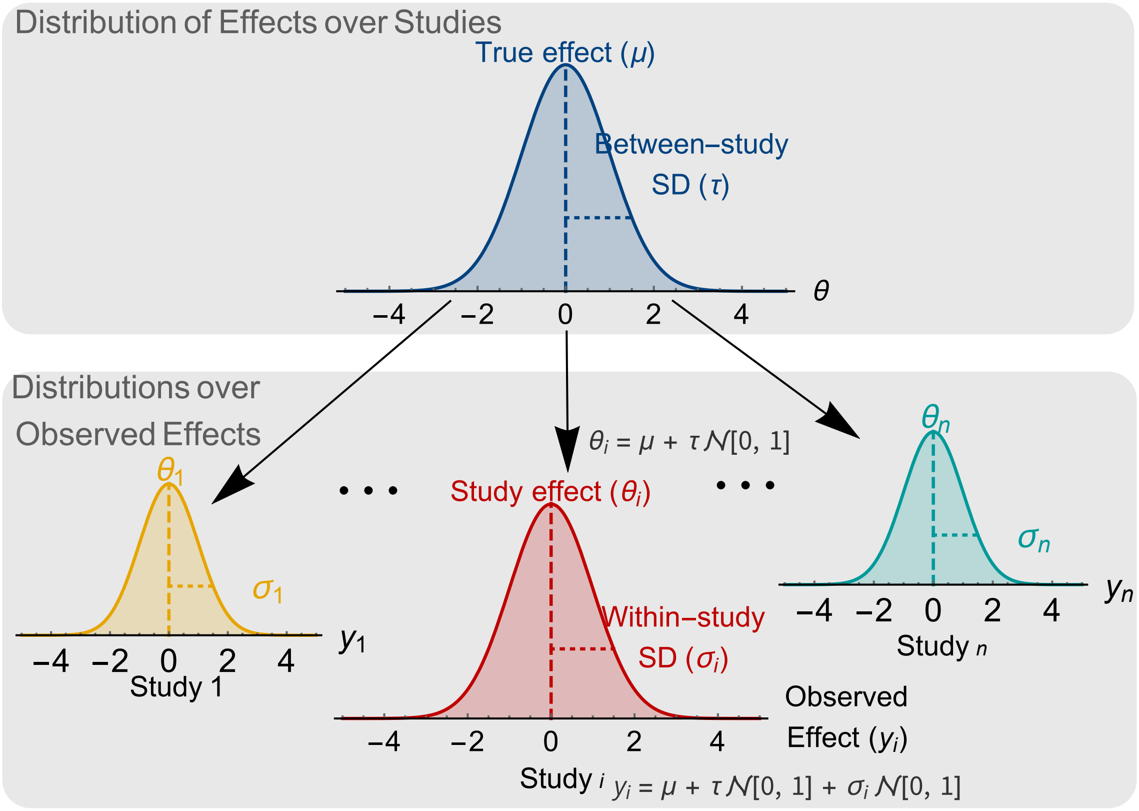


Figure 3.5: Illustration of the mixed-effects model for meta-analysis. The observed effects for n individual studies $(y_i)_{i=1}^{n}$ are each distributed according to study-level Gaussians with effects $\boldsymbol{\theta} = (\theta_i)_{i=1}^{n}$ and study-level standard deviations $\boldsymbol{\sigma} = (\sigma_i^2)_{i=1}^{n}$. The study-level effects $\boldsymbol{\theta}$ are modeled as distributed according to an isotropic Gaussian distribution over studies, $\forall i \in \{1, 2, \ldots, n\}$ $\theta_i \sim \mathcal{N}(\theta_i|\mu, \tau^2)$, where μ is the true (summary) effect, and τ^2 is the heterogeneity of study effects.

$$\mathrm{Var}(\gamma_0, \gamma_n) = \frac{1}{12} (\gamma_n - \gamma_0)^2 . \tag{3.56}$$

Recalling that $|\gamma_n - \gamma_0| = n|\gamma_i - \gamma_{i-1}|$, one can use Equation 3.56 to easily show that $\mathrm{Var}(\gamma_0, \gamma_n) \neq n\mathrm{Var}(\gamma_{i-1}, \gamma_i)$ and thus that variance does not satisfy the replication principle.

Conversely, if we compute heterogeneity using the continuous Rényi heterogeneity,

$$\Pi_q(p) = \left(\int_a^b p^q(x) \, \mathrm{d}x \right)^{\frac{1}{1-q}} , \tag{3.57}$$

then we will satisfy Axiom 7. Expressing Equation 3.57 for the uniform distribution yields

$$\Pi_q(\gamma_{i-1}, \gamma_i) = \gamma_i - \gamma_{i-1}, \tag{3.58}$$

which is expected since Rényi heterogeneity measures the size of the base of support. It can be easily shown to satisfy the replication principle:

$$\Pi_q(\gamma_0, \gamma_n) = n\Pi_q(\gamma_{i-1}, \gamma_i) \tag{3.59}$$

$$\left(\int_{\gamma_0}^{\gamma_n} (\gamma_n - \gamma_0)^q \, \mathrm{d}x \right)^{\frac{1}{1-q}} = n \left(\int_{\gamma_{i-1}}^{\gamma_i} (\gamma_i - \gamma_{i-1})^q \, \mathrm{d}x \right)^{\frac{1}{1-q}} \tag{3.60}$$

$$\gamma_n - \gamma_0 = n(\gamma_i - \gamma_{i-1}). \tag{3.61}$$

One can also show that meta-analytic heterogeneity, if computed under the Rényi formalism—can be decomposed into within and between-group components and thus expressed in the units of numbers equivalent ("the effective range of distinct study effects"). This decomposition is described further in Chapter 4 in the context of a deep neural network with Gaussian latent variables.

3.3.4 A Note on Heterogeneity Indices for Time-Series and Dynamical Systems

We briefly discuss measurement of heterogeneity in time-series data by indices often known as "complexity" measures. Psychiatric studies have employed geometric indices (such as the *Largest Lyapunov Exponent* and *recurrence plot analysis*) [161, 162], entropic indices (such as *Kolmogorov-Sinai* or *metric entropy* [163], *approximate entropy* [164], *sample* and *multiscale entropies* [13, 15, 165], and *Lempel-Ziv complexity* [166, 167]), and various *fractal dimension* indices largely to electrophysiological data, although some studies have evaluated functional neuroimaging [168] and other time-series [13, 15, 169]. Numerous clinical and technical reviews of these indices exist [65, 164, 170–175], so we merely note that numbers equivalent can also be of use in this domain. For example, the Shannon entropy of a time series' normalized power spectrum, also known as *spectral entropy* [171], can be easily converted to the "effective number of typical frequencies" using Equation 10; reporting such a measure in terms of the effective number of frequency bands makes interpretation and criticism more clear. If one reports that a time series of mood recordings contains an effective number of three frequency bands, we may more readily appraise whether such information is useful, and how so. With such clear units, one may decide that indices expressing the "effective number of trajectories" or "effective number of 'mood states"' might be more desirable.

Many conditions have been studied under this paradigm using various modalities [173, 174, 176]. For instance, our group has investigated the temporal dynamics of mood in patients with bipolar disorder. The overall complexity of mood fluctuations is ostensibly reduced among probands and their unaffected relatives [13, 15], with increases observed within 60 days of a mood episode [169]. Unfortunately, on the whole it can be difficult to interpret time-series complexity between studies, since the large number of indices (each with their own units), experimental conditions, data modalities, and disorders can interact to yield various—dare we say *heterogeneous*—conclusions.

3.3.5 Limitations of Non-categorical Heterogeneity Indices

Non-categorical heterogeneity indices are predominantly based on RQE [110]. Unfortunately, the requirement of selecting a distance measure *a priori* introduces problems comparing RQE across datasets with different distance metrics. Moreover, for real-world datasets, standard methods of measuring distance will likely fail to respect data's true underlying geometry. This problem will be shared by dendrogram-based methods and clustering-based approaches that demand pre specification of a distance measure.

The units of RQE-based heterogeneity indices are also not clearly appropriate for thinking about heterogeneity, although one may correctly argue that heterogeneous systems have larger overall amounts of pairwise distance between their elements [177]. Plainly, these indices violate the replication principle which leads to unintuitive scaling behaviours

[112, 113]. Although the numbers equivalent transformations of RQE have addressed this problem to some extent, they have further limitations that virtually preclude their application to research problems in the psychiatric domain. First, they continue to require prespecified categories on the data as well as prespecified distance measures. Second, they have idiosyncratic limitations—such as insensitivity to distance under equally abundant categories [111]—that would be problematic in psychiatric use cases.

We showed above that meta-analytic heterogeneity is at present quantified by variance, which fails to satisfy the replication principle.

Time series complexity measures, too, can be difficult to interpret and synthesize. In many cases, time-series complexity measures based on numbers equivalent could simplify interpretation. In the case of longitudinal self-ratings of mood, for example, reporting heterogeneity as "the effective number of mood states" could meaningfully improve the broader clinical interpretability of such results. However, no such study has heretofore reported time-series heterogeneity in numbers equivalent, and so its evaluation in that context remains an interesting future direction.

3.4 Discussion and Conclusions

This paper defined heterogeneity as the degree to which a system diverges from perfect conformity, and measures it by the effective size of a system's event space. We have highlighted assumptions, strengths, limitations, and psychiatric use-cases for various measures. A large number of indices have been discovered (and rediscovered) independently for different data types and fields of study. Although we believe each index is valuable in describing unique properties of heterogeneous systems, their large variety of units and differences in mathematical properties impede both (A) their synthesis across studies and (B) their broader interpretability. However, we demonstrated that *numbers equivalent* measures of heterogeneity—known in different fields as the Rényi heterogeneity, Hill numbers, or Hannah-Kay indices—are cross-cutting measures that can potentially express the heterogeneity of a system as the size of an equally heterogeneous uniform event space. These measures satisfy most heterogeneity axioms (including the replication principle) and are standard measures of ecological biodiversity yet remain relatively absent from the psychiatric literature. In this section, we re-highlight some of the roadblocks to their psychiatric implementation and future directions of research.

One obstacle for implementation of numbers equivalent heterogeneity measures in psychiatry is conceptual in nature. Heterogeneous systems have many correlated properties that, in the absence of precise definition, could easily be mistaken for heterogeneity itself: they have more sampling uncertainty and information, lower probability of sampling identical pairs, lower modal probabilities, higher variance, less inequality in their probability distributions, and larger event spaces. If one cares simply about "more vs. less" heterogeneity, then any of these properties will be suitable indices. However, if one is interested in *"how much* more/less" heterogeneity exists (such as when comparing groups), then only numbers equivalent measures will show appropriate behaviour under pooling or decomposition. The utility of such measures, including their easily understandable units, must be appreciated through real-world applications.

The chief technical obstacle for adopting numbers equivalent measures in psychiatric research is their limitations when applied to non-categorical data. Existing non-categorical numbers equivalent measures satisfy the replication principle [90], but they still require imposition of *a priori* stratification on the data, and assumption of a distance metric (see also their further limitations in Chapter 4). Both limitations preclude adoption in translational psychiatric research. First, if psychiatric science had reliable and valid strata to impose on some data, then we might not have such concern with heterogeneity in the first place. Second, the types of high-dimensional data often used in modern psychiatric research might lie on latent spaces whose geometries do not admit application of pre-defined distance functions [115]. In such systems, existing non-categorical numbers equivalent measures may fail to accurately measure heterogeneity.

Numbers equivalent heterogeneity measures can be relevant for modern translational psychiatric research, but existing indices (particularly for non-categorical systems) must be adapted to suit the nature of our data and questions. We must do away with the requirement for *a priori* data stratification and distance function specification. It will also be interesting to study if, how, and under what circumstances existing measures of meta-analytic heterogeneity and time-series complexity should be expressed in numbers equivalent. Finally, it would be of interest to investigate whether numbers equivalent heterogeneity measures are indeed more broadly understandable or easier to synthesize across studies.

Chapter 4

Representational Rényi Heterogeneity[1]

Abstract. A discrete system's heterogeneity is measured by the Rényi heterogeneity family of indices (also known as Hill numbers or Hannah–Kay indices), whose units are the numbers equivalent. Unfortunately, numbers equivalent heterogeneity measures for non-categorical data require a priori (A) categorical partitioning and (B) pairwise distance measurement on the observable data space, thereby precluding application to problems with ill-defined categories or where semantically relevant features must be learned as abstractions from some data. We thus introduce representational Rényi heterogeneity (RRH), which transforms an observable domain onto a latent space upon which the Rényi heterogeneity is both tractable and semantically relevant. This method requires neither a priori binning nor definition of a distance function on the observable space. We show that RRH can generalize existing biodiversity and economic equality indices. Compared with existing indices on a beta-mixture distribution, we show that RRH responds more appropriately to changes in mixture component separation and weighting. Finally, we demonstrate the measurement of RRH in a set of natural images, with respect to abstract representations learned by a deep neural network. The RRH approach will further enable heterogeneity measurement in disciplines whose data do not easily conform to the assumptions of existing indices.

4.1 Introduction

Measuring heterogeneity is of broad scientific importance, such as in studies of biodiversity (ecology and microbiology) [22, 50], resource concentration (economics) [94], and consistency of clinical trial results (biostatistics) [28], to name a few. In most of these cases, one measures the heterogeneity of a discrete system equipped with a probability mass function.

Discrete systems assume that all observations of a given state are identical (zero distance), and that all pairwise distances between states are permutation invariant. This assumption is violated when relative distances between states are important. For example, an ecosystem is not biodiverse if all species serve the same functional role [2]. Although species are categorical labels, their pairwise differences in terms of ecological functions differ and thus violate the discrete space assumptions. Mathematical ecologists have thus developed

[1] Nunes A, Alda M, Bardouille T, and Trappenberg T. Representational Rényi Heterogeneity. *Entropy* 2020, 22, 417.

heterogeneity measures for non-categorical systems, which they generally call "functional diversity indices" [45, 90, 111, 113, 114, 147]. These indices typically require a priori discretization and specification of a distance function on the observable space.

The requirement for defining the state space a priori is problematic when the states are incompletely observable: that is, when they may be noisy, unreliable, or invalid. For example, consider sampling a patient from a population of individuals with psychiatric disorders and assigning a categorical state label corresponding to his or her diagnosis according to standard definitions [69]. Given that psychiatric conditions are not defined by objective biomarkers, the individual's diagnostic state will be uncertain. Indeed, many of these conditions are inconsistently diagnosed across raters [178], and there is no guarantee that they correspond to valid biological processes. Alternatively, it is possible that variation within some categorical diagnostic groups is simply related to diagnostic "noise," or nuisance variation, but that variation within other diagnostic groups constitutes the presence of sub-strata. Appropriate measurement of heterogeneity in such disciplines requires freedom from the discretization requirement of existing non-categorical heterogeneity indices.

Pre-specified distance functions may fail to capture semantically relevant geometry in the raw feature space. For example, the Euclidean distance between Edmonton and Johannesburg is relatively useless since the straight-line path cannot be traversed. Rather, the appropriate distances between points must account for the data's underlying manifold of support. Representation learning addresses this problem by learning a latent embedding upon which distances are of greater semantic relevance [115]. Indeed, we have observed superior clustering of natural images embedded on Riemannian manifolds [48] (but also see Shao et al. [179]), and preservation of semantic hierarchies when linguistic data are embedded on a hyperbolic space [180].

Therefore, we seek non-categorical heterogeneity indices without requisite a priori definition of categorical state labels or a distance function. The present study proposes a solution to these problems based on the measurement of heterogeneity on learned latent representations, rather than on raw observable data. Our method, representational Rényi heterogeneity (RRH), involves learning a mapping from the space of observable data to a latent space upon which an existing measure (the Rényi heterogeneity [91], also known as the Hill numbers [17] or Hannah–Kay indices [18]) is meaningful and tractable.

The paper is structured as follows. Section 4.2 introduces the original categorical formulation of Rényi heterogeneity and various approaches by which it has been generalized for application on non-categorical spaces [111, 112, 114]. Limitations of these indices are highlighted, thereby motivating Section 4.3, which introduces the theory of Representational Rényi Heterogeneity (RRH), which generalizes the process for computing many indices of biodiversity and economic equality. Section 4.4 provides an illustration of how RRH may be measured in various analytical contexts. We provide an exact comparison of RRH to existing non-categorical heterogeneity indices under a tractable mixture of beta distributions. To highlight the generalizability of our approach to complex latent variable models, we also provide an evaluation of RRH applied to the latent representations of a handwritten image dataset [181] learned by a variational autoencoder [182]. Finally, in Section 4.5 we provide a summary of our findings and discuss avenues for future work.

4.2 Existing Heterogeneity Indices

4.2.1 Rényi Heterogeneity in Categorical Systems

There are many approaches to derive Rényi heterogeneity [17, 18, 91]. Here, we loosely follow the presentation of Eliazar & Sokolov [97] by using the metaphor of repeated sampling from a discrete system X with event space $\mathcal{X} = \{1, 2, \dots, n\}$ and probability distribution $\mathbf{p} = (p_i)_{i=1,2,\dots,n}$. The probability that $q \in \mathbb{N}_{>1}$ independent and identically distributed (i.i.d.) realizations of X, sampled with replacement, will be identical is

$$\mathbb{P}_X \left[x_1 = x_2 = \dots = x_q \right] = \sum_{i=1}^{n} p_i^q. \tag{4.1}$$

Now let X_* be an idealized reference system with a uniform probability distribution over n_* categorical states, $\mathbf{p}_* = (n_*^{-1})_{i=1,2,\dots,n_*}$, and let $(x_{*1}, x_{*2}, \dots, x_{*q})$ be a sample of q i.i.d. realizations of X_* such that

$$\mathbb{P}_X \left[x_1 = x_2 = \dots = x_q \right] = \mathbb{P}_{X_*} \left[x_{*1} = x_{*2} = \dots = x_{*q} \right] = \sum_{i=1}^{n_*} n_*^{-q}. \tag{4.2}$$

We call X_* an "idealized" categorical system because its probability distribution is uniform, and it is a "reference" system for X in that the probability of drawing homogeneous samples of q observations from both systems is identical. Substituting Equation 4.2 into Equation 4.1 and solving for n_* yields the Rényi heterogeneity of order q,

Table 4.1: Relationships between Rényi heterogeneity and various diversity or inequality indices for a system X with event space $\mathcal{X} = \{1, 2, \dots, n\}$ and probability distribution $\mathbf{p} = (p_i)_{i=1,2,\dots,n}$. The function $\mathbb{1}[\cdot]$ is an indicator function that evaluates to 1 if its argument is true or to 0 otherwise.

Index	**Expression**
Observed richness [70]	$\Pi_0(\mathbf{p}) = \sum_{i=1}^{n} \mathbb{1}[p_i > 0]$
Perplexity [43]	$\Pi_1(\mathbf{p}) = \exp\{-\sum_{i=1}^{n} p_i \log p_i\}$
Inverse Simpson concentration [22]	$\Pi_2(\mathbf{p}) = \left(\sum_{i=1}^{n} p_i^2\right)^{-1}$
Berger-Parker Diversity Index [42, 183]	$\Pi_\infty(\mathbf{p}) = (\max_i p_i)^{-1}$
Rényi entropy [91]	$R_q(\mathbf{p}) = \log \Pi_q(\mathbf{p})$
Shannon entropy [25]	$H(\mathbf{p}) = \log \Pi_1(\mathbf{p})$
Tsallis entropy [26]	$T_q(\mathbf{p}) = \frac{1}{q-1}\left(1 - \Pi_q(\mathbf{p})^{1-q}\right)$
Simpson concentration [87]	$\mathrm{Simpson}(\mathbf{p}) = (\Pi_2(\mathbf{p}))^{-1}$
Gini-Simpson index [24]	$\mathrm{GSI}(\mathbf{p}) = 1 - \mathrm{Simpson}(\mathbf{p})$
Generalized entropy index [93, 94]	$\mathrm{GEI}(\mathbf{p}) = \frac{1}{q(q-1)}\left[\left(\frac{1}{n}\Pi_q(\mathbf{p})\right)^{1-q} - 1\right]$

$$\Pi_q(\mathbf{p}) = \left(\sum_{i=1}^{n} p_i^q\right)^{\frac{1}{1-q}} = n_*, \tag{4.3}$$

whose units are the numbers equivalent of system X [19, 20, 22, 59], insofar as n_* is the number of states in an "equivalent" (idealized reference) system X_*. Thus far, we have restricted the parameter q to take integer values greater than 1 solely to facilitate this intuitive derivation in a concise fashion. However, the elasticity parameter q in Equation 4.3 can be any real number (but $q \neq 1$), although in the context of heterogeneity measurement only $q \geq 0$ are used [22, 97]. Despite Equation 4.3 being udefined at $q = 1$ directly, L'Hôpital's rule can be used to show that the limit $q \to 1$ exists, wherein it corresponds to the exponential of Shannon's entropy [25, 59], known as perplexity [43].

Equation 4.3 is the exponential of Rényi's entropy [91], and is alternatively known as the Hill numbers in ecology [17, 22], Hannah–Kay indices in economics [18], and generalized inverse participation ratio in physics [97]. Interestingly, it generalizes or can be transformed into several heterogeneity indices that are commonly employed across scientific disciplines (Table 4.1).

Properties of the Rényi Heterogeneity

Equation 4.3 satisfies several properties that render it a preferable measure of heterogeneity. These have been detailed elsewhere [18, 22, 23, 42, 59, 97], but we focus on three properties that are of particular relevance for the remainder of this paper.

First, Π_q satisfies the principle of transfers [85, 86] which states that any equality-increasing transfer of probability between states must increase the heterogeneity. The maximal value of Π_q is attained if and only if $p_i = p_j$ for all $(i, j) \in \{1, 2, \dots, n\}$. This property follows from Schur-concavity of Equation 4.3 [18].

Second, Π_q satisfies the replication principle [21–23], which is equivalent to stating that Equation 4.3 scales linearly with the number of equally probable states in an idealized categorical system [97]. More formally, consider a set of systems $X_1, X_2, \dots, X_N$ with probability distributions $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_N$ over respective discrete event spaces $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_N$. These systems are also assumed to satisfy the following properties:

1. Event spaces are disjoint: $\mathcal{X}_i \cap \mathcal{X}_j = \emptyset$ for all $(i, j) \in \{1, 2, \dots, N\}$ where $i \neq j$
2. All systems have equal heterogeneity: $\Pi_q(\mathbf{p}_1) = \Pi_q(\mathbf{p}_2) = \cdots = \Pi_q(\mathbf{p}_i) = \cdots = \Pi_q(\mathbf{p}_N)$

The replication principle states that if we combine $X_1, X_2, \dots, X_N$ into a pooled system X with probability distribution $\bar{\mathbf{p}}$, then

$$\Pi_q(\bar{\mathbf{p}}) = N \Pi_q(\mathbf{p}_i) \tag{4.4}$$

must hold (see Appendix A.1 for proof that Rényi heterogeneity satisfies the replication principle).

The replication principle suggests that Equation 4.3 satisfies a property known as decomposability, in that the heterogeneity of a pooled system can be decomposed into that arising from variation within and between component subsystems. However, we require that this property be satisfied when either (A) subsystems' event spaces are overlapping, or (B) subsystems do not have equal heterogeneity. The decomposability property will be particularly important for Section 4.3, and so we detail it further in Section 4.2.1.

Decomposition of Categorical Rényi Heterogeneity

Consider a system X defined by pooling subsystems $X_1, X_2, \dots, X_N$ with potentially overlapping event spaces $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_N$, respectively. The event space of the pooled system is defined as

$$\mathcal{X} = \cup_{i=1}^{N} \mathcal{X}_i = \{1, 2, \dots, n\} . \tag{4.5}$$

Furthermore, we define the matrix $\mathbf{P} = (p_{ij})_{i=1,2,\dots,N}^{j=1,2,\dots,n}$ whose i^{th} row is the probability of system X_i being observed in each state $j \in \{1, 2, \dots, n\}$.

It may be the case that some subsystems comprise a larger proportion of X than others. For instance, if the probability distribution for subsystem X_i was estimated based on a larger sample size than that of X_j, one may want to weight the contribution of X_i higher. Thus, we define a column vector of weights $\mathbf{w} = (w_i)_{i=1,2,\dots,N}$ over the N subsystems such that $\sum_{i=1}^{N} w_i = 1$ and $w_i \geq 0$ for all i. The probability distribution over states in the pooled system X may thus be computed as $\bar{\mathbf{p}} = \sum_{i=1}^{N} w_i \mathbf{p}_i$, from which the definition of pooled heterogeneity follows:

$$\Pi_q^{\mathrm{P}}\left(\mathbf{P}, \mathbf{w}\right) = \left[\sum_{j=1}^{n} \left(\sum_{i=1}^{N} w_i p_{ij}\right)^q\right]^{\frac{1}{1-q}} . \tag{4.6}$$

One can interpret $\Pi_q^{\mathrm{P}}\left(\mathbf{P}, \mathbf{w}\right)$ as the effective number of states in the pooled categorical system X.

Jost [59] showed that the within-group heterogeneity, which is the effective number of unique states arising from individual component systems, can be defined as

$$\Pi_q^{\mathrm{W}}\left(\mathbf{P}, \mathbf{w}\right) = \left[\frac{\sum_{i=1}^{N} w_i^q \left(\sum_{j=1}^{n} p_{ij}^q\right)}{\sum_{k=1}^{N} w_k^q}\right]^{\frac{1}{1-q}} , \tag{4.7}$$

For example, in the case where all subsystems have disjoint event spaces with heterogeneity equal to constant ν, then they each contribute ν unique states to the pooled system X.

Deriving the between-group heterogeneity $\Pi_q^{\mathrm{B}}\left(\mathbf{P}, \mathbf{w}\right)$, is thus straightforward. If the effective total number of states in the pooled system is $\Pi_q^{\mathrm{P}}\left(\mathbf{P}, \mathbf{w}\right)$, and the effective number of unique states contributed by distinct subsystems is $\Pi_q^{\mathrm{W}}\left(\mathbf{P}, \mathbf{w}\right)$, then

$$\Pi_q^{\mathrm{B}}\left(\mathbf{P}, \mathbf{w}\right) = \frac{\Pi_q^{\mathrm{P}}\left(\mathbf{P}, \mathbf{w}\right)}{\Pi_q^{\mathrm{W}}\left(\mathbf{P}, \mathbf{w}\right)} \tag{4.8}$$

is the effective number of completely distinct subsystems in the pooled system X. A word of caution is warranted. If we require that within-group heterogeneity is a lower bound on pooled heterogeneity [184], then Jost [59] showed that Equation 4.8 will hold (A) at any value of q when weights are equal (i.e., $w_i = 1/N$ for all $i \in \{1, 2, \dots, N\}$), or (B) only at $q = 0$ and $q = 1$ if weights are unequal.

Limitations of Categorical Rényi Heterogeneity

The chief limitation of Rényi heterogeneity (Equation 4.3) is its assumption that all states in a system X (with event space $\mathcal{X} = \{1, 2, \dots, n\}$ and probability distribution $\mathbf{p} = (p_i)_{i=1,2,\dots,n}$) are categorical. More formally, the dissimilarity between a pair of observations $(x, y) \in \mathcal{X}$ from this system is defined by the discrete metric

$$d^*(x, y) = 1 - \delta_{xy}, \tag{4.9}$$

where δ_{xy} is Kronecker's delta, which takes a value of 1 if $x = y$ and 0 otherwise. Since the discrete metric assumption is an idealization, we have continued to use the asterisk to qualify an arbitrary distance function $d(\cdot, \cdot)$ as categorical in nature. The resulting expected pairwise distance matrix between states in X is

$$\mathbf{D}^* = \left[d^*(i, j)\right]_{i=1,2,\dots,n}^{j=1,2,\dots,n} = \mathbf{1}\,\mathbf{1}^\top - \mathbf{I}, \tag{4.10}$$

where $\mathbf{1} = (1)_{i=1,2,\dots,n}$ is a column vector of ones, and $\mathbf{I} = (\delta_{ij})_{i=1,2,\dots,n}^{j=1,2,\dots,n}$ is the $n \times n$ identity matrix.

Clearly, many systems of interest in the real world are not categorical. For example, although we may label a sample of organisms according to their respective species, there may be differences between these taxonomic classes that are relevant to the functioning of the ecosystem as a whole [2]. It is also possible that no valid and reliable set of categorical labels is known a priori for a system whose event space is naturally non-categorical.

4.2.2 Non-Categorical Heterogeneity Indices

Consider a system X with probability distribution $\mathbf{p} = (p_i)_{i=1,2,\dots,n}$ defined over event space $\mathcal{X} = \{1, 2, \dots, n\}$ and equipped with dissimilarity function $d_X(\cdot, \cdot)$. We assume that d_X is more general than the discrete metric (Equation 4.9), and further still need not be a true (metric) distance. For such systems, there are three heterogeneity indices whose units are numbers equivalent, and respect the replication principle [90, 111–114]. Much like our derivation of the Rényi heterogeneity in Section 4.2.1, these indices quantify the heterogeneity of a non-categorical system as the number of states in an idealized reference system, but differ primarily in how the idealized reference is defined. We begin with a discussion of the Numbers-Equivalent Quadratic Entropy (Section 4.2.2), followed by the Functional Hill Numbers (Section 4.2.2) and the Leinster–Cobbold index [114] (Section 4.2.2).

Numbers Equivalent Quadratic Entropy

Rao [110] introduced the diversity index commonly known as Rao's quadratic entropy (RQE),

$$Q_1\left(\mathbf{D}, \mathbf{p}\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij} p_i p_j \tag{4.11}$$

where $\mathbf{D}$ is an $n \times n$ matrix where $D_{ij} = d_X(i, j)$ for states $(i, j) \in \mathcal{X}$.

Ricotta & Szeidl [112] assume that $D_{ij} = 1$ means that states i and j are maximally dissimilar (i.e., categorically different), and that $D_{ij} = 0$ means $i = j$, which occurs when $\mathcal{X}$ is a categorical system. An arbitrary dissimilarity matrix $\mathbf{D}$ can be rescaled to respect this assumption by applying the following transformation:

$$\tilde{\mathbf{D}} = \frac{\mathbf{D} - \min_{ij} D_{ij}}{\max_{ij} D_{ij} - \min_{ij} D_{ij}}. \tag{4.12}$$

Under this transformation, Ricotta & Szeidl [112] search for an idealized categorical reference system X_* with event space $\mathcal{X}_* = \{1, 2, \dots, n_*\}$, probability distribution $\mathbf{p}_* = (n_*^{-1})_{i=1,2,\dots,n_*}$, and RQE equal to that of X. For a column vector of ones, $\mathbf{1} = (1)_{i=1,2,\dots,n_*}$, and the identity matrix $\mathbf{I} = (\delta_{ij})_{i=1,2,\dots,n_*}^{j=1,2,\dots,n_*}$, this is

$$Q_1\left(\tilde{\mathbf{D}}, \mathbf{p}\right) = Q_1\left(\mathbf{1}\mathbf{1}^\top - \mathbf{I}, \mathbf{p}_*\right). \tag{4.13}$$

Expanding the right-hand side, we have

$$Q_1\left(\tilde{\mathbf{D}},\mathbf{p}\right)=\sum_{i=1}^{n_*}\sum_{j=1}^{n_*}n_*^{-2}\left(1-\delta_{ij}\right)=1-\frac{1}{n_*}. \tag{4.14}$$

Recalling that $\Pi_q\left(\mathbf{p}_*\right)=n_*$ and substituting into Equation 4.14 yields

$$\Pi_q\left(\mathbf{p}_*\right)=\left[1-Q_1\left(\tilde{\mathbf{D}},\mathbf{p}\right)\right]^{-1}, \tag{4.15}$$

which establishes the units of $\left[1-Q_1\left(\tilde{\mathbf{D}},\mathbf{p}\right)\right]^{-1}$ as numbers equivalent.

For consistency, we require that $\Pi_q\left(\mathbf{p}_*\right)=\Pi_q\left(\mathbf{p}\right)$ if $\tilde{\mathbf{D}}$ were categorical. This only holds at $q=2$:

$$\left[1-Q_1\left(\tilde{\mathbf{D}},\mathbf{p}\right)\right]^{-1}=\left[1-\sum_{i=1}^{n}\sum_{j=1}^{n}p_ip_j\left(1-\delta_{ij}\right)\right]^{-1}=\left(\sum_{i=1}^{n}p_i^2\right)^{-1}=\Pi_2\left(\mathbf{p}_*\right). \tag{4.16}$$

Based on this result, Ricotta & Szeidl [112] define the numbers equivalent quadratic entropy $\hat{Q}_e$ as

$$\hat{Q}_e\left(\tilde{\mathbf{D}},\mathbf{p}\right)=\left(1-Q_1\left(\tilde{\mathbf{D}},\mathbf{p}\right)\right)^{-1}. \tag{4.17}$$

This can be interpreted as the inverse Simpson concentration of an idealized categorical reference system whose average pairwise distance between states is equal to $Q_1\left(\tilde{\mathbf{D}},\mathbf{p}\right)$.

Functional Hill Numbers

Chiu & Chao [111] derived the Functional Hill Numbers, denoted F_q, based on a similar procedure to that of Ricotta & Szeidl [112]. However, whereas $\hat{Q}_e$ uses a purely categorical system as the idealized reference, F_q requires only that

$$Q_1\left(\mathbf{D},\mathbf{p}\right)=\sum_{i=1}^{n_*}\sum_{j=1}^{n_*}Q_1\left(\mathbf{D},\mathbf{p}\right)p_{*i}p_{*j}=\sum_{i=1}^{n_*}\sum_{j=1}^{n_*}Q_1\left(\mathbf{D},\mathbf{p}\right)n_*^{-2}, \tag{4.18}$$

which means that the idealized reference system is one for which the between-state distance matrix is set to $Q_1\left(\mathbf{D},\mathbf{p}\right)$ everywhere (or to 0 along the leading diagonal and $Q_1\left(\mathbf{D},\mathbf{p}\right)n_*/(n_*-1)$ on the off diagonals).

Chiu & Chao [111] generalized Rao's quadratic entropy to include the elasticity parameter $q \geq 0$

$$Q_q\left(\mathbf{D}, \mathbf{p}\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij} \left(p_i p_j\right)^q, \tag{4.19}$$

and sought to find n_* for the idealized reference system satisfying Equation 4.18 and the following:

$$Q_q\left(\mathbf{D}, \mathbf{p}\right) = \sum_{i=1}^{n_*} \sum_{j=1}^{n_*} Q_1\left(\mathbf{D}, \mathbf{p}\right) \left(\frac{1}{n_*} \frac{1}{n_*}\right)^q. \tag{4.20}$$

Solving Equation 4.20 for n_* yields the functional Hill numbers of order q:

$$F_q\left(\mathbf{D}, \mathbf{p}\right) = \left(\frac{Q_q\left(\mathbf{D}, \mathbf{p}\right)}{Q_1\left(\mathbf{D}, \mathbf{p}\right)}\right)^{\frac{1}{2(1-q)}} = n_*, \tag{4.21}$$

which is the effective number of states in an idealized categorical reference system whose distance function is scaled by a factor of $Q_1\left(\mathbf{D}, \mathbf{p}\right) n_*/(n_* - 1)$.

Leinster–Cobbold Index

The index derived by Leinster & Cobbold [114], denoted L_q, is distinct from $\hat{Q}_e$ and F_q in two ways. First, for a given system X, the L_q is not derived based on finding an idealized reference system X_* whose average between-state dissimilarity is equal to that of X. Second, it does not use a dissimilarity matrix; rather, it uses a measure of similarity or affinity.

The Leinster–Cobbold index may be derived by simple extension of Equation 4.3. Assuming X has state space $\mathcal{X} = \{1, 2, \dots, n\}$ with probability distribution $\mathbf{p} = (p_i)_{i=1,2,\dots,n}$, we note that

$$\Pi_q\left(\mathbf{p}\right) = \left(\sum_{i=1}^{n} p_i^q\right)^{\frac{1}{1-q}} = \left[\sum_{i=1}^{n} p_i \left(\mathbf{I}\mathbf{p}\right)_i^{q-1}\right]^{\frac{1}{1-q}}. \tag{4.22}$$

Here, $\mathbf{I}$ is the $n \times n$ identity matrix representing the pairwise similarities between states in X. The Leinster–Cobbold index generalizes $\mathbf{I}$ to be any $n \times n$ similarity matrix $\mathbf{S}$, yielding the following formula:

$$L_q\left(\mathbf{S}, \mathbf{p}\right) = \left[\sum_{i=1}^{n} p_i \left(\sum_{j=1}^{n} S_{ij} p_j\right)^{q-1}\right]^{\frac{1}{1-q}}. \tag{4.23}$$

The similarity matrix can be obtained from a dissimilarity matrix by the transformation $S_{ij} = e^{-uD_{ij}}$, where $u \geq 0$ is a scaling factor. When $u = 0$, then $\mathbf{S}$ is 1 everywhere. Conversely, when $u \to \infty$, then $\mathbf{S}$ approaches $\mathbf{I}$. The Leinster–Cobbold index can thus be interpreted as an effective number if the states are in an idealized reference system (i.e., one with uniform probabilities over states) whose topology is also governed by the similarity matrix $\mathbf{S}$.

Limitations of Existing Non-Categorical Heterogeneity Indices

We illustrate several limitations of the $\hat{Q}_e$, F_q, and L_q indices using a simple 3-state system X with event space $\mathcal{X} = \{1, 2, 3\}$ over which we specify a probability distribution

$$\mathbf{p}(\kappa) = \begin{cases} (1, 0, 0)^\top & \kappa = 0 \\ \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)^\top & \kappa = 1 \\ (0, 0, 1)^\top & \kappa = \infty \\ \left(\frac{1}{1+\sqrt{\kappa}+\kappa}, \frac{\sqrt{\kappa}}{1+\sqrt{\kappa}+\kappa}, \frac{\kappa}{1+\sqrt{\kappa}+\kappa}\right)^\top & \text{Otherwise} \end{cases} \tag{4.24}$$

where $0 \leq \kappa$ is a parameter that smoothly varies the level of inequality. When $\kappa = 1$ the distribution is perfectly even (Figure 4.1A). Since an undirected graph of the system is arranged in a triangle with height h and base b, we also specify the following parametric distance matrix,

$$\mathbf{D}(h, b) = \begin{pmatrix} 0 & b & \sqrt{\frac{b^2}{4} + h^2} \\ b & 0 & \sqrt{\frac{b^2}{4} + h^2} \\ \sqrt{\frac{b^2}{4} + h^2} & \sqrt{\frac{b^2}{4} + h^2} & 0 \end{pmatrix}, \tag{4.25}$$

which allows us to smoothly vary the level of dissimilarity between states in X. Importantly, Equation 4.25 allows us to generate distance matrices that are either metric (when $h < b\sqrt{3}/2$; Definition 1) or ultrametric (when $h \geq b\sqrt{3}/2$; Definition 2). This is illustrated in Figure 4.1B.

Definition 1 (Metric distance)**.** A function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ on a set $\mathcal{X}$ is a metric if and only if all of the following conditions are satisfied for all $(x, y, z) \in \mathcal{X}$:

1. Non-negativity: $d(x, y) \geq 0$

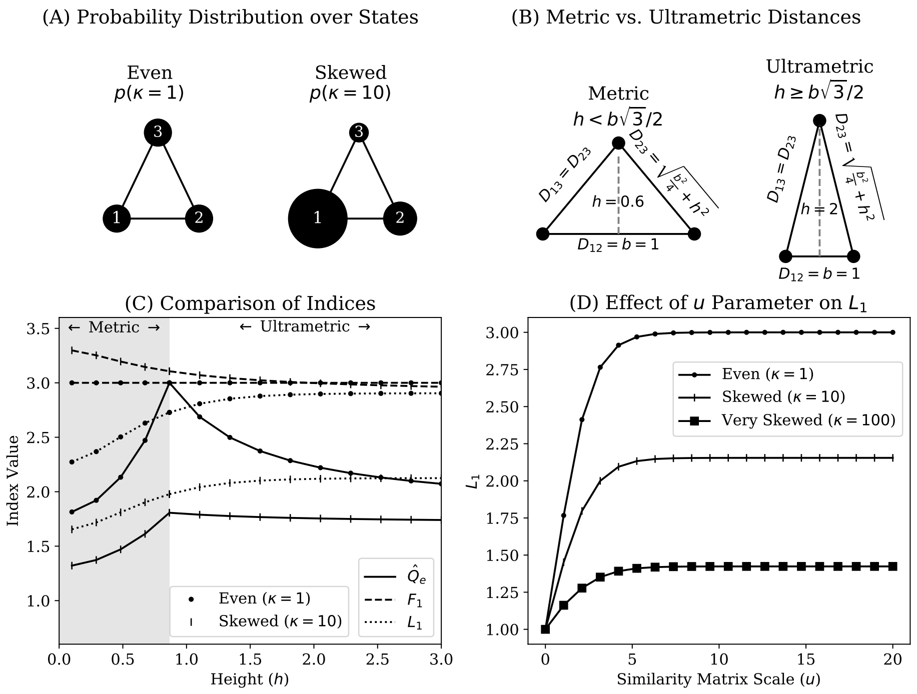


Figure 4.1: Illustration of simple three-state system under which we compare existing non-categorical heterogeneity indices. **Panel A** depicts a three state system X as an undirected graph, with node sizes corresponding to state probabilities governed by Equation 4.24. As $0 \leq \kappa$ diverges further from $\kappa = 1$, the probability distribution over states becomes more unequal. **Panel B** visually represents the parametric pairwise distance matrix $\mathbf{D}(h, b)$ of Equation 4.25 (h is height, b is base length, D_{ij} is distance between states i and j). In the examples shown in Panels B and C, we set $b = 1$. Specifically, we provide visual illustration of settings for which the distance function on X is a metric (Definition 1; when $h < b\sqrt{3}/2$) or ultrametric (Definition 2; when $h \geq b\sqrt{3}/2$). **Panel C** compares the numbers equivalent quadratic entropy (solid lines marked $\hat{Q}_e$; Section 4.2.2), functional Hill numbers (at $q = 1$, dashed lines marked F_1; Section 4.2.2), and the Leinster–Cobbold Index (at $q = 1$, dotted lines marked L_1; Section 4.2.2) for reporting the heterogeneity of X. The y-axis reports the value of respective indices. The x-axis plots the height parameter for the distance matrix $\mathbf{D}(h, 1)$ (Equation 4.25 and Panel B). The range of h at which $\mathbf{D}(h, 1)$ is only a metric is depicted by the gray shaded background. The range of h shown with a white background is that for which $\mathbf{D}(h, 1)$ is ultrametric. For each index, we plot values for a probability distribution over states that is perfectly even ($\kappa = 1$; dotted markers) or skewed ($\kappa = 10$; vertical line markers). **Panel D** shows the sensitivity of the Leinster–Cobbold index (L_1; y-axis) to the scaling parameter $0 \leq u$ (x-axis) used to transform a distance matrix into a similarity matrix ($S_{ij} = e^{-uD_{ij}}$). This is shown for three levels of skewness for the probability distribution over states (no skewness at $\kappa = 1$, dotted markers; significant skewness at $\kappa = 10$, vertical line markers; extreme skewness at $\kappa = 100$, square markers).

2. Identity of indiscernibles: $d(x, y) = 0 \iff x = y$

3. Symmetry: $d(x, y) = d(y, x)$

4. Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$

Definition 2 (Ultrametric distance)**.** A function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ on a set $\mathcal{X}$ is ultrametric if and only if, for all $(x, y, z) \in \mathcal{X}$, criteria 1-3 for a metric are satisfied (Definition 1), in addition to the ultrametric triangle inequality:

$$d(x, z) \leq \max \{d(x, y), d(y, z)\} \tag{4.26}$$

Figure 4.1C compares the $\hat{Q}_e$, F_q, and L_q indices when applied to X across variation in between-state distances (via Equation 4.25) and skewness in the probability distribution over states (Equation 4.24). With respect to the numbers equivalent quadratic entropy ($\hat{Q}_e$; Section 4.2.2), we note that its behavior is categorically different with respect to whether the distance matrix is ultrametric. That is $\hat{Q}_e$ increases with the triangle height parameter h (Equation 4.25) until it passes the ultrametric threshold, after which it decreases monotonically with h. The behavior of $\hat{Q}_e$ is sensible in the ultrametric range. When the distance matrix is scaled, as in Equation 4.12, pulling one of the three states in X further away from the remaining two should function similarly to progressively merging the latter states. Thus, the behavior of $\hat{Q}_e$ is highly sensitive to whether a given distance matrix is ultrametric (which will often not be the case in real-world applications).

With respect to F_q, a notable benefit in comparison to $\hat{Q}_e$ is that F_q behaves consistently regardless of whether distance is ultrametric. However, Figure 4.1 shows other drawbacks. First, we can see that F_q becomes insensitive to $\mathbf{D}(h, 1)$ when $\mathbf{p}(\kappa)$ is perfectly even (shown analytically in Appendix A.1). Second, F_q can paradoxically estimate a greater number of states than the theoretical maximum allows. That this occurs when the state probability distribution is more unequal violates the principle of transfers [18, 42, 85, 86] (Section 4.2.1). This is made more problematic since Figure 4.1C shows it occurs when one state is being pushed closer to the others (i.e., with smaller values of h). To summarize, the functional Hill numbers are estimating more states than are really present despite the reduction in between-state distances and greater inequality in the probability mass function.

Figure 4.1C shows that the Leinster-Cobbold index compares favorably to F_q because the former does not lose sensitivity to dissimilarity when $\mathbf{p}(\kappa)$ is perfectly even. However, Figure 4.1D shows that the Leinster-Cobbold index is particularly sensitive to the form of similarity transformation. In the present case, the maximal value of the L_q gradually approaches 3 as u grows (and only when $u \to \infty$ does it reach 3), while progressively losing sensitivity to distance. As mentioned by Leinster & Cobbold [114], the choice of u or other similarity transformation is dependent on the importance assigned to functional differences between states. However, it is not clear how a given similarity transformation (e.g., u), and therefore the idealized reference system of L_q, should be validated.

Above all of the idiosyncratic limitations of existing numbers equivalent heterogeneity indices, we must highlight two basic assumptions they all share. First, they continue to assume that some valid and reliable categorical partitioning on X is known a priori. Second, they assume that a distance function specified a priori describes semantically relevant geometry of the system in question. These two limitations are not independent, since an unreliable categorical partitioning of the state space will lead to erroneous estimates of the pairwise distances between states. Thus, we seek an approach for measuring heterogeneity that has neither these limitations, nor those shown above to be specific to the other numbers equivalent heterogeneity indices for non-categorical systems.

4.3 Representational Rényi Heterogeneity

In this section, we propose an alternative approach to the indices of Section 4.2.2 that we call representational Rényi heterogeneity (RRH). It involves transforming X into a representation Z, defined on an unobservable or latent event space $\mathcal{Z}$, that satisfies two criteria:

1. The representation Z captures the semantically relevant variation in X
2. Rényi heterogeneity can be directly computed on Z

Satisfaction of the first criterion can only be ascertained in a domain-specific fashion. Since Z is essentially a model of X, investigators must justify that this model is appropriate for the scientific question at hand. For example, an investigator may evaluate the ability of X to be reconstructed from representation Z under cross-validation. The second criterion simply means that the transformation of $X \to Z$ must specify a probability distribution on $\mathcal{Z}$ upon which the Rényi heterogeneity can be directly computed.

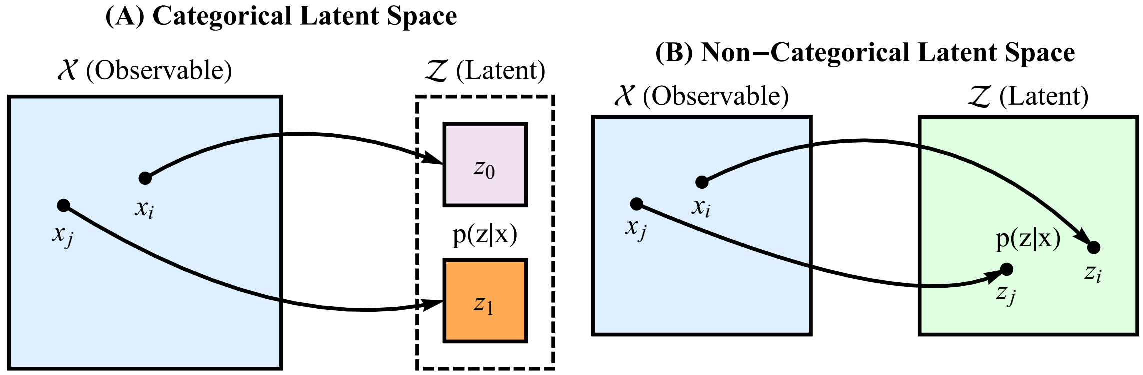

(A) Categorical Latent Space
$\mathcal{X}$ (Observable)
$\mathcal{Z}$ (Latent)
x_i
x_j
z_0
p(z|x)
z_1
(B) Non−Categorical Latent Space
$\mathcal{X}$ (Observable)
$\mathcal{Z}$ (Latent)
x_i
x_j
p(z|x)
z_j
z_i

Figure 4.2: Graphical illustration of the two main approaches for computing representational Rényi heterogeneity. In both cases, we map sampled points on an observable space $\mathcal{X}$ onto a latent space $\mathcal{Z}$, upon which we apply the Rényi heterogeneity measure. The mapping is illustrated by the curved arrows, and should yield a posterior distribution over the latent space. **Panel A** shows the case in which the latent space is categorical (for example, discrete components of a mixture distribution on a continuous space). **Panel B** illustrates the case in which the latent space has non-categorical topology. A special case of the latter mapping may include probabilistic principal components analysis. When the latent space is continuous, we must derive a parametric form for the Rényi heterogeneity.

Figure 4.2 illustrates the basic idea of RRH. However, the specifics of this framework differ based on the topology of the representation Z. Thus, the remainder of this section discusses the following approaches:

A. Application of standard Rényi heterogeneity (Section 4.2.1) when Z is a categorical representation

B. Deriving parametric forms for Rényi heterogeneity when Z is a non-categorical representation

4.3.1 Rényi Heterogeneity on Categorical Representations

Let X be a system defined on an observable space $\mathcal{X}$ that is non-categorical and n_x-dimensional. Consider the scenario in which the semantically relevant variation in X is categorical: for instance, images of different object categories stored in raw form as real-valued vectors. An investigator may be interested in measuring the effective number of states in X with respect to this categorical variation. This requires transforming X into a semantically relevant categorical representation Z (such as one does in unsupervised clustering) upon which Equation 4.3 can be applied.

Assume we have a large random sample of N points $\mathbf{X} = (\mathbf{x}_i)_{i=1,2,\dots,N}$ from system X. We can conceptualize each discrete observation $\mathbf{x}_i$ in this sample as the single point in the event space of a perfectly homogeneous subsystem X_i. When pooled, the subsystems $\{X_i\}_{i=1,2,\dots,N}$ constitute X. The contribution weights of each subsystem to X as a whole are denoted $\mathbf{w} = (w_i)_{i=1,2,\dots,N}$, where $\sum_{i=1}^{N} w_i = 1$ and $w_i \geq 0$.

We now specify a vector-valued function $\mathbf{f} : \mathcal{X} \to \mathcal{P}(\mathcal{Z})$ such that $\mathbf{x} \mapsto \mathbf{f}(\mathbf{x}) = [f_j(\mathbf{x})]_{j=1,2,\dots,n_z}$ is a mapping from n_x-dimensional coordinates on the observable space, $\mathbf{x} \in \mathcal{X}$, onto an n_z-dimensional discrete probability distribution over $\mathcal{Z} = \{1, 2, \dots, n_z\}$. Thus, $\mathbf{f}(\mathbf{x}_i)$ can be conceptualized as mapping subsystem X_i onto its categorical representation Z_i. After defining $\mathbf{f}$, the effective number of states in the latent representation of X_i can be computed as

$$\Pi_q\left(\mathbf{x}_i\right) = \left(\sum_{j=1}^{n_z} f_j^q(\mathbf{x}_i)\right)^{\frac{1}{1-q}}. \tag{4.27}$$

When $\Pi_q\left(\mathbf{x}_i\right) = 1$, then $\mathbf{f}$ assigns $\mathbf{x}$ to a single category with perfect certainty. Conversely, when $\Pi_q\left(\mathbf{x}_i\right) = n_z$, then either $\mathbf{x}_i$ belongs to all categorical states with equal probability, or $\mathbf{f}$ is maximally uncertain about the mapping of point $\mathbf{x}_i$.

Mapping all points $\mathbf{X}$ onto the categorical latent space yields a collection of subsystems $\{Z_i\}_{i=1,2,\dots,N}$, which generate Z when pooled. Using Equation 4.6, we can compute the effective number of total states in Z as the pooled heterogeneity:

$$\Pi_q^{\mathrm{P}}\left(\mathbf{X}, \mathbf{w}\right) = \left[\sum_{j=1}^{n_z}\left(\sum_{i=1}^{N} w_i f_j(\mathbf{x}_i)\right)^q\right]^{\frac{1}{1-q}}, \tag{4.28}$$

Unfortunately, $\Pi_q^{\mathrm{P}}\left(\mathbf{X}, \mathbf{w}\right)$ counts some heterogeneity that is due to uncertainty in the model (i.e., that quantified by Equation 4.27). We, therefore, compute the effective number of states in Z per point $\mathbf{x} \in \mathcal{X}$ using the within-group heterogeneity formula (Equation 4.7):

$$\Pi_q^{\mathrm{W}}\left(\mathbf{X}, \mathbf{w}\right) = \left[\frac{\sum_{i=1}^{N} w_i^q \left(\sum_{j=1}^{n_z} f_j^q(\mathbf{x}_i)\right)}{\sum_{k=1}^{N} w_k^q}\right]^{\frac{1}{1-q}}. \tag{4.29}$$

Finally, the effective number of states (points) in X—with respect to the categorical variation modeled by Z—can then be computed using the between-group heterogeneity formula (Equation 4.8):

$$\Pi_q^{\mathrm{B}}\left(\mathbf{X},\mathbf{w}\right) = \frac{\Pi_q^{\mathrm{P}}\left(\mathbf{X},\mathbf{w}\right)}{\Pi_q^{\mathrm{W}}\left(\mathbf{X},\mathbf{w}\right)}. \tag{4.30}$$

Example 3 demonstrates that current methods of measuring biodiversity and wealth concentration can be viewed as special cases of categorical RRH.

Example 3 (Classical measurement of biodiversity and economic equality as categorical RRH)**.** Definitions necessary for this example are shown in Table 4.2. The traditional analysis of species diversity and economic equality can be recovered from an RRH-based formulation when $\mathbf{f}$ is assumed to be deterministic and $\mathbf{w} = \left(N^{-1}\right)_{i=1,2,\dots,N}$. In this case within-group heterogeneity can be shown to reduce to 1:

$$\begin{aligned}\Pi_q^{\mathrm{W}}\left(\mathbf{X},\mathbf{w}\right) &= \left[\sum_{i=1}^{N} \frac{N^{-q}}{\sum_{k=1}^{N} N^{-q}} \left(\sum_{j=1}^{n_z} f_j^q(\mathbf{x}_i)\right)\right]^{\frac{1}{1-q}} \\ &= \left[\sum_{i=1}^{N} N^{-1}\left(1\right)\right]^{\frac{1}{1-q}} \\ &= 1.\end{aligned} \tag{4.31}$$

Thus, we have

$$\begin{aligned}\Pi_q^{\mathrm{B}}\left(\mathbf{X},\mathbf{w}\right) &= \Pi_q^{\mathrm{P}}\left(\mathbf{X},\mathbf{w}\right) \\ &= \left[\sum_{j=1}^{n_z}\left(\sum_{i=1}^{N} N^{-1} f_j(\mathbf{x}_i)\right)^q\right]^{\frac{1}{1-q}} \\ &= \left[\sum_{j=1}^{n_z}\left(\frac{N_j}{N}\right)^q\right]^{\frac{1}{1-q}},\end{aligned} \tag{4.32}$$

which yields the categorical Rényi heterogeneity (Hill numbers for biodiversity analysis and Hannah–Kay indices in the economic setting [17, 18]), and by extension many diversity indices to which it is connected (Table 4.1). Thus, traditional analysis of species biodiversity and economic equality are special cases of representational Rényi heterogeneity where the representation is specified by a mapping onto degenerate distributions over categorical labels. The only differences lie in the definition of observable and latent spaces, and the representational models.

In the case of biodiversity analysis, the model $\mathbf{f}$ in real-world practice may simply be a human expert assigning species labels to a sample of organisms from a field study. In the economic setting,

Table 4.2: Definitions in formulation of classical biodiversity and economic equality analysis as categorical representational Rényi heterogeneity. Superscripted indexing on $\mathbf{x} = (x_i)^{i=1,\dots,n_x}$ denotes that this is a row vector.

	Analytical Context	
Symbol	**Biodiversity**	**Economic Equality**
X	Ecosystem, whose observation yields an organism denoted by vector $\mathbf{x} = (x_i)^{i=1,\dots,n_x} \in \mathcal{X}$	A system of resources, whose observation yields an asset denoted by vector $\mathbf{x} = (x_i)^{i=1,\dots,n_x} \in \mathcal{X}$
$\mathcal{X} \subseteq \mathbb{R}^{n_x}$	n_x-dimensional feature space of organisms in the ecosystem	n_x-dimensional feature space of assets in the economy, whose topology is such that the "economic" or monetary value is equal at each coordinate $\mathbf{x} \in \mathcal{X}$
$\mathcal{Z} = \{\mathbf{z} \in \{0,1\}^{n_z} : \sum_{i=1}^{n_z} z_i = 1\}$	n_z-dimensional space of one-hot species labels	n_z-dimensional space of one-hot labels over wealth-owning agents
$\mathbf{f} : \mathcal{X} \to \mathcal{P}(\mathcal{Z})$	A model that performs the mapping $\mathbf{x} \mapsto \mathbf{f}(\mathbf{x})$ of organisms to discrete probability distributions over $\mathcal{Z}$	A model that performs the mapping $\mathbf{x} \mapsto \mathbf{f}(\mathbf{x})$ of assets to discrete probability distributions over $\mathcal{Z}$
$N_i \in \mathbb{N}_+$	The number of organisms observed belonging to species $i \in \{1, \dots, n_z\}$	The number of equal valued assets belonging to agent $i \in \{1, \dots, n_z\}$
$N = \sum_{i=1}^{n_z} N_i$	The total number of organisms observed	The total quantity of assets observed
$\mathbf{X} = (x_{ij})_{i=1,\dots,N}^{j=1,\dots,n_x}$	A sample of N organisms	A sample of N assets
$\mathbf{w} = (w_i)_{i=1,\dots,N}$	Sample weights, such that $w_i \geq 0$ and $\sum_{i=1}^{N} w_i = 1$	

one may speculate that $\mathbf{f}$ would essentially reduce to contracts specifying ownership of assets, whose value is deemed by market forces.

4.3.2 Rényi Heterogeneity on Non-Categorical Representations

In Section 4.3.1, we dealt with instances in which semantically relevant variation in X is categorical, such as when object categories are embedded in images stored as real-valued vectors. Here, we consider scenarios in which the semantically relevant information in an observable system X is non-categorical: for instance, where a piece of text contains information about semantic concepts best represented as real-valued "word vectors" [185, 186]. Measuring the effective number of distinct states in X with respect to this continuous variation requires transforming X into a semantically relevant continuous representation Z upon which procedures analogous to those of Section 4.3.1 may be undertaken.

Let Z be defined on an n_z-dimensional event space $\mathcal{Z} \subseteq \mathbb{R}^{n_z}$ over which there exists a family of parametric probability distributions $\mathcal{P}(\mathcal{Z})$ of a form chosen by the experimenter. Let $f : \mathcal{X} \to \mathcal{P}(\mathcal{Z})$ be a model that performs the mapping $\mathbf{x} \mapsto f(\cdot|\mathbf{x})$ from a point $\mathbf{x} \in \mathcal{X}$ on the observable space to a probability density on $\mathcal{Z}$. For example, if $\mathcal{P}(\mathcal{Z})$ is the family of multivariate Gaussians, then $f(\mathbf{z}|\mathbf{x}_i) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the Gaussian mean and covariance functions at $\mathbf{x}_i$, respectively. Given a sample $\mathbf{X} = (\mathbf{x}_i)_{i=1,2,\dots,N}$, as in Section 4.3.1, we compute the continuous analogue of Equation 4.27 as follows

$$\Pi_q\left(\mathbf{x}_i\right) = \left(\int_{\mathcal{Z}} f^q(\mathbf{z}|\mathbf{x}_i)\,\mathrm{d}\mathbf{z}\right)^{\frac{1}{1-q}}. \tag{4.33}$$

This formula yields the effective size of the domain of a uniform distribution on $\mathbb{R}^{n_z}$ whose Rényi heterogeneity is equal to $\Pi_q\left(\mathbf{x}_i\right)$ (proof is given in Appendix A.1). Thus, it is possible for $\Pi_q\left(\mathbf{x}_i\right)$ to be less than 1, though it will remain non-negative.

Similar to the procedure in Section 4.3.1, we now define a continuous version of the within-observation heterogeneity

$$\Pi_q^{\mathrm{W}}\left(\mathbf{X}, \mathbf{w}\right) = \left[\sum_{i=1}^{N} \frac{w_i^q}{\sum_{j=1}^{N} w_j^q} \int_{\mathcal{Z}} f^q(\mathbf{z}|\mathbf{x}_i)\,\mathrm{d}\mathbf{z}\right]^{\frac{1}{1-q}}, \tag{4.34}$$

which estimates the effective size of the latent space occupied per observable point $\mathbf{x} \in \mathcal{X}$.

In order to compute the pooled heterogeneity $\Pi_q^{\mathrm{P}}\left(\mathbf{X}, \mathbf{w}\right)$, the experimenter must specify the form of the pooled distribution, here denoted $\bar{f}_{\mathbf{w}}$. The conceptually most simple approach is non-parametric, using a model average,

$$\bar{f}_{\mathbf{w}}\left(\mathbf{z}|\mathbf{X}\right) = \sum_{i=1}^{N} w_i f(\mathbf{z}|\mathbf{x}_i), \tag{4.35}$$

whereby the pooled heterogeneity would be

$$\Pi_q^{\mathrm{P}}\left(\mathbf{X}, \mathbf{w}\right) = \left[\int_{\mathcal{Z}} \left(\sum_{i=1}^{N} w_i f(\mathbf{z}|\mathbf{x}_i)\right)^q \mathrm{d}\mathbf{z}\right]^{\frac{1}{1-q}}. \tag{4.36}$$

The integral in Equation 4.36 may often be analytically intractable and potentially difficult to solve accurately in high dimensions with numerical methods. Furthermore, some areas of $\mathcal{Z}$ may be assigned low probability by $f(\mathbf{z}|\mathbf{x}_i)$ for all $i \in \{1, 2, \dots, N\}$. This is not a problem as the sample $\mathbf{X}$ becomes infinitely large. However, with finite samples, it may be the case that some representational states in $\mathcal{Z}$ are unlikely simply because we have not

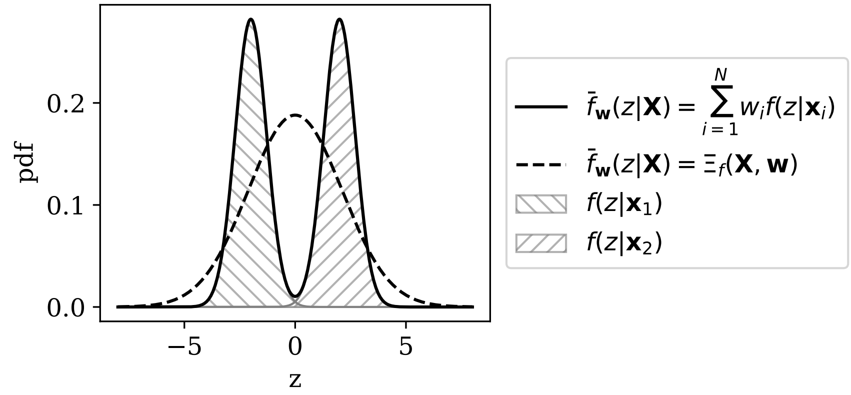


Figure 4.3: Illustration of approaches to computing the pooled distribution on a simple representational space $\mathcal{Z} = \mathbb{R}$. In this example, two points on the observable space, $(\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{X}$, are mapped onto the latent space via model $f(\cdot|\mathbf{x}_i)$ for $i \in \{1, 2\}$, which indexes univariate Gaussians over $\mathcal{Z}$ (depicted as hatched patterns for $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively). A pooled distribution computed non-parametrically by model-averaging (Equation 4.35) is depicted as the solid black line. The parametrically pooled distribution (see Example 4) is depicted as the dashed black line. The parametric approach implies the assumption that further samples from $\mathcal{X}$ would yield latent space projections in some regions assigned low probability by $f(z|\mathbf{x}_1)$ and $f(z|\mathbf{x}_2)$.

sampled from the corresponding regions of $\mathcal{X}$. An alternative to Equation 4.35 is therefore to specify a parametric pooled distribution

$$\bar{f}_{\mathbf{w}}\left(\cdot|\mathbf{X}\right) = \Xi_f\left(\mathbf{X}, \mathbf{w}\right), \tag{4.37}$$

where Ξ_f is a deterministic function that combines $f(\cdot|\mathbf{x}_i)$ for $i \in \{1, 2, \ldots, N\}$ into a valid probability density on $\mathcal{Z}$. In this case, the pooled Rényi heterogeneity is simply

$$\Pi_q^{\mathrm{P}}\left(\mathbf{X}, \mathbf{w}\right) = \left(\int_{\mathcal{Z}} \bar{f}_{\mathbf{w}}^{q}(\mathbf{z}|\mathbf{X})\, \mathrm{d}\mathbf{z}\right)^{\frac{1}{1-q}}. \tag{4.38}$$

Using either Equation 4.36 or 4.38 as the pooled heterogeneity and Equation 4.34 as the within-group heterogeneity, the effective number of distinct states in X—with respect to the non-categorical representation Z—can then be computed using Equation 4.30.

Figure 4.3 demonstrates the difference between the parametric and non-parametric approaches to pooling for non-categorical RRH, and Example 4 demonstrates one approach to parametric pooling for a mixture of multivariate Gaussians.

Example 4 (Parametric pooling of multivariate Gaussian distributions)**.** Let $\mathbf{X} = (\mathbf{x}_i)_{i=1,2,\dots,N}$ be a sample of n_x-dimensional vectors from a system X with event space $\mathcal{X} \subseteq \mathbb{R}^{n_x}$. Let Z be a latent representation of X with n_z-dimensional event space $\mathcal{Z} = \mathbb{R}^{n_z}$. Let

$$f(\mathbf{z}|\mathbf{x}_i) = \mathcal{N}\left(\mathbf{z}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \tag{4.39}$$

be a model that returns a multivariate Gaussian density with mean $\boldsymbol{\mu}_i$ and covariance $\boldsymbol{\Sigma}_i$ given point $\mathbf{x}_i \in \mathcal{X}$. Finally, let $\mathbf{w} = (w_i)_{i=1,2,\dots,N}$ be weights assigned to each sample in $\mathbf{X}$ such that $w_i \geq 0$ and $\sum_{i=1}^{N} w_i = 1$.

If one assumes that the pooled distribution over $\mathcal{Z}$ given the set of components $f(\mathbf{z}|\mathbf{x}_1)$, $f(\mathbf{z}|\mathbf{x}_2)$, $\dots$, $f(\mathbf{z}|\mathbf{x}_N)$ is itself a multivariate Gaussian,

$$\bar{f}_{\mathbf{w}}\left(\mathbf{z}|\mathbf{X}\right) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \tag{4.40}$$

with $n_z \times 1$ pooled mean,

$$\boldsymbol{\mu}_* = \sum_{i=1}^{N} w_i \boldsymbol{\mu}_i \tag{4.41}$$

and $n_z \times n_z$ pooled covariance matrix

$$\boldsymbol{\Sigma}_* = -\boldsymbol{\mu}_* \boldsymbol{\mu}_*^\top + \sum_{i=1}^{N} w_i \left[\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top\right], \tag{4.42}$$

then the pooled heterogeneity Π_q^{P} is therefore simply the Rényi heterogeneity of a multivariate Gaussian,

$$\Pi_q\left(\boldsymbol{\Sigma}\right) = \begin{cases} \text{Undefined} & q = 0 \\ (2\pi e)^{\frac{n_z}{2}} \sqrt{|\boldsymbol{\Sigma}|} & q = 1 \\ (2\pi)^{\frac{n_z}{2}} \sqrt{|\boldsymbol{\Sigma}|} & q = \infty \\ (2\pi)^{\frac{n_z}{2}} q^{\frac{n_z}{2(q-1)}} \sqrt{|\boldsymbol{\Sigma}|} & \text{Otherwise} \end{cases} \tag{4.43}$$

evaluated at $\boldsymbol{\Sigma}_*$. The derivation is provided in Appendix A.1 [187]. Equation 4.43 at $\boldsymbol{\Sigma}_*$ is interpreted as the effective size of space $\mathcal{Z}$ occupied by the complete latent representation of X under model f.

The within-group heterogeneity can be obtained for the set of components $[f(\mathbf{z}|\mathbf{x}_i)]_{i=1,2,\dots,N}$ by solving Equation 4.34 for the Gaussian densities, yielding:

$$\Pi_q^{\mathrm{W}}\left(\mathbf{\Sigma}_{1:N}, \mathbf{w}\right) = \begin{cases} \text{Undefined} & q = 0 \\ \exp\left\{\frac{1}{2}\left(n_z + \sum_{i=1}^N w_i \log\left|2\pi\mathbf{\Sigma}_i\right|\right)\right\} & q = 1 \\ 0 & q = \infty \\ (2\pi)^{\frac{n_z}{2}}\left(\sum_{i=1}^N \frac{\bar{w}_i^q \left|\mathbf{\Sigma}_i\right|^{\frac{1}{2}}}{q^{\frac{n_z}{2}}}\right)^{\frac{1}{1-q}} & \text{Otherwise} \end{cases}, \tag{4.44}$$

where we denote $\mathbf{\Sigma}_{1:N} = \{\mathbf{\Sigma}_i\}_{i=1,2,\dots,N}$ for parsimony, and $\bar{w}_i = w_i\left(\sum_{j=1}^N w_j^q\right)^{-1/q}$. Equation 4.44 estimates the effective size of the n_z-dimensional representational space occupied per state $\mathbf{x} \in \mathcal{X}$.

The effective number of states in X with respect to the continuous representation Z is thus the between-group heterogeneity Π_q^{B} which can be computed as the ratio $\Pi_q\left(\mathbf{\Sigma}_*\right)/\Pi_q^{\mathrm{W}}\left(\mathbf{\Sigma}_{1:N}, \mathbf{w}\right)$. The properties of this decomposition—specifically the conditions under which $\Pi_q^{\mathrm{B}} \geq 1$ (Lande's requirement [59, 184])—are discussed further elsewhere [187].

4.4 Empirical Applications of Representational Rényi Heterogeneity

In this section, we demonstrate two applications of RRH under assumptions of categorical (Section 4.4.1) and continuous (Section 4.4.2) latent spaces. First, Section 4.4.1, uses a simple closed-form system consisting of a mixture of two beta distributions on the (0,1) interval to give exact comparisons of the behavior of RRH against that of existing non-categorical heterogeneity indices (Section 4.2.2). This experiment provides evidence that existing non-categorical heterogeneity indices can demonstrate counterintuitive behavior under various circumstances. Second, Section 4.4.2 demonstrates that RRH can yield heterogeneity measurements that are sensible and tractably computed, even for highly complex mappings $f : \mathcal{X} \to \mathcal{P}(\mathcal{Z})$. There, we use a deep neural network to compute the effective number of observations in a database of handwritten images with respect to compressed latent representations on a continuous space.

4.4.1 Comparison of Heterogeneity Indices Under a Mixture of Beta Distributions

Consider a system X with event space $\mathcal{X}$ on the open interval $(0, 1)$, containing an embedded, unobservable, categorical structure represented by the latent system Z with event space

$\mathcal{Z} = \{1, 2\}$. The systems' collective behavior is governed by the joint distribution of a beta mixture model (BMM),

$$p(x, z) = \mathbb{1}[z = 1](1 - \theta_1)\mathrm{Beta}_{\theta_2,\theta_3}\left(x\right) + \mathbb{1}[z = 2]\theta_1\mathrm{Beta}_{\theta_3,\theta_2}\left(x\right), \tag{4.45}$$

where $\mathrm{Beta}_{\alpha,\beta}\left(x\right)$ is the probability density function for a beta distribution with shape parameters α, β, and $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ are parameters. The indicator function $\mathbb{1}[\cdot]$ evaluates to 1 if its argument is true, and to 0 otherwise. The prior distribution is

$$p(z) = \mathbb{1}[z = 1](1 - \theta_1) + \mathbb{1}[z = 2]\theta_1, \tag{4.46}$$

and marginal probability of observable data is as follows (see Figure 4.4 for illustrations):

$$p(x) = (1 - \theta_1)\mathrm{Beta}_{\theta_2,\theta_3}\left(x\right) + \theta_1\mathrm{Beta}_{\theta_3,\theta_2}\left(x\right). \tag{4.47}$$

To facilitate exact comparisons between heterogeneity indices, below, let us assume we have a model $f : \mathcal{X} \to \mathcal{P}(\mathcal{Z})$ that maps an observation $x \in \mathcal{X}$ onto a degenerate distribution over $\mathcal{Z}$:

$$f_{\boldsymbol{\theta}}(z|x) = \mathbb{1}[z = 1]\mathbb{1}[x \leq \tau(\boldsymbol{\theta})] + \mathbb{1}[z = 2]\mathbb{1}[x > \tau(\boldsymbol{\theta})]. \tag{4.48}$$

The subscripting of $f_{\boldsymbol{\theta}}$ denotes that the model is optimized such that the threshold $0 \leq \tau(\boldsymbol{\theta}) \leq 1$ is the solution to

$$p\left(z = 1|x = \tau(\boldsymbol{\theta})\right) = p\left(z = 2|x = \tau(\boldsymbol{\theta})\right), \tag{4.49}$$

which is

$$\tau(\boldsymbol{\theta}) = \begin{cases} \left[\left(\theta_1^{-1} - 1\right)^{\frac{1}{2(\theta_2-\theta_3)}}\left(1 - \theta_1\right)^{\frac{1}{2(\theta_2-\theta_3)}}\theta_1^{-\frac{1}{2(\theta_2-\theta_3)}} + 1\right]^{-1} & \theta_2 - \theta_3 \neq 0 \\ 0 & \left((\theta_2 = \theta_3) \wedge (\theta_1 > \frac{1}{2})\right) \\ 1 & \text{Otherwise} \end{cases} \tag{4.50}$$

Under this model, the categorical RRH at point $x \in \mathcal{X}$ is

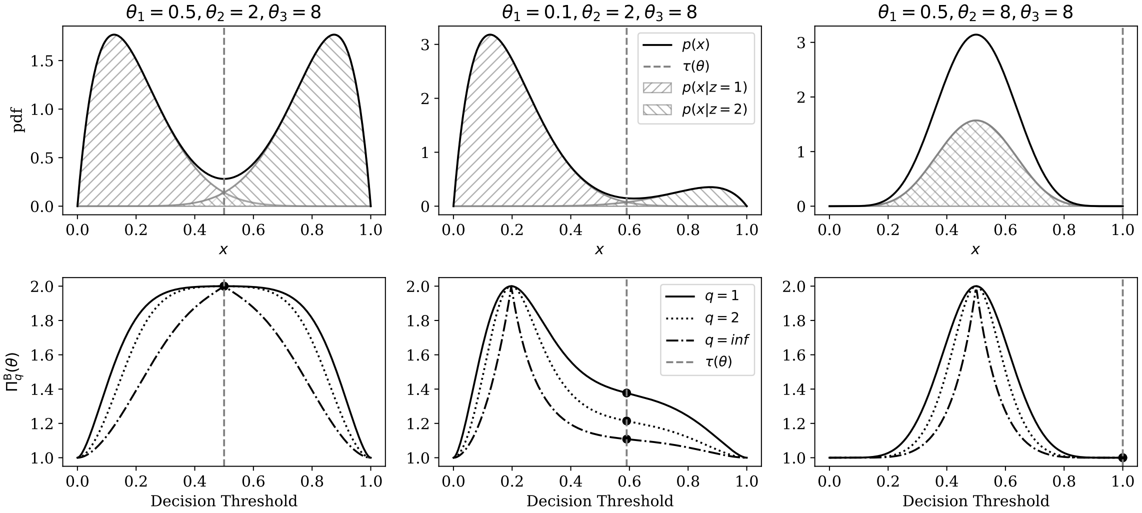


Figure 4.4: Demonstration of data-generating distribution (top row; Equations 4.45-4.47), and relationship between the representational model's decision threshold (Equations 4.48, and 4.50) and categorical representational Rényi heterogeneity (bottom row). The optimal decision boundary (Equation 4.50) is shown as a gray vertical dashed line in all plots. Each column depicts a specific parameterization of the data-generating system (parameters are stated above the top row). **Top Row:** Probability density functions for data-generating distributions. Shaded regions correspond to the two mixture components. Solid black lines denote the marginal distribution (Equation 4.47). The x-axis represents the observable domain, which is the (0,1) interval. **Bottom Row:** Effect of varying categorical representational Rényi heterogeneity (RRH) for $q \in \{1, 2, \infty\}$ across different category assignment thresholds for the beta-mixture models shown in the top row. Varying levels of decision boundary are plotted on the x-axis. The y-axis shows the resulting between-observation RRH. Black dots highlight the RRH computed at the optimal decision boundary.

$$\Pi_q\left(x\right) = \left(\sum_{i=1}^{2} f_{\boldsymbol{\theta}}^q\left(z=i|x\right)\right)^{\frac{1}{1-q}} = \left(\mathbb{1}^q\left[x \leq \tau(\boldsymbol{\theta})\right] + \mathbb{1}^q\left[x > \tau(\boldsymbol{\theta})\right]\right)^{\frac{1}{1-q}} = 1. \qquad (4.51)$$

The expected value of $f_{\boldsymbol{\theta}}(z = 2|x)$ with respect to the data generating distribution (Equation 4.47) is

$$\begin{aligned} \bar{f}_{\boldsymbol{\theta}}(z=2) &= \mathbb{E}_{x\sim p(x)}\left[f_{\boldsymbol{\theta}}(z=2|x)\right] \\ &= \int_0^1 p(x)\mathbb{1}\left[x > \tau(\boldsymbol{\theta})\right]\,\mathrm{d}x \\ &= \int_{\tau(\boldsymbol{\theta})}^1 p(x)\,\mathrm{d}x \\ &= (1-\theta_1)I_x^1\left(\theta_2,\theta_3\right) + \theta_1 I_x^1\left(\theta_3,\theta_2\right), \end{aligned} \qquad (4.52)$$

where $I_{x_0}^{x_1}(a, b)$ is the generalized regularized incomplete beta function. Equation 4.52 implies that $\bar{f}_{\boldsymbol{\theta}}(z = 1) = 1 - \bar{f}_{\boldsymbol{\theta}}(z = 2)$. The pooled heterogeneity is thus expressed as a function of $\boldsymbol{\theta}$ as follows:

$$\Pi_q^{\mathrm{P}}\left(\boldsymbol{\theta}\right) = \begin{cases} \sum_{i=1}^{2}\mathbb{1}[\bar{f}_{\boldsymbol{\theta}}(z=i) > 0] & q=0 \\ \exp\left\{-\sum_{i=1}^{2}\bar{f}_{\boldsymbol{\theta}}(z=i)\log\bar{f}_{\boldsymbol{\theta}}(z=i)\right\} & q=1 \\ \left(\max_i \bar{f}_{\boldsymbol{\theta}}(z=i)\right)^{-1} & q=\infty \\ \left(\sum_{i=1}^{2}\bar{f}_{\boldsymbol{\theta}}^q(z=i)\right)^{\frac{1}{1-q}} & \text{Otherwise} \end{cases}. \qquad (4.53)$$

As a function of $\boldsymbol{\theta}$, the within-group heterogeneity is

$$\begin{aligned} \Pi_q^{\mathrm{W}}\left(\boldsymbol{\theta}\right) &= \left[\int_0^1 \frac{p^q(x)}{\int_0^1 p^q(u)\,\mathrm{d}u}\left(\sum_{i=1}^{2} f_{\boldsymbol{\theta}}(z=i|x)\right)^q\,\mathrm{d}x\right]^{\frac{1}{1-q}} \\ &= \left[\int_0^1 \frac{p^q(x)}{\int_0^1 p^q(u)\,\mathrm{d}u}(1)\,\mathrm{d}x\right]^{\frac{1}{1-q}} \\ &= 1, \end{aligned} \qquad (4.54)$$

and therefore the between-group heterogeneity is $\Pi_q^{\mathrm{B}}\left(\boldsymbol{\theta}\right) = \Pi_q^{\mathrm{P}}\left(\boldsymbol{\theta}\right)$.

Analytic expressions for the existing non-categorical heterogeneity indices $\hat{Q}_e$ (Equation 4.17), F_q (Equation 4.21), and L_q (Equation 4.23) were computed as "best-case" scenarios, as follows. First, the probability distributions over states for all expressions was the true prior distribution (Equation 4.46). Distance matrices—and by extension, the similarity matrix for L_q—were computed using the closed-form expectation of the absolute distance between two beta-distributed random variables (see Appendix A.2 and the Supplementary Materials).

Figure 4.5 compares the categorical RRH against $\hat{Q}_e$, F_q, and L_q for BMM distributions of varying degrees of separation, and across different mixture component weights ($0.5 \leq \theta_1 < 1$). Without significant loss of generality, we show only those comparisons at $q = 1$ (which excludes the numbers equivalent quadratic entropy), and $q = 2$.

The most salient differences between these indices occur when the BMM mixture components completely overlap (i.e., at $\theta_2 = \theta_3$). The RRH correctly identifies that there is effectively only one component, regardless of mixture weights. Only the Leinster–Cobbold index showed invariance to the mixture weights when $\theta_2 = \theta_3$, but it could not correctly identify that data were effectively unimodal.

The other stark difference arose when the mixture components were furthest apart (here when $\theta_2 = 5$ and $\theta_3 = 20$). At this setting, the functional Hill numbers showed a paradoxical increase in the heterogeneity estimate as the prior distribution on components was skewed. The Leinster–Cobbold index was appropriately concave throughout the range of prior weights, but it never reached a value of 2 at its peak (as expected based on the predictions outlined in Section 4.2.2). Conversely, the RRH was always concave and reached a peak of 2 when both mixture components were equally probable.

4.4.2 Representational Rényi Heterogeneity is Scalable to Deep Learning Models

In this example, the observable system X is that of images of handwritten digits defined on an event space $\mathcal{X} = [0, 1]^{784}$ of dimension $n_x = 784$ (the black and white images are flattened from 28×28 pixel matrices into 784-dimensional vectors). Our sample $\mathbf{X} = (x_{ij})_{i=1,2,\dots,N}^{j=1,2,\dots,784}$ from this space is the familiar MNIST training dataset [181] (Figure 4.6), which consists of $N = 60,000$ images roughly evenly distributed across digits $\{0, 1, \dots, 9\}$, and where approximately 10% of all images come from each class. We assume each image carries equal importance, given by a weight vector $\mathbf{w} = (N^{-1})_{i=1,2,\dots,N}$. We are interested in measuring the heterogeneity of X with respect to a continuous latent representation Z defined on event space $\mathcal{Z} = \mathbb{R}^2$. In the present example, this space is simply the continuous 2-dimensional

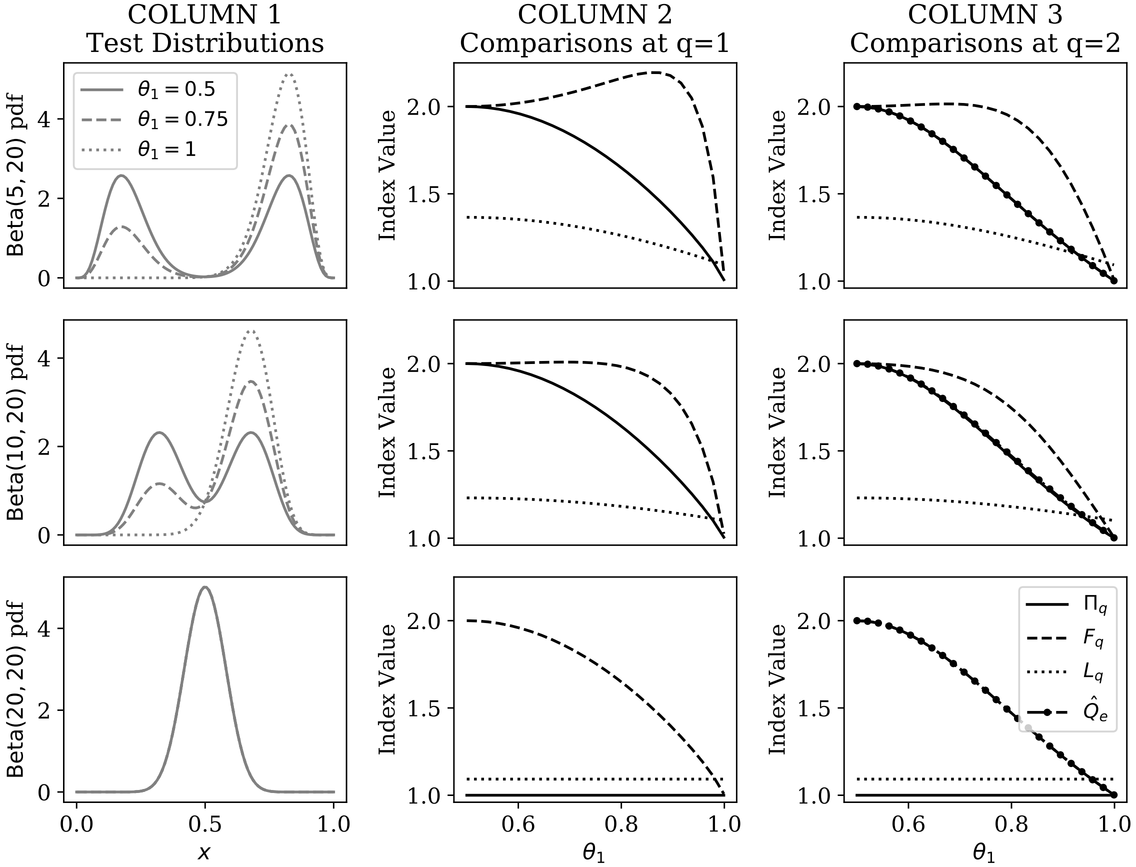


Figure 4.5: Comparison of categorical representational Rényi heterogeneity (Π_q), the functional Hill numbers (F_q), the numbers equivalent quadratic entropy ($\hat{Q}_e$), and the Leinster–Cobbold index (L_q) within the beta mixture model. Each row of plots corresponds to a given separation between the beta mixture components. **Column 1** illustrates the beta mixture distributions upon which indices were compared. The x-axis plots the domain of the distribution (open interval between 0 and 1). The y-axis shows the corresponding probability density. Different line styles in Column 1 provides visual examples of the effect of changing the θ_1 parameter over the range [0.5,1]. **Column 2** compares Π_q (solid line), F_q (dashed line), and L_q (dotted line), each at elasticity $q = 1$. The x-axis shows the value of the $0.5 \leq \theta_1 < 1$ parameter at which the indices were compared. Index values are plotted along the y-axis. **Column 3** compares the indices shown in Column 2, as well as $\hat{Q}_e$ (dot-dashed line).

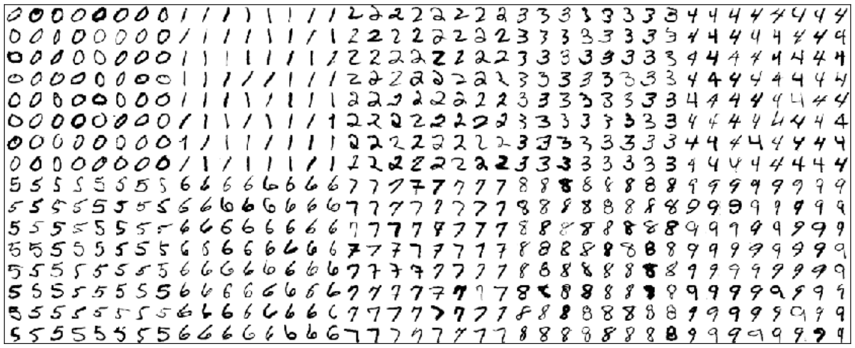


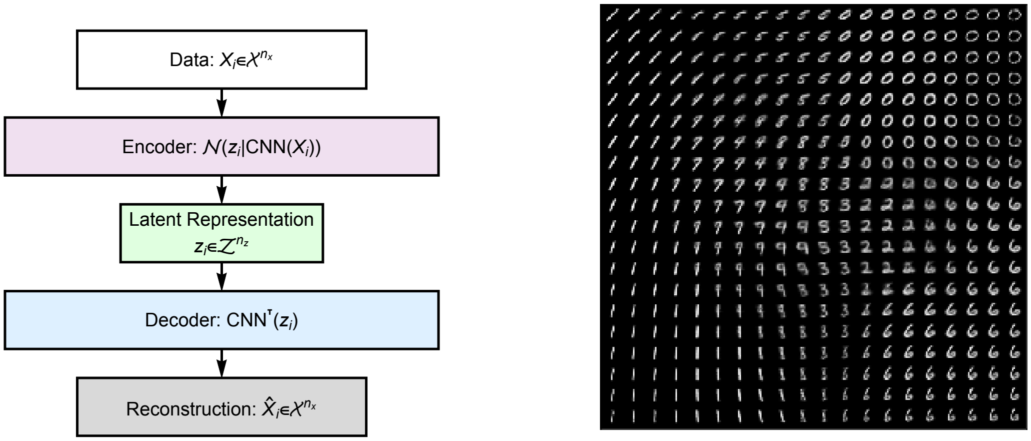
Figure 4.6: Sample images from the MNIST dataset [181].

compression of an image that best facilitates its reconstruction. We choose a dimensionality of $n_z = 2$ for the latent space in order to facilitate a pedagogically useful visualization of the latent feature representation, below. Unlike Section 4.4.1, in the present case we have no explicit representation of the true marginal distribution over the data, $p(\mathbf{x})$.

Having defined the observable and latent spaces, measuring RRH now requires defining a model $f : \mathcal{X} \to \mathcal{P}(\mathcal{Z})$ that maps a (flattened) image vector $\mathbf{x}_i \in \mathcal{X}$ onto a probability distribution over the latent space. Our chosen model is the encoder module of a pre-trained convolutional variational autoencoder (cVAE) provided by the Smart Geometry Processing Group at University College London (Figure 4.7) [182, 188]:

$$f_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_i) = \mathcal{N}\left(\mathbf{z}|\mathbf{m}(\mathbf{x}_i), \mathbf{C}(\mathbf{x}_i)\right) \tag{4.55}$$

where $\boldsymbol{\phi}$ are the encoder's parameters, which specify a convolutional neural network (CNN) whose output layer returns a 2×1 mean vector $\mathbf{m}(\mathbf{x}_i)$ and a 2×1 log-variance vector $\mathbf{s}(\mathbf{x}_i)$ given $\mathbf{x}_i$. For simplicity, we denote the latter as the 2×2 diagonal covariance matrix $\mathbf{C}(\mathbf{x}_i) = \left(e^{s_j(\mathbf{x}_i)}\delta_{jk}\right)_{j=1,2}^{k=1,2}$. Further details of the cVAE and its training can be found in Kingma and Welling [182, 188], although the specific implementation in this paper was a pre-trained implementation by the Smart Geometry Processing Group at University College London. Briefly, the cVAE learns to generate a compressed latent representation (via encoder $f_{\boldsymbol{\phi}}$, which is an approximate posterior distribution) that contains enough information about the input $\mathbf{x}_i$ to facilitate its reconstruction by a "decoder" module. The objective function is a lower bound on the model evidence $p(\mathbf{x})$, which if maximized is equivalent to minimizing the Kullback–Leibler divergence between the approximate and true (but unknown) posteriors $f_{\boldsymbol{\phi}}$ and $p(\mathbf{z}|\mathbf{x})$, respectively.


Data: $X_i \in \mathcal{X}^{n_x}$
Encoder: $\mathcal{N}(z_i|\mathrm{CNN}(X_i))$
Latent Representation
$z_i \in \mathcal{Z}^{n_z}$
Decoder: $\mathrm{CNN}^\top(z_i)$
Reconstruction: $\hat{X}_i \in \mathcal{X}^{n_x}$

(a) Schematic of the model architecture.

(b) Visualization of the two-dimensional latent space.

Figure 4.7: **Panel A:** Illustration of the convolutional variational autoencoder (cVAE) [182]. The computational graph is depicted from top to bottom. An n_x-dimensional input data $\mathbf{X}_i$ (white rectangle) is passed through an encoder (in our experiment this is a convolutional neural network, CNN) which parameterizes an n_z-dimensional multivariate Gaussian over the coordinates $\mathbf{z}_i$ for the image's embedding on the latent space $\mathcal{Z} = \mathbb{R}^{n_z}$. The latent embedding can then be passed through a decoder (blue rectangle) which is a neural network employing transposed convolutions (here denoted $\mathrm{CNN}^\top$) to yield a reconstruction $\hat{\mathbf{X}}_i$ of the original input data. The objective function for this network is a variational lower bound on the model evidence of the input data (see Kingma & Welling [182] for details). **Panel B:** Depiction of the latent space learned by the cVAE. This model was a pre-trained model from the Smart Geometry Processing Group at University College London.

The continuous RRH under the model in Equation 4.55 for a single example $\mathbf{x}_i \in \mathcal{X}$ can be computed by merely evaluating the Rényi heterogeneity of a multivariate Gaussian (Equation 4.43 in Example 4) for the covariance matrix given by $\mathbf{C}(\mathbf{x}_i)$. This is interpreted as the effective area of the 2-dimensional latent space consumed by representation of $\mathbf{x}_i$.

Since the handwritten digit images belong to groups of "Zeros, Ones, Twos, . . . , Nines," this section will call the quantity Π_q^{W} the within-observation heterogeneity (rather than the "within-group" heterogeneity) in order to avoid its interpretation as measuring the heterogeneity of a group of digits. Rather, it is interpreted as the effective area of latent space consumed by representation of a single observation $\mathbf{x} \in \mathcal{X}$ on average. It is computed by evaluation of Equation 4.44 at $\mathbf{C}(\mathbf{X}) = \{\mathbf{C}(\mathbf{x}_i)\}_{i=1,2,\dots,N}$, given uniform weights on samples.

Finally, to compute the pooled heterogeneity Π_q^{P}, we use the parametric pooling approach detailed in Example 4, wherein the pooled distribution is a multivariate Gaussian with mean and covariance given by Equations 4.41 and 4.42, respectively. The pooled heterogeneity is then merely Equation 4.43 evaluated at $\mathbf{C}_*(\mathbf{X})$, and represents the total amount of area in the latent space consumed by the representation of X under $f_{\boldsymbol{\phi}}$. The effective number of observations in X with respect to the continuous latent representation Z is, therefore, given by the between-observation heterogeneity:

$$\Pi_q^{\mathrm{B}}\left(\mathbf{C}(\mathbf{X}), \mathbf{w}\right) = \frac{\Pi_q^{\mathrm{P}}\left(\mathbf{C}_*(\mathbf{X})\right)}{\Pi_q^{\mathrm{W}}\left(\mathbf{C}(\mathbf{X}), \mathbf{w}\right)}. \tag{4.56}$$

Equation 4.56 gives the effective number of observations in X because it uses the entire sample $\mathbf{X}$ (of course, assuming $\mathbf{X}$ provides adequate coverage of the observable event space). However, one could compute the effective number of observations in a subset of $\mathbf{X}$, if necessary. Let $\mathbf{X}^{(j)} = (\mathbf{x}_k)_{k=1,2,\dots,N_j}$ be the subset of N_j points in $\mathbf{X}$ found in the observable subspace $\mathcal{X}_j \subset \mathcal{X}$ (such as the subspace of MNIST digits corresponding to a given digit class). Given corresponding weights $\mathbf{w}^{(j)} = \left(N_j^{-1}\right)_{k=1,2,\dots,N_j}$, Equation 4.56 is then simply

$$\Pi_q^{\mathrm{B}}\left(\mathbf{C}(\mathbf{X}^{(j)}), \mathbf{w}^{(j)}\right) = \frac{\Pi_q^{\mathrm{P}}\left(\mathbf{C}_*(\mathbf{X}^{(j)})\right)}{\Pi_q^{\mathrm{W}}\left(\mathbf{C}(\mathbf{X}), \mathbf{w}^{(j)}\right)}. \tag{4.57}$$

Figure 4.8 shows the effective number of observations in the subsets of MNIST images belonging to each image class, under the continuous representation learned by the cVAE. One can appreciate that the MNIST class of "Ones" (in the training set) has the smallest

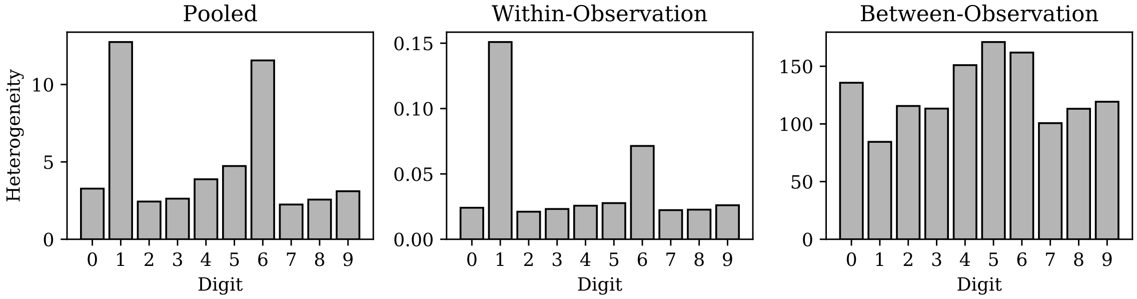
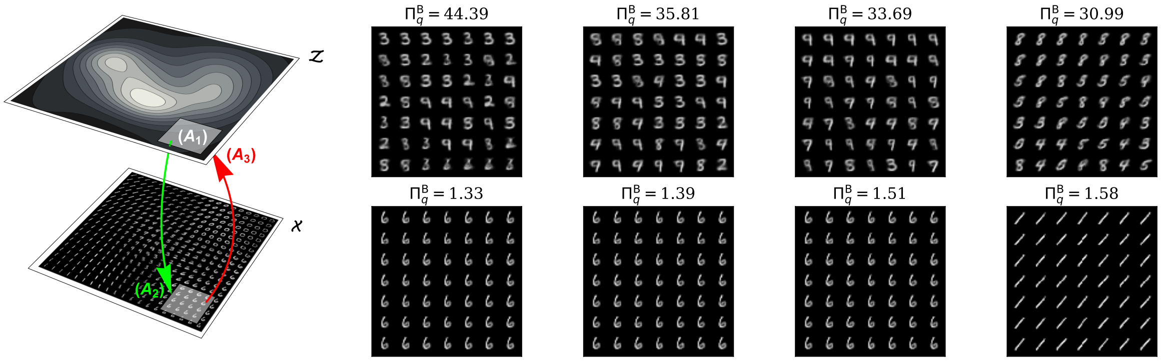


Figure 4.8: Heterogeneity for the subset of MNIST training data belonging to each digit class respectively projected onto the latent space of the convolutional variational autoencoder (cVAE). The leftmost plot shows the pooled heterogeneity for each digit class (the effective total area of latent space occupied by encoding each digit class). The middle plot shows the within-observation heterogeneity (the effective total area of latent space per encoded observation of each digit class, respectively). The rightmost plot shows the between-observation heterogeneity (the effective number of observations per digit class). Recall that Rényi heterogeneity on a continuous distribution gives the effective size of the domain of an equally heterogeneous uniform distribution on the same space, which explains why the within-observation heterogeneity values here are less than 1.

effective number of observations. Subjective visual inspection of the MNIST samples in Figure 4.6 may suggest that the Ones are indeed relatively more homogeneous as a group than the other digits (this claim is given further objective support in Appendix A.3 based on deep similarity metric learning [189, 190]).

Figure 4.9 demonstrates the correspondence of between-observation heterogeneity (i.e., the effective number of observations) and the visual diversity of different samples from the latent space of our cVAE model. For each image in the MNIST training dataset, we computed the effective location of its latent representation: $\mathbf{m}(\mathbf{x}_i)$ for $i \in \{1, 2, \ldots, N\}$. For each of these image representations, we defined a "neighborhood" including the 49 other images whose latent coordinates were closest in Euclidean distance (which is sensible on the latent space given the Gaussian prior). For all such neighbourhoods defined, we then reconstructed the corresponding images on $\mathcal{X}$, whose between-observation heterogeneity was then computed using Equation 4.57. Figure 4.9b shows the estimated effective number of observations for the latent neighborhoods with the greatest and least heterogeneity. One can appreciate that neighborhoods with Π_q^{B} close to 1 include images with considerably less diversity than neighborhoods with Π_q^{B} closer to the upper limit of 49. These data suggest that the between-observation heterogeneity—which is the effective number of observations in X



(a) Illustration of analysis.

(b) Heterogeneity of patches in the latent space.

Figure 4.9: Visual illustration of MNIST image samples corresponding to different levels of representational Rényi heterogeneity under the convolutional variational autoencoder (cVAE). **Panel (a)** illustrates the approach to this analysis. Here, the surface $\mathcal{Z}$ shows hypothetical contours of a probability distribution over the 2-dimensional latent feature space. The surface $\mathcal{X}$ represents the observable space, upon which we have projected an "image" of the latent space $\mathcal{Z}$ for illustrative purposes. We first compute the expected latent locations $\mathbf{m}(\mathbf{x}_i)$ for each image $\mathbf{x}_i \in \mathcal{X}$. **($\mathbf{A_1}$)** We then define the latent neighbourhood of image $\mathbf{x}_i$ as the 49 images whose latent locations are closest to $\mathbf{m}(\mathbf{x}_i)$ in Euclidean distance. **($\mathbf{A_2}$)** Each coordinate in the neighbourhood of $\mathbf{m}(\mathbf{x}_i)$ is then projected onto a corresponding patch on the observable space of images. **($\mathbf{A_3}$)** These images are then projected as a group back onto the latent space, where Equation 4.57 can be applied, given equal weights over images, to compute the effective number of observations in the neighbourhood of $\mathbf{x}_i$. **Panel (b)** plots the most and least heterogeneous neighbourhoods so that we may compare the estimated effective number of observations with the visually appreciable sample diversity.

with respect to the latent features learned by a cVAE—can indeed correspond to visually appreciable sample diversity.

4.5 Discussion

This paper introduced representational Rényi heterogeneity, a measurement approach that satisfies the replication principle [21–23] and is decomposable [59] while requiring neither a priori (A) categorical partitioning nor (B) specification of a distance function on the input space. Rather, the experimenter is free to define a model that maps observable data onto a semantically relevant domain upon which Rényi heterogeneity may be tractably computed, and where a distance function need not be explicitly manipulated. These properties facilitate heterogeneity measurements for several new applications. Compared to state-of-the-art comparator indices under a beta mixture distribution, RRH more reliably quantified the number of unique mixture components (Section 4.4.1), and under a deep generative model of image data, RRH was able to measure the effective number of distinct images with respect to latent continuous representations (Section 4.4.2). In this section, we further synthesize our conclusions, discuss their implications, and highlight open questions for future research.

The main problem we set out to address was that all state of the art numbers equivalent heterogeneity measures (Section 4.2.2) require a priori specification of a distance function and categorical partitioning on the observable space. To this end, we showed that RRH does not require categorical partitioning of the input space (Section 4.3). Although our analysis under the two-component BMM assumed that the number of components was known, RRH was the only index able to accurately identify an effectively singular cluster (i.e., where mixture components overlapped; Figure 4.5). We also showed that the categorical RRH did not violate the principle of transfers [85, 86] (i.e., it was strictly concave with respect to mixture component weights), unlike the functional Hill numbers (Figure 4.5). Future studies should extend this evaluation to mixtures of other distributional forms in order to better characterize the generalizability of our conclusions.

Sections 4.3.1 and 4.3.2 both showed that RRH does not require specification of a distance function on the observable space. Instead, one must specify a model that maps the observable space onto a probability distribution over the latent representation. This is beneficial since input space distances are often irrelevant or misleading. For example, latent representations of image data learned by a convolutional neural network will be robust to translations of the inputs since convolution is translation invariant. However, pairwise distances on the observable space will be exquisitely sensitive to semantically irrelevant translations of input data. Furthermore, semantically relevant information must often be learned from raw data using hierarchical abstraction. Ultimately, when (A) pre-defined distance metrics are sensitive to noisy perturbations of the input space, or (B) the relevant semantic content of some input data is best captured by a latent abstraction, the RRH measure will be particularly useful. However, we emphasize that RRH does not limit the number of assumptions required when measuring heterogeneity, but rather shifts the assumptions onto the domain-specific mapping function, rather than to the heterogeneity measure itself. In RRH, heterogeneity is always the effective number of states in a system with respect to a latent/transformed representation. The assumptions are now primarily related to definition of that representation. To this end, a representational model's stability and appropriateness must be validated by the investigator using domain-specific methods.

The requirement of specifying a representational model $f : \mathcal{X} \to \mathcal{P}(\mathcal{Z})$ implies the additional problem of model selection. In Section 4.3, we noted that the determination of whether a model is appropriate must be made in a domain-specific fashion. For instance, the method by which ecologists assign species labels prior to measurement of species diversity implies the use of a mapping from the observable space of organisms to a degenerate distribution over species labels (Example 3). In Section 4.4.2, we used the encoder module of a cVAE (a generative model based on a convolutional neural network architecture [182, 188]) to represent images as 2-dimensional real-valued vectors in order to demonstrate our ability to capture variation in digits' written forms (see Figures 4.7B and 4.9). Someone concerned with measuring heterogeneity of image batches in terms of the digit-class distribution could choose a categorical latent representation corresponding to the digit classes (this would return the effective number of digit classes per sample). Regardless, the model used to map between observations and the latent space should be validated using either explanatory power (e.g., maximization of a lower bound on the model evidence), generalizability (e.g., out of sample predictive power), or another approach that is justifiable within the investigator's scientific domain of interest.

In addition to the results of empirical applications of RRH in Section 4.4, we were also able to show that RRH generalizes the process by which species diversity and indices of economic equality are computed (Example 3). In doing so, we are able to clarify some of the assumptions inherent in those indices. Specifically, that assignment of species or ownership labels (in ecological and economic settings, respectively) corresponds to mapping from an observable space, such as the space of organisms' identifiable features or the space of economic resources, onto a degenerate distribution over the categorical labels (Table 4.2). It is possible that altering the form of that mapping may yield new insights about ecological and economic diversity.

In conclusion, we have introduced an approach for measuring heterogeneity that requires neither (A) categorical partitioning nor (B) distance measure on the observable space. Our RRH method enables measurement of heterogeneity in disciplines where categorical entities are unreliably defined, or where relevant semantic content of some data is best captured by a hierarchical abstraction. Furthermore, our approach includes many existing heterogeneity indices as special cases, while facilitating clarification of many of their assumptions. Future work should evaluate the RRH in practice and under a broader array of models.

Chapter 5

Prediction of Lithium Response Using Clinical Data[1]

Abstract. Promptly establishing maintenance therapy could reduce morbidity and mortality in patients with bipolar disorder. Using a machine learning approach, we sought to evaluate whether lithium responsiveness (LR) is predictable using clinical markers. Using the largest existing sample of direct interview-based clinical data from lithium treated patients (n=1266, 34.7% responders), collected across 7 sites, internationally, we trained a random forest model to classify LR—as defined by the previously validated Alda scale—against 180 clinical predictors. Under appropriate cross-validation procedures, LR was predictable in the pooled sample with an area under the receiver operating characteristic curve of 0.80 (95% CI 0.78-0.82) and a Cohen's kappa of 0.46 (0.4-0.51). The model demonstrated a particularly low false positive rate (specificity 0.91 [0.88-0.92]). Features related to clinical course and the absence of rapid cycling appeared consistently informative. Clinical data can inform out-of-sample LR prediction to a potentially clinically relevant degree. Despite the relevance of clinical course and the absence of rapid cycling, there was substantial between-site heterogeneity with respect to feature importance. Future work must focus on improving classification of true positives, better characterizing between- and within-site heterogeneity, and further testing such models on new external datasets.

5.1 Introduction

Bipolar disorder (BD) is a severe neuropsychiatric disorder for which lithium treatment has been a mainstay for over 60 years [191]. Treatment selection currently depends on empirical trials, yet only 30% of patients treated with lithium will be fully responsive in the long term [192]. This trial and error approach may further compound the approximate 9-10 years between symptomatic onset until treatment with a mood stabilizer [193]. If a given patient will ultimately be a lithium responder, it would be of interest to predict this early in order to stabilize him or her expediently. Conversely, if the patient is unlikely to be a lithium responder, prediction as such could avoid exposure to lithium's non-trivial side-effects.

[1] Nunes A, Ardau R, Berghöfer A, Bocchetta A, Chillotti C, Deiana V, Garnham J, Grof E, Hajek T, Manchia M, Müller-Oerlinghausen B, Pinna M, Pisanu C, O'Donovan C, Severino G, Slaney C, Suwalska A, Zvolsky P, Cervantes P, Del Zompo M, Grof P, Rybakowski JK, Tondo L, Trappenberg T, and Alda M. Prediction of Lithium Response using Clinical Data. *Acta Psychiatrica Scandinavica.* 2019;In Press.

Great efforts have pursued biological predictors of lithium response (LR). Although genetic studies have found some promising associations [30], the ability of models trained on genomic data to predict LR in previously unseen patients remains unclear. Recently, interesting neurophysiological markers related to hyperexcitability of neurons re-programmed from patient-derived stem-cells have emerged [194, 195]. However, these methods are resource and time intensive, and further testing is required to incorporate them into routine clinical practice, since existing studies include very small sample sizes of highly selected patients. It is therefore of interest to search for inexpensive markers that may be readily available to clinicians at little to no additional cost beyond that of the necessary clinical interview.

Information from clinical interview is readily available to all clinicians, and associations between clinical factors and LR have been demonstrated. In the most recent meta-analysis of LR in BD, Hui et al. [196] found that good LR was associated with the mania-depression polarity sequence, absence of psychotic symptoms, shorter duration of illness prior to lithium initiation, family history of BD, and later onset of illness. Other correlates have included low-rates of psychiatric comorbidity [197], clinical course characterized by clear episodes marked with good inter-episode functioning [36], and family history of LR [198] (although this did not reach meta-analytic statistical significance in Hui et al. [196]). The LR phenotype is often deemed "classical" in terms of its resemblance to the original Kraepelinian descriptions [191]. However, Hui et al. [196] and Kessing [199], in his accompanying editorial, note that the existing body of literature on clinical predictors of LR warrants a cautious interpretation: there appears to be substantial meta-analytic heterogeneity, sample sizes have tended to be small, and response has often not been defined using validated instruments. Therefore, the predictive utility of clinical information is promising but warrants further scrutiny.

Hui et al. [196] identified the lack of multivariate analyses as an important gap in the literature that limits our understanding of clinical prediction of LR. This is in large part a result of nearly all studies using the orthodox statistical paradigm, which generally admits application of neither high-dimensional nor non-linear models. Moreover, we add that the discovery of "statistically significant" correlates of LR in associational models does not necessarily translate into a model with strong out-of-sample predictive power. The machine learning (ML) paradigm is uniquely well posed to address this problem. Specifically, the evaluation of model performance is entirely concerned with a model's predictive performance on previously unobserved data. As a consequence, the models typically used in ML make simultaneous use of many predictor variables and the (potentially small) relationships between them.

One previous study attempted to predict LR using a ML approach in 192 patients from the Bipolar CHOICE trial [200, 201]. Unfortunately, the sample size and duration of treatment were limited. Moreover, their outcome incorporated no information about individual change in symptoms over the treatment period, and their predictors could not capture many of the essential features of the previously demonstrated "classical" phenotype. Finally, their model performance was assessed in terms of the proportion of variance in the Clinical Global Impressions-Bipolar version (CGI-BP) explained, which does not clearly answer the question of most relevance to patients and psychiatrists: "what is the probability that this patient will respond to lithium?"

The present study evaluated the capacity of a ML approach based on clinical interview-based data to predict LR in BD. Our study was performed on the largest-ever cohort of lithium-treated bipolar patients (n=1266) sourced from 7 international specialist clinics, with a minimum treatment duration of 1 year, and using a validated response scale [202]. We hypothesized that clinical information would indeed offer predictive performance in exceedance of chance, and that the features most relevant to predictive performance would be reflective of the "classical" bipolar phenotype.

5.2 Material and Methods

5.2.1 Data Collection

Seven cohorts of men and women with bipolar I or bipolar II disorder were included in our dataset (Table 5.1). To be eligible, patients had to be treated with lithium as their principal mood stabilizer for a minimum of one year. Clinical assessments followed a strict procedure. After providing informed consent, participants were interviewed using one of the structured or semi-structured interviews (SADS-L, SCID or DIGS). Clinical diagnosis was confirmed by DSM-IV criteria. We also used available medical records, narrative summaries of all interviews, and details such as baseline assessments, clinical course, response to treatment, treatment adherence, psychiatric and medical comorbidities, history of suicidal behaviour, and symptom profiles in OPCRIT format [203].

Table 5.1: Description of constituent datasets. *Abbreviations*: number of patients (N), lithium responders (LR+), Cagliari (Centro Bini; CB), Cagliari (University; CU), International Group for the Study of Lithium (IGSLi), Maritimes (MAR), Ontario (ON), Poznan (POZ).

Sample	**N (LR+)**	**Description**
CB	324 (21%)	Patients followed at the Mood Disorder Lucio Bini Center in Cagliari, Italy. Clinical data collection and response assessment was done by two psychiatrists.
CU	206 (29%)	Patients in the long term treatment program at the Lithium Clinic of the Unit of the Clinical Pharmacology Center, University Hospital of Cagliari, Italy. Clinical data collection and response assessment was done by three psychiatrists and three clinical psychopharmacologists.
IGSLi	70 (100%)	Patients recruited for a genetic study of lithium responsive bipolar disorder. [204] By design of that study, all patients were lithium responders. Clinical data collection and response assessment was done by three psychiatrists.
MAR	343 (20%)	Patients followed by the Mood Disorders program at the Nova Scotia Health Authority and the Maritime Bipolar Registry. Clinical data collection and response assessment was done by two psychiatrists and two research nurses working in pairs.
MTL	95 (16%)	Patients followed by the Mood Disorders Program at the McGill University Health Centre. Clinical data collection and response assessment was done by one psychiatrist.
ON	117 (84%)	Patients from our earlier studies of lithium responsive bipolar disorder, [204, 205] which, like the IGSLi sample, explains the greater proportion of responders. Clinical data collection and response assessment was done by three psychiatrists (including MA, who is now in the Maritimes).
POZ	111 (53%)	Patients followed longitudinally by the Psychiatry Department at the University of Poznan, Poland. Clinical data collection and response assessment was done by two psychiatrists.

For uniform evaluation of treatment response, we used all available information including data from clinical records, diagnostic interviews, and prospective follow-up assessed by NIMH Life-Chart Method converted to a score on a scale previously validated and adopted by a number of research groups including the genome-wide association study of LR by the ConLiGen consortium [30, 206]. This is also known as the Alda score, which has a range of 0 to 10, with scores of 7 and higher considered a good response. All centres underwent inter-rater reliability training in preparation for the ConLiGen study, and achieved a weighted kappa of 0.75 (for dichotomous response) and an intraclass correlation of 0.96

for the total score (a continuous measure of response; data extracted from Manchia et al. [202]).

5.2.2 Statistical Analysis

Demographic Statistics

All data were anonymized and aggregated. Demographic statistics were studied within and between sites. Continuous variables were compared between using two-sample permutation tests of independence using the `perm` package in R [207]. Categorical variables were compared using the randomization chi-squared test. Demographic descriptive statistics are presented in tabular format for lithium responders vs. lithium non-responders in the aggregate sample. We also present comparisons across alternative stratifications in Appendix B.1.

Classification Analyses

Our study is split into four phases: (A) analysis of data pooled across sites (henceforth the aggregate analysis), (B) analysis of data within sites (henceforth the site-level analysis), (C) an analysis in which we attempted to classify patients from one site with a model trained on data from all other sites (henceforth the predict-one-site-out analysis), and (D) a leave-one-site-out analysis in which we repeat the aggregate analysis $n_{sites} = 7$ times, each run leaving data from one of the sites out. The predict-one-site-out analysis evaluates the degree to which a given site's "signal" is present in the remainder of the data, whereas the leave-one-site-out analysis evaluates the degree to which a given site's data contributes to the aggregate performance.

After pruning for variables with missingness $> 40\%$, 138 predictors remained in the dataset, all of which would be available to clinicians prior to lithium prescription. Lithium response was defined as the dichotomized Alda score.

The random forest classifier (RFC) [208] pools classification and regression trees in order to reduce their variance. We used the RFC model included in SciKit-Learn v.0.20.2, [209] with the number of estimators set to 100 *a priori*. Sensitivity analysis (Appendix B.2) showed that our results were not improved by hyperparameter optimization or alternative architectures.

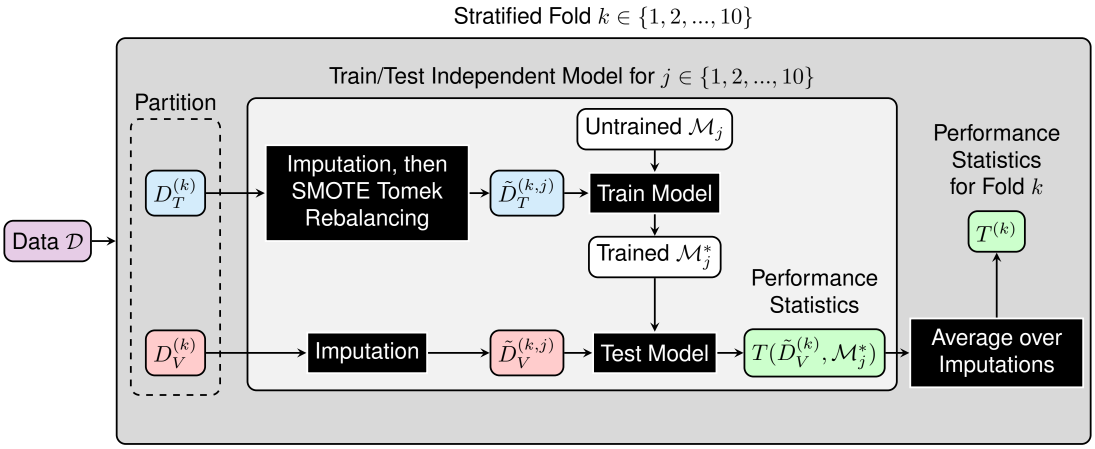

Stratified Fold $k \in \{1, 2, ..., 10\}$
Train/Test Independent Model for $j \in \{1, 2, ..., 10\}$
Partition
Data $\mathcal{D}$
$D_T^{(k)}$
$D_V^{(k)}$
Imputation, then SMOTE Tomek Rebalancing
$\tilde{D}_T^{(k,j)}$
Untrained $\mathcal{M}_j$
Train Model
Trained $\mathcal{M}_j^*$
Imputation
$\tilde{D}_V^{(k,j)}$
Test Model
Performance Statistics
$T(\tilde{D}_V^{(k)}, \mathcal{M}_j^*)$
Average over Imputations
Performance Statistics for Fold k
$T^{(k)}$

Figure 5.1: Graphical illustration of classification protocol. For each fold k of the stratified K-fold cross-validation procedure, we began by partitioning the data $\mathcal{D}$ into training $D_T^{(k)}$ and validation $D_V^{(k)}$ subsets. Within each fold, for $j \in \{1, 2, \ldots, 10\}$ imputations of the training partition, denoted $\tilde{D}_T^{(k)}$, we learned an independent instance of the classifier—where $\mathcal{M}_j^*$ denotes the trained model—and tested it on the imputed validation partition of fold k, returning performance statistics $T(\tilde{D}_V^{(k)}, \mathcal{M}_j^*)$. Note that imputation was done within training and testing sets, respectively. Averaging over the independent imputation runs results in the final performance statistics for fold k, denoted as $T^{(k)}$. In the site-level analysis, the data $\mathcal{D}$ simply consisted of the data for a specific site in our dataset. For the predict-one-site-out analysis, there were $n_{sites} = 7$ folds, where each site was designated as the validation set in one (and only one) fold.

Figure 5.1 summarizes the analysis. For the aggregate and site-level analyses, model criticism was done under stratified cross-validation (10 folds in the aggregate analysis). Given significant missingness in these data, within each fold we sampled missing values tenfold and uniformly over the respective variables' domains. The sampling distribution on missing values was uninformative and independent of observed data, which was the most conservative approach to marginalization of the missing values. Appendix B.3 reports a mixed-effects meta-regression showing that prediction errors in the aggregate analysis were unrelated to missingness.

To minimize the risk of simply predicting the prevalent class, training partitions were rebalanced using the *Synthetic Minority Oversampling Technique* (SMOTE) with a Tomek link function in the imbalanced-learn Python package (v.0.4.3) [210, 211]. To provide a multifaceted view of model capacity, classification performance was measured using area under the receiver operating characteristic curve (ROC-AUC), accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and Cohen's Kappa. We also evaluated the Brier score loss, which is the mean squared error between models' probabilistic predictions and observed classes. Lower Brier scores are indications of improvements in both predictive power and confidence. A Brier score close to 0.25 suggests that the model's predictions tend to rely on a hard threshold close to 50%, rather than being informative about the broader range of response probabilities.

The site-level analysis was identical to the aggregate analysis, with the exception that the number of folds was computed site-wise, such that each validation partition would have 2 cases. Similarly, in the predict-one-site-out analysis, there were $n_{sites} = 7$ folds, such that each site's data served as the validation set in one (and only one) fold.

Feature importance values were extracted for the RFC model and their expectations computed over folds. Unfortunately, feature importance values from RFC models cannot describe whether a variable is associated with increases or decreases of the probability of LR. However, to evaluate the consistency of features' informativeness, we repeated the entire analysis above with the target variable now defined as lithium non-response (Alda score < 4); these results are shown in Appendix B.4. We also repeated the entire analysis using classifiers trained only on two of the more salient variables emerging from our analysis (Appendix B.5).

5.2.3 Role of Funding Source

Funding agencies had role in neither design, analysis, nor interpretation of results, nor composition or review of the manuscript.

5.3 Results

5.3.1 Demographic Statistics

Table 5.2 presents demographic statistics. A total of 439/1266 patients (34.7%) were full lithium responders and 827 (65.3%) were non-responders. Cagliari (Centro Bini), Montreal, and the Maritimes each had approximately 80% non-responders. The Cagliari (University) sample had 29% responders, the IGSLi sample consisted entirely of lithium responders, and the Ottawa/Hamilton sample contained 84% responders. The Poznan sample was roughly balanced (53% responders). There were statistically significant univariate differences between responders and non-responders across many variables, which for parsimony we will not list here, since our central analysis is primarily concerned with multivariate prediction (the reader is referred to Table 5.2).

Table 5.2: Demographic descriptive statistics stratified by lithium response. Abbreviations: N (number or count), "with" (w/) Li(+) (lithium responder), Li(-) (lithium non-responder), BD (bipolar disorder), BD-I (bipolar I disorder), BD-II (bipolar II disorder), NOS (not otherwise specified), MDD (major depressive disorder), SZA (schizoaffective disorder), FDR (first degree relative), SDR (second degree relative) GAF (global assessment of functioning scale), SA (suicide attempt) SI (suicidal ideation), SES (socioeconomic status), UI (unemployment insurance). Normally distributed variables are represented as mean (standard deviation), while non-normally distributed variables are represented as median [interquartile range, IQR]. Categorical variables are represented as count (percentage), with all unique categories listed; where a categorical variable has no subheadings identifying the categories, it is implicitly a binary variable where the count (percentage) refers to the affirmative response of the variable.

Variable	**Li(-)**	**Li(+)**	**p**
N	827	439	
Male (%)	313 (37.8)	176 (40.1)	0.465
Age (y)	43.50 [32.42, 54.96]	50.45 [39.20, 63.34]	0.002
Diagnosis (%)			0.029
BD-I	540 (65.4)	290 (66.1)	
BD-II	218 (26.4)	124 (28.2)	
BD NOS	1 (0.1)	0 (0.0)	
MDD Recurrent	4 (0.5)	8 (1.8)	
MDD Single	2 (0.2)	1 (0.2)	
Not 1° mood disorder	1 (0.1)	0 (0.0)	
SZA	60 (7.3)	16 (3.6)	
Age of onset (y)	21.00 [17.00, 28.00]	25.00 [19.00, 33.00]	0.002
Onset of depression (y)	24.00 [18.15, 33.00]	28.00 [20.00, 36.00]	0.004
Onset of mania (y)	27.00 [21.00, 37.00]	30.00 [22.00, 38.75]	0.026
Onset of hypomania (y)	30.00 [22.00, 42.00]	35.74 [25.00, 44.50]	0.01

Continued on next page...

Variable	Li(-)	Li(+)	p
Polarity at first episode (%)			0.005
Biphasic DM	7 (1.9)	11 (4.7)	
Biphasic MD	23 (6.3)	12 (5.1)	
Hypomania	43 (11.7)	20 (8.5)	
Major depression	220 (60.1)	116 (49.4)	
Mania	54 (14.8)	52 (22.1)	
Minor depression	13 (3.6)	22 (9.4)	
Mixed	3 (0.8)	1 (0.4)	
Periodic rapid cycling	3 (0.8)	1 (0.4)	
Clinical course (%)			<0.001
Chronic	62 (7.8)	37 (13.5)	
Chronic deteriorating	16 (2.0)	2 (0.7)	
Chronic fluctuating	215 (26.9)	54 (19.6)	
Completely episodic	269 (33.6)	138 (50.2)	
Continuous cycling	32 (4.0)	8 (2.9)	
Episodic + residual	193 (24.1)	35 (12.7)	
Single episode	13 (1.6)	1 (0.4)	
N Lifetime manias	3.00 [1.00, 7.00]	2.00 [1.00, 4.00]	0.002
N Lifetime depressions	4.00 [2.00, 9.00]	3.00 [2.00, 6.00]	0.002
N Lifetime mixed	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.002
N Lifetime multiphasic	0.00 [0.00, 0.00]	0.00 [0.00, 1.00]	0.002
Total lifetime episodes	8.00 [5.00, 16.00]	6.00 [5.00, 11.00]	0.002
Rapid cycling (%)			<0.001
Never	273 (68.2)	222 (96.1)	
Only on Antidepressants	17 (4.2)	1 (0.4)	
Spontaneous	110 (27.5)	8 (3.5)	
Rapid mood switching (%)	64 (34.8)	8 (8.5)	<0.001
Lifetime psychosis (%)			<0.001
In episodes congruent	219 (40.6)	73 (32.3)	
In episodes incongruent	72 (13.3)	15 (6.6)	
Never	240 (44.4)	137 (60.6)	
Outside mood episodes	9 (1.7)	1 (0.4)	

Continued on next page...

Variable	Li(-)	Li(+)	p
GAF at last assessment	70.00 [60.00, 80.00]	90.00 [80.00, 95.00]	0.002
Total ALDA score	3.00 [0.00, 5.00]	8.00 [7.00, 9.00]	0.002
N episodes on Li	3.00 [1.00, 6.00]	0.00 [0.00, 1.00]	0.002
N episodes pre Li	4.00 [2.00, 7.00]	5.00 [3.00, 8.00]	0.052
N SA	0.00 [0.00, 1.00]	0.00 [0.00, 0.00]	0.026
N Potentially lethal SA	0.00 [0.00, 1.00]	0.00 [0.00, 0.00]	0.006
Age at first SA (y)	30.30 [20.00, 40.00]	36.00 [22.00, 41.55]	0.288
Mood disorder in FDR	321 (59.1)	153 (51.7)	0.047
Any FDR w/ BD	215 (33.4)	128 (34.8)	0.676
N FDR w/ BD-I	0.00 [0.00, 0.00]	0.00 [0.00, 1.00]	0.282
N FDR w/ BD-II	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.372
N FDR w/ MDD	0.00 [0.00, 1.00]	0.00 [0.00, 1.00]	0.3
N FDR w/ SZA	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	1
N FDR w/ SCZ	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.018
N FDR w/ Anxiety disorder	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.214
N FDR unaffected	1.00 [0.00, 3.00]	0.00 [0.00, 1.00]	0.004
N FDR completed suicide	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.272
N FDR SA	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.682
N SDR completed suicide	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.274
N SDR attempted suicide	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.34
Mood polarity at SA (%)			0.273
Biphasic MD	0 (0.0)	1 (4.8)	
Major depression	112 (87.5)	19 (90.5)	
Mania	10 (7.8)	1 (4.8)	
Minor depression	1 (0.8)	0 (0.0)	
Mixed	4 (3.1)	0 (0.0)	
Rapid cycling	1 (0.8)	0 (0.0)	
Lifetime Hx SI	266 (51.8)	87 (42.9)	0.047
Mood episode related SI (%)			0.21
No	1 (0.7)	1 (3.1)	
Sometimes not always	6 (4.3)	0 (0.0)	

Continued on next page...

Variable	**Li(-)**	**Li(+)**	**p**
Yes	134 (95.0)	31 (96.9)	
Social anxiety disorder (%)	103 (16.3)	37 (15.5)	0.831
Panic disorder (%)	170 (26.7)	73 (22.1)	0.115
Generalized anxiety disorder (%)	159 (25.2)	54 (22.3)	0.376
OCD (%)	54 (8.4)	15 (4.5)	0.04
Substance use disorder (%)	226 (27.9)	76 (20.5)	0.008
ADHD (%)	69 (14.3)	51 (28.8)	<0.001
Learning disability (%)	71 (14.7)	45 (25.4)	0.001
Primary insomnia (%)	79 (16.4)	17 (9.5)	0.026
Personality disorder (%)	77 (17.5)	31 (19.4)	0.619
Diabetes mellitus (%)	43 (10.1)	10 (6.9)	0.322
Hypertension (%)	90 (21.4)	51 (35.9)	0.001
Menstrual abnormality (%)	59 (26.9)	10 (13.9)	0.028
Thyroid disease (%)	113 (29.3)	28 (21.1)	0.074
Head injury (%)	125 (36.1)	39 (33.6)	0.662
Migraine (%)	97 (26.0)	16 (11.9)	0.001
SES (%)			<0.001
Disabled	89 (14.7)	11 (4.7)	
Other	59 (9.7)	25 (10.7)	
Retired	71 (11.7)	48 (20.5)	
Social assistance	51 (8.4)	11 (4.7)	
Student	22 (3.6)	3 (1.3)	
UI	38 (6.3)	10 (4.3)	
Unknown	9 (1.5)	7 (3.0)	
Work full-time	220 (36.3)	89 (38.0)	
Work part-time	47 (7.8)	30 (12.8)	
Marital status (%)			0.001
Divorced	101 (12.9)	24 (8.9)	
Married	382 (48.7)	148 (54.6)	
Single	261 (33.2)	71 (26.2)	
Widowed	41 (5.2)	28 (10.3)	

5.3.2 Classification Between and Within Sites

Table 5.3 reports classification performance across the four analysis strategies. Classification was best in the aggregate analysis, with an accuracy of 0.77 (95% CI [0.75-0.79]), ROC-AUC of 0.8 (0.78-0.82), sensitivity of 0.53 (0.48-0.57), specificity of 0.9 (0.88-0.92), PPV of 0.74 (0.69-0.79), and NPV of 0.78 (0.77-0.80). Cohen's kappa for agreement of predicted and ground truth classes was 0.46 (0.4-0.51).

Site-level ROC-AUC performance was similar to the aggregate analysis. The Maritime group achieved 0.79 (0.74-0.84), Montreal 0.73 (0.56-0.91), Poznan 0.66 (0.6-0.72), and Cagliari (University) achieved 0.66 (0.6-0.72). However, under Cohen's kappa, which is more conservative under class imbalance, site-level results were inferior to those in the aggregated data. Kappa in the Poznan site was 0.24 (0.16-0.33), 0.22 (0.13-0.31) in the Maritimes, and 0.1 (0.03-0.16) for Cagliari (University). Under the Brier score, performance in the Maritimes sample (0.15 [0.13,0.16]) was superior to that observed in Poznan (0.24 [0.23, 0.25]).

The Maritimes data were most robust to the predict-one-site-out analysis, with a reduction in Kappa (to 0.16 [0.12, 0.19]) that was within the margin of error in the site-level protocol (0.22 [0.13, 0.31]).

Exclusion of IGSLi data (which consisted only of lithium responders) under leave-one-site-out reduced the PPV from the full aggregate performance of 0.74 (95% CI 0.69, 0.79) to 0.65 (0.6, 0.7), in addition to reducing the sensitivity from 0.53 (0.48, 0.57) to 0.47 (0.42, 0.52), and the Kappa score from 0.46 (0.4, 0.5) to 0.38 (0.32, 0.44). Exclusion of the Ontario sample (84% responders) reduced the sensitivity to 0.42 (0.35, 0.49), the PPV to 0.66 (0.61, 0.71), and the Kappa to 0.36 (0.3, 0.41). Sensitivity improved slightly to 0.61 (0.58, 0.64) with exclusion of Cagliari (Centro Bini), albeit without substantial improvement in the overall Kappa (mean 0.51 [0.47, 0.55]). Exclusion of the other datasets had negligible effect on classification performance.

5.3.3 Variable Importance

Variable importance results are plotted in Figure 5.2 for the RFC trained on the aggregated data sample, as well as the sites with the most robust Cohen's kappa results in the site-level analyses. Completely episodic clinical course was the most important predictor of LR in both the aggregated and Maritime samples. When the aggregate analysis was repeated excluding

Table 5.3: Performance for aggregate (ALL), site-level (Site), predict-one-site-out (POSO), and leave-one-site-out (LOSO) analyses. *Abbreviations:* Area under receiver operating characteristic curve (AUC), positive and negative predictive values (PPV, NPV).

Study	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC	Kappa	Brier
All Sites (Pooled or "Aggregate" Analysis)								
ALL	0.77 (0.75,0.79)	0.53 (0.48, 0.57)	0.9 (0.88, 0.92)	0.74 (0.69, 0.79)	0.78 (0.77, 0.80)	0.80 (0.78, 0.82)	0.46 (0.4, 0.51)	0.16 (0.15, 0.17)
Cagliari (University)								
POSO	0.71 (0.71, 0.71)	0.01 (0.01, 0.02)	1.0 (0.99, 1.0)	0.43 (0.18, 0.69)	0.71 (0.71, 0.71)	0.58(0.56, 0.6)	0.01 (0.0, 0.02)	0.20 (0.20, 0.21)
LOSO	0.79 (0.77,0.8)	0.58 (0.55,0.61)	0.9 (0.89, 0.91)	0.77 (0.74, 0.8)	0.79 (0.78, 0.81)	0.83 (0.81, 0.85)	0.51 (0.47, 0.55)	0.15 (0.15, 0.16)
Site	0.68 (0.65,0.71)	0.2 (0.15, 0.25)	0.88 (0.84, 0.92)	0.29 (0.22, 0.36)	0.73 (0.71, 0.75)	0.66 (0.6, 0.72)	0.1 (0.03, 0.16)	0.20 (0.19, 0.21)
Cagliari (Centro Bini)								
POSO	0.62 (0.6, 0.64)	0.23 (0.19, 0.28)	0.73 (0.69, 0.76)	0.18 (0.17, 0.2)	0.78 (0.78, 0.79)	0.47 (0.46, 0.48)	-0.04 (-0.06, -0.02)	0.23 (0.23, 0.24)
LOSO	0.77 (0.76,0.79)	0.61 (0.58, 0.64)	0.88 (0.85, 0.91)	0.78 (0.73, 0.82)	0.78 (0.77, 0.79)	0.84 (0.82, 0.86)	0.51 (0.48, 0.54)	0.16 (0.15, 0.17)
Site	0.76 (0.75,0.77)	0.04 (0.02, 0.05)	0.96 (0.95, 0.96)	0.06 (0.03, 0.1)	0.79 (0.78, 0.79)	0.5 (0.46, 0.55)	-0.01 (0, 0.01)	0.19 (0.18, 0.19)
International Group for the Study of Lithium (IGSLi)								
POSO	0.98 (0.97, 0.98)	0.98 (0.97, 0.98)	-	1.0 (1.0, 1.0)	0.0 (0.0, 0.0)	-	0.0 (0.0, 0.0)	0.61 (0.59, 0.62)
LOSO	0.76 (0.73,0.78)	0.47 (0.42, 0.52)	0.88 (0.86, 0.91)	0.65 (0.6, 0.7)	0.79 (0.77, 0.81)	0.77 (0.74, 0.8)	0.38 (0.32, 0.44)	0.25 (0.24, 0.26)
Site	-	-	-	-	-	-	-	-
Maritimes								
POSO	0.77 (0.76, 0.78)	0.23 (0.19, 0.27)	0.9 (0.89, 0.92)	0.38 (0.34, 0.41)	0.82 (0.82, 0.83)	0.68 (0.66, 0.69)	0.16 (0.12, 0.19)	0.17 (0.17, 0.18)
LOSO	0.77 (0.76,0.79)	0.58 (0.55, 0.61)	0.9 (0.88, 0.92)	0.8 (0.76, 0.84)	0.76 (0.75, 0.78)	0.79 (0.78, 0.81)	0.5 (0.47, 0.54)	0.23 (0.22, 0.24)
Site	0.8 (0.77, 0.82)	0.28 (0.19, 0.36)	0.93 (0.9, 0.96)	0.35 (0.25, 0.46)	0.84 (0.82, 0.86)	0.79 (0.74, 0.84)	0.22 (0.13, 0.31)	0.15 (0.13, 0.16)
Montreal								
POSO	0.84 (0.84, 0.85)	0.01 (-0.01, 0.04)	1.0 (0.99, 1.0)	0.1 (-0.09, 0.29)	0.84 (0.84, 0.85)	0.75 (0.73, 0.76)	0.02 (-0.02, 0.06)	0.13 (0.13, 0.13)
LOSO	0.76 (0.75,0.78)	0.52 (0.48, 0.56)	0.9 (0.88, 0.92)	0.76(0.71, 0.8)	0.77 (0.75, 0.79)	0.8 (0.77, 0.82)	0.45 (0.41, 0.5)	0.24 (0.22, 0.25)
Site	0.84 (0.81,0.88)	0.09 (0.02, 0.16)	0.99 (0.96, 1.0)	0.16 (0.02, 0.31)	0.85 (0.83, 0.87)	0.73 (0.56, 0.91)	0.09 (0, 0.19)	0.13 (0.11, 0.15)
Ontario								
POSO	0.8 (0.79, 0.81)	0.94 (0.93, 0.94)	0.11 (0.08, 0.14)	0.84 (0.84, 0.85)	0.25 (0.19, 0.32)	0.6 (0.58, 0.63)	0.06 (0.02, 0.1)	0.15 (0.14, 0.15)
LOSO	0.76 (0.74,0.78)	0.42 (0.35, 0.49)	0.9 (0.88, 0.93)	0.66 (0.61, 0.71)	0.79 (0.77, 0.81)	0.76 (0.74, 0.79)	0.36 (0.3, 0.41)	0.23 (0.22, 0.23)
Site	0.83 (0.82, 085)	0.99 (0.98, 0.99)	0.03 (0, 0.06)	0.84 (0.83, 0.85)	0.17 (0, 0.34)	0.52 (0.43, 0.62)	0.02 (0, 0.05)	0.16 (0.14, 0.18)
Poznan								
POSO	0.48 (0.47, 0.49)	0.04 (0.02, 0.05)	0.98 (0.97, 0.99)	0.73 (0.59, 0.87)	0.47 (0.47, 0.48)	0.54 (0.49, 0.59)	0.02 (0.01, 0.03)	0.27 (0.27, 0.28)
LOSO	0.79 (0.78, 0.8)	0.48 (0.44, 0.53)	0.94 (0.93, 0.95)	0.8 (0.77, 0.83)	0.79 (0.77, 0.8)	0.81 (0.79, 0.83)	0.47 (0.43, 0.51)	0.16 (0.16, 0.17)
Site	0.62 (0.58,0.67)	0.71 (0.66, 0.77)	0.53 (0.47, 0.59)	0.62 (0.58, 0.67)	0.66 (0.59, 0.72)	0.66 (0.6, 0.72)	0.24 (0.16, 0.33)	0.24 (0.23, 0.24)

the Maritimes data, the absence of chronic clinical course was the most important predictor. However, in both cases, four out of the six most informative features were related to clinical course. Clinical course variables were also disproportionately represented among the most informative when we attempted to predict lithium non-response (Appendix B.4). In the Poznan sample, the proportion of life affected with primary insomnia was the most important predictor, while clinical course was not among the most contributory features. Within all of the aforementioned analyses, the absence of rapid cycling was always within the 10 most informative features. Appendix B.5 shows that excluding all variables except clinical course and rapid cycling from the model preserves much of the classification performance in the aggregate analysis (kappa 0.35 [0.28,0.38]), while improving performance within the Maritime (kappa 0.40 [0.15,0.61]) and Montreal site-level analyses (kappa 0.27 [0.23,0.32]), and leaving the Poznan site-level results relatively unchanged (kappa 0.27 [0.05,0.50]).

5.4 Discussion

We have demonstrated that LR in BD can be predicted using only clinically available information. Variables characterizing the pre-treatment clinical course were among the most relevant predictors, but there was substantial between-sample heterogeneity. Our study motivates further work toward clarifying the phenotypic picture of canonical “lithium responders” and “non-responders,” respectively.

Analysis on the pooled sample yielded the best classification performance (Kappa 0.46 [0.4-0.51]), with relatively balanced PPV (0.74 [0.69-0.79]) and NPV (0.78 [0.77, 0.8]). This was not likely the result of any single site’s data, since classification performance remained relatively stable in a leave-one-site-out analysis (Kappa range 0.36-0.51). In contrast, site level classification retained moderate performance for the Maritime and Poznan samples (with Kappa values of 0.22 [0.13, 0.31] and 0.24 [0.16, 0.33], respectively), and to a lesser extent the Cagliari (University; Kappa 0.1 [0.03, 0.16]) sample.

There was likely substantial between-site heterogeneity in our data. At the site-level, performance was best within the Maritime (Cohen’s Kappa 0.22 [0.13, 0.31]) and Poznan (Cohen’s Kappa 0.24 [0.16, 0.33]) sites’ data. Yet, their most informative features were, respectively, completely episodic clinical course and proportion of life affected with primary insomnia. Completely episodic clinical course was the most informative feature in the aggregate analysis, but when the Maritimes sample was removed—which did not diminish

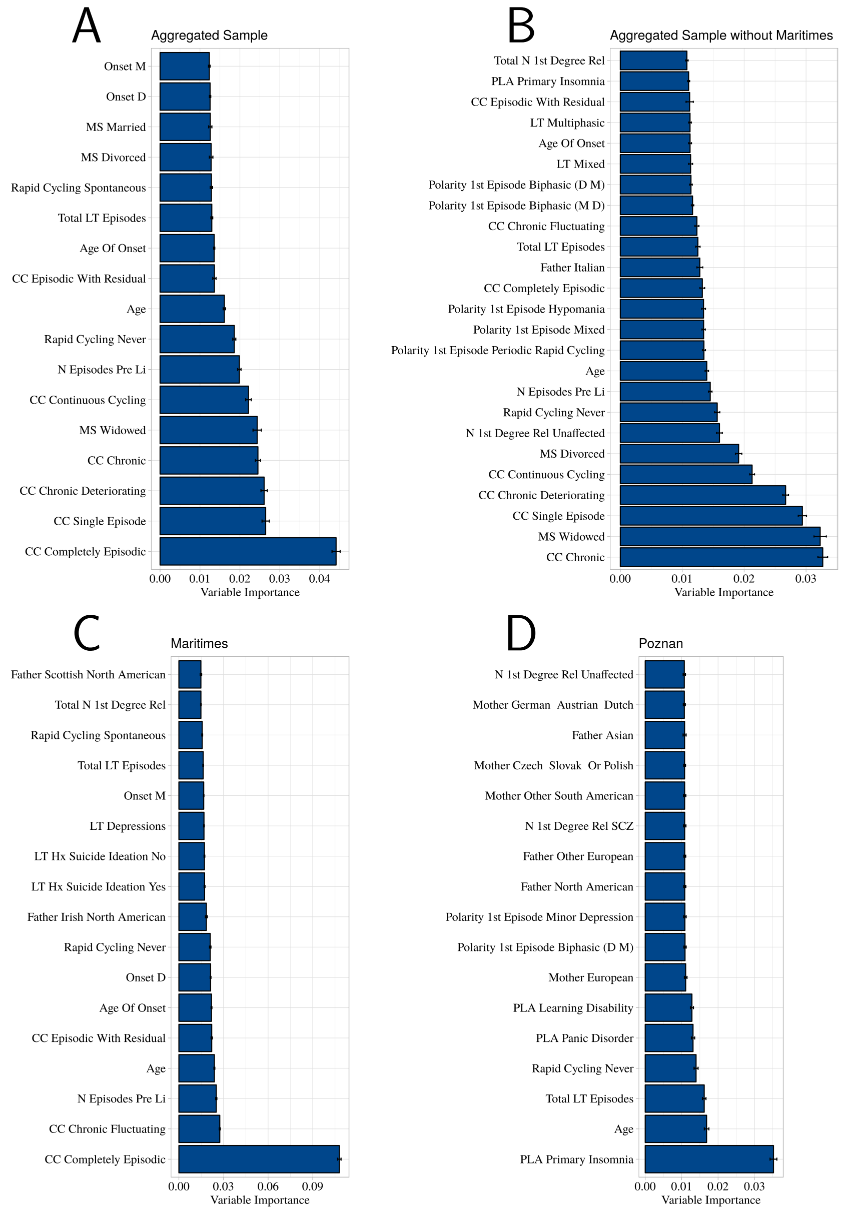


Figure 5.2: Variable importance across (A) Aggregate dataset, (B) Aggregate dataset excluding the Maritimes data, (C) Maritimes site-level data, and (D) Poznan site level data. Due to space constraints, only those variables with coefficients above the overall mean were included in these plots. Notwithstanding, only bars that strongly deviate from the height of others should be considered "important." Bars are variable importance means over the stratified cross-validation folds, and error bars are standard errors. Variable importance in random forest classifiers do not indicate the direction of a variable's influence (i.e. whether a feature is associated with response or non-response). *Abbreviations*: lifetime (LT), clinical course (CC), global assessment of functioning (GAF), marital status (MS), proportion of life affected (PLA), schizophrenia (SCZ).

classifier performance (Cohen's Kappa 0.5 [0.47, 0.54])—the completely episodic clinical course was no longer the most salient feature. Notwithstanding, 4 of the top 6 important features were still clinical course related, and clinical course alone was shown to have non-trivial predictive validity (Appendix B.5). Classification within the Maritimes sample also showed a lower Brier score than Poznan (0.15 [0.13, 0.16] vs. 0.24 [0.23, 0.24]), potentially underscoring that the Maritimes' feature importance results may confer stronger predictive ability. Moreover, when repeating the analysis for the purpose of predicting *non-response* (Alda score ≤ 3; Appendix B.4), features concerning clinical course and rapid cycling were also the most salient, suggesting consistency across the phenotypic continuum. Thus, the features characteristic of LR differed between sites, although no one site was predominantly responsible for the overall strength in classification performance. This between-site heterogeneity may limit the benefits of increased sample size [33], although some conflicting evidence exists [34]. Further work should focus on development of methods for better understanding the between-site heterogeneity in these data, potentially facilitating characterization of clinical features most reliably associated with lithium response and non-response.

Within the most robustly performing site, completely episodic clinical course was the most informative predictor of LR. This supports Grof's [36] highlighting of inter-episode remission quality as a central phenotypic element associated with LR. Other features of the classical phenotype, such as family history [198, 212], did not arise as important predictors. It is possible that the family history variables included in our dataset did not sufficiently capture this feature of the LR phenotype. This may also be due to relationships between episodicity and family history variables. For example, in our sample, those patients with a completely episodic course may have disproportionately more relatives with BD-I (Table B.15). Clarifying these relationships in a principled fashion will require further studies.

That the absence of rapid cycling was consistently observed within the 10 most informative variables across analyses would also agree with previous meta-analytically supported findings [196, 213, 214]. The relatively consistent importance of absence of rapid cycling in our study is interesting in light of the more variable importance of completely episodic clinical course, since these two variables were highly correlated. Indeed, 93% of those with a completely episodic course also had no rapid cycling, and the majority of those without rapid cycling (33.9%) had a completely episodic course (see Tables B.16 and B.17).

Disentangling these relationships in light of the potentially nonlinear feature combinations in ML models is an important future line of work.

The clinical implications of our results are of foremost interest. When engaging in a medication trial, it is critical to plan the trial duration. Too short of a period of time may preclude what could ultimately be a clinical response; too long of a trial would offer unnecessary exposure to side effects and slowing of a search for the right medication. On the aggregated data, our study suggests that LR can be predicted with a low rate of false positives (specificity 0.9, 95% CI 0.88-0.92). If such a model were to classify a patient as a lithium responder, one might foreseeably consider a longer treatment trial with the expectation of eventual response. Unfortunately, our model's sensitivity was comparatively less strong (0.53 [0.48-0.57]), and so the prediction of non-response would be a more difficult scenario for clinical decision-making. Improving the true positive rate of LR prediction is thus an important subsequent goal for our work. That being said, when we attempted to predict the most unequivocal *non-responders* (i.e. those with the lowest Alda scores, ≤ 3), we attained a Cohen's kappa of 0.63 (0.55-0.7), with an ROC-AUC of 0.88 (0.86-0.91), PPV of 0.82 (0.79-0.85), and NPV of 0.86 (0.81-0.91). In totality, these results together suggest two things. First, they support the notion that clinical data indeed are useful for prediction of LR and non-response, since the prediction strength is relatively symmetrical with respect to the extremes of the Alda score distribution. Second, these results suggest that work toward improving our model's detection of lithium responders should largely focus on better discriminating those individuals whose Alda scores range from 4 to 6. Such individuals could represent an intermediate phenotype (partial responders), or their response classification may have been less certain due to other factors (compliance related issues, use of other drugs, etc.), or simply inaccuracy. The systematic evaluation of these possibilities will be an important set of future experiments. Such experiments will be important for better understanding the predictive relationships captured by our model, thereby improving interpretability and facilitating transfer to clinical practice. At this stage, however, our model is not ready for clinical translation, pending better explication of the underlying predictive relationships and between-site heterogeneity.

Our study has several limitations. First, data are not sourced from a randomized clinical trial. However, it would be difficult to gather a sample of sufficient size and with the duration of clinical follow up as in our study. Second, since data were sourced from specialist clinics, our results may not translate into the general psychiatric practice; that being said, randomized clinical trial populations are usually even more narrowly selected, and our overall balance of responders and non-responders ($\approx$ 1/3 responders) is likely reflective of real-world practice, although registry studies using non-hospitalization as a response definition estimate lower response rates [192, 215, 216]. Moreover, the features comprising these data are all obtainable by careful assessment in general psychiatric practice at no additional cost to assessment as usual. Third, our results evaluate prediction of complete response to lithium monotherapy, and so future work should evaluate the prediction of whether individuals may respond to lithium in combination with other agents. Fourth, while complex patterns of missingness in the data did not bias our results, they may have reduced predictive capacity through information loss. Fifth, interpretation of the exact feature combinations relevant to prediction in our model is difficult. However, several other models examined *post hoc* showed inferior performance compared to the RFC, and so their feature importance coefficients would not be relevant. Finally, the data herein included neither biological nor behavioural data, which remains an important line of future research.

Despite these limitations, our study has several important strengths that advance the state of knowledge concerning LR in BD. In terms of raw sample size, our study is exceeded only by the Scandinavian registry analyses (n=3762 [215] and n=4714 [216]), whose LR definition was time to treatment failure. As discussed by Hui et al. [196], such a definition may bias studies toward false negatives, since those who experience a relapse may otherwise have a reduction in symptom frequency over the long term. In contrast, the Alda score captures the potential reduction in symptom frequency/severity (A score), as well as incorporating a “causality factor” (B score) which penalizes the overall score where confounding information is present. These scores were collected from patients followed prospectively at mood disorders clinics, by clinicians who underwent reliability training. This, in conjunction with the sheer breadth of predictors available (180 in total), renders the present analysis the largest of its kind. Moreover, given that our data were collected from 7 sites internationally, including 4 from ethnically diverse Canadian cities, our study likely includes one of the most heterogeneous samples yet reported in the LR literature. Finally, the most important strength of our study is its direct evaluation of variables’ classification power *out of sample*, which provides more robust evidence of the degree to which LR can be predicted in real-world settings.

Chapter 6

Asymmetrical Reliability of the Alda Score Favours a Dichotomous Representation of Lithium Responsiveness[1]

Abstract. The Alda score is commonly used to quantify lithium responsiveness in bipolar disorder. Most often, this score is dichotomized into "responder" and "non-responder" categories, respectively. This practice is often criticized as inappropriate, since continuous variables are thought to invariably be "more informative" than their dichotomizations. We therefore investigated the degree of informativeness across raw and dichotomized versions of the Alda score, using data from a published study of the scale's inter-rater reliability (n=59 raters of 12 standardized vignettes each). After learning a generative model for the relationship between observed and ground truth scores (the latter defined by a consensus rating of the 12 vignettes), we show that the dichotomized scale is more robust to inter-rater disagreement than the raw 0-10 scale. Further theoretical analysis shows that when a measure's reliability is stronger at one extreme of the continuum—a scenario which has received little-to-no statistical attention, but which likely occurs for the Alda score ≥ 7—dichotomization of a continuous variable may be more informative concerning its ground truth value, particularly in the presence of noise. Our study suggests that research employing the Alda score of lithium responsiveness should continue using the dichotomous definition, particularly when data are sampled across multiple raters.

6.1 Introduction

The Alda score is a validated index of lithium responsiveness commonly used in bipolar disorder (BD) research [202]. This scale has two components. The first is the "A" subscale that provides an ordinal score (from 0 to 10, inclusive) of the overall "response" in a therapeutic trial of lithium. The second component is the "B" subscale that attempts to qualify the degree to which any improvement was causally related to lithium. The total Alda score is computed based on these two subscale scores, and takes integer values between 0 and 10. Many studies that employ the Alda score as a target variable dichotomize it, such that individuals with scores ≥ 7 are classified as "responders," and those with scores < 7 are "non-responders."

[1] Nunes A, Trappenberg T, and Alda M. Asymmetrical reliability of the Alda Score favours a dichotomous representation of lithium responsiveness. *PLOS ONE* 15(1): e0225353.

A common criticism that arises from this practice is that continuous variables should not be discretized by virtue of "information loss." Indeed, discretizing continuous variables is widely viewed as an inappropriate practice [217–227]. However, the practice remains common across many areas of research, including our group's work on lithium responsiveness in BD [35]. The primary justification for using the dichotomized Alda score as the lithium responsiveness definition has been based on the inter-rater reliability study by Manchia et al. [202], who showed that a cut-off of 7 had strong inter-rater agreement (weighted kappa 0.66). Furthermore, using mixture modeling, they also found that the empirical distribution of Alda scores supports the discretized definition. Therefore, there exist competing arguments regarding the appropriateness of dichotomizing lithium response. Resolving this dispute is critical, since the operational definition of lithium responsiveness is a concept upon which a large body of research will depend.

Although the Manchia et al. [202] analysis provides some justification for using a dichotomous lithium response definition, it does not dispel the argument of discretization-induced information loss entirely. However, there is some intuitive reason to believe that discretization is, at least pragmatically, the best approach to defining lithium response using the Alda score. First, the Alda score remains inherently subjective to some degree and is not based on precise biological measurements; an individual whose "true" Alda score is 6, for example, could have observed scores that vary widely across raters. Second, it is possible that responders may be more reliably identified than non-responders. For example, unambiguously "excellent" lithium response is a phenomenon that undoubtedly exists in naturalistic settings [36, 37], and for which the space of possible Alda scores is substantially smaller than for non-responders; that is, an Alda score of 8 can be obtained in far fewer ways than an Alda score of 5. As such, we hypothesize that agreement on the Alda score is higher at the upper end of the score range, and that this asymmetric agreement is a scenario in which dichotomization of the score is more informative than the raw measure. To evaluate this, we present both empirical re-analysis of the ConLiGen study by Manchia et al. [202], and analyses of simulated data with varying levels of asymmetrical inter-rater reliability.

6.2 Materials and Methods

6.2.1 Data

Detailed description of data and collection procedures is found in Manchia et al. [202]. Samples included in our analysis are detailed in Table 6.1, including the number of raters included across sites, and the average ratings obtained at each of those sites across the 12 assessment vignettes. As a gold standard, we used ratings that were assigned to each case vignette using a consensus process at the Halifax site (scores are noted in the first row of Table 6.1). The lithium responsiveness inter-rater reliability data are available in the online supplemental material.

Table 6.1: Number of raters and mean scores across sites. The total number of raters (n_r) was 59.

		Case Vignette											
Site	n_r	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Gold standard		8	9	6	7	9	3	5	9	3	9	5	1
Halifax	9	8.4	8.6	6.6	6.9	9.2	3	3.9	8.8	3.1	9.1	4.7	1.2
NIMH	4	7.8	8.2	6.2	7	8.8	3.2	4	8.5	2.2	8.5	3.2	1.8
Poznan	2	9	8.5	6.5	5.5	9	4	7.5	9	5	8	4.5	4.5
Dresden	2	8.5	7.5	6	5	8.5	1.5	6	9	3.5	8.5	4	1.5
Japan	4	8	8.2	4.8	6.5	8.5	2	3	8.5	1	8.2	4.5	1.5
Wuerzburg	2	7.5	7.5	4	6.5	8	1.5	3	9	0	7	3	0.5
Cagliari	3	7.7	9	4.3	7	5.7	4	1.3	9	0.7	7.3	4	2
San Diego	2	7.5	8.5	7.5	7	9	5	7.5	8.5	3.5	8.5	6	3.5
Boston	2	8.5	8.5	6	7	9	3	3.5	8.5	1.5	9	4	1
Gottingen	2	9.5	9	4	6	9	1	1	9	1.5	9	4	3
Berlin	1	7	9	4	6	9	2	3	8	0	7	0	2
Taipeh	1	8	8	5	8	9	5	6	9	4	9	8	1
Prague	1	7	9	4	8	9	3	6	9	3	9	6	1
Johns Hopkins	7	8	8.7	5.3	5.9	8.3	2.7	2.4	9.1	2	8.3	4.4	1.1
Mayo	6	8	8.2	6	8	9	4.2	3	9	4.2	8.8	3.7	0.3
Brasil	3	8	8.3	5.3	6.3	8.7	2	4	9	4.3	8	4.7	0.7
Medellin	4	7.5	9	5.5	6.5	5	2.5	4	7.2	4.8	8.8	1.2	2
Geneve	3	7.7	8.7	6.7	5.3	9.7	5	6	8.7	1.3	9	3.7	0.3

6.2.2 Empirical Analysis of the Alda Score

In this analysis, we seek to evaluate whether discretization of the Alda score under the existing inter-rater reliability values preserves *more* mutual information (MI) between the observed and ground truth labels than does the raw scale representation. To accomplish this, we first develop a probabilistic formulation of raters' score assignments based on a Multinomial-Dirichlet model, which we describe here.

Let $n_i^{(k)} \in \mathbb{N}_+$ denote the number of raters who assigned an Alda score $i \in \mathcal{A}$, with $\mathcal{A} = \{0,\ 1,\ ...,\ 10\}$ to an individual whose gold standard Alda score is $k \in \mathcal{A}$. The vector of rating counts for the gold standard score k is is $\mathbf{n}^{(k)} = \left(n_i^{(k)}\right)_{i \in \mathcal{A}}$. The probability of $\mathbf{n}^{(k)}$ is multinomial with parameter vector $\boldsymbol{\theta}^{(k)} = \left(\theta_i^{(k)}\right)_{i \in \mathcal{A}}$, which is itself Dirichlet distributed $\boldsymbol{\theta}^{(k)} \sim \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is a pseudocount denoting the prior expectation of the number of ratings received for each score $i \in \mathcal{A}$. In the present analysis, we assume that $\boldsymbol{\alpha}$ is equal across all scores in $\mathcal{A}$, and thus we denote it simply as a scalar $\boldsymbol{\alpha} = \alpha$; this has the effect of increasing the uncertainty of $\boldsymbol{\theta}^{(k)}$ (i.e. the ratings become more "noisy").

The posterior of $\boldsymbol{\theta}^{(k)}$ given $\mathbf{n}^{(k)}$ and α is Dirichlet with parameters $\boldsymbol{\alpha}' = \left\{\alpha + n_i^{(k)}\right\}_{i=0}^{10}$, and its *maximum a posteriori* (MAP) estimate is

$$\widehat{\boldsymbol{\theta}}_\alpha \left(\mathbf{n}^{(k)}\right) = \left\{ \frac{\alpha + n_i^{(k)} - 1}{\sum_{j=0}^{10} \alpha + n_j^{(k)} - 1} \right\}_{i=0}^{10}, \tag{6.1}$$

which can be viewed as the conditional distribution over scores $\mathcal{A}$ for any given rater when the gold standard is k. In cases where no assessment vignette had a gold standard rating of k, we assumed that

$$\mathbf{n}^{(k)} = \begin{cases} \frac{1}{2}\left(\mathbf{n}^{(k-1)} + \mathbf{n}^{(k+1)}\right) & 0 < k < 10 \\ \mathbf{n}^{(k+1)} & k = 0 \\ \mathbf{n}^{(k-1)} & k = 10 \end{cases} \tag{6.2}$$

The dichotomized Alda scores are defined as $T = \{\delta[i \geq \tau] : \forall i \in \mathcal{A}\}$, where τ is the dichotomization threshold (set at $\tau = 7$ for the Alda score), and where $\delta[\cdot]$ is an indicator function that evaluates to 1 if the argument is true, and 0 otherwise. Given threshold τ (Responders $\geq \tau$ and Non-responders $< \tau$), the dichotomous counts are represented as follows

$$\begin{aligned} c_0^{(0)} &= \textstyle\sum_{k=0}^{\tau-1}\sum_{i=0}^{\tau-1} n_i^{(k)} && \text{Observed} < \tau\text{, Gold Standard} < \tau \\ c_0^{(1)} &= \textstyle\sum_{k=\tau}^{10}\sum_{i=0}^{\tau-1} n_i^{(k)} && \text{Observed} < \tau\text{, Gold Standard} \geq \tau \\ c_1^{(0)} &= \textstyle\sum_{k=0}^{\tau-1}\sum_{i=\tau}^{10} n_i^{(k)} && \text{Observed} \geq \tau\text{, Gold Standard} < \tau \\ c_1^{(1)} &= \textstyle\sum_{k=\tau}^{10}\sum_{i=\tau}^{10} n_i^{(k)} && \text{Observed} \geq \tau\text{, Gold Standard} \geq \tau \end{aligned} \tag{6.3}$$

with $\mathbf{c}^{(k)} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_k)$, and $\boldsymbol{\phi}_k \sim \mathrm{Dir}(\boldsymbol{\phi}|\xi)$, where ξ is a pseudocount for the number of dichotomized ratings assigned to each of non-responders and responders. We can thus estimate the conditional distribution over observed dichotomized response ratings as

$$\widehat{\boldsymbol{\phi}}_\xi\left(\mathbf{c}^{(k)}\right) = \left\{ \frac{\xi + c_0^{(k)} - 1}{2\xi - 2 + c_0^{(k)} + c_1^{(k)}}, \frac{\xi + c_1^{(k)} - 1}{2\xi - 2 + c_0^{(k)} + c_1^{(k)}} \right\} \tag{6.4}$$

Mutual Information of Raw and Dichotomized Alda Score Representations

Let

$$x_o \sim p(x_o|x_*) = \mathrm{Categorical}\left(\widehat{\boldsymbol{\theta}}_\alpha\left(\mathbf{n}^{(x_*)}\right)\right) \tag{6.5}$$

denote a given observed *raw* Alda score assigned to a case with ground truth score of $x_* \in \mathcal{A}$. Given uniform priors on the true classes, the joint distribution is

$$p(x_o, x_*) = p(x_o|x_*)p(x_*) = \left\{ \frac{1}{11}\widehat{\boldsymbol{\theta}}_\alpha\left(\mathbf{n}^{(x_*=k)}\right) \right\}_{k=0,1,\ldots,10}. \tag{6.6}$$

For the binarized classes, we have a prior of $p(y_* = 1) = \frac{4}{11}$, and the joint distribution is thus

$$p(y_o, y_*) = p(y_o|y_*)p(y_*) = \left\{ p(y_* = k)\, \widehat{\boldsymbol{\phi}}_\xi\left(\mathbf{c}^{(y_*=k)}\right) \right\}_{k\in\{0,1\}}. \tag{6.7}$$

The MI for these distributions can be computed as functions of the prior pseudocounts α and ξ:

$$I_\alpha[x_o||x_*] = \sum_{x_o}\sum_{x_*} p(x_o, x_*) \log \frac{p(x_o,\ x_*)}{p(x_o)p(x_*)} \tag{6.8}$$

$$I_\xi[y_o||y_*] = \sum_{y_o}\sum_{y_*} p(y_o, y_*) \log \frac{p(y_o, y_*)}{p(y_o)p(y_*)} \tag{6.9}$$

for the raw and dichotomized Alda scores, respectively. We can express the MI of the raw and dichotomized Alda score distributions both in terms of α, such that both distributions have an equivalent total "concentration" when $\xi = 11\alpha/2$. This is equivalent to saying that our prior assumption about the uncertainty of the raw and dichotomized distributions assumes the same number of a priori ratings.

Our primary hypothesis—that the dichotomized Alda score is more informative with greater observation uncertainty—is evaluated by determining whether $I_\xi[y_o||y_*]$ exceeds $I_\alpha[x_o||x_*]$ as we increase the *a priori* observation noise (α and ξ).

6.2.3 Theoretical Modeling of Dichotomization under Asymmetrical Reliability

The previous experiment regarding dichotomization of the raw Alda score did not fully capture the effect of dichotomization of a continuous variable, since the raw Alda score is still discrete (albeit with a larger domain of support). Thus, we sought to investigate whether dichotomization of a truly continuous, though asymmetrically reliable, variable would show a similar pattern of preserving MI and statistical power under higher levels of observation noise and agreement asymmetry.

Synthetic Datasets

The simplest synthetic dataset generated was merely a sample of regularly spaced points across the [0,10] interval in both the x and y directions. This dataset was merely used to conduct a "sanity check" that our methods for computing MI correctly identified a value of 0. This was necessary since data with uniform random noise over the same interval will only yield MI of 0 in the limit of large sample sizes.

The main synthetic dataset accepted "ground truth" values $x \in [0,\ 10]$ and yielded "observed" values $y \in [0,\ 10]$ based on the following formula for the i^{th} sample:

$$y_i = \omega\, f(x_i)\ +\ (1-\omega)\ \text{Uniform}(0,\ 10), \tag{6.10}$$

where $0 \leq \omega \leq 1$ is a parameter governing the degree to which observed values are coupled to the ground truth based on $f(x_i)$ (data are entirely uniform random noise when $\omega = 0$, and come entirely from $f(x_i)$ when $\omega = 1$). The function $f(x_i)$ governing the agreement between ground truth and observed is essentially a 1:1 correspondence between x and y to which we add noise along the diagonal based on a uniform random variate $\widetilde{U}(-\sigma,\ \sigma)$ with width σ.

We simulated two forms of diagonal spread. The first is constant across all values $x \in [0,\ 10]$, which we call the *symmetrical* case, and which is represented by a parameter $\beta = 1$. The other is an *asymmetrical* case (represented as $\beta = 0$), in which the agreement between x and y is not constant across the $[0, 10]$ range. Overall, the function $f(x_i)$ is defined as

$$f(x_i)\ = \beta\, R_{(0,\ 10)} \left(x_i + \frac{\widetilde{U}(-\sigma,\ \sigma)}{1 + e^{-0.75\ x_i + 5}} \right) \ +\ (1 - \beta)\, R_{(0,\ 10)} \left(x_i + \widetilde{U}(-\sigma,\ \sigma) \right),\ \tag{6.11}$$

where $R_{(l,\ u)}(\cdot)$ is a function to ensure that all points remain within the $[l, u]$ interval in both axes. In the asymmetrical case, $R_{(l,\ u)}(\cdot)$ reflects points at the $[0, 10]$ bounds. In the symmetrical case, the data are all simply rescaled to lie in the $[0, 10]$ interval.

Demonstration of the simulated synthetic data are shown in Figure 1. Every synthetically generated dataset included 750 samples, and for notational simplicity, we denote the k^{th} synthetic dataset (given parameters β, ω, σ) as $D^{(k)}_{\beta,\omega,\sigma} = \left(x^{(k)}_j,\ y^{(k)}_j \right)_{j=1,2,\ldots,750}$.

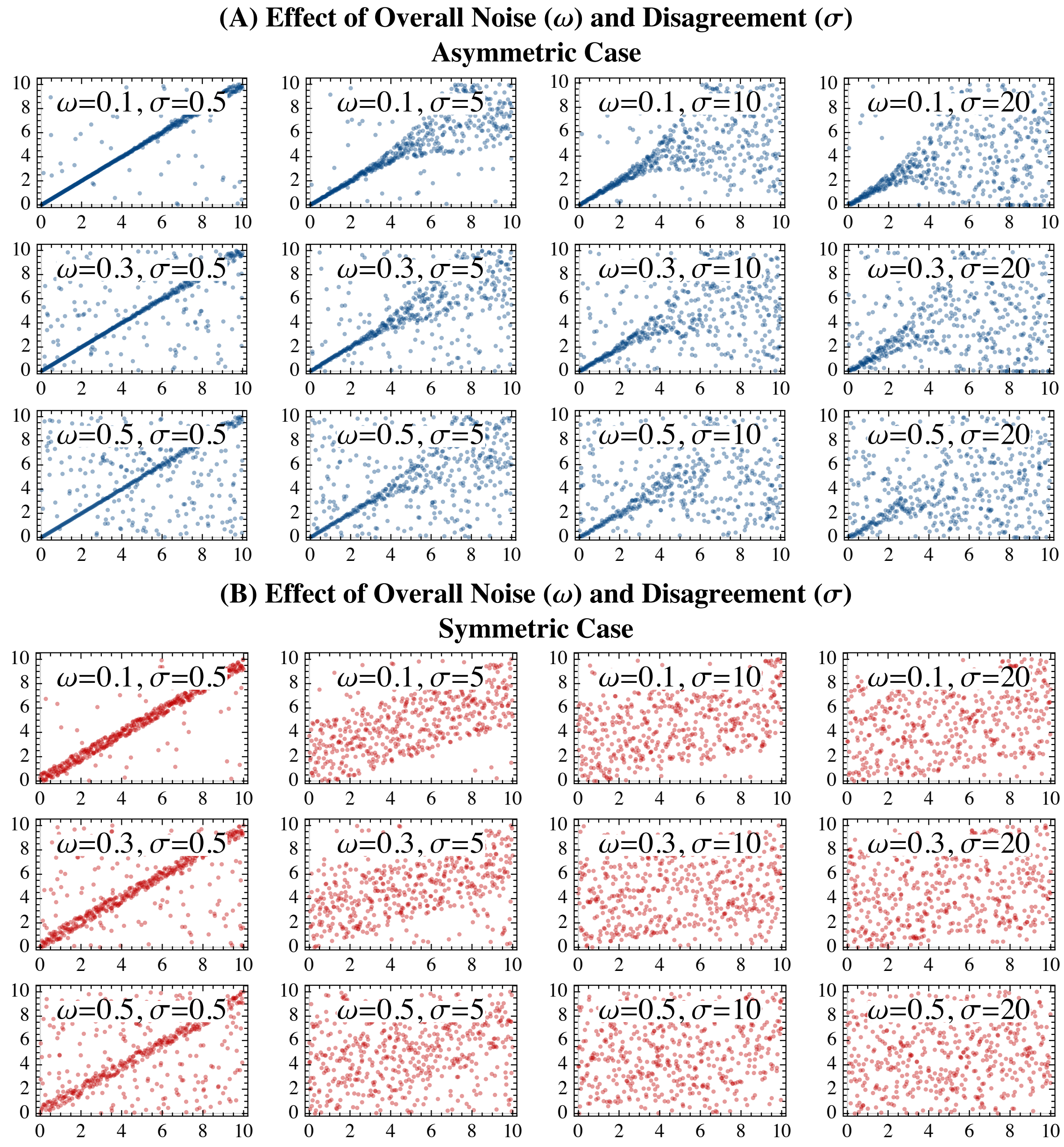
Computation of Mutual Information for Continuous and Discrete Distributions

Mutual information was computed for both continuous and dichotomized probability distributions on the data. Mutual information for the continuous distribution was computed by first performing Gaussian kernel density estimation (using Scott's method for bandwidth selection) on the simulated dataset, and then approximating the following integral using Markov chain Monte-Carlo sampling:

$$I_{\text{KDE}}[y||x] = \int \int p(x,\ y) \log \frac{p(x,\ y)}{p(x)\, p(y)}\, \mathrm{d}x\, \mathrm{d}y \tag{6.12}$$

Conversely, discrete MI was computed by first creating a 2-dimensional histogram by binning data based on a dichotomization threshold τ. Data that lie below the dichotomization threshold are denoted 1, and those that lie above the threshold are represented as 0. Based on this joint distribution, the dichotomized MI is

$$I_\tau[y||x] = \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x) p(y)}. \tag{6.13}$$



(A) Effect of Overall Noise (ω) and Disagreement (σ)
Asymmetric Case

(B) Effect of Overall Noise (ω) and Disagreement (σ)
Symmetric Case

Figure 6.1: Demonstration of the synthetic agreement data across differences in the parameter ranges and presence of asymmetry. The x-axes all represent the ground truth value of the variable, and the y-axes represent the “observed” values. Data are depicted based on different values of a uniform noise parameter ($0 \leq \omega \leq 1$) that governs what proportion of the data is merely uniform noise over the interval [0, 10], and a disagreement parameter ($\sigma \geq 0$), which governs the variance around the diagonal line. **Panel A** (upper three rows, shown in blue) depicts the synthetic data in which there was asymmetrical levels of agreement across the score domain. **Panel B** (lower three rows, shown in red) depict synthetic data in which there was symmetrical agreement over the score domain.

Note that continuous MI will remain constant across τ.

Statistical Power of Classical Tests of Association

Association between the observed (y) and ground truth (x) data can be measured using Pearson's correlation coefficient (ρ) when data are left as continuous, or using Fisher's exact test when data are dichotomized. The statistical power of the hypothesis that $\rho \neq 0$ given dataset $D^{(k)}_{\beta,\omega,\sigma}$ with $N^{(k)}$ observations and two-tailed statistical significance threshold α—which here is not the same α used as a Dirichlet concentration in Section 6.3.1—can be easily shown to equal

$$\text{power}_\rho(D^{(k)}_{\beta,\omega,\sigma};\ \alpha = 0.05) = \Phi\left(|\zeta(\rho)|\sqrt{N^{(k)} - 3} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right), \tag{6.14}$$

where $\Phi(\cdot)$ and $\Phi^{-1}(\cdot)$ are the cumulative distribution function and quantile functions for a standard normal distribution, and $\zeta(\cdot)$ is Fisher's Z-transformation

$$\zeta(\rho) = \frac{1}{2}\log\frac{1+\rho}{1-\rho}. \tag{6.15}$$

Under a dichotomization of $D^{(k)}_{\beta,\omega,\sigma}$ with threshold τ association between the ground truth and observed data can be evaluated using a (two-tailed) Fisher's exact test, whose alternative hypothesis is that the odds ratio (η) of the dichotomized data does *not* equal 1. The null-hypothesis has a Fisher's noncentral hypergeometric distribution,

$$\Lambda_o = \text{FisherHypergeometricDistribution}\left(N^{(k)}_{\delta[y<\tau]},\ N^{(k)}_{\delta[x<\tau]},\ N^{(k)},\ \eta = 1\right) \tag{6.16}$$

where $N^{(k)}$ is the total number of observations in sample k, and $N^{(k)}_{\delta[x<\tau]}$ and $N^{(k)}_{\delta[y<\tau]}$ are the number of ground truth and observed data, respectively, that fall below the dichotomization threshold τ. Under the alternative hypothesis, this distribution has an odds ratio parameter estimated from the data:

$$\Lambda_a = \text{FisherHypergeometricDistribution}\left(N^{(k)}_{\delta[y<\tau]},\ N^{(k)}_{\delta[x<\tau]},\ N^{(k)},\ \widehat{\eta}\right). \tag{6.17}$$

The statistical power of Fisher's exact test under this setup and a two-tailed significance threshold of α is

$$\mathrm{fp}\left(D^{(k)}_{\beta,\omega,\sigma},\ \tau;\alpha\right) = \delta\left[\hat{\eta} < 1\right]\left[1 - S_{\Lambda_a}\left(S^{-1}_{\Lambda_o}\left(1 - \frac{\alpha}{2}\right)\right)\right] + \delta\left[\hat{\eta} \geq 1\right] S_{\Lambda_a}\left(S^{-1}_{\Lambda_o}\left(\frac{\alpha}{2}\right)\right) \tag{6.18}$$

where $S_{\Lambda_a}(\cdot)$ and $S^{-1}_{\Lambda_o}(\cdot)$ are the survival functions of the alternative hypothesis and the inverse survival function of the null hypothesis, respectively.

Evaluation of Mutual Information

The central aspect of this analysis is comparison of the dichotomized and continuous MI across values of the dichotomization threshold τ, global noise ω, asymmetry parameter β, and diagonal spread σ. Under all cases, we expect that increases in the global noise parameter ω will reduce the MI. We also expect that with symmetrical reliability (i.e. $\beta = 0$), the dichotomized MI will be lower than the continuous MI across all thresholds. However, as the degree of asymmetry in the reliability increases, we expect the dichotomized MI to exceed the continuous MI (i.e. as σ increases when $\beta = 1$). Finally, as a sanity check, we expect that both continuous and dichotomized MI will be approximately 0 when applied to a grid of points regularly spaced over the [0,10] interval.

Evaluation of Effects on Statistical Power of Classical Tests of Association

Statistical power of the Pearson correlation coefficient and Fisher's exact test were computed across symmetrical ($\beta = 0$) and asymmetrical ($\beta = 1$) conditions of the synthetic dataset described above. Owing to the greater computational efficiency of these calculations (compared to the MI), the diagonal spread parameter was varied more densely ($\sigma \in \{1, 2, ..., 20\}$). The power of Fisher's exact test was evaluated at two dichotomization thresholds: a median split at $\tau = 5$ and a "tail split" at $\tau = 3$. We evaluated three global noise settings ($\omega \in \{0.3,\ 0.5,\ 0.7\}$). At each experimental setting, we computed the aforementioned power levels for 100 independent synthetic datasets. Results are presented using the mean and 95% confidence intervals of the power estimates over the 100 runs under each condition. We expect that the Fisher's exact test under a "tail split" dichotomization (not a median split) will yield greater statistical power in the presence of asymmetrical reliability, greater diagonal spread, and higher global noise. However, under the symmetrically reliable condition, we expect that the statistical power will be greater for the continuous test of association.

6.2.4 Materials

Mutual information experiments were conducted in Mathematica v. 12.0.0 (Wolfram Research, Inc.; Champaign, IL). Experiments evaluating the statistical power under classical tests of continuous and dichotomous association were conducted in the Python programming language. Data and code for analyses are also provided as online supplementary materials.

6.3 Results

6.3.1 Empirical Evaluation of the Alda Score of Lithium Response

Histograms of the observed Alda scores for each of the gold standard vignette values are depicted in Figure 6.2. Resulting joint distributions of the gold standard vs. observed Alda scores (in both the raw or dichotomized representations) are shown in Figure 6.3 (Panels A-C) across varying levels of observation noise. Figure 6.3D plots the MI for the raw and dichotomized Alda scores across increasing levels of the observation noise parameter α (recalling that $\xi = 11\alpha/2$). Beyond an observation noise of approximately $\alpha > 3.52$, one can see that the dichotomized lithium response definition retains greater MI between the true and observed labels, compared to the raw representation.

6.3.2 Discrete vs. Continuous Mutual Information in Asymmetrically Reliable Data

Figure 6.4 shows the results of the experiment on synthetic data. Under agreement levels that are constant across the (x, y) domains, one can observe that MI of dichotomized representations of the variables are generally lower than their continuous counterparts. However, under asymmetrical reliability (i.e. where agreement between x and y decreases as x increases), we see that MI is higher for the dichotomized, rather than the continuous, representations. In particular, as the level of agreement asymmetry increased (i.e. for higher values of σ), the best dichotomization thresholds decreased.

6.3.3 Statistical Power of Classical Associative Tests

Figure 6.5 plots the statistical power of null-hypothesis tests using continuous and dichotomized representations of the synthetic dataset. As expected, under conditions of

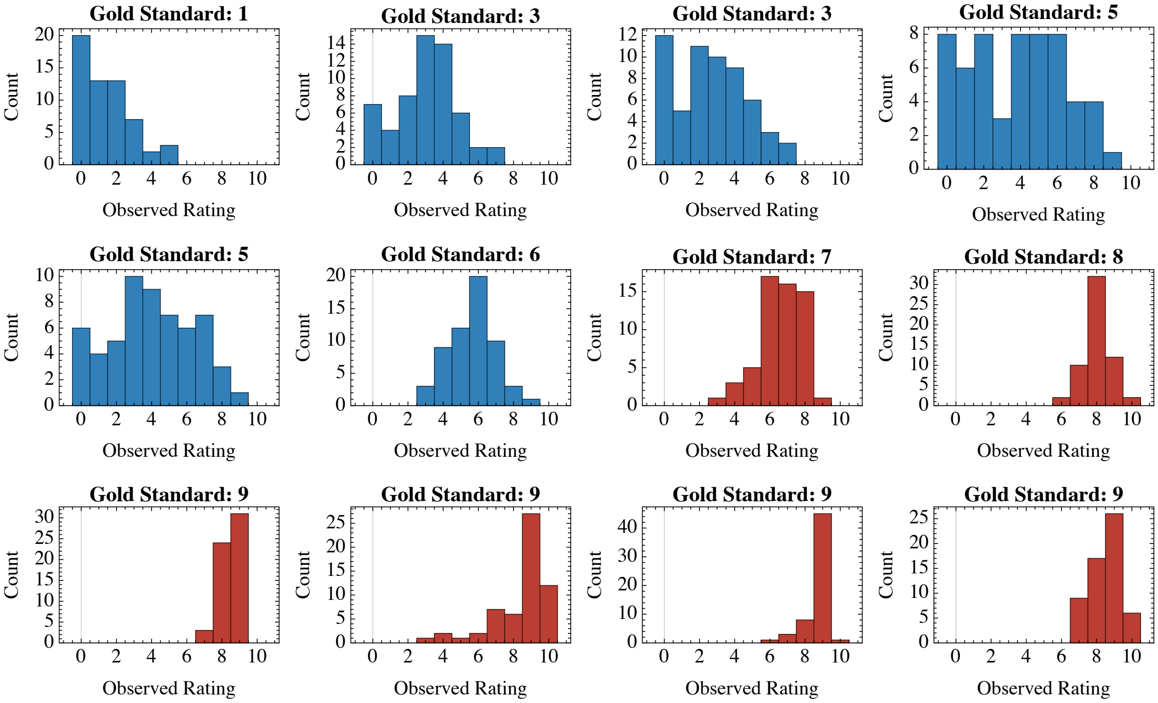


Figure 6.2: Histograms of ratings for each value of the ground truth Alda score available in the first wave dataset from Manchia et al. [202]. Each histogram represents the distribution of ratings ($n_r = 59$) for a single one of twelve assessment vignettes. The gold standard (“ground truth”) Alda score, obtained by the Halifax consensus sample, is depicted as the title for each histogram. Plots in blue are those for vignettes with gold standard Alda scores less than 7, which would be classified as “non-responders” under the dichotomized setting. Vignettes with gold standard Alda scores ≥ 7 are shown in red, and represent the dichotomized group of lithium responders.

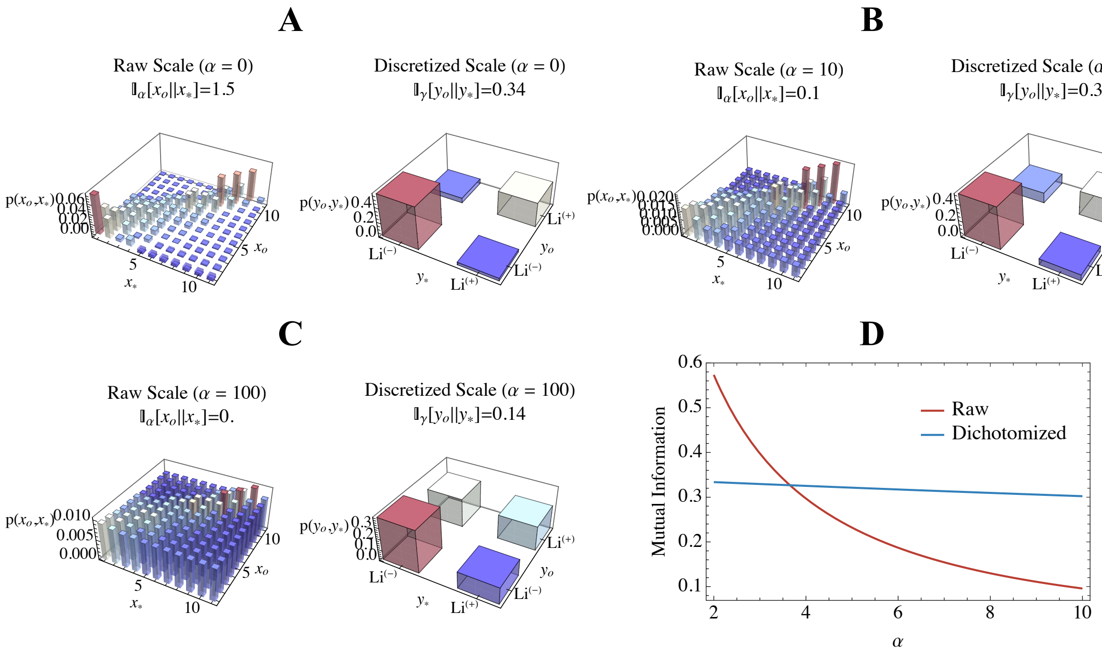


Figure 6.3: Mutual information between gold standard and observed Alda scores in relation to the observation noise (α) and whether the scale is in its raw or dichotomized form (lithium responder [Li(+)] is Alda score ≥ 7; non-responder [Li(-)] is Alda score <7). **Panels A-C** show the inferred joint distributions of the observed (x_o for raw, y_o for discrete) and gold standard (x_* for raw, y_* for discrete) values at different levels of observation noise ($\alpha \in \{0, 10, 100\}$). **Panel D** plots the mutual information for the raw (red) and discrete (blue) settings of the Alda score across increasing values of α. Recall that here we set $\xi = 11\alpha/2$.

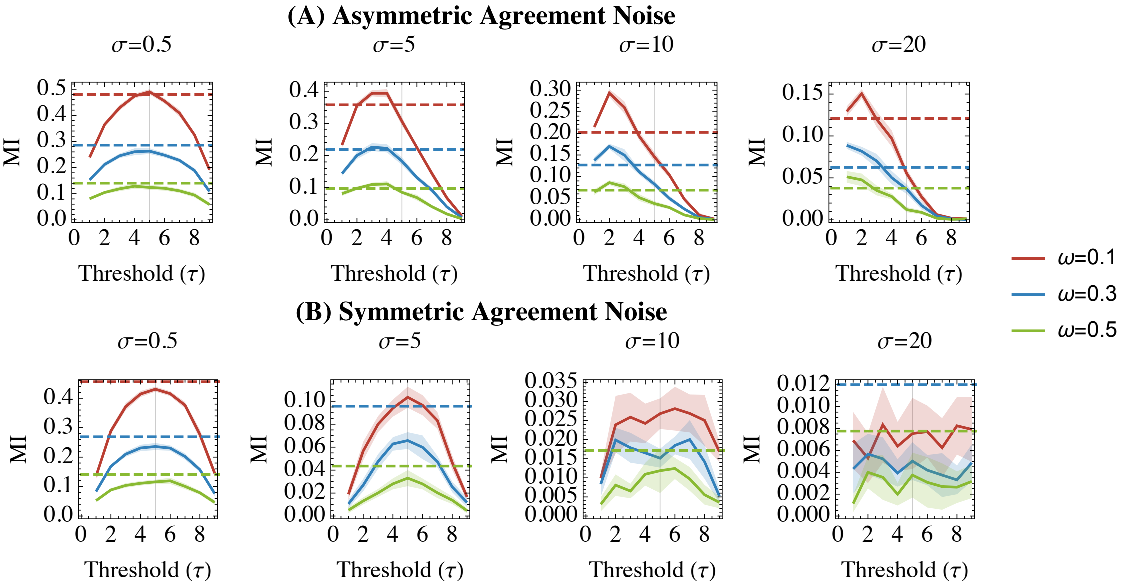


Figure 6.4: Mutual information (MI) for dichotomized (solid lines) and continuous (dashed lines) distributions on synthetic data with asymmetrical (upper row, **Panel A**) and symmetrical (lower row, **Panel B**) properties with respect to agreement. X-axes represent the dichotomization thresholds at which we recalculate the dichotomized MI. Mutual information is depicted on the y-axes. Plot titles indicate the different diagonal spread (σ) parameters used to synthesize the synthetic datasets. Solid lines (for dichotomized MI) are surrounded by ribbons depicting the 95% confidence intervals over 10 runs at each combination of parameters $(\tau, \omega, \beta, \sigma)$.

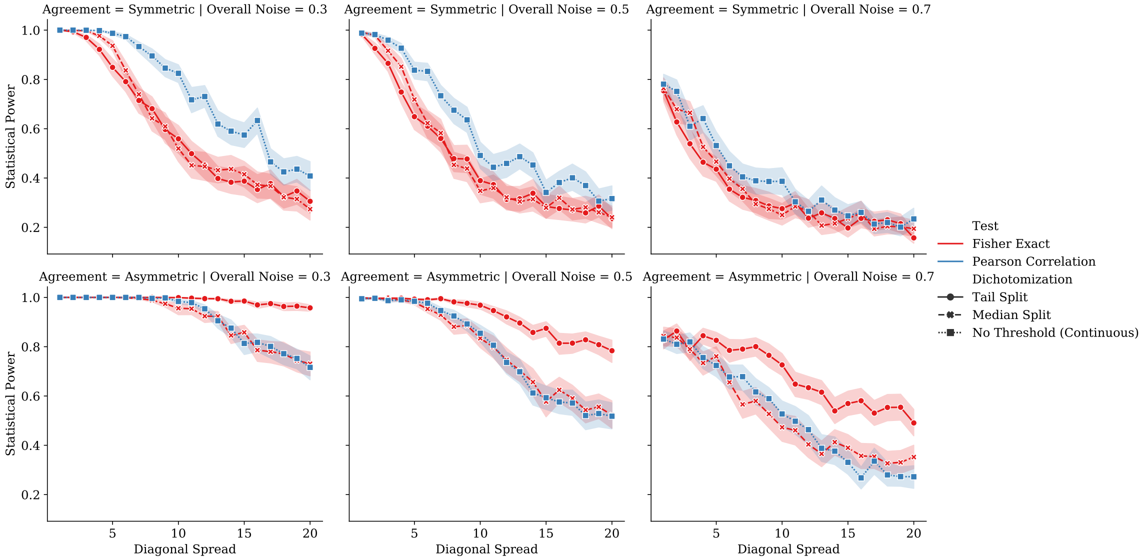


Figure 6.5: Statistical power achieved with the Pearson coefficient (a continuous measure of association; blue lines) and Fisher's exact test (a measure of association between dichotomized variables; red lines) for synthetic data with symmetrical (upper row) and asymmetrical (lower row) properties with respect to agreement. Columns correspond to the level of uniform "overall" noise (ω) added to the data, representing prior uncertainty. X-axes represent the diagonal spread (σ), and the y-axes represent the test's statistical power for the given sample size and estimated effect sizes. Data subjected to Fisher's exact test were dichotomized at either a threshold of 5 (the "Median Split," denoted by '+' markers in red) or 3 (the "Tail Split," denoted by the dot markers in red). For all series, dark lines denote means and the ribbons are 95% confidence intervals over 100 runs.

symmetrical reliability, the continuous test of association (Pearson correlation) retains greater statistical power as the degree of diagonal spread increases, although this difference lessens at very high levels of diagonal spread or overall (uniform) noise. However, under conditions of asymmetrical reliability, dichotomizing data according to a "tail split" (here a threshold of $\tau = 3$) preserves greater statistical power than either a median split ($\tau = 5$) or continuous representation; this relationship was present even at high levels of diagonal spread and overall uniform noise.

6.4 Discussion

The present study makes two important contributions. First, using a sample of 59 ratings obtained using standardized vignettes compared to a consensus-defined gold standard [202], we showed that the dichotomized Alda score has a higher MI between the observed and gold standard ratings than does the raw scale (which ranges from 0-10). Those data suggested that the Alda score's reliability is asymmetrical, with greater inter-rater agreement at the upper extreme. Secondly, therefore, using synthetic experiments we showed that asymmetrical inter-rater reliability in a score's range is the likely cause of this relationship. Our results do not argue that lithium response is itself a categorical natural phenomenon. Rather, using the dichotomous definition as a target variable in supervised learning problems likely confers greater robustness to noise in the observed ratings.

Some have argued that the existence of categorical structure in one's data [224], or evidence of improved reliability under a dichotomized structure [228], are potentially justifiable rationales for dichotomization of continuous variables. These claims are generally stated only briefly, and with less quantitative support than the more numerous mathematical treatments of the problems with dichotomization [224, 225, 228, 229]. However, these more rigorous quantitative analyses typically involve assumptions of symmetrical or Gaussian distributions of the underlying variables in the context of generalized linear modeling (although Irwin & McClelland [225] demonstrated that median splits of asymmetric and bimodal beta distributions is also deleterious). These analyses have led to vigorous generalized denunciation of variable dichotomization across several disciplines, but our current work offers important counterexamples to this narrative [225, 226].

The Alda score is more broadly used as a target variable in both predictive and associative analyses, and not as a predictor variable, which is an important departure from most analyses against dichotomization. Since there is no valid and reliable biomarker of lithium response, these cases must rely on the Alda score-based definition of lithium response as a "ground truth" target variable. In the case of predicting lithium response, where these ground truth labels are collected from multiple raters across different international sites, variation in lithium response scoring patterns across centres might further accentuate the extant between-site heterogeneity.

To this end, inter-individual differences in subjective rating scales may be more informative about the raters than the subjects, and one may wish to use dichotomization to discard this nuisance variance [223, 224, 228]. Doing so means that one turns regression supervised by a dubious target into classification with a more reliable (although coarser) target. Appropriately balancing these considerations may require more thought than adopting a blanket prohibition on dichotomization or some other form of preprocessing.

An important criticism of continuous variable dichotomization is that it may impede comparability of results across studies, both in terms of diminishing power and inflating heterogeneity [229]. However, this is more likely a problem when dichotomization thresholds are established on a study-by-study basis, without considering generalizability from the outset. These arguments do not necessarily apply to the Alda score, since the threshold of 7 has been established across a large consortium with support from both reliability and discrete mixture analysis [202], and is the effective standard split point for this scale [30].

Our study thus provides a unique point of support for the dichotomized Alda score insofar as we show that the retention of MI and frequentist statistical power is likely due to asymmetrical reliability across the range of scores. Our analyses show that there is a range of Alda scores (those identifying good lithium responders; scores ≥ 7) for which scores correspond more tightly to a consensus-defined gold standard in a large scale international consortium. In particular, we showed that this dichotomization will be more robust to increases in the prior uncertainty (i.e. the overall level of background "noise" in the relationship between true/observed scores). This feature is important since the sample of raters included in the Alda score's calibration study [202] was relatively small and consisted of individuals involved in ConLiGen centres. It is reasonable to suspect that assessment of Alda score reliability in broader research and clinical settings would add further disagreement-based noise to the inter-rater reliability data. At present, use of the dichotomized scale could confer some robustness to that uncertainty.

More generally our study showed that if reliability of a measure is particularly high at one tail of its range, then a "tail split" dichotomization can outperform even the continuous representation of the variable. This presents an important counterexample to previous authors, such as Cohen [220], Irwin & McClelland [225], and MacCallum et al. [224] who argued that "tail splits" are still worse than median splits. While our study reaffirms these claims in the case of measures whose reliability is constant over the domain (see Figure 4B and the upper row of Figure 5), our analysis of the asymmetrically reliable scenario yields opposite conclusions, favouring a "tail split" dichotomization over both median splits and continuous representations. Tail split dichotomization was particularly robust when data were affected by both asymmetrical reliability *and* high degrees of uniform noise over the variable's range. Together, these results suggest that dichotomization/categorization of a continuous measurement may be justifiable when its relationship to the underlying ground truth variable is noisy everywhere except at an extreme.

Our study has several limitations. First, our sample size for the re-analysis of the Alda score reliability was relatively small, and sourced from highly specialized raters involved in lithium-specific research. However, one may consider this sample as representative of the "best case scenario" for the Alda score's reliability. It is likely that further expansion of the subject population would introduce more noise into the relationship between ground truth and observed Alda scores. It is likely that most of this additional disagreement would be observed for lower Alda scores, since (A) there are simply more potential item combinations that can yield an Alda score of 5 than an Alda score of 9, for example, and (B) unambiguously excellent lithium response is a phenomenon so distinct that some question whether lithium responsive BD may constitute a unique diagnostic entity [38, 40]. Thus, we believe that our sample size for the reliability analysis is likely sufficient to yield the present study's conclusions.

Our study is also limited by the fact that theoretical analysis was largely simulation-based, and thus cannot offer the degree of generalizability obtained through rigorous mathematical proof. Nonetheless, our study offers sufficient evidence—in the form of a counterexample—to show that there exist scenarios in which dichotomization is statistically superior to preserving a variable's continuous representation. Furthermore, we used well controlled experiments to isolate asymmetrical reliability as the cause of dichotomization's superiority across simulated conditions.

6.5 Conclusion

In conclusion, we have shown that a dichotomous representation of the Alda score for lithium responsiveness is more robust to noise arising from inter-rater disagreement. The dichotomous Alda score is therefore likely a better representation of lithium responsiveness for multi-site studies in which lithium response is a target or dependent variable. Through both re-analysis of the Alda score's real-world inter-rater reliability data and careful theoretical simulations, we were able to show that asymmetrical reliability across the score's domain was the likely cause for superiority of the dichotomous definition. Our study is not only important for future research on lithium response, but other studies using subjective and potentially unreliable measures as dependent variables. Practically speaking, our results suggest that it might be better to classify something we can all agree upon than to regress something upon which we can not.

Chapter 7

Exemplar Scoring Identifies Genetically Separable Phenotypes of Lithium Responsive Bipolar Disorder[1]

Abstract. Predicting lithium response (LiR) in bipolar disorder (BD) could expedite effective pharmacotherapy, but phenotypic heterogeneity of bipolar disorder has complicated the search for genomic markers. We thus sought to determine whether patients with “exemplary phenotypes”—those whose clinical features are reliably predictive of LiR and non-response (LiNR)—are more genetically separable than those with less exemplary phenotypes. We applied machine learning methods to clinical data collected from people with BD (n=1266 across 7 international centres; 34.7% responders) to compute an “exemplar score,” which identified a subset of subjects whose clinical phenotypes were most robustly predictive of LiR/LiNR. For subjects whose genotypes were available (n=321), we evaluated whether responders/non-responders with exemplary phenotypes could be more accurately classified based on genetic data than those with non-exemplary phenotypes. We showed that the best LiR exemplars had later illness onset, completely episodic clinical course, absence of rapid cycling and psychosis, and few psychiatric comorbidities. The best exemplars of LiR and LiNR were genetically separable with an area under the receiver operating characteristic curve of 0.88 (IQR [0.83, 0.98]), compared to 0.66 [0.61, 0.80] (p=0.0032) among the poor exemplars. Variants in the Alzheimer’s amyloid secretase pathway, along with G-protein coupled receptor, muscarinic acetylcholine, and histamine H1R signaling pathways were particularly informative predictors. In sum, the most reliably predictive clinical features of LiR and LiNR patients correspond to previously well-characterized phenotypic spectra whose genomic profiles are relatively distinct. Future work must enlarge the sample for genomic classification and include prediction of response to other mood stabilizers.

[1] Nunes A, Stone W, Ardau R, Berghöfer A, Bocchetta A, Chillotti C, Deiana V, Degenhardt F, Forstner AJ, Menzies JS, Grof E, Hajek T, Manchia M, McMahon FJ, Müller-Oerlinghausen B, Nöthen MM, Pinna M, Pisanu C, O’Donovan C, Rietschel MDC, Rouleau G, Schulze T, Severino G, Slaney CM, Squassina A, Suwalska A, Turecki G, Zvolsky P, Cervantes P, del Zompo M, Grof P, Rybakowski J, Tondo L, Trappenberg T, and Alda M. Exemplar Scoring Identifies Genetically Separable Phenotypes of Lithium Responsive Bipolar Disorder. *Submitted manuscript*.

7.1 Introduction

Bipolar disorder (BD) is a severe lifelong illness characterized by recurrent manias, depressions, and a relatively high suicide risk [230, 231]. Mood stabilizer initiation occurs approximately a decade after symptom onset, on average [193], and the trial-and-error process of pharmacological optimization for BD may lengthen this time. However, by predicting individuals' mood-stabilizer response, this burden of untreated illness may be reduced.

Clinical data are currently the best lithium response predictors. Responders often have a completely episodic course with full inter-episode remissions, absence of rapid cycling, and family history of fully remitting BD (particularly the lithium responsive type) in a first degree relative [36, 198]. This has motivated the search for strong genomic predictors of lithium response, but they remain elusive [30].

In large multi-site studies, lithium responder and non-responder groups may be too heterogeneous to classify robustly. However, it is possible that within this pooled group of heterogeneous subjects there exist more distinct "exemplars" of each phenotype, whose clinical profiles are consistent across sites, and who may be genomically more distinct. Our paper is thus motivated by two questions. First, can clinical presentation identify exemplars of lithium response and non-response? Second, are clinical exemplars of lithium response and non-response more genetically separable than their less exemplary counterparts?

Using the largest clinical database on lithium treatment in BD, we developed a method for rating the degree to which a subject is an exemplar of lithium response or non-response, respectively (an exemplar score). We hypothesized that the clinical differences between the best exemplars of lithium response and non-response would be reflective of factors previously associated with the "classical" bipolar phenotype. Finally, on a subset of subjects who were genotyped, we hypothesized that clinically exemplary responders and non-responders would be more accurately separable by application of a machine learning (ML) classifier to their genomic data (compared to their counterparts with low exemplar scores).

7.2 Methods

Our analysis is split into two parts. In Part 1, we use a multi-centre database of clinical variables in order to derive a score that identifies subjects whose clinical phenotypes reliably predict lithium response/non-response. Part 2 uses a separate set of genomic data collected from a subset of subjects included in the clinical data from Part 1. In Part 2, we compare the ability to classify lithium response using those genetic data when they are stratified according to subjects' *clinical* exemplar scores.

7.2.1 Part 1: Scoring and Characterization of Clinical Exemplars

Data Collection

Clinical data collection procedures were described in Nunes et al. [35]. Data consisted of 180 variables recorded prior to instituting lithium maintenance therapy in 1266 people with BD across 7 sites internationally (Table 5.1). Response was evaluated after a minimum treatment duration of 1 year. Lithium response was defined as a score of ≥ 7 on the previously validated Alda scale [202].

Exemplar Scoring Based on Clinical Predictors

Subjects who are most exemplary of their clinical phenotype should be classified accurately by models trained on data from any given site. Our overall exemplar scoring protocol thus involves (1) obtaining out-of-sample predictions of every subject's class based on models trained on each individual site's data, then (2) summarizing accuracy and level of agreement with which each subject was classified into a single value known as the exemplar score (Figure 7.1).

The Clinical Exemplar Score

Let $(\mathbf{x}_{ij}, y_{ij}) \in \mathcal{X}$ denote phenotypic data from subject $i \in \{1, 2, \dots, n_j\}$, where $\mathbf{x}_{ij}$ is a vector of clinical features, $y_{ij} \in \{0, 1\}$ denotes whether the patient is a lithium responder, and n_j is the number of patients in the sample from site $j \in \{1, 2, \dots, S\}$. A pair $(\mathbf{x}, y)$ can thus be viewed as a set of coordinates on the (observable) phenotypic space $\mathcal{X}$. Data are sampled from S sites, each of which can be considered to sample a subdomain of the phenotypic space $\mathcal{X}^{(j)} \subseteq \mathcal{X}$. These site-wise subdomains are not necessarily disjoint. Indeed, if they were disjoint, the sites' data would share nothing in common.

Now let $\mathcal{M}_j$ denote a classifier learned on training data from site j. Given a new set of clinical features, $\mathbf{x}'$, the classifier predicts the probability that the corresponding patient is a

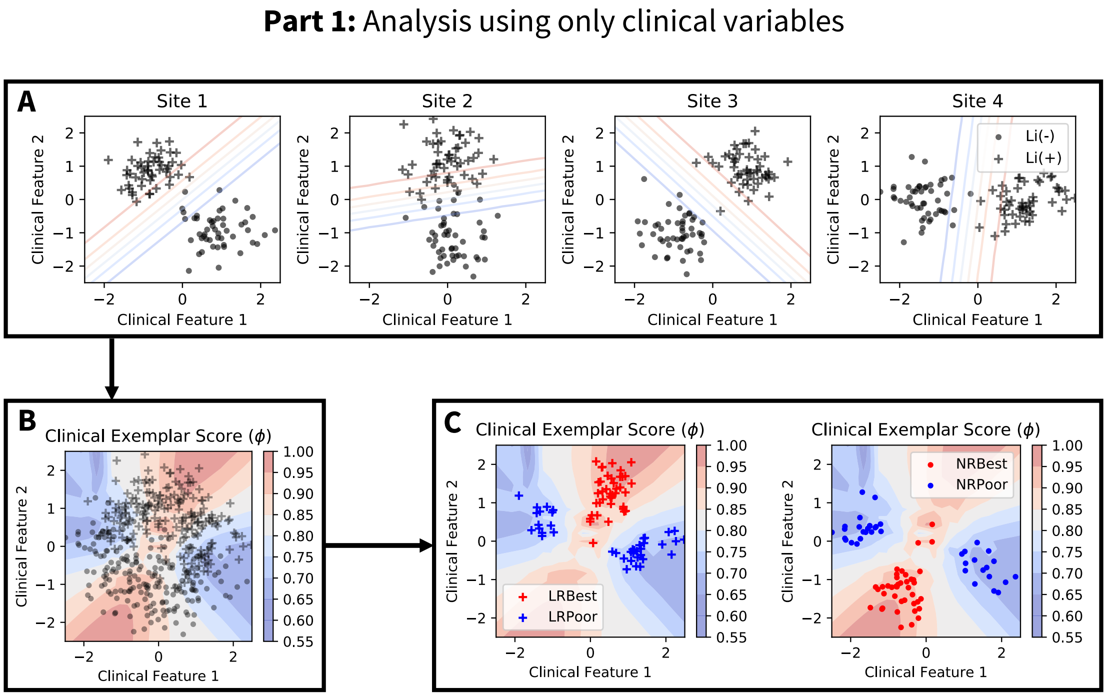


Figure 7.1: Hypothetical illustration of the clinical exemplar scoring analysis. Note that this part of the analysis is performed using the clinical feature dataset alone. **Panel A**: Demonstration of heterogeneity in the relationship between lithium responsiveness (depicted as "Li(+)" for responders and "Li(-)" for non-responders) and clinical features across four hypothetical sites. A classifier trained on data from each individual site may yield different discriminative functions. **Panel B**: Points demonstrate the aggregated dataset ("+" and "-" are responders and non-responders, respectively). Contours demonstrate regions of clinical feature space in which site-level classifiers (from Panel A) agree with high accuracy on the predicted class. An exemplar score can be computed for each subject in the clinical dataset by (1) holding his data out of the training set, (2) predicting his lithium responsiveness using site-level classifiers trained on the remaining subjects, then (3) using the site-wise prediction results to compute the exemplar score. **Panel C**: Stratification of the clinical dataset according to lithium responsiveness and exemplar score quartile. The "LRBest" and "NRBest" exemplars are those responders and non-responders with exemplar scores above the 75th percentile, respectively. The "LRPoor" and "NRPoor" exemplars are those responders and non-responders with exemplar scores below the 25th percentile, respectively. This stratification can be used to evaluate the clinical features that differentiate good from poor exemplars of lithium response and non-response, respectively.

lithium responder: that is, $\hat{p}'_j = \mathcal{M}_j\left(\mathbf{x}'\right)$. We denote the accuracy score of this prediction as

$$\tilde{f}_j\left(\mathbf{x}', y'\right) = 1 - \left|y' - \mathcal{M}_j\left(\mathbf{x}'\right)\right|. \tag{7.1}$$

Recall from Chapter 4 that representational Rényi heterogeneity consists of measuring heterogeneity on a latent or transformed space onto which observable data are mapped. To apply this in the present case, where we have defined our observable space, $\mathcal{X}$, we must now devise an appropriate transformed space upon which the Rényi heterogeneity will be both meaningful and tractable. Hence, we recall from Chapter 5 that the heterogeneity deemed relevant presented in terms of differences in classification models across sites. Most starkly, we noted that the informative features for lithium response prediction varied between the best performing sites. In other words, depending on which site's data are used for training, one might learn quite different (and perhaps even contradictory) relationships between clinical features and lithium responsiveness. In the limit where data from each site encodes completely different relationships between clinical features and lithium response, then each classifier $\mathcal{M}_j$ will behave distinctly (they will tend to disagree). In terms of numbers equivalent, we would say that in such a case there is an effective number of S distinct classifiers. Conversely, if the phenotypic domains of all sites overlap completely, then all classifiers $\mathcal{M}_j$ will tend to make similar predictions, which would correspond to an effective number of one classifier.

Let the accuracy of classifier $\mathcal{M}_j$ in predicting the relationship $\mathbf{x} \to y$ be a measure of that model's informativeness at point $(\mathbf{x}, y)$. We can thus define $\mathcal{T}$ as a categorical space representing an index on "the most informative classifier." We illustrate the mapping $f : \mathcal{X} \to \mathcal{T}$ in Figure 7.2. A probability distribution over $\mathcal{T}$ can be computed using a normalization of Equation 7.1:

$$f\left(\mathbf{x}, y\right) = \left\{ \frac{1 - \left|y - M_j\left(\mathbf{x}\right)\right|}{\sum_{k=1}^{S}\left(1 - \left|y - M_k\left(\mathbf{x}\right)\right|\right)} \right\}_{j=1}^{S}. \tag{7.2}$$

The quantity $f_j\left(\mathbf{x}, y\right)$ can be taken to represent the probability that a classifier trained on data from site j is the most informative about the $\mathbf{x} \to y$ mapping in that particular region of $\mathcal{X}$. With this, we can compute the representational Rényi heterogeneity at $(\mathbf{x}, y)$ as follows:

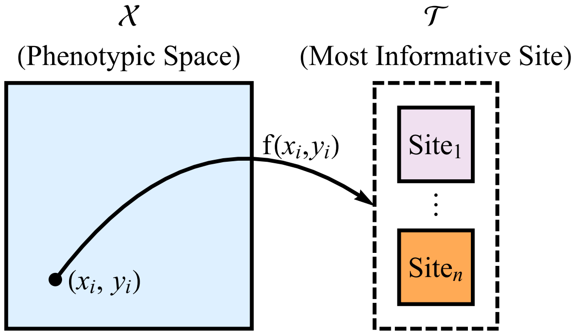


Figure 7.2: Representation of the mapping from phenotypic space $\mathcal{X}$ onto the representation of "most informative site-level model" ($\mathcal{T}$). The transformation function is the normalized accuracy score for a classification model trained on each site's data individually (Equation 7.2).

$$\Pi_q\left(\mathbf{x}, y\right) = \left(\sum_{j=1}^{S} f_j^q\left(\mathbf{x}, y\right)\right)^{\frac{1}{1-q}}. \tag{7.3}$$

If the models $\mathcal{M}_{j=1,2,\dots,S}$ differ only in their training data (i.e. they have the same architecture, optimization routine, and hyperparameters) then the units of Equation 7.3 are "the effective number of informative sites."

Recall that we defined a "clinical exemplar" as a subject whose phenotype $(\mathbf{x}, y)$ is reliably predicted accurately across all sites. In other words, regardless of the differences between sites' data, all sites would agree in their predictions of the exemplars' phenotypes. More formally, clinical exemplars must have high values of $\Pi_q\left(\mathbf{x}, y\right)$ (all sites are similarly informative). However, to identify more specifically the exemplars of lithium response and non-response, we cannot solely rely on $\Pi_q\left(\mathbf{x}, y\right)$, since that value may be high, despite sites' prediction accuracies being low.

Let $t_* = \max_j \tilde{f}_j\left(\mathbf{x}, y\right)$ denote the maximal accuracy score obtained in classification at $(\mathbf{x}, y)$. We take this value to represent the degree to which a subject with that phenotype can be clearly associated with one class or another. An interesting case occurs where both t_* and $\Pi_q\left(\mathbf{x}, y\right)$ are high, suggesting the point $(\mathbf{x}, y)$ is an exemplar of the regions of $\mathcal{X}$ that are reliably well classified across sites. Conversely, if $t_* \approx 0.5$ and $\Pi_q\left(\mathbf{x}, y\right)$ is high, then that point is exemplary of a region of $\mathcal{X}$ of which all sites are uncertain. When t_* is low and $\Pi_q\left(\mathbf{x}, y\right)$ is high, then $(\mathbf{x}, y)$ is exemplary of a region of $\mathcal{X}$ that reliably misleads all sites' classifiers.

In the present study, we are concerned with identifying only those subjects with high values of both t_* and $\Pi_q(\mathbf{x}, y)$, since they exemplify the most canonical "phenotypes" of lithium response and non-response, respectively. We accomplish this by combining t_* and $\Pi_q(\mathbf{x}, y)$ into a single index we call the *exemplar score*. The exemplar score for the phenotypic point $(\mathbf{x}, y)$ is defined as

$$\phi_i = \sqrt{\frac{\tilde{\Pi}_q^2(\mathbf{x}, y) + (t^*)^2}{2}}, \tag{7.4}$$

where $\tilde{\Pi}_q(\mathbf{x}, y)$ is a standardization of the Rényi heterogeneity to the [0,1] interval (the same scale as t^*):

$$\tilde{\Pi}_q(\mathbf{x}, y) = \frac{\Pi_q(\mathbf{x}, y) - 1}{S - 1} \tag{7.5}$$

In the present study, we define the "best exemplars" as subjects whose exemplar scores (within their lithium response classes) were in the top 25%. Poor exemplars were those subjects whose phenotypes were in the lower quartile of exemplar scores within their response classes.

The Predict Every Subject Out (PESO) Protocol

The predict every subject out (PESO) protocol is a method by which we can compute exemplar scores for each subject in the dataset while (A) ensuring that subject is not included in the training data and (B) having each model train on only that site's data. All classifiers in our data were random forests, (RFC) [208] under the same specifications as in Nunes et al. [35] (100 estimators; SciKit Learn implementation; [209]). Similar to that study, missing data were marginalized by sampling from uninformative priors on respective variables' domains [35]. A schematic of the protocol is shown in Figure 7.3.

For each site in the clinical predictors dataset, the PESO analysis protocol begins with a Leave-One-Out cross-validation run to obtain out-of-sample predictions for each of that site's constituent subjects. We then train an RFC on that site's data and predict lithium response in all other sites' subjects. Each subject is thus mapped onto our categorical space $\mathcal{T}$, upon which we can measure their exemplar scores.

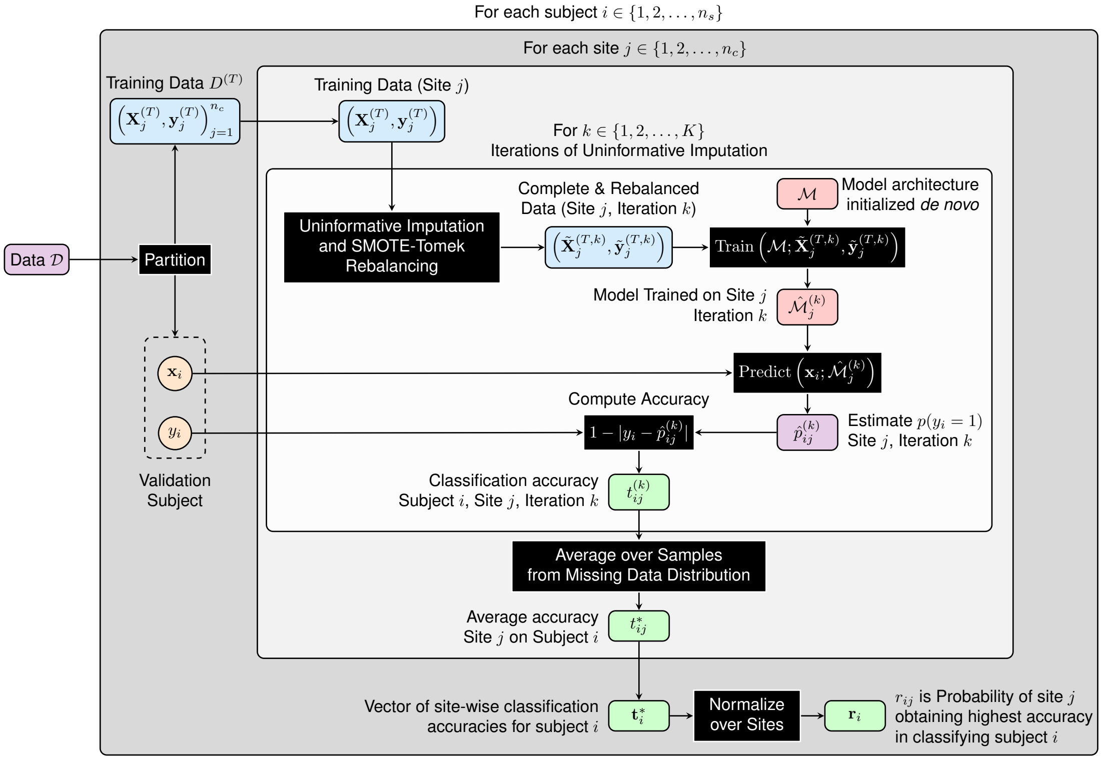

For each subject $i \in \{1, 2, \ldots, n_s\}$
For each site $j \in \{1, 2, \ldots, n_c\}$
Training Data $D^{(T)}$
$\left(\mathbf{X}_j^{(T)}, \mathbf{y}_j^{(T)}\right)_{j=1}^{n_c}$
Training Data (Site j)
$\left(\mathbf{X}_j^{(T)}, \mathbf{y}_j^{(T)}\right)$
For $k \in \{1, 2, \ldots, K\}$
Iterations of Uninformative Imputation
Complete & Rebalanced
Data (Site j, Iteration k)
$\mathcal{M}$
Model architecture
initialized *de novo*
Data $\mathcal{D}$
Partition
Uninformative Imputation
and SMOTE-Tomek
Rebalancing
$\left(\tilde{\mathbf{X}}_j^{(T,k)}, \tilde{\mathbf{y}}_j^{(T,k)}\right)$
$\mathrm{Train}\left(\mathcal{M}; \tilde{\mathbf{X}}_j^{(T,k)}, \tilde{\mathbf{y}}_j^{(T,k)}\right)$
Model Trained on Site j
Iteration k
$\hat{\mathcal{M}}_j^{(k)}$
$\mathbf{x}_i$
$\mathrm{Predict}\left(\mathbf{x}_i; \hat{\mathcal{M}}_j^{(k)}\right)$
Compute Accuracy
y_i
$1 - |y_i - \hat{p}_{ij}^{(k)}|$
$\hat{p}_{ij}^{(k)}$
Estimate $p(y_i = 1)$
Site j, Iteration k
Validation
Subject
Classification accuracy
Subject i, Site j, Iteration k
$t_{ij}^{(k)}$
Average over Samples
from Missing Data Distribution
Average accuracy
Site j on Subject i
t_{ij}^{*}
Vector of site-wise classification
accuracies for subject i
$\mathbf{t}_i^{*}$
Normalize
over Sites
$\mathbf{r}_i$
r_{ij} is Probability of site j
obtaining highest accuracy
in classifying subject i

Figure 7.3: Illustration of the algorithm for the predict every subject out protocol.

Comparison of Clinical Characteristics of the Best and Worst Exemplars

Univariate clinical feature differences were compared between the best exemplars of lithium response and non-response (“LRBest” and “NRBest,” respectively; the upper exemplar score quartile per class), and the corresponding poor exemplars (“LRPoor” and “NRPoor,” respectively; the lower exemplar score quartile per class). Continuous variables were compared using the two-sample permutation test of independence and categorical variables were compared using the randomization chi-square test (with 10,000 replications owing to multiple comparison corrections). The significance threshold was adjusted for 116 comparisons: $\alpha_C = 0.05/116 = 0.0004$.

7.2.2 Part 2: Biological Validation hrough Genomic Classification

Figure 7.4 illustrates Part 2 of the present study, wherein we compare the genetic prediction of lithium response between subjects whose clinical profiles are exemplary and non-exemplary, respectively. After comparing genomic classification performance between the “Best” and “Poor” exemplar strata, respectively, we submit the genomic classifiers’ coefficients to gene enrichment analysis. This part of our study uses genomic data from subjects in the Consortium on Lithium Genetics GWAS cohort [30] who also had detailed clinical information collected for Part 1 of the present study.

Data Collection

Genomic data, obtained as part of the ConLiGen GWAS [30], were available for 321 of the subjects whose clinical data were analyzed in Part 1 of our study. In the Supplementary Materials, we show that there was no population stratification in this subsample, particularly in comparison to the broader ConLiGen sample. We restricted the data to only the 47,465 SNPs for which complete data were available across all ConLiGen sites. Preprocessing and quality control were done according to the Hou et al. [30] protocol.

Genomic Classification Analysis

For genotyped subjects, we compared the performance of a classifier applied to (A) all 321 subject’s genomic data, (B) the worst exemplars’ genomic data, and (C) the best exemplars’

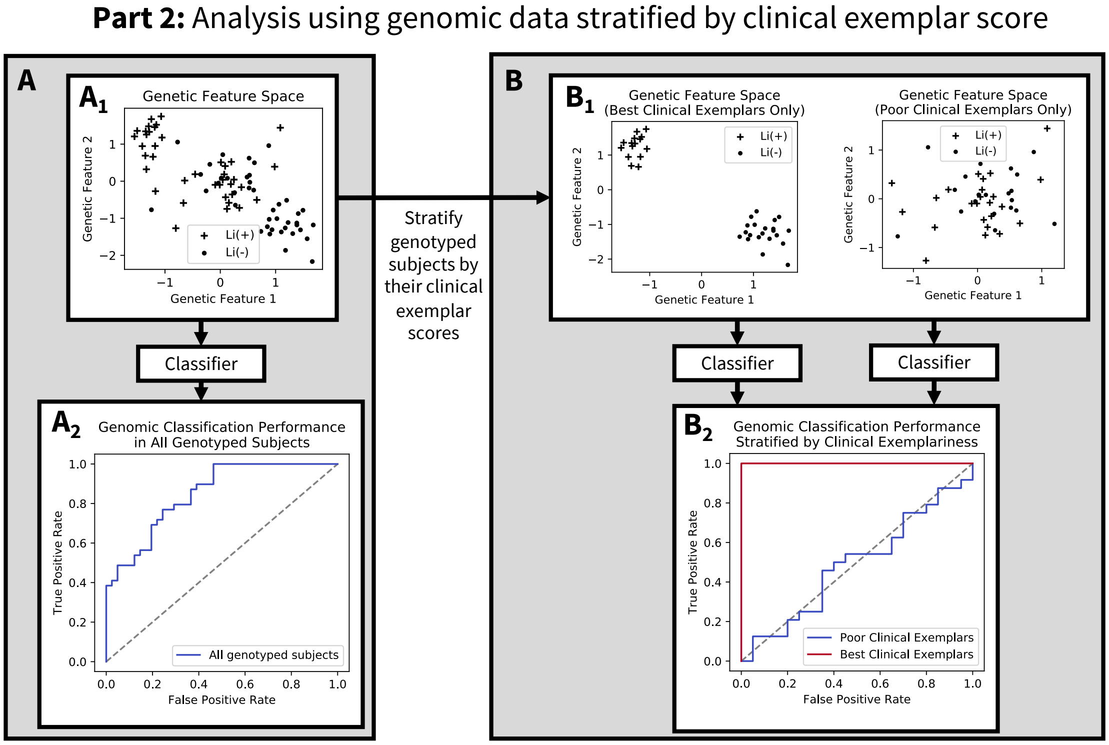

Part 2: Analysis using genomic data stratified by clinical exemplar score
A
A1
Genetic Feature Space
Genetic Feature 1
Genetic Feature 2
Li(+)
Li(-)
Classifier
A2
Genomic Classification Performance in All Genotyped Subjects
True Positive Rate
False Positive Rate
All genotyped subjects
Stratify genotyped subjects by their clinical exemplar scores
B
B1
Genetic Feature Space (Best Clinical Exemplars Only)
Genetic Feature Space (Poor Clinical Exemplars Only)
Classifier
Classifier
B2
Genomic Classification Performance Stratified by Clinical Exemplariness
Poor Clinical Exemplars
Best Clinical Exemplars

Figure 7.4: Hypothetical illustration of Part 2 of this study's analysis, which evaluates the degree to which stratification of genomic data by corresponding subjects' *clinical* exemplar scores can improve genomic classification performance. **Panel A**: Subjects' genotypes lie on a genotypic feature space (shown in Panel A_1 as a simplified 2 dimensional plane). Panel A_2 shows a hypothetical ROC curve for these aggregated data. **Panel B**: Each genotyped subject has an exemplar score computed from Part 1 of the present study. Recall that the exemplar score merely identifies the degree to which a subject's clinical profile (i.e. symptoms, family history, comorbidities, etc.) is reliably predictive of lithium responsiveness. Panel B_1 shows that the aggregated genotyped sample can then be stratified into the "Best" clinical exemplars (subjects with top 25% of clinical exemplar scores within each of the responder and non-responder groups, respectively), and the "Poor" clinical exemplars (those with the lowest 25% of clinical exemplar scores in each responsiveness class). We then apply classifiers to the genomic data in each of these "Best Exemplar" and "Poor Exemplar" strata, respectively, and compare classification performance (Panel B_2). The hypothetical receiver operating characteristic curve in Panel B_2 reflects our hypothesis, that genetic classification of lithium response will be superior among the subgroup of Best clinical exemplars.

genomic data. We employed L2-penalized logistic regression (C=1 set *a priori*). Model criticism was performed under stratified-10-fold cross-validation.

Our primary outcome was the average cross-validated Matthews correlation coefficient (MCC), which is conservative under class imbalance. Classification performance differences were compared between conditions using the Kruskal-Wallis test. Where a statistically significant difference was observed (at $\alpha = 0.05$), pairwise comparisons were done with the Mann-Whitney U tests (at threshold $\alpha_C = 0.05/3 = 0.017$). We secondarily report accuracy, area under the receiver operating characteristic curve (ROC-AUC), Cohen's kappa, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

In the model trained on the best exemplars, we indexed variants whose logistic regression coefficients agreed in sign across all cross-validation folds, then applied a statistical enrichment test to the nearest associated genes using the PANTHER classification system v. 14.1 [232]. To evaluate the relationship between exemplar strata and enriched pathways, we repeated this analysis using logistic regression coefficients from the poor exemplar group. The threshold for statistical significance was set at $\alpha_{FDR} = 0.05$, where FDR indicates correction for false discovery rate. Further gene set analysis details are provided in Appendix C.1.

7.3 Results

7.3.1 Part 1: Scoring and Characterization of Clinical Exemplars

Accuracy Distributions in the Predict Every Subject Out Analysis

A classifier trained on data from the Maritimes site achieved the highest mean overall accuracy (0.59, 95% confidence interval, CI, [0.58, 0.6]; Figure 7.5), which appeared largely driven by that site's ability to accurately classify its own subjects (0.69 [0.66, 0.71]), and those from Montreal (0.71 [0.67, 0.75]). However, Figure 7.5 shows that site-level models' accuracy distributions were highly variable in shape and modality, suggesting heterogeneous classification behaviour between sites.

Characteristics of the Best and Poor Exemplars

Within the clinical dataset of Part 1, there were 110 individuals in LRBest and LRPoor groups, and 207 individuals in the NRBest and NRPoor groups (Table 2). The LRBest group

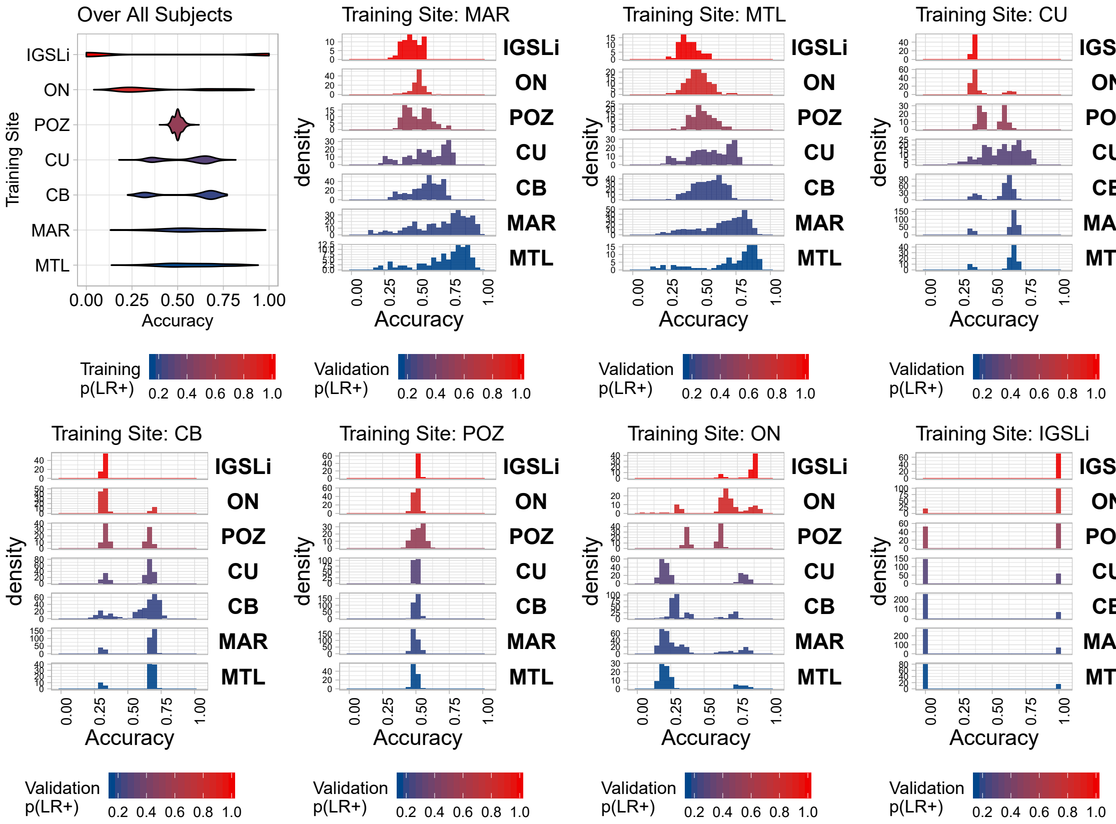


Figure 7.5: Accuracy distributions for models evaluated under the predict every subject out (PESO) regime. The violin plot at the upper leftmost corner shows the accuracy distributions for each site model evaluated over all subjects in the dataset, with the densities colored according to the proportion of lithium responders in the training site's data. The remaining subplots show accuracy histograms for training site models (specified in the titles) stratified across out-of-sample sites. For the site-wise histograms, color indicates the responder/non-responder balance in the respective validation site. *Abbreviations*: Lithium responder (LR+), Cagliari (Centro Bini; CB), Cagliari (University; CU), International Group for the Study of Lithium (IGSLi), Maritimes (MAR), Ontario (ON), Poznan (POZ).

came predominantly from IGSLi (53.6%) and Ontario (21.8%), and most NRBest subjects were from Maritimes (72.5%) and Montreal (25.1%).

Table 7.1: Clinical characteristics of exemplars, by lithium responsiveness. Characteristics of the best (upper 25% of exemplar scores) and poor (lower 25% of exemplar scores) exemplars of lithium response (LiR) and non-response (LiNR), respectively. Categorical data are presented as count (%), whereas normally distributed continuous variables are presented as mean (standard deviation), and non-normal continuous variables are presented as median [interquartile range]. Abbreviations: Calgiari (University; CU), Cagliari (Centro Bini; CB), International Group for the Study of Lithium (IGSLi), Maritimes (MAR), Montreal (MTL), Ontario (ON), Poznan (POZ), bipolar disorder (BD), major depressive disorder (MDD), antidepressants (AD), schizoaffective disorder (SZA), global assessment of functioning (GAF), lithium (Li), suicide attempts (SA), first degree relatives (FDR), second degree relatives (SDR), schizophrenia (SCZ), suicidal ideation (SI), history (Hx), generalized anxiety disorder (GAD), obsessive compulsive disorder (OCD), attention deficit hyperactivity disorder (ADHD), hypertension (HTN), socioeconomic status (SES).

	Best Exemplars			**Poor Exemplars**		
	LiNR	**LiR**	**p**	**LiNR**	**LiR**	**p**
n	207	110		207	110	
Male (%)	76 (36.7)	46 (41.8)	0.398	79 (38.2)	34 (30.9)	0.215
Age (y)	42.4 [32.1, 51.8]	54.2 [42.4, 65.5]	<1e-3	45.9 [36.4, 57.8]	59.7 [44.4, 66.1]	<1e-3
Centre (%)			-			1e-3
CU	4 (1.9)	21 (19.1)		74 (35.7)	1 (0.9)	
CB	1 (0.5)	0 (0.0)		70 (33.8)	11 (10.0)	
IGSLi	0 (0.0)	59 (53.6)		0 (0.0)	8 (7.3)	
MAR	150 (72.5)	6 (5.5)		38 (18.4)	16 (14.5)	
MTL	52 (25.1)	0 (0.0)		11 (5.3)	2 (1.8)	
ON	0 (0.0)	24 (21.8)		6 (2.9)	21 (19.1)	
POZ	0 (0.0)	0 (0.0)		8 (3.9)	51 (46.4)	
Diagnosis (%)			0.124			0.047
BD I	139 (67.1)	71 (64.5)		136 (65.7)	66 (60.0)	
BD II	62 (30.0)	33 (30.0)		51 (24.6)	36 (32.7)	
MDD Recurrent	0 (0.0)	3 (2.7)		3 (1.4)	4 (3.6)	
MDD Single				0 (0.0)	1 (0.9)	
SZA	6 (2.9)	3 (2.7)		17 (8.2)	3 (2.7)	
Age of onset (y)	19. [16., 24.]	28 [21., 36.]	<1e-3	22.5 [18., 32.25]	27.5 [18.25, 35.]	0.166
Onset D (y)	20. [16., 25.]	30 [23., 37.]	<1e-3	28 [20., 38.]	30 [20.50, 37.50]	0.775
Onset M (y)	25. [21., 32.]	30 [26., 40.]	1e-3	29.3 [22., 36.5]	32 [28., 39.7]	0.009
Onset m (y)	26.5 [21., 38.5]	38 [25.5, 45.5]	0.003	32.49 (14.59)	38.13 (12.16)	0.060
Polarity episode 1 (%)			0.0002			0.011

Continued on next page...

	Best Exemplars			Poor Exemplars		
	LiNR	**LiR**	**p**	**LiNR**	**LiR**	**p**
Biphasic (D-M)	4 (2.0)	5 (5.8)		3 (5.8)	1 (2.4)	
Biphasic (M-D)	13 (6.6)	4 (4.7)		2 (3.8)	2 (4.8)	
Hypomania	19 (9.7)	8 (9.3)		10 (19.2)	3 (7.1)	
Major depression	142 (72.4)	42 (48.8)		20 (38.5)	30 (71.4)	
Mania	13 (6.6)	16 (18.6)		16 (30.8)	4 (9.5)	
Minor depression	5 (2.6)	11 (12.8)		1 (1.9)	2 (4.8)	
Clinical course (%)			1e-3			1e-3
Chronic	14 (6.8)	0 (0.0)		8 (4.1)	21 (25.3)	
Chronic deteriorating	2 (1.0)	0 (0.0)		3 (1.5)	2 (2.4)	
Chronic fluctuating	90 (43.5)	0 (0.0)		11 (5.6)	34 (41.0)	
Completely episodic	7 (3.4)	27 (100.0)		146 (74.1)	15 (18.1)	
Continuous cycling	1 (0.5)	0 (0.0)		7 (3.6)	2 (2.4)	
Episodic + residual	93 (44.9)	0 (0.0)		22 (11.2)	9 (10.8)	
N LT manias	3. [1., 7.]	2. [0., 3.]	1e-3	3. [1., 6.]	2. [1., 3.]	0.021
N LT depressions	5. [3., 15.]	3. [2., 6.]	<1e-3	4. [2., 8.]	4. [2., 6.]	0.030
N LT mixed	0. [0., 1.]	0. [0., 0.]	<1e-3	0. [0., 0.]	0. [0., 0.]	0.403
N LT multiphasic	0. [0., 1.]	0. [0., 2.]	1e-3	0. [0., 0.]	0. [0., 0.]	0.184
Total N LT episodes	9. [5., 24.50]	6. [5., 10.]	<1e-3	8. [5., 15.]	5. [4., 9.]	0.005
Rapid cycling (%)			1e-3			0.701
Never	92 (47.2)	59 (98.3)		56 (93.3)	80 (96.4)	
Only on AD	7 (3.6)	0 (0.0)		2 (3.3)	1 (1.2)	
Spontaneous	96 (49.2)	1 (1.7)		2 (3.3)	2 (2.4)	
Rapid mood switch (%)	47 (63.5)	0 (0.0)	0.061	6 (21.4)	1 (1.8)	0.005
LT psychosis (%)			1e-3			0.002
Episodic congruent	83 (42.8)	5 (16.7)		51 (38.9)	15 (20.0)	
Episodic incong.	36 (18.6)	0 (0.0)		8 (6.1)	1 (1.3)	
Never	72 (37.1)	25 (83.3)		70 (53.4)	59 (78.7)	
Outside of episodes	3 (1.5)	0 (0.0)		2 (1.5)	0 (0.0)	
GAF last assessment	70. [55., 75.]	90 [90., 95.]	<1e-3	75 [60., 86.25]	87.5 [80., 90.]	0.013
Li total score	2. [0., 4.]	8. [8., 10.]	<1e-3	3. [1., 5.]	8. [7., 9.]	<1e-3
N episodes on Li	4. [1.25, 10.]	0. [0., 1.75]	0.012	2. [1., 4.]	1. [0., 1.50]	1e-3
N episodes pre Li	4. [3., 12.]	5. [4., 15.75]	0.144	4. [3., 7.]	4. [3., 6.]	0.775
N SA	0. [0., 1.]	0. [0., 0.]	0.003	0. [0., 0.]	0. [0., 0.]	0.155
N significant SA	1. [0., 1.]	0. [0., 0.]	0.0003	0. [0., 0.]	0. [0., 1.]	0.005
Age at SA1 (%)	26. [17., 35.]	20 [18., 36.]	0.752	36.16 (13.87)	33.79 (12.08)	0.670
FDR mood d/o (%)	99 (55.3)	31 (35.2)	0.003	76 (73.8)	22 (40.0)	0.0002
FDR BD (%)	44 (21.7)	9 (10.1)	0.021	62 (51.2)	42 (39.3)	0.080
N FDR BD-I	0. [0., 0.]	0. [0., 0.]	0.003	0. [0., 1.]	0. [0., 1.]	0.055
N FDR BD-II	0. [0., 0.]	0. [0., 0.]	0.716	0. [0., 0.]	0. [0., 0.]	0.899
N FDR Unipolar D	1. [0., 1.]	0. [0., 1.]	0.005	0. [0., 1.]	0. [0., 1.]	0.550
N FDR SZA	0. [0., 0.]	0. [0., 0.]	0.721	0. [0., 0.]	0. [0., 0.]	0.767
N FDR SCZ	0. [0., 0.]	0. [0., 0.]	0.051	0. [0., 0.]	0. [0., 0.]	0.212
N FDR Anxiety	0. [0., 0.]	0. [0., 0.]	0.001	0. [0., 0.]	0. [0., 0.]	0.323
N FDR Unaffected	0. [0., 1.]	0. [0., 0.]	0.0004	3.50 [0., 7.]	0. [0., 0.]	<1e-3
N FDR Suicide	0. [0., 0.]	0. [0., 0.]	0.681	0. [0., 0.]	0. [0., 0.]	0.865

Continued on next page...

	Best Exemplars			Poor Exemplars		
	LiNR	**LiR**	**p**	**LiNR**	**LiR**	**p**
N FDR SA	0. [0., 0.]	0. [0., 0.]	0.222	0. [0., 0.]	0. [0., 0.]	0.073
N SDR Suicide	0. [0., 0.]	0. [0., 0.]	0.366	0. [0., 0.]	0. [0., 0.]	0.668
N SDR SA	0. [0., 0.]	0. [0., 0.]	0.266	0. [0., 0.]	0. [0., 0.]	0.686
Mood at SA (%)			0.338			1
Major depression	74 (91.4)	3 (75.0)		0 (0.0)	0 (0.0)	
Mania	3 (3.7)	1 (25.0)		3 (16.7)	0 (0.0)	
Minor depression	1 (1.2)	0 (0.0)		0 (0.0)	0 (0.0)	
Mixed	2 (2.5)	0 (0.0)		0 (0.0)	0 (0.0)	
Rapid cycling	1 (1.2)	0 (0.0)		0 (0.0)	0 (0.0)	
LT Hx SI (%)	114 (61.3)	18 (34.0)	0.001	61 (44.2)	11 (40.7)	0.826
SI episodic (%)			1			-
No	1 (0.9)	0 (0.0)		0 (0.0)	0 (0.0)	
Sometimes	6 (5.7)	0 (0.0)		0 (0.0)	0 (0.0)	
Yes	99 (93.4)	2 (100.0)		9 (100.0)	8 (100.0)	
Social anxiety d/o (%)	54 (26.6)	0 (0.0)	0.001	8 (4.5)	26 (35.6)	1e-3
Panic d/o (%)	57 (27.9)	2 (2.1)	1e-3	28 (15.5)	43 (48.9)	1e-3
GAD (%)	84 (41.2)	1 (3.6)	1e-3	13 (7.3)	39 (52.0)	1e-3
OCD (%)	29 (14.1)	2 (2.1)	0.003	1 (0.6)	8 (9.2)	0.0004
Substance abuse (%)	78 (37.9)	2 (2.0)	1e-3	43 (21.0)	39 (41.5)	0.001
ADHD (%)	11 (5.5)	0 (0.0)	1	11 (10.6)	45 (60.8)	1e-3
Learning d/o (%)	9 (4.5)	0 (0.0)	1	11 (10.6)	38 (51.4)	1e-3
Primary Insomnia (%)	35 (17.5)	0 (0.0)	0.380	7 (6.7)	9 (11.8)	0.287
Personality d/o (%)	38 (19.1)	0 (0.0)	0.375	3 (3.4)	23 (31.9)	1e-3
Diabetes mellitus (%)	20 (10.3)	0 (0.0)	0.600	6 (8.3)	5 (7.5)	1
HTN (%)	22 (11.4)	2 (20.0)	0.610	17 (23.6)	35 (53.0)	0.001
Menstrual d/o (%)	39 (34.2)	3 (60.0)	0.348	8 (26.7)	2 (4.7)	0.014
Thyroid disease (%)	55 (29.3)	2 (33.3)	1	18 (32.1)	8 (11.9)	0.008
Head injury (%)	48 (27.0)	1 (20.0)	1	17 (34.0)	24 (39.3)	0.698
Migraine (%)	44 (23.5)	2 (33.3)	0.622	11 (19.3)	9 (13.8)	0.474
SES (%)			1e-3			1e-3
Disabled	65 (36.3)	1 (3.4)		6 (3.4)	3 (4.0)	
Other	12 (6.7)	8 (27.6)		23 (13.2)	0 (0.0)	
Retired	8 (4.5)	7 (24.1)		25 (14.4)	22 (29.3)	
Social assistance	32 (17.9)	2 (6.9)		4 (2.3)	3 (4.0)	
Unemployment ins.	18 (10.1)	0 (0.0)		7 (4.0)	3 (4.0)	
Unknown	2 (1.1)	1 (3.4)		1 (0.6)	0 (0.0)	
Work full-time	30 (16.8)	10 (34.5)		96 (55.2)	29 (38.7)	
Work part-time	12 (6.7)	0 (0.0)		12 (6.9)	15 (20.0)	
Marital status (%)			1e-3			0.049
Divorced	47 (23.3)	2 (6.7)		16 (8.1)	9 (11.0)	
Married	84 (41.6)	19 (63.3)		118 (59.6)	51 (62.2)	
Single	67 (33.2)	2 (6.7)		51 (25.8)	11 (13.4)	
Widowed	4 (2.0)	7 (23.3)		13 (6.6)	11 (13.4)	

The LRBest group showed a later age of onset (median 28y, interquartile range, IQR

[21, 36]) compared to NRBest (median 19, IQR [16, 24]; p<0.00001).

The LRBest subjects for whom clinical course information was available all showed a completely episodic course, whereas NRBest courses were mainly chronic fluctuating (43.5%) and episodic with residual symptoms (44.9%). These differences were statistically significant at the omnibus level (p=0.0001). Interestingly, differences in clinical course between LRPoor and NRPoor were opposite in direction to those observed among best exemplars. NRPoor subjects had predominantly completely episodic clinical courses (74.1%), whereas LRPoor subjects exhibited predominantly chronic fluctuating (41%) and chronic (25.3%) courses, with only 18.1% being completely episodic (omnibus p=0.0001).

The complete absence of rapid cycling was reported in 98.3% of LRBest, and in only 47.2% of NRBest (p=0.0001). The remaining majority of the NRBest subjects (49.2%) reported having experienced spontaneous rapid cycling. The occurrence of rapid cycling was no different between LRPoor and NRPoor groups.

The occurrence of lifetime psychosis differed between LRBest and NRBest, with a total of 42.8% of the non-responders reporting episodic and mood congruent psychosis (compared to only 16.7% of responders; p=0.0001). Non-responders also reported incongruent episodic psychosis in 18.6% of cases, with only 37.1% of non-responders reporting an absence of psychosis altogether. In contrast, 83.3% of the best exemplars of lithium response reported a complete absence of lifetime psychosis.

The LRBest group had a lower rate of panic disorder (2.1% vs. 27.9%; p=0.0001), generalized anxiety disorder (3.6% vs 41.2%; p=0.00025), and substance abuse (2% vs. 37.9%; p=0.0001) than NRBest. There was also a general trend toward lower rates of psychiatric comorbidity in LRBest compared to the NRBest group. Social anxiety disorder was present in 0% of lithium responders but 27.9% of non-responders (p=0.0007). Responders also had relatively lower rates of obsessive-compulsive disorder (2.1%) compared to non-responders (14.1%; p=0.0025). These findings were largely reversed when looking at the poor exemplars. LRPoor subjects had higher rates of social anxiety disorder (35.6% vs 4.5%; p=0.0001), panic disorder (48.9% vs 15.5%; p=0.0001), generalized anxiety disorder (52% vs 7.3%; p=0.0001), substance abuse (41.5% vs 21.0%; p=0.0005), attention deficit hyperactivity disorder (60.8% vs. 10.6%; p=0.0001), learning disability (51.4% vs 10.6%; p=0.0001), and personality disorder (31.9% vs 3.4%; p=0.0001) compared to the NRPoor subjects.

7.3.2 Part 2: Biological Validation through Genomic Classification

Recall that the genomic data for this element of the analysis are derived from a single site in the ConLiGen data. In Appendix C.2, we demonstrate relative lack of genomic population stratification in this subset, with a comparison to the broader ConLiGen sample.

Genomic Classification among the Best and Poor Exemplars

Genotyped subjects overlapped with clinical data from the Maritimes (n=129; 40%), Montreal (n=74; 23%), Ontario (n=62; 19%), and IGSLi (n=56; 17%), although in the ConLiGen GWAS [30], they were all classified as from the Maritimes (Dalhousie University). Most clinical differences reflect those reported in Section 7.3.1 and thus are reported in Table C.1.

Genomic classification results are presented in Figure 7.6 and in tabular fashion in Table C.2. The median MCC for classification of the Best exemplars was 0.58 (IQR [0.41, 0.77]), which was greater than classification analyses with either the poor exemplars (0.29 [0.06, 0.5]; p=0.0043), or the entire dataset (0.32 [0.2, 0.44]; p=0.002). The ROC-AUC for classification of lithium response in the Best exemplars was 0.88 [0.83, 0.98], which was greater than that of the model trained only on poor exemplars (0.66 [0.61, 0.80]; p=0.0032) or the whole dataset (0.7 [0.62, 0.75]; p=0.001).

Figure 7.7 shows pathway analysis results for the best exemplars. Enriched pathways involved (A) muscarinic acetylcholine receptor types 1 and 3 signaling (mAChR1/3; 27 genes, false discovery rate FDR=0.017), (B) Alzheimer disease-amyloid secretase (30 genes, FDR=0.034), (C) heterotrimeric G-protein coupled receptor Gq/Go α signaling (GPCRq/o-α; 53 genes, FDR=0.04), and (D) histamine H1R mediated signaling (H1R; 27 genes, FDR=0.039). Complete gene set analysis results are shown in Table C.3. Enrichment studies in the gene ontology "cellular component" and "biological function" categories are shown in Tables C.4 and C.5.

7.4 Discussion

Individuals who are most phenotypically representative of lithium response and non-response may be more genetically distinct than their less exemplary counterparts, particularly in genes related to GPCRq/o-α, mAChR1/3 or H1R signaling, and the Alzheimer's amyloid-secretase pathway. Exemplars also showed distinct clinical profiles that are consistent with past

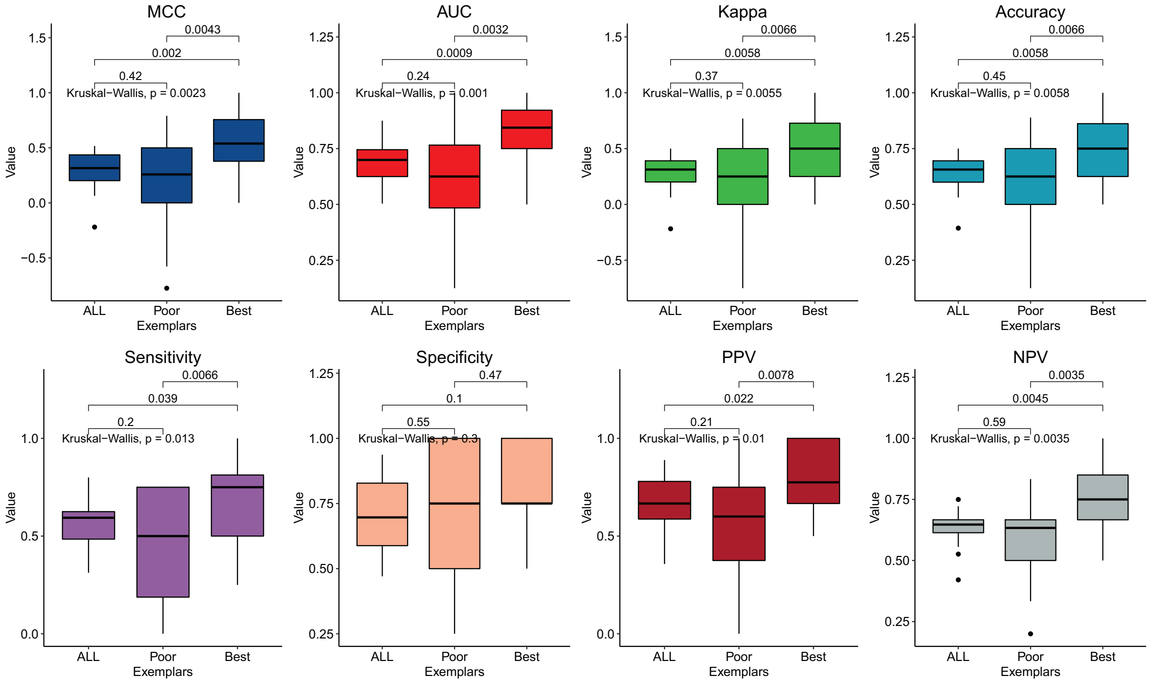
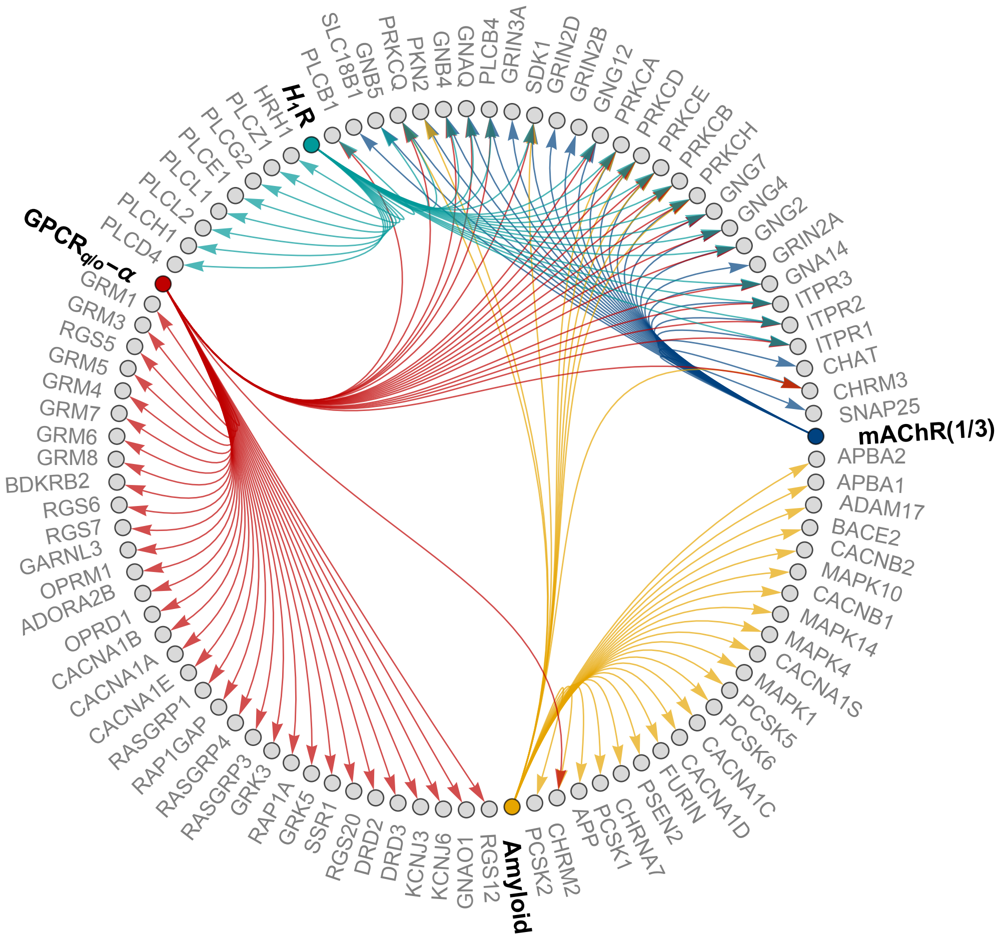


Figure 7.6: Genomic classification results. Results of classifying lithium response based on the genomic data of all subjects (“ALL”; n=321), the poor exemplars (<25th percentile of exemplar score; n=81), and the best exemplars (>75th percentile of exemplar score; n=79). Boxes are defined by the interquartile range (IQR), with the median shown as the black centered line. Whiskers are 1.5 times the IQR. Each panel shows the results for a different classification performance metric. *Abbreviations*: Matthews’ correlation coefficient (MCC), area under the receiver operating characteristic curve (AUC), Cohen’s kappa (Kappa), positive predictive value (PPV), negative predictive value (NPV).



Pathway	N_{genes}	±	**p**	**FDR**
🟦 Muscarinic AChR$_{1,3}$ signaling pathway	27	+	0.0001	0.017
🟨 Alzheimer disease–amyloid secretase pathway	30	+	0.0005	0.034
🟥 Heterotrimeric GPCR$_{q/o}$ α signaling pathway	53	+	0.0008	0.041
🟩 Histamine H_1 R signaling pathway	27	+	0.001	0.039

Figure 7.7: Gene enrichment in the best exemplars. Results of the statistical enrichment test using the logistic regression coefficients from the classifier trained on the best exemplar. Individual genes are shown in gray, with pathway nodes (and edges) colored according to the pathway identity. Pathway names are shown in bold along the perimeter of the graph. *Abbreviations*: acetylcholine receptor (AChR), G-protein coupled receptor (GPCR), histamine H1 receptor (H1R), false discovery rate (FDR).

phenotypic research on lithium responders. Since clinical exemplars are more genetically separable, our study confers a measure of biological validity upon the practice of detailed clinical evaluation, whose predictive utility we have previously demonstrated [35].

One of our most important findings was characterization of the LRBest group as individuals with (A) a predominantly completely episodic clinical course, (B) low levels of psychiatric comorbidity, (C) later age of onset, (D) a general absence of rapid cycling, and (E) either absence of psychosis or limitation to mood congruent intra-episodic form. The first two findings are likely the strongest since we observe the opposite pattern among the LRPoor and NRPoor groups. Notwithstanding, all of these elements support past evidence on the clinical phenotype of lithium responsive bipolar disorder. For instance, Passmore et al. [233] found that lithium responders generally had a more episodic course of illness, whereas lamotrigine responders were more likely to have experienced rapid cycling, a higher rate of psychiatric comorbidity, and an earlier age of onset. A later age of onset in lithium responders has been demonstrated in meta-analysis [196, 234]. Absence of rapid cycling has also been associated with good lithium response by Backlund et al [213] and Tondo et al. [214]. Finally, Kleindienst & Greil [235] found that carbamazepine responders were more likely to have had mood incongruent psychosis than lithium responders, while the updated meta-analysis by Hui et al. found an association between absence of psychotic symptoms and lithium responsiveness [196]. Aside from not including family history related variables (potentially an artifact of related variable definitions), the clinical picture of the exemplary lithium responder that emerges from our study largely aligns with that noted by several authors, such as Grof [36], Gershon & Malhi [37], and Alda [236].

Recently, Kendler [237] reminded us that the utility of biological tests, such as the electrocardiograms and troponin assays used to detect myocardial infarction, is generally contingent upon the clinician's identification of candidate patients whose presentations are clinically consistent with the illness being targeted. The present study, which shows that refinement of a clinical sample into those whose phenotypes are clinically most exemplary of the target syndrome, provides strong data-driven support for Kendler's statement. Further still, we have noted that the clinical picture of the exemplary lithium responders (and non-responders) has been hypothesized for some time, and our study now provides biological support for the predictive validity of these phenotypic hypotheses. Specifically, we were able to genomically classify the best clinical exemplars of lithium response and non-response with a ROC-AUC of 0.88 (IQR [0.83,0.98]), whereas poor exemplars of these classes could only be discriminated with a ROC-AUC of 0.66 (IQR [0.61,0.80]; p=0.0032). If there was no biologically mediated information in the exemplary phenotype of lithium response (and non-response), then this difference would not have been observed.

Variants most informative in discrimination of the best exemplars showed enrichment of genes involved in the heterotrimeric GPCRq/o-α, mAChR1/3 or H1R signaling, and the Alzheimer's amyloid-secretase pathway. Lithium response and BD have long been associated with GPCR signaling [238]. In particular, lithium may affect signaling in both the Go-alpha pathway (at least via adenylate cyclase) and the Gq-alpha pathway (via effects on 1,4,5-triphosphate and protein kinase C, PKC) [39, 41, 194, 239, 240]. Interestingly, our results imply that differences in GPCR signaling may be segregated according to medication responsiveness. Enrichment in the Alzheimer's amyloid-secretase pathway is interesting given the growing interest in the effects of lithium on Alzheimer's pathology. Alterations in cholinergic and histaminergic systems have figured less prominently in the biological literature on BD and lithium response. However, note that Figure 7.7 shows that many genes enriched in the cholinergic and histaminergic systems were also enriched in the GPCR and Alzheimer's amyloid pathways (which comparatively have more individual genetic associations). It is possible that alterations in the cholinergic and histaminergic systems may be subcomponents of the broader differences in the GPCR and Alzheimer's amyloid systems. In future work, it would be of interest to characterize a more fine-grained "gradient" of genetic differences across the spectrum of exemplar scores, and to further evaluate the significance of cholinergic and histaminergic system enrichment in our study.

One limitation of our study includes the relatively low sample size for the genomic analysis. Future work could endeavor to obtain further genotypic information for individuals in our clinical database, or detailed clinical information for individuals in our genomic database. As features, our study also only used those SNPs that overlapped across genotyping platforms in the ConLiGen dataset. Unfortunately, however, the number of fully imputed variants was on the order of millions, which would be analytically intractable in the present context. Filtering-based feature selection approaches in our present study would be (A) too computationally expensive across these millions of variants and (B) require much larger sample sizes since they must be repeated within each training partition. We also had no dominant a priori biological rationale for limiting the data to a restricted subset, since, as our results later confirmed, these biological systems may differ between exemplar strata. Ultimately, we chose the set of completely genotyped SNPs that overlapped across ConLiGen sites in order to facilitate the potential conceptual generalizability of our pathway analysis results, in particular. That is, since the pathways detected were based on variants that are broadly genotyped, these results could potentially be extended to other ConLiGen sites, should the corresponding clinical variables become available.

Our study is also limited by its focus on lithium response, at the exclusion of other mood stabilizers. It is therefore possible our lithium responders are simply those with a more generally responsive form of BD. The only way to prove specificity would be to obtain data showing a single subject's non-response to other mood stabilizers and response to lithium. That being said, there is evidence that excellent response to lithium may be exclusive to that medication [194]. After further validity checks on larger samples of genomic data in lithium responders and non-responders, it will be of great interest to examine exemplar-based genomic classification of mood stabilizer response more broadly. Such work could potentially advance the development of joint clinical-biological prediction models for mood stabilizer response.

Chapter 8

Discussion

Abstract. We have developed a method for heterogeneity measurement, *representational Rényi heterogeneity* (RRH), that is interpretable, flexible, and useful, particularly for computational psychiatric research. In addition to, and as a result of our primary contribution, we have also made significant advancements toward improving the understanding and management of bipolar disorder (BD). This argument is synthesized in light of the evidence presented in Chapters 2-7. This thesis' main contributions are reviewed, including (A) our outlining the desiderata for a heterogeneity measure suitable for psychiatric research, (B) introduction of the representational Rényi heterogeneity framework, and (C) our demonstration of its utility in deriving an exemplar score, which allowed us to (D) identify canonical clinical phenotypes of lithium responsiveness in BD upon whom (E) strong results were obtained in treatment response prediction with genomic markers. We also comment on the ancillary statistical contribution made in Chapter 6, wherein we demonstrate that asymmetrical reliability across the domain of a noisy measurement creates a situation in which dichotomization of a continuous variable is appropriate. This latter finding is also significant for the large body of studies on lithium responsiveness in humans with BD. Throughout our discussion, we touch upon the importance of our findings, their limitations and immediate opportunities, and open questions for longer term research.

8.1 One Measure for Many Systems

Heterogeneity is the degree to which a system diverges from a state of perfect internal conformity. It is important across many scientific fields, whose perspectives on it differ primarily with respect to their systems of interest. In the present work, our focus has been the development of a heterogeneity measure that can be applicable to psychiatric research. This introduces a number of challenges, most significantly that the concept of heterogeneity has not been operationalized. For instance, consider that ecologists are interested in biodiversity, which is merely the heterogeneity of the distribution of species or biological functions in an ecosystem. Here, the system of interest is a community of organisms whose event space includes a set of categorical labels (species classifications) with or without associated data on functional traits. As another example, economists are interested in wealth inequality, which is analogous to the heterogeneity of wealth ownership. Here, one's system is the set of individual wealth-owning entities, and the abundance function is a count of all resources in each entity's possession. Thus, in ecology and economics, operational definitions of heterogeneity follow quite naturally from the definition of their systems of interest (see Chapter 4, Example 3 and Table 4.2). Such clarity of system definition does not exist in psychiatric research at present, which may be an important factor in delaying our discovery of appropriate heterogeneity measurement methods. Another reason why heterogeneity in psychiatric research has not been viewed as a unified concept with heterogeneity measurement in ecology and economics is that heterogeneity's impact may differ between fields. In ecology, biodiversity may impact ecosystem function [2]. In economics, heterogeneity of wealth ownership may impact sociopolitical functions [3, 4]. In psychiatry, the heterogeneity of clinical conditions may compromise diagnostic and treatment effect estimation, among other things [10]. Notwithstanding, heterogeneity is in all cases a single statistical phenomenon: the degree to which a system diverges from a state of perfect conformity.

8.2 Desiderata for a Heterogeneity Measure Suitable for Psychiatric Research

Our first task was thus to develop a better, more precise definition of what it means to measure heterogeneity in computational psychiatry. To this end, we conducted a broad survey of more than a century of research on heterogeneity measures, with further specific focus on how they have been applied in psychiatry. We hypothesized that heterogeneity is not uniquely defined across fields, but rather that it is a statistical concept that can be broadly applied to systems that may themselves be defined differently. Indeed, the ecological and economic definitions of heterogeneity merely swap organism sample counts for wealth amounts, and species labels for proofs of ownership. Ultimately, we found that psychiatric research studies tend to view heterogeneity as comprising either *deviance* (the degree to which a system's configurations differ from each other) or *multimodality* (the number and degree to which a system's configurations cluster together or form categories).

The common property to both deviance and multimodality is that adding them to any given system will tend to increase the number of unique observations that system will yield. Consider that the evolution of a diffusion process from a single point in a chamber of finite volume results in a greater number of unique configurations of particles as time progresses (i.e. entropy generates deviance over time). If one considers the relevant state space as categorical, even though those categories may be inferred by the clustering of continuous data, then adding clusters corresponds to increasing the number of distinct categorical system configurations. Thus, both deviance and multimodality are forms in which a system's number of configurations (the size of its event space) may increase.

Our review found that a single family of measures, the *Rényi heterogeneity* [17, 18, 91], precisely measures the effective number of configurations a system may occupy. This measure furthermore has useful properties, which most importantly include satisfaction of the principle of transfers (Axiom 6; [85, 86]), the replication principle (Axiom 7; [21–23, 90]) and decomposability [59, 93]. The principle of transfers must be satisfied since the diversity of system configurations must increase as the system's sampling probability diffuses further over the space of configurations (i.e. an "equality increasing" transfer), with heterogeneity maximized when all configurations are equally likely. The replication principle is critical to ensure that the physical analogy of sizes and volumes are being used in the formulation of heterogeneity [22]. Finally, decomposability is important because any system with multiple configurations can be broken down into subsets of those configurations, which are heterogeneous in themselves; thus, the operation of splitting and recombining of these subsets must result in a conservation of heterogeneity.

Rényi heterogeneity also accounts for the ways in which other fields have viewed heterogeneity as comprised of inequality and set size (adding more equally probable events increases heterogeneity). The Rényi heterogeneity fails, however, when we are trying to measure heterogeneity as a combination of both deviance and multimodality, such as when our data are drawn from non-categorical spaces. Ecologists have developed several approaches for this problem but they unfortunately all rely on (A) being able to group data on the observable space into categories, and (B) assuming that pairwise distances between these categories with respect to their observable non-categorical features is relevant and can be calculated using a closed form expression.

What happens if we do not have the knowledge or ability to group data on the observable space *a priori*? This is in fact the typical scenario encountered in psychiatric research. While we have a basic system for categorization known as the *Diagnostic and Statistical Manual of Mental Disorders* [69], these are merely symptom checklists that have dubious validity and poor reliability across diagnosticians [178]. Furthermore, what happens when distances on the observable space are misleading? This, too, is a common scenario in psychiatry.

For instance, consider that small imperceptible translations or noise corruption of an image can cause its self-distance on the observable space to deviate from 0 despite the intrinsic semantic content of that image remaining intact. For instance, using the dataset from Haxby et al. [241] in the `nilearn` package for the Python programming language, the reader can verify for him or herself that variation in neural activation to semantically equivalent stimuli occur between testing sessions for the same subjects.

Thus, our review (Chapters 2 and 3) suggested that a suitable measure of heterogeneity for computational psychiatric research must capture both deviance and multimodality, without requiring *a priori* knowledge of a grouping structure or pairwise distance measure on the observable space.

8.3 Representational Rényi Heterogeneity

Representational Rényi heterogeneity (RRH) was introduced in Chapter 4 is an approach by which heterogeneity can be measured such that (A) deviance and multimodality are both captured, and (B) the requirements of categorization and pairwise distance measurement on the observable space are avoided. To be clear, RRH is not itself a novel measure of heterogeneity *per sé*, but rather a conceptual framework within which one may reorganize the assumptions involved in heterogeneity measurement such that the statistical measure's assumptions remain relatively constant across applications. Rather, the primary differences in assumptions between studies will be related to definition of the event space whose size is being measured.

Representational Rényi heterogeneity involves transforming the space of observable data into one upon which Rényi heterogeneity (Equation 4.3) is both tractable and semantically relevant. In doing so, we can inherit the properties of interpretability associated with Rényi heterogeneity discussed above (with further details in Chapter 2, 3, and 4). It will satisfy the replication principle, allowing us to exploit the fact that most people can reason intuitively about sizes. It will consequently scale linearly with growth in the event space. We showed that this property also holds for continuous distributions in Chapter 4 and Appendix A.1.

Furthermore, the units of RRH will remain as numbers equivalent [20], which allows us to make the domain-specific assumptions more clear. Specifically, when we report the "effective number of X," we must be clear about what X is, and it must be relevant to the domain of study at hand. This benefit could be made no clearer than when our example of measuring mood state heterogeneity (see Chapter 2) was criticized in peer-review because our assumption of only three mood states (i.e. mania, depression, and euthymia) was overly simplistic and could not truly capture the heterogeneity of a mood disorder. We agreed wholeheartedly with this criticism, and in fact view it as a strong endorsement of our approach: for the reviewer was able to provide such detailed and relevant criticism precisely because Rényi heterogeneity requires assumptions about the system's event space to be made clear. We believe a great strength of the RRH approach is that the necessary assumptions are shifted further toward the scientifically relevant aspects (such as how one defines the event space of mood states), rather than toward statistical aspects (such as whether observable space distances were defined using an optimal metric). Applied scientists are likely to find greater utility in discussions regarding the former, rather than the latter set of assumptions.

There are other numbers equivalent measures of heterogeneity for non-categorical spaces, namely the numbers equivalent Rao's quadratic entropy ($\hat{Q}_e$), the functional Hill numbers (F_q), and the Leinster-Cobbold index (L_q). In Chapter 4, we showed that RRH yielded better interpretability and flexibility than these approaches. In the case where some continuous data are mapped onto a categorical space (we used a beta mixture distribution in our example; Section 4.3.1), we found that $\hat{Q}_e$ required even further restrictions on the distance matrix: namely that it must be ultrametric. The functional Hill numbers unfortunately showed conditions under which it increased while the distribution over mixture components became more uneven. This violated the principle of transfers (Axiom 6). Furthermore, F_q places particular importance on the class distributions, since when class probabilities are equal, F_q loses all sensitivity to dissimilarities (Appendix A.1). The L_q avoided these particular pitfalls by neither violating the principle of transfers, nor relying on ultrametric distances. In fact, RRH and L_q were similar in one respect, insofar as when data in a mixture distribution had effectively one mode (i.e. when components overlapped), these measures were insensitive to the mixture component weights (Figure 4.5). However, we note that RRH required neither pre-specification of a categorical grouping structure nor a pairwise distance matrix on the input space, unlike the existing measures. Chapters 2 and 3 suggested these properties were important for a heterogeneity measure's applicability in psychiatric research.

Despite not requiring a pairwise distance metric specified on the observable space, we showed that the RRH could be applied to measure the heterogeneity of semantic content of natural images (Section 4.4.2). Note that it is trivial to show that a distance metric on the space of observable images would be sensitive to imperceptible and semantically relevant perturbations such as translations or subtle changes in pixel intensity. However, in Section 4.4.2 we learned a hierarchical abstraction of the semantic features of handwritten digit images using a convolutional variational autoencoder (cVAE; [182]) which is translation invariant and relatively more robust than raw pairwise distance measurement. Hence, much potentially irrelevant variation on the observable space can be abstracted away by the model such that one can focus on measuring the heterogeneity of the semantically relevant features.

For many of the same reasons that RRH is interpretable, it is also flexible. One needs only a map from the observable space onto a transformed representation that can be tractably submitted to Equation 4.3, and has scientifically relevant interpretation in terms of numbers equivalent (or effective hypervolume of the event space in the continuous setting). Indeed, many existing heterogeneity indices are special cases of RRH (Figure 8.1). Moreover, we showed in Chapter 4 (Example 3, Table 4.2) that existing approaches to the measurement of biodiversity and economic equality are special cases of RRH. In each of these cases, the core interpretation of heterogeneity as the effective size of the system's event space does not change. The core differences in assumption relate primarily to how the event space is defined. In so doing, we have maintained a consistent interpretability of heterogeneity as a measurement while enabling its application across various domains. An associated benefit of this approach is that the assumptions around definition of the event space must be made clear, and are therefore more easily submitted to critical appraisal.

We note that differences between RRH, $\hat{Q}_e$, F_q, and L_q largely result from their different concepts of the *idealized reference system* (Section 4.2). Thus, given the unification of categorical Rényi heterogeneity across disciplines under the RRH framework, and the absence of counterintuitive behaviours under various experimental tests, we would not recommend use of the other indices unless (A) their specific notions of idealized reference systems are appropriate for the investigator's domain-specific problem, or (B) there already exists significant experience with the particular index for a given dataset, such that comparisons can be easily made to existing measurements as references. That being said, one must apply RRH carefully. We recommend the following steps:

1. Define the observable space $\mathcal{X}$ clearly

2. Define the latent/unobservable space $\mathcal{Z}$ clearly and justify the chosen topological

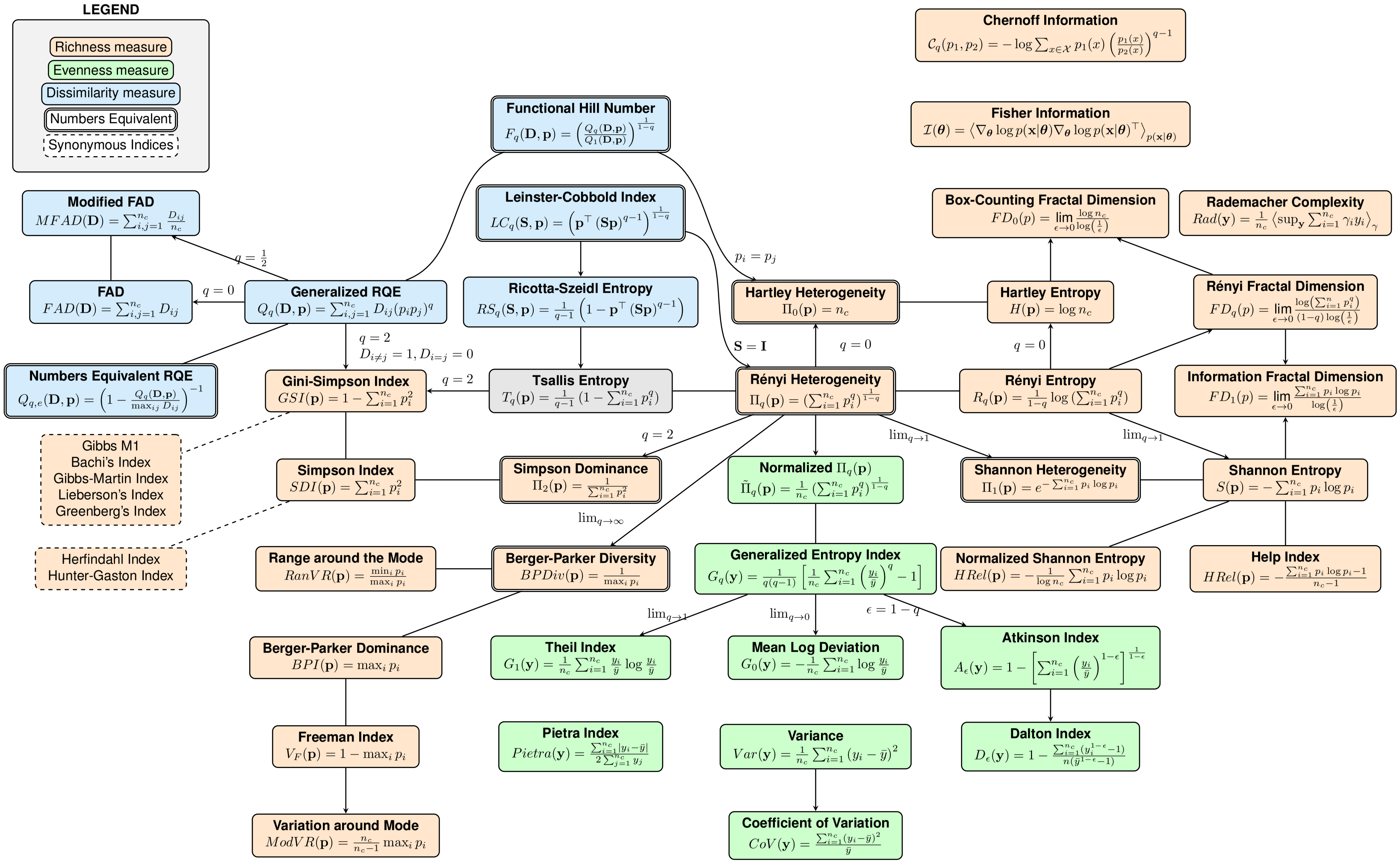

LEGEND
Richness measure
Evenness measure
Dissimilarity measure
Numbers Equivalent
Synonymous Indices
Functional Hill Number $F_q(\mathbf{D},\mathbf{p}) = \left(\frac{Q_q(\mathbf{D},\mathbf{p})}{Q_1(\mathbf{D},\mathbf{p})}\right)^{\frac{1}{1-q}}$
Chernoff Information $\mathcal{C}_q(p_1,p_2) = -\log\sum_{x\in\mathcal{X}} p_1(x)\left(\frac{p_1(x)}{p_2(x)}\right)^{q-1}$
Fisher Information $\mathcal{I}(\boldsymbol{\theta}) = \langle\nabla_{\boldsymbol{\theta}}\log p(\mathbf{x}|\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\log p(\mathbf{x}|\boldsymbol{\theta})^\top\rangle_{p(\mathbf{x}|\boldsymbol{\theta})}$
Modified FAD $MFAD(\mathbf{D}) = \sum_{i,j=1}^{n_c}\frac{D_{ij}}{n_c}$
Leinster-Cobbold Index $LC_q(\mathbf{S},\mathbf{p}) = \left(\mathbf{p}^\top(\mathbf{Sp})^{q-1}\right)^{\frac{1}{1-q}}$
Box-Counting Fractal Dimension $FD_0(p) = \lim_{\epsilon\to0}\frac{\log n_c}{\log(\frac{1}{\epsilon})}$
Rademacher Complexity $Rad(\mathbf{y}) = \frac{1}{n_c}\langle\sup_{\mathbf{y}}\sum_{i=1}^{n_c}\gamma_i y_i\rangle_\gamma$
$q=\frac{1}{2}$
$p_i = p_j$
FAD $FAD(\mathbf{D}) = \sum_{i,j=1}^{n_c} D_{ij}$
$q=0$
Generalized RQE $Q_q(\mathbf{D},\mathbf{p}) = \sum_{i,j=1}^{n_c} D_{ij}(p_ip_j)^q$
Ricotta-Szeidl Entropy $RS_q(\mathbf{S},\mathbf{p}) = \frac{1}{q-1}\left(1-\mathbf{p}^\top(\mathbf{Sp})^{q-1}\right)$
Hartley Heterogeneity $\Pi_0(\mathbf{p}) = n_c$
Hartley Entropy $H(\mathbf{p}) = \log n_c$
Rényi Fractal Dimension $FD_q(p) = \lim_{\epsilon\to0}\frac{\log(\sum_{i=1}^{n} p_i^q)}{(1-q)\log(\frac{1}{\epsilon})}$
$q=2$
$D_{i\neq j}=1, D_{i=j}=0$
$\mathbf{S}=\mathbf{I}$
$q=0$
$q=0$
Numbers Equivalent RQE $Q_{q,e}(\mathbf{D},\mathbf{p}) = \left(1-\frac{Q_q(\mathbf{D},\mathbf{p})}{\max_{ij}D_{ij}}\right)^{-1}$
Gini-Simpson Index $GSI(\mathbf{p}) = 1-\sum_{i=1}^{n_c}p_i^2$
$q=2$
Tsallis Entropy $T_q(\mathbf{p}) = \frac{1}{q-1}(1-\sum_{i=1}^{n_c}p_i^q)$
Rényi Heterogeneity $\Pi_q(\mathbf{p}) = (\sum_{i=1}^{n_c}p_i^q)^{\frac{1}{1-q}}$
Rényi Entropy $R_q(\mathbf{p}) = \frac{1}{1-q}\log(\sum_{i=1}^{n_c}p_i^q)$
Information Fractal Dimension $FD_1(p) = \lim_{\epsilon\to0}\frac{\sum_{i=1}^{n_c}p_i\log p_i}{\log(\frac{1}{\epsilon})}$
Gibbs M1
Bachi's Index
Gibbs-Martin Index
Lieberson's Index
Greenberg's Index
$q=2$
$\lim_{q\to1}$
$\lim_{q\to1}$
Simpson Index $SDI(\mathbf{p}) = \sum_{i=1}^{n_c}p_i^2$
Simpson Dominance $\Pi_2(\mathbf{p}) = \frac{1}{\sum_{i=1}^{n_c}p_i^2}$
Normalized $\Pi_q(\mathbf{p})$ $\tilde{\Pi}_q(\mathbf{p}) = \frac{1}{n_c}(\sum_{i=1}^{n_c}p_i^q)^{\frac{1}{1-q}}$
Shannon Heterogeneity $\Pi_1(\mathbf{p}) = e^{-\sum_{i=1}^{n_c}p_i\log p_i}$
Shannon Entropy $S(\mathbf{p}) = -\sum_{i=1}^{n_c}p_i\log p_i$
$\lim_{q\to\infty}$
Herfindahl Index
Hunter-Gaston Index
Range around the Mode $RanVR(\mathbf{p}) = \frac{\min_i p_i}{\max_i p_i}$
Berger-Parker Diversity $BPDiv(\mathbf{p}) = \frac{1}{\max_i p_i}$
Generalized Entropy Index $G_q(\mathbf{y}) = \frac{1}{q(q-1)}\left[\frac{1}{n_c}\sum_{i=1}^{n_c}\left(\frac{y_i}{\bar{y}}\right)^q - 1\right]$
Normalized Shannon Entropy $HRel(\mathbf{p}) = -\frac{1}{\log n_c}\sum_{i=1}^{n_c}p_i\log p_i$
Help Index $HRel(\mathbf{p}) = -\frac{\sum_{i=1}^{n_c}p_i\log p_i - 1}{n_c-1}$
$\lim_{q\to1}$
$\lim_{q\to0}$
$\epsilon = 1-q$
Berger-Parker Dominance $BPI(\mathbf{p}) = \max_i p_i$
Theil Index $G_1(\mathbf{y}) = \frac{1}{n_c}\sum_{i=1}^{n_c}\frac{y_i}{\bar{y}}\log\frac{y_i}{\bar{y}}$
Mean Log Deviation $G_0(\mathbf{y}) = -\frac{1}{n_c}\sum_{i=1}^{n_c}\log\frac{y_i}{\bar{y}}$
Atkinson Index $A_\epsilon(\mathbf{y}) = 1-\left[\sum_{i=1}^{n_c}\left(\frac{y_i}{\bar{y}}\right)^{1-\epsilon}\right]^{\frac{1}{1-\epsilon}}$
Freeman Index $V_F(\mathbf{p}) = 1-\max_i p_i$
Pietra Index $Pietra(\mathbf{y}) = \frac{\sum_{i=1}^{n_c}|y_i-\bar{y}|}{2\sum_{j=1}^{n_c}y_j}$
Variance $Var(\mathbf{y}) = \frac{1}{n_c}\sum_{i=1}^{n_c}(y_i-\bar{y})^2$
Dalton Index $D_\epsilon(\mathbf{y}) = 1-\frac{\sum_{i=1}^{n_c}(y_i^{1-\epsilon}-1)}{n(\bar{y}^{1-\epsilon}-1)}$
Variation around Mode $ModVR(\mathbf{p}) = \frac{n_c}{n_c-1}\max_i p_i$
Coefficient of Variation $CoV(\mathbf{y}) = \frac{\sum_{i=1}^{n_c}(y_i-\bar{y})^2}{\bar{y}}$

Figure 8.1: Graph depicting relationships between a select subset of potential heterogeneity indices.

properties. For instance, if a categorical latent space is chosen, the implicit assumption is that observable states cluster into categories.

3. Define a mapping $f : \mathcal{X} \to \mathcal{P}(\mathcal{Z})$ that is relevant to the domain-specific problem, and demonstrate its statistical validity using approaches such as maximization of model evidence (an explanatory measure) or cross-validated accuracy (a measure of predictive or generalization power)

4. Provide justification for the weighting across samples $\mathbf{w} = (w_i)_{i=1,2,\dots,N}$

5. Unless there are specific contraindications, set the elasticity parameter to unity ($q = 1$)

6. Wherever possible, use bootstrapping to estimate confidence intervals on the Rényi heterogeneity

8.4 The Utility of Representational Rényi Heterogeneity in Psychiatric Research and Contributions to the Applied Research Domain

A heterogeneity measure is useful for psychiatry if it (A) captures multimodality and deviance, (B) does not require a priori categorical partitioning or specification of a distance function on the observable space, (C) is interpretable and flexible across scientific questions, and (D) demonstrably contributes to solution of a meaningful problem related to heterogeneity and psychiatric research. Points A-C have already been discussed, and so we now focus on element (D), concerning the ecological validity of RRH.

Representational Rényi heterogeneity has contributed to solution of a meaningful problem in psychiatric research. Specifically, while data sampled from different clinical sources is necessary to obtain sufficient sample sizes in large-scale projects, it introduces heterogeneity caused by factors unrelated to the clinical phenotype: that is, heterogeneity unrelated to natural variation in a clinical condition's intrinsic features. This may impact the generalizability of any model learned on pooled data. Such a case was demonstrated in Chapter 5. Using clinical-interview based information from 1266 people with BD, we showed that lithium response could be predicted in the pooled data based primarily on variables related to patients' long-term pattern of relapse and remission of mood episodes (mean cross-validated ROC-AUC 0.8 95% CI [0.78, 0.82]; Kappa 0.46 [0.4, 0.51]). Patients with a

Table 8.1: Conceptual partitioning of causes of heterogeneity in medical datasets pooled across multiple sources.

	Intrinsic	**Extrinsic**
Shared	Variation related to the underlying biology of the condition itself.	Variation related to data collection practices, instrument reliability, or other factors that are (A) independent of biological differences between subjects and (B) affect all data sources to roughly the same degree.
Unique	Source-wise variation in the underlying biology of the condition. For instance, this may relate to differences in geographic expression of a disease.	Variation unrelated to the condition that varies by site. For instance, site-specific diagnostic biases or measurement error.

completely episodic course—whose manias and depressions are distinct episodes with full recovery in between—and those with fewer than four episodes per year (i.e. without rapid cycling), were more likely to be lithium responders. These variables were also the most important features in the classifier trained only on data from Halifax, Nova Scotia, wherein classification was relatively accurate (ROC-AUC 0.79 [0.74, 0.84]; Kappa 0.22 [0.13, 0.31]) and well calibrated (Brier score loss 0.15 [0.13, 0.16]). Conversely, the only other classifier with substantial above chance accuracy at the site level (Poznan, Kappa 0.24 [0.16,0.33]; ROC-AUC 0.66 [0.60, 0.72]) prioritized completely different variables (such as whether subjects had a diagnosis of primary insomnia). However, the classifier trained on Poznan was poorly calibrated (Brier score loss 0.24 [0.23, 0.24]; also see Figure 7.5). Although these results reinforced the utility of clinical interviews for prediction of lithium response in BD, there remain significant generalizability concerns since different classification functions are learned from different sites' data.

Heterogeneity of clinical disorders and multisite research can be broken down along intrinsic-extrinsic and shared-unique dimensions. Intrinsic heterogeneity is that natural variation in the condition itself, whereas extrinsic heterogeneity is related to factors outside of the disorder, such as data collection methods, the reliability of scales, and so forth. Shared heterogeneity sources are those that affect all data sources equally while unique heterogeneity sources are those that will affect one site more than the others (Table 8.1).

The heterogeneity sources most relevant for the primary medical research question are concentrated in the shared-intrinsic domain. In the classification problem at hand, we are specifically interested in the heterogeneity of the phenotypic space $\mathcal{X}$ indexed by clinical features and lithium response targets $(\mathbf{x}, y)$ respectively.

To isolate the subspace of $\mathcal{X}$ whose size is primarily reflective of shared-intrinsic heterogeneity, we begin with the assumption that the ground truth presentation of BD is geographically invariant. That is, BD should be the same biological disease regardless of where the patient comes from. Thus, we assume that exclusion of heterogeneity caused by all source-wise variations (intrinsic or extrinsic) is resulting in minimal loss of information concerning the core variation in BD.

The intrinsic components of shared heterogeneity must also be isolated. This is difficult primarily because one must be able to clearly identify all extrinsic sources of heterogeneity for which control is desired. While procedures for this step are an important aim for future developments, our work focused on two aspects. First, we ensured that heterogeneity due to unreliability of the target variable was minimized (this was done Chapter 6). Second, if shared intrinsic heterogeneity is reflective of true underlying biological variation, then subjects whose phenotypes are reflective of shared-intrinsic heterogeneity should be separable based on biological markers. This validation was established using our classification experiment and subsequent genetic enrichment studies in Chapter 7.

To the first point, Chapter 6 demonstrated that the target variable (lithium response defined by a dichotomization of the Alda score) is more robust to uncertainty in raters' judgements than the raw score (on the 0-10 scale) or symmetrical dichotomization of the score (that is, the Alda score split at 5). Recall that a classifier $\mathcal{M}$ maps clinical variables $\mathbf{x}$ onto an observable label, y. However, this label is only an estimate of subject's "true" lithium responsiveness, y^*. That is, $y = g(y^*)$. Any uncertainty in the function $g(\cdot)$ will result in heterogeneity of the classification function $\mathcal{M}(\mathbf{x}) \mapsto y$. By isolating shared heterogeneity we can eliminate that component of noise in $g(\cdot)$ related to variations in the data source. However, to minimize the heterogeneity related to shared variation in $g(\cdot)$, we argued in favour of dichotomization of the Alda score (the existing *de facto* standard definition of lithium response in the literature). Chapter 6 supported this controversial assertion with experiments that showed a more robust mutual information and statistical power when "tail split" (dichotomization away from the median) is done on a variable that is highly reliable at one extreme (that is, "asymmetrical reliability"). Our findings are controversial because most statistical literature on this matter argues strongly against dichotomization [217–227, 229]. Yet we provide data from controlled simulations showing that this asymmetrical reliability (a condition to our knowledge which is not been examined in the statistical body of evidence) is the reason why dichotomization is favourable in our case. Ultimately, the use of the dichotomized Alda score maximized the probability that variation in the target variable was due to core variation in the BD phenotype proper, rather than due to inter-rater differences. In our data, there were no other clear means by which to minimize shared extrinsic heterogeneity.

The exemplar scoring method introduced in Chapter 7 is the first approach, to our knowledge, of isolating areas of the phenotypic domain that are representative of data from the maximum number of sources. Existing approaches to deal with between site heterogeneity and medical ML have focused primarily on various cross-validation approaches [34, 35, 242]. However, none of these approaches can identify phenotypic subspaces that are robust to heterogeneity caused by pooling multi-source data. The exemplar score approach did just this, by mapping the phenotypic space onto a categorical domain representing the most informative data source for each phenotypic configuration in our clinical dataset. The Rényi heterogeneity computed on this transformed space gives the degree to which the mapping $\mathbf{x} \to y$ at a given point on the phenotypic space is captured by all sites' data. By restricting the data set to only those points whose $\mathbf{x} \to y$ relationships are agreed upon by most sites' classifiers, then one essentially isolates the phenotypic subspace most reflective of shared intrinsic heterogeneity. If this subspace is further restricted to only those $(\mathbf{x}, y)$ points that could be accurately classified, then one can isolate the best exemplars of the intrinsic relationship between clinical features and lithium responsiveness. Two validation steps taken in Chapter 7 suggested that we accomplished the isolation of shared-intrinsic heterogeneity.

First, we examined the clinical profiles of those identified as the "best" exemplars (top 25% of exemplar scores) and "poor" exemplars (bottom 25% of exemplar scores) of lithium responsiveness for agreement with symptom profiles that have been consistently identified over decades of research [36–38, 236]. Indeed, we found that factors related to age of onset [196], clinical course [36, 233], rapid cycling [213, 214], history of lifetime psychosis [235], and the degree of psychiatric comorbidity [233], were the most salient differences between the best exemplars of lithium response and nonresponse. The primary departure from the existing literature relates to the absence of family history variables amongst this salient set [198]. This may have been related to representation of family history variables, but it is also possible that geographic variations in family size could obscure the salience of this variable. For example, it is possible that an individual with BD may have a significant genetic etiology, but that cultural norms are associated with a small family size, thereby precluding the observation of strong evidence of heritability. Notwithstanding, the ultimate point is that the shared intrinsic heterogeneity in clinical features is primarily along variables for which there are already consistent associations with lithium response.

Second, and most importantly, since shared intrinsic heterogeneity should be primarily driven by underlying biological variation in BD, the best exemplars of lithium response and nonresponse should be separable using some biomarker(s). Chapter 7 validated this claim, since a logistic regression classifier trained on genomic data could separate lithium response among the best exemplars with an ROC-AUC 0.88 (interquartile range, IQR, 0.83-0.98), and the Matthews correlation coefficient (MCC) of 0.58 (IQR 0.41-0.77). Conversely, among the poor exemplars, response could be classified with only in AUC of 0.66 (IQR 0.61-0.80). To further evaluate the construct validity of the exemplar set, we examine whether the most informative variants in the genetic classification were involved in pathways associated with BD and lithium response.

Among variants used to separate the *poor* exemplars of lithium responsiveness, there were no enriched biological pathways. Conversely, features identified as salient by a classifier trained on the best exemplars' data clustered into pathways previously implicated in BD and lithium physiology. These pathways involved G-protein coupled receptor, histamine H1R, and cholinergic signalling pathways [39, 41, 238–240]. We also noted enrichment of the Alzheimer's amyloid precursor protein signalling pathway, which is particularly interesting since there is a growing body of evidence that lithium may slow progression of Alzheimer's dementia [243–250] (but also see Dunn, Holmes, & Mullee [251]).

Together, the above results show that using an approach that rested centrally on the RRH formulation solved a demonstrable problem in psychiatric ML research, concerning the impact of between-source heterogeneity on development of clinical prediction models. Combined with the fact that RRH in general is capable of capturing both deviance and multimodality, and requires neither *a priori* categorization nor imposition of a distance function on the input space, we conclude that RRH demonstrates theoretical and empirical utility for applications in psychiatric research.

8.5 Limitations and Future Directions

To recap, we have shown that our heterogeneity measurement framework, RRH, is interpretable, flexible, and demonstrably useful, particularly for psychiatric research applications. In this section, we discuss several limitations of the research herein, some immediate opportunities for advancement, and longer-term open questions. More specific limitations and future directions corresponding to each study have already been presented in Chapters 2-7, so our focus here is on the most salient points and broader themes.

Despite having conducted a broad review of work on heterogeneity measures, our synthesis presented in Chapter 3 is relatively streamlined in the presentation of technical measures. This was necessary by virtue of constraints on the publication length since there are, by our count thus far, more than 200 potential candidates for heterogeneity indices in the published literature (many of which are special cases or re-discoveries of indices from other disciplines). Furthermore, an extensive body of work by Jost [22, 23, 59, 92] and both his colleagues [90, 111–114, 149, 252] and predecessors in [17, 20] ecology have converged on the Rényi heterogeneity family (Hill numbers [17]) as the "true" measure of diversity. The moniker "true diversity" owes to these indices' satisfaction of the replication principle (Axiom 7; [21–23]) in addition to other axioms (Chapter 3). In the same vein we have found that many of the most prominent heterogeneity indices may also be derived from the Rényi family (Figure 8.1). For these reasons, we believe that the focused approach taken in Chapter 3 was sufficient to illustrate the necessary points primarily as they related to identifying desiderata for a heterogeneity measure useful for psychiatric research. At present, we are working on a book length compendium of heterogeneity measures from which Figure 8.1 is an excerpt.

Our initial presentation of RRH was designed to highlight limitations of existing noncategorical heterogeneity measures and their relationships to RRH in some canonical problems. The first example was a simplistic and analytically tractable beta mixture distribution which, although elementary, offered us the certainty in comparisons by virtue of analyses in closed form. The second was a hierarchical generative model learned on MNIST images. This is arguably the simplest data set used in modern ML research. We also used only one model (a convolutional VAE) as our transformation of observable data on to a latent space. However, the simplicity and familiarity of MNIST was important for the demonstrations in Chapter 4, since there are only a few and well defined image classes in that dataset. Furthermore, the relative levels of internal heterogeneity and between-class differences are reasonably appreciable to the naked eye (this contention was further supported empirically in Chapter 4). Notwithstanding, future work should investigate RRH in deep models of data such as music (as an investigation of stylistic heterogeneity) and language, using various architectures.

Although we believe that the RRH approach captures properties of heterogeneity that are intuitive to most researchers, and is therefore broadly interpretable, this assertion can be tested empirically. That is, it would be of interest to identify whether humans rank collections of objects by their heterogeneity in a fashion predicted by RRH. For instance, consider ranking the following sequences of strings based on subjective perception of their heterogeneity:

$$\underbrace{\mathrm{AAABBB}}_{1}, \underbrace{\mathrm{AAAAAA}}_{2}, \underbrace{\mathrm{AABBCC}}_{3}$$

We would predict that most individuals, like RRH, would rank these as follows: $2 < 1 < 3$. Indeed, the Rényi heterogeneity of these sequences would equal 2,1, and 3, respectively. Perhaps more interestingly, though, one might consider the relationship between human ranking of image set heterogeneity and the associated samples' RRH under a generative model (for example, ranking the heterogeneity in batches of MNIST samples). Such an experiment could evaluate whether, and with what fidelity, RRH captures heterogeneity in the same ordinal fashion as human intuition.

Another limitation of the study in Chapter 4 is that the parametric form for Rényi heterogeneity on the non-categorical space was merely a Gaussian mixture. However, the benefit of this model is its simplicity and ubiquity. In fact, the Gaussian mixture formulation of Rényi heterogeneity facilitated application of RRH to a deep generative model (Section 4.4.2), and is the *de facto* standard model in statistical meta-analysis (Section 3.3.3). Indeed, we are interested in further developing numbers equivalent meta-analytic heterogeneity (which would be reflected in units of "the effective range of distinct study effects"), which under the mixed-effects model uses the Gaussian RRH (Section 3.3.3). Development of such a statistic would require abstraction of data from existing meta-analyses, along with identification of a gold standard with which to compare the numbers equivalent method. However, future work must also develop parametric forms for the Rényi heterogeneity of other noncategorical distributions, since they could capture some properties not detected by the Gaussian implementation.

Limitations and future directions specific to our applied work on lithium response prediction are detailed further in Chapters 5-7, but here we discuss broader limitations and future directions that our research program will undertake, particularly in relation to application of the exemplar scoring method. Foremost is the fact that our database on lithium genomics overlaps with only 321 subjects in our clinical features data set. This leaves more than 2000 genomic samples untested using the exemplar scoring method. We thus seek some way to attribute exemplar scores to each of the remaining subjects in the genomic data set. There are two possible approaches to this problem.

One approach is to use the genomic data from the initial 321 subjects (those whose genomic data overlapped with the clinical dataset) to learn a genetic prediction model for the exemplar score. This genetic prediction model for the exemplar score could then be applied to the remaining subjects in our genomic dataset. Genomic prediction of lithium response could then be applied on subsamples stratified by the predicted exemplar scores.

It is also possible that the best exemplars of lithium response and non-response are simply more likely to have a biologically robust diagnosis of BD. It is possible that many individuals diagnosed with BD and included in these datasets may not truly have that condition, and this may complicate the associations with lithium responsiveness and non-responsiveness. For instance, if someone diagnosed as having BD due to a manic episode was actually manic secondary to undetected drug use, they may have been labeled as a lithium responder due to resolution of symptoms that would have happened after drug abstinence, regardless of lithium treatment. As another example, patients who simply have volatile moods and impulsivity (for other reasons) might be mistakenly diagnosed as having BD and show lack of lithium response. These subjects could potentially share features with both lithium responders and non-responders, despite not truly having BD in the first place. Our group is developing an experiment to address this question using genomic data (W. Stone, personal communication).

We seek to obtain multi-source genomic data for people with BD, schizophrenia, major depression, and healthy controls, upon which we could repeat the exemplar scoring procedure (these genomic data will be requested through the Psychiatric Genomics Consortium [31]). This would identify the "best genetic exemplars" of each condition. A diagnostic classification model would then be trained on this subset of best genetic exemplars, then used to predict the diagnoses of each subject in our dataset on lithium response genetics (from the Consortium on Lithium Genetics; [30, 202]). We hypothesize that genetic classification accuracy of lithium response would be highest among subjects whose genomes are also more predictive of BD, rather than of the other conditions. Such a result would provide evidence that response to lithium may be a biologically relevant dimension of BD in general.

8.6 Conclusions

This thesis has introduced a framework for the measurement of heterogeneity, RRH, that is interpretable, flexible, and useful, particularly for applications in psychiatric research. Future directions include compilation of a more comprehensive volume on heterogeneity measurement, extension of RRH to additional models and applications (such as meta-analytic RRH and heterogeneity of computational cognitive models), and the validation of RRH against human heterogeneity ranking. This novel approach has enabled resolution of an important problem concerning the effects of between-site heterogeneity in large-scale medical ML research, and enabled us to obtain strong results in genomic classification of treatment response in psychiatry. Further applications of the exemplar scoring method to larger and multi-domain data (i.e. mixed clinical, neuroimaging, and genomic data) could improve the efficiency of large scale multi-site medical ML studies, and further advance our understanding of the biological causes of psychiatric disorders.

Bibliography

[1] Eliazar II. A tour of inequality. *Annals of Physics.* 2018;389:306–332.

[2] Hooper DU, Chapin FS, Ewel JJ, et al. Effects of biodiversity on ecosystem functioning: A consensus of current knowledge. *Ecological Monographs.* 2005;75:3–35.

[3] Dabla-Norris E, Kochhar K, Suphaphiphat N, Ricka F, Tsounta E. Causes and Consequences of Income Inequality: A Global Perspective. Tech. Rep. 13International Monetary Fund 2015.

[4] Laakso M, Taagepera R. “Effective” number of parties: A Measure with Application to West Europe. *Comparative Political Studies.* 1979;12:3–27.

[5] Lloyd S. Measures of Complexity: A Nonexhaustive List. *IEEE Control Systems.* 2001;21:7–8.

[6] Pearson K, Lee A, Bramley-Moore L. Mathematical Contributions to the Theory of Evolution VI. Genetic (Reproductive) Selection: Inheritance of Fertility in Man, and of Fecundity in Thoroughbred Racehorses. *Phil. Trans. R. Soc. A.* 1899;192:257–330.

[7] Yule GU. Notes on the Theory of Association of Attributes in Statistics. *Biometrika.* 1903;2:121.

[8] Simpson EH. The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 1951;13:238–241.

[9] Blyth CR. On Simpson’s Paradox and the Sure-Thing Principle. *Journal of the American Statistical Association.* 1972;67:364.

[10] Lombardo MV, Lai M, Baron-Cohen S. Big data approaches to decomposing heterogeneity across the autism spectrum. *Molecular psychiatry.* 2019.

[11] Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology.* 2018;15:81–94.

[12] Thompson AJ, Banwell BL, Barkhof F, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurology.* 2018;17:162–173.

[13] Ortiz A, Bradler K, Garnham J, Slaney C, Alda M. Nonlinear dynamics of mood regulation in bipolar disorder. *Bipolar Disorders.* 2014;17:139–49.

[14] Ortiz A, Alda M. The perils of being too stable: Mood regulation in bipolar disorder. *Journal of Psychiatry and Neuroscience.* 2018;43:363–365.

[15] Ortiz A, Bradler K, Garnham J, Slaney C, McLean S, Alda M. Nonlinear dynamics of mood regulation in unaffected first-degree relatives of bipolar disorder patients. *Journal of Affective Disorders.* 2019;243:274–279.

[16] Alnæs D, Kaufmann T, Van Der Meer D, et al. Brain Heterogeneity in Schizophrenia and Its Association with Polygenic Risk. *JAMA Psychiatry.* 2019;76:739–748.

[17] Hill MO. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology.* 1973;54:427–432.

[18] Hannah L, Kay JA. *Concentration in Modern Industry: Theory, measurement and the U.K. experience*. London, UK: The MacMillan Press, Ltd. 1977.

[19] Adelman MA. Comment on the ”H” Concentration Measure as a Numbers-Equivalent *The Review of Economics and Statistics.* 1969;51:99-101.

[20] Patil GP, Taillie C. Diversity as a Concept and its Measurement. *Journal of the American Statistical Association.* 1982;77:548–561.

[21] MacArthur RH. Patterns of species diversity. *Biological Reviews.* 1965;40:510–533.

[22] Jost L. Entropy and diversity. *Oikos.* 2006;113:363–375.

[23] Jost L. Mismeasuring biological diversity: Response to Hoffmann and Hoffmann (2008). *Ecological Economics.* 2009;68:925–928.

[24] Gini C. *Variabilità e mutabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche.* Bologna, Italy: C. Cuppini 1912.

[25] Shannon CE. A mathematical theory of communication. *The Bell System Technical Journal.* 1948;27:379–423.

[26] Tsallis C. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics.* 1988;52:479–487.

[27] DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials.* 1986;7:177–188.

[28] Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ : British Medical Journal.* 2003;327:557–560.

[29] Thompson P, Jahanshad N, Ching CRK, et al. ENIGMA and Global Neuroscience: A Decade of Large-Scale Studies of the Brain in Health and Disease across more than 40 Countries. *PsyArXiv.* 2019.

[30] Hou L, Heilbronner U, Degenhardt F, et al. Genetic variants associated with response to lithium treatment in bipolar disorder: A genome-wide association study *The Lancet.* 2016;387:1085–1093.

[31] Sullivan PF. The Psychiatric GWAS Consortium: Big science comes to psychiatry *Neuron.* 2010;68:182–186.

[32] Nunes A. Two common questions about machine learning methods in psychiatric applications. *Bipolar Disorders.* 2019.

[33] Schnack HG, Kahn RS. Detecting neuroimaging biomarkers for psychiatric disorders: Sample size matters. *Frontiers in Psychiatry.* 2016;7.

[34] Nunes A, Schnack HG, Ching CRK, others . Using structural MRI to identify bipolar disorders – 13 site machine learning study in 3020 individuals from the ENIGMA Bipolar Disorders Working Group *Molecular Psychiatry.* 2018.

[35] Nunes A, Ardau R, Berghöfer A, et al. Prediction of Lithium Response using Clinical Data *Acta Psychiatrica Scandinavica.* 2019;In Press.

[36] Grof P. Responders to long-term lithium treatment in *Lithium in Neuropsychiatry: The Comprehensive Guide* (Bauer M, Grof P, Muller-Oerlinghausen B. , eds.):157–178UK: Informa Healthcare 2006.

[37] Gershon S, Chengappa KNR, Malhi GS. Lithium specificity in bipolar illness: A classic agent for the classic disorder *Bipolar Disorders.* 2009;11:34–44.

[38] Alda M. Who are excellent lithium responders and why do they matter? *World Psychiatry.* 2017;16:319–320.

[39] Bezchlibnyk Y, Young LT. The Neurobiology of Bipolar Disorder: Focus on Signal Transduction Pathways and the Regulation of Gene Expression *Canadian journal of psychiatry. Revue canadienne de psychiatrie.* 2002;47:135–148.

[40] Alda M. Lithium in the treatment of bipolar disorder: pharmacology and pharmacogenetics *Molecular Psychiatry.* 2015;20:661.

[41] Saxena A, Scaini G, Bavaresco DV, et al. Role of Protein Kinase C in Bipolar Disorder: A Review of the Current Literature *Molecular Neuropsychiatry.* 2017;3:108–124.

[42] Daly A, Baetens J, De Baets B. Ecological Diversity: Measuring the Unmeasurable. *Mathematics.* 2018;6:119.

[43] Eliazar I. How random is a random vector? *Annals of Physics.* 2015;363:164–184.

[44] Marquand AF, Kia SM, Zabihi M, Wolfers T, Buitelaar JK, Beckmann CF. Conceptualizing mental disorders as deviations from normative functioning. *Molecular Psychiatry.* 2019.

[45] Mouchet MA, Villéger S, Mason NWH, Mouillot D. Functional diversity measures: An overview of their redundancy and their ability to discriminate community assembly rules. *Functional Ecology.* 2010;24:867–876.

[46] Nunes A, Trappenberg T, Alda M. The Meaning and Measure of Heterogeneity *Manuscript under peer-review.* 2019.

[47] Brugger SP, Howes OD. Heterogeneity and Homogeneity of Regional Brain Structure in Schizophrenia: A Meta-analysis. *JAMA psychiatry.* 2017;74:1104–1111.

[48] Arvanitidis G, Hansen LK, Hauberg S. Latent Space Oddity: on the Curvature of Deep Generative Models. in *International Conference on Learning Representations*:1–15 2018.

[49] Olbert CM, Gala GJ, Tupler LA. Quantifying heterogeneity attributable to polythetic diagnostic criteria: Theoretical framework and empirical application. *Journal of Abnormal Psychology.* 2014;123:452–462.

[50] Prehn-Kristensen A, Zimmermann A, Tittmann L, et al. Reduced microbiome alpha diversity in young patients with ADHD. *PLoS ONE.* 2018;13:e0200728.

[51] Donohue J, O'Malley AJ, Horvitz-Lennon M, Taub AL, Berndt ER, Huskamp HA. Changes in Physician Antipsychotic Prescribing Preferences, 2002–2007. *Psychiatric Services.* 2014;65:315–322.

[52] Berndt ER, Gibbons RS, Kolotilin A, Taub AL. The heterogeneity of concentrated prescribing behavior: Theory and evidence from antipsychotics. *Journal of Health Economics.* 2015;40:26–39.

[53] American Psychiatric Association . *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5).* American Psychiatric Publishing 2013.

[54] Zimmerman M, Ellison W, Young D, Chelminski I, Dalrymple K. How many different ways do patients meet the diagnostic criteria for major depressive disorder? *Comprehensive Psychiatry.* 2015;56:29–34.

[55] Park Seon-Cheol, Kim Jae-Min, Jun Tae-Youn, et al. How many different symptom combinations fulfil the diagnostic criteria for major depressive disorder? Results from the CRESCEND study. *Nordic Journal of Psychiatry.* 2017;71:217–222.

[56] Marquand AF, Wolfers T, Mennes M, Buitelaar J, Beckmann CF. Beyond Lumping and Splitting: A Review of Computational Approaches for Stratifying Psychiatric Disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging.* 2016;1:433–447.

[57] Beijers L, Wardenaar KJ, Loo HM, Schoevers RA. Data-driven biological subtypes of depression: systematic review of biological approaches to depression subtyping. *Molecular Psychiatry.* 2019;24:888–900.

[58] McCluskey L, Falcone D. Familial amyotrophic lateral sclerosis in *UpToDate*Waltham, MA: UpToDate Inc 2019.

[59] Jost L. Partitioning Diversity into Independent Alpha and Beta Components. *Ecology.* 2007;88:2427–2439.

[60] Young G, Lareau C, Pierre B. One Quintillion Ways to Have PTSD Comorbidity: Recommendations for the Disordered DSM-5. *Psychological Injury and Law.* 2014;7:61–74.

[61] Lieberman DZ, Peele R, Razavi M. Combinations of DSM-IV-TR Criteria Sets for Bipolar Disorders. *Psychopathology.* 2008;41:35–38.

[62] Farmer AE, McGuffin P, Spitznagel EL. Heterogeneity in schizophrenia: A cluster-analytic approach. *Psychiatry Research.* 1983;8:1–12.

[63] Putnam K, Rubinow D, Meltzer-Brody S, et al. Heterogeneity of postpartum depression: A latent class analysis. *The Lancet Psychiatry.* 2015;2:59–67.

[64] Stewart SE, Rosario MC, Brown TA, et al. Principal Components Analysis of Obsessive-Compulsive Disorder Symptoms in Children and Adolescents. *Biological Psychiatry.* 2007;61:285–291.

[65] Rapp PE, Schmah T. Complexity measures in molecular psychiatry. *Molecular Psychiatry.* 1996.

[66] Marquand AF, Rezek I, Buitelaar J, Beckmann CF. Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. *Biological Psychiatry.* 2016;80:552–561.

[67] Lenferink LIM, Eisma MC. 37,650 ways to have "persistent complex bereavement disorder" yet only 48 ways to have "prolonged grief disorder" *Psychiatry Research.* 2018;261:88–89.

[68] Østergaard SD, Jensen SOW, Bech P. The heterogeneity of the depressive syndrome: when numbers get serious. *Acta Psychiatrica Scandinavica.* 2011;124:495–496.

[69] American Psychiatric Association . *Cautionary Statement for Forensic Use of DSM-5.* American Psychiatric Publishing 2015.

[70] Gotelli NJ, Chao A. *Measuring and Estimating Species Richness, Species Diversity, and Biotic Similarity from Sampling Data.*;5. Elsevier Ltd. 2013.

[71] Krebs CJ. Species Diversity Measures. in *Ecological Methodology*:532–5963rd (in preparation) ed. 2014.

[72] Chao A. Nonparametric Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics.* 1984;11:265–270.

[73] Rong H, Hui Xie X, Zhao J, et al. Similarly in depression, nuances of gut microbiota: Evidences from a shotgun metagenomics sequencing study on major depressive disorder versus bipolar disorder with current major depressive episode patients. *Journal of Psychiatric Research.* 2019;113:90–99.

[74] Bird SM, King R. Multiple Systems Estimation (or Capture-Recapture Estimation) to Inform Public Policy. *Annual Review of Statistics and Its Application.* 2018;5:95–118.

[75] Corrao G, Bagnardi V, Vittadini G, Favilli S. Capture-recapture methods to size alcohol related problems in a population. *Journal of Epidemiology & Community Health.* 2000;54:603–610.

[76] Domingo-Salvany A, Hartnoll RL, Maguire A, Suelves JM, Antó JM. Use of Capture-Recapture to Estimate the Prevalence of Opiate Addiction in Barcelona, Spain, 1989. *American Journal of Epidemiology.* 1995;141:567–574.

[77] Harrison MJ, O'Hare AE, Campbell H, Adamson A, McNeillage J. Prevalence of autistic spectrum disorders in Lothian, Scotland: An estimate using the "capture-recapture" technique. *Archives of Disease in Childhood.* 2006;91:16–19.

[78] Jones HE, Welton NJ, Ades AE, et al. Problem drug use prevalence estimation revisited: heterogeneity in capture–recapture and the role of external evidence. *Addiction.* 2016;111:438–447.

[79] Fisher N, Turner SW, Pugh R, Taylor C. Estimated numbers of homeless and homeless mentally ill people in north east Westminster by using capture-recapture analysis. *BMJ.* 1994;308:27–30.

[80] Hay G, Gannon M, MacDougall J, Eastwood C, Williams K, Millar T. Capture—recapture and anchored prevalence estimation of injecting drug users in England: national and regional estimates. *Statistical Methods in Medical Research.* 2009;18:323–339.

[81] Kake TR, Arnold R, Ellis P. Estimating the Prevalence of Schizophrenia Among New Zealand Māori: A Capture–Recapture Approach. *Australian & New Zealand Journal of Psychiatry.* 2008;42:941–949.

[82] Hope VD, Hickman M, Tilling K. Capturing crack cocaine use: Estimating the prevalence of crack cocaine use in London using capture-recapture with covariates. *Addiction.* 2005;100:1701–1708.

[83] Hay G, McKeganey N. Estimating the prevalence of drug misuse in Dundee, Scotland: an application of capture-recapture methods. *Journal of Epidemiology & Community Health.* 2008;50:469–472.

[84] Krebs CJ. Estimating Abundance in Animal and Plant Populations. in *Ecological Methodology*:24–773rd ed. 2016.

[85] Pigou AC. *Wealth and Welfare.* London, England: MacMillan and Co., Ltd 1912.

[86] Dalton H. The Measurement of the Inequality of Incomes. *The Economic Journal.* 1920;30:348.

[87] Simpson EH. Measurement of Diversity. *Nature.* 1949;163:688.

[88] Herfindahl OC. *Concentration in the Steel Industry.* PhD thesisColumbia UniversityNew York, NY 1950.

[89] Kessler RC, Berglund P, Chiu WT, et al. The US National Comorbidity Survey Replication (NCS-R): Design and field procedures. *International Journal of Methods in Psychiatric Research.* 2004;13:69–92.

[90] Botta-Dukát Z. The generalized replication principle and the partitioning of functional diversity into independent alpha and beta components. *Ecography.* 2018;41:40–50.

[91] Rényi A. On measures of information and entropy. *Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability.* 1961;114:547–561.

[92] Jost L. The relation between evenness and diversity. *Diversity.* 2010;2:207–232.

[93] Shorrocks AF. The Class of Additively Decomposable Inequality Measures. *Econometrica.* 1980;48:613–625.

[94] Cowell F. *Measuring Inequality.* Oxford, UK: Oxford University Press2nd ed. 2011.

[95] Lorenz MO. Methods of Measuring the Concentration of Wealth. *Publications of the American Statistical Association.* 1905;9:202–219.

[96] Pietra G. Delle relazioni fra indici di variabilit'a, note I e II. *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti.* 1915;74:775–804.

[97] Eliazar II, Sokolov IM. Measuring statistical evenness: A panoramic overview. *Physica A: Statistical Mechanics and its Applications.* 2012;391:1323–1353.

[98] Williams RFG, Doessel DP. Private psychiatry and Medicare: Regional equality of access in Australia. *Journal of Mental Health.* 2009;18:242–252.

[99] Roick C, Heider D, Kilian R, Matschinger H, Toumi M, Angermeyer MC. Factors contributing to frequent use of psychiatric inpatient services by schizophrenia patients. *Social Psychiatry and Psychiatric Epidemiology.* 2004;39:744–751.

[100] Lewis EN, Nash KC, Kelleher KJ. Lorenz curves: A new model for the distribution of psychiatric services. *Journal of Child and Family Studies.* 2003;12:475–482.

[101] Pottegård A, Bjerregaard BK, Glintborg D, Kortegaard LS, Hallas J, Moreno SI. The use of medication against attention deficit/hyperactivity disorder in Denmark: A drug use study from a patient perspective. *European Journal of Clinical Pharmacology.* 2013;69:589–598.

[102] Gjerden P, Bramness JG, Slørdal L. The use and potential abuse of anticholinergic antiparkinson drugs in Norway: A pharmacoepidemiological study. *British Journal of Clinical Pharmacology.* 2009;67:228–233.

[103] Peckham AM, Fairman KA, Sclar DA. Prevalence of Gabapentin Abuse: Comparison with Agents with Known Abuse Potential in a Commercially Insured US Population. *Clinical Drug Investigation.* 2017;37:763–773.

[104] Schjerning O, Pottegård A, Damkier P, Rosenzweig M, Nielsen J. Use of Pregabalin - A Nationwide Pharmacoepidemiological Drug Utilization Study with Focus on Abuse Potential. *Pharmacopsychiatry.* 2016;49:155–161.

[105] Heip C. A New Index Measuring Evenness. *Journal of the Marine Biological Association of the United Kingdom.* 1974;54:555–557.

[106] Pielou EC. The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology.* 1966;13:131–144.

[107] Theil H. *Economics and Information Theory.* Amsterdam: North Holland 1967.

[108] Atkinson AB. On the Measurement of Inequality. *Journal of Economic Theory.* 1970;2:244–263.

[109] Jaccard P. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines *Bulletin de la Société Vaudoise des Sciences Naturelles.* 1901;37:241 – 272.

[110] Rao CR. Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology.* 1982;21:24–43.

[111] Chiu CH, Chao A. Distance-based functional diversity measures and their decomposition: A framework based on hill numbers. *PLoS ONE.* 2014;9.

[112] Ricotta C, Szeidl L. Diversity partitioning of Rao's quadratic entropy. *Theoretical Population Biology.* 2009;76:299–302.

[113] Chao A, Chiu CH, Jost L. Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annual Review of Ecology, Evolution, and Systematics.* 2014;45:297–324.

[114] Leinster T, Cobbold CA. Measuring diversity: The importance of species similarity. *Ecology.* 2012;93:477–489.

[115] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2013;35:1798–1828.

[116] Cornwell WK, Schwilk DW, Ackerly DD. A trait-based test for habitat filtering: Convex hull volume *Ecology.* 2006;87:1465–1471.

[117] Barber CB, Dobkin DP, Huhdanpaa H. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software.* 1996;22:469–483.

[118] Castle DJ, Sham PC, Wessely S, Murray RM. The subtyping of schizophrenia in men and women: a latent class analysis. *Psychological Medicine.* 1994;24:41–51.

[119] Geisler D, Walton E, Naylor M, et al. Brain structure and function correlates of cognitive subtypes in schizophrenia. *Psychiatry Research: Neuroimaging.* 2015;234:74–83.

[120] Sun H, Lui S, Yao L, et al. Two patterns of white matter abnormalities in medication-naive patients with first-episode schizophrenia revealed by diffusion tensor imaging and cluster analysis. *JAMA Psychiatry.* 2015;72:678–686.

[121] Dollfus S, Everitt B, Ribeyre JM, Assouly-Besse F, Sharp C, Petit M. Identifying subtypes of schizophrenia by cluster analyses. *Schizophrenia Bulletin.* 1996;22:545–555.

[122] Kendler KS, Karkowski LM, Walsh D. The structure of psychosis: Latent class analysis of probands from the Roscommon family study. *Archives of General Psychiatry.* 1998;55:492–509.

[123] Murray V, McKee I, Miller PM, et al. Dimensions and classes of psychosis in a population cohort: a four-class, four-dimension model of schizophrenia and affective psychoses. *Psychological Medicine.* 2005;35:499–510.

[124] Dawes SE, Jeste DV, Palmer BW. Cognitive profiles in persons with chronic schizophrenia. *Journal of Clinical and Experimental Neuropsychology.* 2011;33:929–936.

[125] Cole VT, Apud JA, Weinberger DR, Dickinson D. Using latent class growth analysis to form trajectories of premorbid adjustment in schizophrenia. *Journal of Abnormal Psychology.* 2012;121:388–395.

[126] Bell MD, Corbera S, Johannesen JK, Fiszdon JM, Wexler BE. Social cognitive impairments and negative symptoms in schizophrenia: Are there subtypes with distinct functional correlates? *Schizophrenia Bulletin.* 2013;39:186–196.

[127] Brodersen KH, Deserno L, Schlagenhauf F, et al. Dissecting psychiatric spectrum disorders by generative embedding. *NeuroImage: Clinical.* 2014;4:98–111.

[128] Fair DA, Bathula D, Nikolas MA, Nigg JT. Distinct neuropsychological subgroups in typically developing youth inform heterogeneity in children with ADHD. *Proceedings of the National Academy of Sciences.* 2012;109:6769–6774.

[129] Karalunas SL, Fair D, Musser ED, Aykes K, Iyer SP, Nigg JT. Subtyping attention-deficit/hyperactivity disorder using temperament dimensions: Toward biologically based nosologic criteria. *JAMA Psychiatry.* 2014;71:1015–1024.

[130] Gates KM, Molenaar PCM, Iyer SP, Nigg JT, Fair DA. Organizing heterogeneous samples using community detection of GIMME-Derived resting state functional networks. *PLoS ONE.* 2014;9:e91322.

[131] Costa Dias TG, Iyer SP, Carpenter SD, et al. Characterizing heterogeneity in children with and without ADHD based on reward system connectivity. *Developmental Cognitive Neuroscience.* 2015;11:155–174.

[132] Van Hulst B. M., De Zeeuw P., Durston S. Distinct neuropsychological profiles within ADHD: A latent class analysis of cognitive control, reward sensitivity and timing. *Psychological Medicine.* 2015;45:735–745.

[133] Mostert JC, Hoogman M, Onnink AMH, et al. Similar Subgroups Based on Cognitive Performance Parse Heterogeneity in Adults With ADHD and Healthy Controls. *Journal of Attention Disorders.* 2018;22:281–292.

[134] Munson J, Dawson G, Sterling L, et al. Evidence for Latent Classes of IQ in Young Children With Autism Spectrum Disorder. *American Journal on Mental Retardation.* 2008;113:439–452.

[135] Sacco R, Lenti C, Saccani M, et al. Cluster analysis of autistic patients based on principal pathogenetic components. *Autism Research.* 2012;5:137–147.

[136] Fountain C, Winter AS, Bearman PS. Six Developmental Trajectories Characterize Children With Autism. *Pediatrics.* 2012;129:e1112–e1120.

[137] Georgiades S, Szatmari P, Boyle M, et al. Investigating phenotypic heterogeneity in children with autism spectrum disorder: a factor mixture modeling approach. *Journal of Child Psychology and Psychiatry.* 2013;54:206–215.

[138] Doshi-Velez F, Ge Y, Kohane I. Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis. *Pediatrics.* 2014;133:e54–e63.

[139] Veatch OJ, Veenstra-VanderWeele J, Potter M, Pericak-Vance MA, Haines JL. Genetically meaningful phenotypic subgroups in autism spectrum disorders. *Genes, Brain and Behavior.* 2014;13:276–285.

[140] Taylor S. Early versus late onset obsessive-compulsive disorder: Evidence for distinct subtypes. *Clinical Psychology Review.* 2011;31:1083–1100.

[141] Grados MA, Mathews CA. Latent Class Analysis of Gilles de la Tourette Syndrome Using Comorbidities: Clinical and Genetic Implications. *Biological Psychiatry.* 2008;64:219–225.

[142] Bulik CM, Sullivan PF, Kendler KS. An empirical study of the classification of eating disorders. *The American journal of psychiatry.* 2000;157:886–895.

[143] Kendler KS, Eaves LJ, Walters EE, Neale MC, Heath AC, Kessler RC. The identification and validation of distinct depressive syndromes in a population-based sample of female twins. *Archives of General Psychiatry.* 1996;53:391–399.

[144] Tokuda T, Yoshimoto J, Shimizu Y, et al. Identification of depression subtypes and relevant brain regions using a data-driven approach. *Scientific Reports.* 2018;8:1–13.

[145] Drysdale AT, Grosenick L, Downar J, et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine.* 2017;23:28–38.

[146] Chekroud AM, Gueorguieva R, Krumholz HM, Trivedi MH, Krystal JH, McCarthy G. Reevaluating the efficacy and predictability of antidepressant treatments: A symptom clustering approach *JAMA Psychiatry.* 2017;74:370–378.

[147] Petchey OL, Gaston KJ. Functional diversity (FD), species richness and community composition. *Ecology Letters.* 2002.

[148] Petchey OL, Gaston KJ. Dendrograms and measuring functional diversity. *Oikos.* 2007.

[149] Chiù C, Jost L, Chao A. Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers. *Ecological Monographs.* 2014;84:21–44.

[150] Horvath S, Dong J. Geometric interpretation of gene coexpression network analysis. *PLoS Computational Biology.* 2008;4.

[151] Chen C, Cheng L, Grennan K, et al. Two gene co-expression modules differentiate psychotics and controls. *Molecular Psychiatry.* 2013;18:1308–1314.

[152] Radulescu E, Jaffe AE, Straub RE, et al. Identification and prioritization of gene sets associated with schizophrenia risk by co-expression network analysis in human brain. *Molecular Psychiatry.* 2018;Nov 26.

[153] Bethlehem RAI, Seidlitz J, Romero-Garcia R, Lombardo MV. Using normative age modelling to isolate subsets of individuals with autism expressing highly age-atypical cortical thickness features. *bioRxiv.* 2018:252593.

[154] Zabihi M, Oldehinkel M, Wolfers T, et al. Dissecting the Heterogeneous Cortical Anatomy of Autism Spectrum Disorder Using Normative Models. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging.* 2019;4:567–578.

[155] Wolfers T, Beckmann CF, Hoogman M, Buitelaar JK, Franke B, Marquand AF. Individual differences v. The average patient: Mapping the heterogeneity in ADHD using normative models. *Psychological Medicine.* 2019.

[156] Kessler D, Angstadt M, Sripada C. Growth charting of brain connectivity networks and the identification of attention impairment in youth. *JAMA Psychiatry.* 2016;73:481–489.

[157] Wolfers T, Doan NT, Kaufmann T, et al. Mapping the Heterogeneous Phenotype of Schizophrenia and Bipolar Disorder Using Normative Models. *JAMA Psychiatry.* 2018;75:1146–1155.

[158] Alexander-Bloch AF, Reiss PT, Rapoport J, et al. Abnormal cortical growth in schizophrenia targets normative modules of synchronized development. *Biological Psychiatry.* 2014;76:438–446.

[159] Ziegler G, Ridgway GR, Dahnke R, Gaser C. Individualized Gaussian process-based prediction and detection of local and global gray matter abnormalities in elderly subjects. *NeuroImage.* 2014;97:333–348.

[160] Huizinga W, Poot DHJ, Vernooij MW, et al. A spatio-temporal reference model of the aging brain. *NeuroImage.* 2018;169:11–22.

[161] Yeragani VK, Radhakrishna Rao KA., Pohl R, Jampala VC, Balon R. Heart rate and QT variability in children with anxiety disorders: A preliminary report. *Depression and Anxiety.* 2001.

[162] Acharya UR, Sudarshan VK, Adeli H, et al. A novel depression diagnosis index using nonlinear features in EEG signals. *European Neurology.* 2015;74:79–83.

[163] Zhao Q, Hu B, Li Y, et al. An Alpha resting EEG study on nonlinear dynamic analysis for schizophrenia. in *International IEEE/EMBS Conference on Neural Engineering, NER*:484–488 2013.

[164] Pincus SM. Approximate entropy as a measure of irregularity for psychiatric serial metrics. *Bipolar Disorders.* 2006;8:430–440.

[165] Leistedt SJJ, Linkowski P, Lanquart JP, et al. Decreased neuroautonomic complexity in men during an acute major depressive episode: Analysis of heart rate dynamics. *Translational Psychiatry.* 2011;1:e27.

[166] Fernández A, López-Ibor MI, Turrero A, et al. Lempel-Ziv complexity in schizophrenia: A MEG study. *Clinical Neurophysiology.* 2011;122:2227–2235.

[167] Fernández A, Hornero R, Gómez C, et al. Complexity analysis of spontaneous brain activity in alzheimer disease and mild cognitive impairment: An MEG study. *Alzheimer Disease and Associated Disorders.* 2010;24:182–189.

[168] Lai MC, Lombardo MV, Chakrabarti B, et al. A shift to randomness of brain oscillations in people with autism. *Biological Psychiatry.* 2010;68:1092–1099.

[169] Glenn T, Whybrow PC, Rasgon N, et al. Approximate entropy of self-reported mood prior to episodes in bipolar disorder. *Bipolar Disorders.* 2006;8:424–429.

[170] MacKay DJC. *Information Theory, Inference, and Learning Algorithms.* Cambridge, England: Cambridge University Press 2003.

[171] Tang L, Lv H, Yang F, Yu L. Complexity testing techniques for time series data: A comprehensive literature review. *Chaos, Solitons and Fractals.* 2015;81:117–135.

[172] Stam CJ. Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field. *Clinical Neurophysiology.* 2005;116:2266–2301.

[173] Paulus MP, Braff DL. Chaos and schizophrenia: Does the method fit the madness? *Biological Psychiatry.* 2003;53:3–11.

[174] Torre-Luque A, Bornas X, Balle M, Fiol-Veny A. Complexity and nonlinear biomarkers in emotional disorders: A meta-analytic study. *Neuroscience and Biobehavioral Reviews.* 2016;68:410–422.

[175] Torre Luque A, Bornas X. Complexity and Irregularity in the Brain Oscillations of Depressive Patients: A Systematic Review. *Neuropsychiatry.* 2017;07:466–477.

[176] Yang AC, Tsai SJ. Is mental illness complex? From behavior to brain. *Progress in Neuro-Psychopharmacology and Biological Psychiatry.* 2013;45:253–257.

[177] Weitzman ML. On Diversity. *The Quarterly Journal of Economics.* 1992;107:363–405.

[178] Regier DA, Narrow WE, Clarke DE, et al. DSM-5 field trials in the United States and Canada, part II: Test-retest reliability of selected categorical diagnoses *American Journal of Psychiatry.* 2013;170:59–70.

[179] Shao H, Kumar A, Thomas Fletcher P. The riemannian geometry of deep generative models in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* 2018.

[180] Nickel M, Kiela D. Poincaré embeddings for learning hierarchical representations in *Advances in Neural Information Processing Systems*;2017-Decem:6339–6348 2017.

[181] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE.* 1998;86:2278–2324.

[182] Kingma DP, Welling M. Auto-Encoding Variational Bayes *ICLR 2014.* 2014.

[183] Berger WH, Parker FL. Diversity of planktonic foraminifera in deep-sea sediments *Science.* 1970;168:1345–1347.

[184] Lande R. Statistics and partitioning of species diversity and similarity among multiple communities *Oikos.* 1996;76:5-13.

[185] Mikolov T, Chen K, Corrado G, Dean J. Distributed representations of words and hrases and their compositionality in *NIPS*:1–9 2013.

[186] Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*:1532–1543 2014.

[187] Nunes A, Alda M, Trappenberg T. On the Multiplicative Decomposition of Heterogeneity in Continuous Assemblages *arXiv preprint arXiv:2002.09734.* 2020.

[188] Kingma DP, Welling M. An Introduction to Variational Autoencoders *Foundations and Trends® in Machine Learning.* 2019;12:307–392.

[189] Bromley J, Guyon I, LeCun Y, Säckinger E, Shah R. Signature verification using a "siamese" time delay neural network in *Advances in Neural Information Processing Systems 6*:737–744 1994.

[190] Hadsell R, Chopra S, LeCun Y. Dimensionality Reduction by Learning an Invariant Mapping in *CVPR (2)*:1735-1742 2006.

[191] Grof P. Sixty years of lithium responders. *Neuropsychobiology.* 2010;62:8–16.

[192] Garnham J, Munro A, Slaney C, et al. Prophylactic treatment response in bipolar disorder: Results of a naturalistic observation study *Journal of Affective Disorders.* 2007;104:185–190.

[193] Drancourt N, Etain B, Lajnef M, et al. Duration of untreated bipolar disorder: Missed opportunities on the long road to optimal treatment *Acta Psychiatrica Scandinavica.* 2013;127:136–144.

[194] Mertens J, Wang QW, Kim Y, et al. Differential responses to lithium in hyperexcitable neurons from patients with bipolar disorder *Nature.* 2015;527:95–99.

[195] Stern S, Santos R, Marchetto MC, et al. Neurons derived from patients with bipolar disorder divide into intrinsically different sub-populations of neurons, predicting the patients' responsiveness to lithium *Molecular Psychiatry.* 2018;23:1453–1465.

[196] Hui TP, Kandola A, Shen L, et al. A systematic review and meta-analysis of clinical predictors of lithium response in bipolar disorder *Acta Psychiatrica Scandinavica.* 2019;140:94–115.

[197] Sportiche S, Geoffroy PA, Brichant-Petitjean C, et al. Clinical factors associated with lithium response in bipolar disorders *Australian and New Zealand Journal of Psychiatry.* 2017;51:524–530.

[198] Grof P, Duffy A, Cavazzoni P, et al. Is response to prophylactic lithium a familial trait? *Journal of Clinical Psychiatry.* 2002;63:942–947.

[199] Kessing LV. Lithium as the drug of choice for maintenance treatment in bipolar disorder *Acta Psychiatrica Scandinavica.* 2019;140:91–93.

[200] Kim TT, Dufour S, Xu C, et al. Predictive modeling for response to lithium and quetiapine in bipolar disorder. *Bipolar Disorders.* 2019:1–9.

[201] Nierenberg AA, McElroy S, Ketter TA, et al. Bipolar CHOICE (Clinical Health Outcomes Initiative in Comparative Effectiveness): A pragmatic six month trial of lithium vs. quetiapine for bipolar disorder. *Journal of Clinical Psychiatry.* 2016;77:90–99.

[202] Manchia M, Adli M, Akula N, et al. Assessment of Response to Lithium Maintenance Treatment in Bipolar Disorder: A Consortium on Lithium Genetics (ConLiGen) Report *PLoS ONE.* 2013;8.

[203] McGuffin P, Farmer A, Harvey I. A polydiagnostic application of operational criteria in studies of psychotic illness: Development and reliability of the OPCRIT system. *Archives of General Psychiatry.* 1991;48:764–770.

[204] Turecki G, Grof P, Cavazzoni P, et al. Evidence for a role of phospholipase C-γ1 in the pathogenesis of bipolar disorder *Molecular Psychiatry.* 1998;3:534–538.

[205] Turecki G, Grof P, Grof E, et al. Mapping susceptibility genes for bipolar disorder: A pharmacogenetic approach based on excellent response to lithium *Molecular Psychiatry.* 2001;6:570–578.

[206] Denicoff KD, Ali SO, Sollinger AB, Smith-Jackson EE, Leverich GS, Post RM. Utility of the daily prospective National Institute of Mental Health Life-Chart Method (NIMH-LCM-P) ratings in clinical trials of bipolar disorder. *Depression and Anxiety.* 2002;15:1–9.

[207] Fay MP, Shaw PA. Exact and Asymptotic Weighted Logrank Tests for Interval Censored Data: The interval R Package *Journal of Statistical Software.* 2010;36:1–34.

[208] Breiman L. Random Forests *Machine Learning.* 2001;45:5–32.

[209] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python *Journal of Machine Learning Research.* 2012;12:2825–2830.

[210] He H, Garcia EA. Learning from Imbalanced Data Sets. *IEEE Transactions on knowledge and data engineering.* 2010;21:1263–1264.

[211] Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning *Journal of Machine Learning Research.* 2017;18:1–5.

[212] Grof P, Alda M, Grof E, Zvolsky P, Walsh M. Lithium response and genetics of affective disorders *Journal of Affective Disorders.* 1994;32:85–95.

[213] Backlund L, Ehnvall A, Hetta J, Isacsson G, Ågren H. Identifying predictors for good lithium response - A retrospective analysis of 100 patients with bipolar disorder using a life-charting method. *European Psychiatry.* 2009;24:171–177.

[214] Tondo L, Hennen J, Baldessarini RJ. Rapid-cycling bipolar disorder: Effects of long-term treatments. *Acta Psychiatrica Scandinavica.* 2003;108:4–14.

[215] Kessing LV, Hellmund G, Andersen PK. Predictors of excellent response to lithium: results from a nationwide register-based study *International Clinical Psychopharmacology.* 2011;26:323–328.

[216] Kessing LB, Vradi E, Andersen PK. Starting lithium prophylaxis early v. late in bipolar disorder *The British Journal of Psychiatry: The Journal of Mental Science.* 2014;205:214–220.

[217] Humphreys LG, Fleishman A. Pseudo-orthogonal and other analysis of variance designs involving individual-differences variables *Journal of Educational Psychology.* 1974;66:464–472.

[218] Humphreys LG. Research on individual differences requires correlational analysis, not ANOVA *Intelligence.* 1978;2:1–5.

[219] Humphreys LG. Doing research the hard way: Substituting analysis of variance for a problem in correlational analysis *Journal of Educational Psychology.* 1978;70:873–876.

[220] Cohen J. The Cost of Dichotomization *Applied Psychological Measurement.* 1983;7:249–253.

[221] Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea *Statistics in Medicine.* 2006;25:127–141.

[222] Altman DG, Royston P. The cost of dichotomising continuous variables *BMJ.* 2006;332:1080.

[223] Rucker DD., McShane BB, Preacher KJ. A researcher's guide to regression, discretization, and median splits of continuous variables *Journal of Consumer Psychology.* 2015;25:666–678.

[224] MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables *Psychological Methods.* 2002;7:19–40.

[225] Irwin JR., McClelland GH. Negative Consequences of Dichotomizing Continuous Predictor Variables *Journal of Marketing Research.* 2003;40:366–371.

[226] Fitzsimons GJ. Death to Dichotomizing *Journal of Consumer Research.* 2008;35:5–8.

[227] Streiner DL. Breaking up is Hard to Do: The Heartbreak of Dichotomizing Continuous Data *The Canadian Journal of Psychiatry.* 2002;47:262–266.

[228] DeCoster J, Iselin AR, Gallucci M. A conceptual and empirical examination of justifications for dichotomization. *Psychological Methods.* 2009;14:349–366.

[229] Hunter JE, Schmidt FL. Dichotomization of continuous variables: The implications for meta-analysis *Journal of Applied Psychology.* 1990;75:334–349.

[230] Chesney E, Goodwin GM, Fazel S. Risks of all-cause and suicide mortality in mental disorders: A meta-review *World Psychiatry.* 2014;13:153–160.

[231] Manchia M, Hajek T, O'Donovan C, et al. Genetic risk of suicidal behavior in bipolar spectrum disorder: analysis of 737 pedigrees. *Bipolar disorders.* 2013;15:496–506.

[232] Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools *Nucleic Acids Research.* 2018;47:D419–D426.

[233] Passmore MJ, Garnham J, Duffy A, et al. Phenotypic spectra of bipolar disorder in responders to lithium versus lamotrigine *Bipolar Disorders.* 2003;5:110–114.

[234] Kleindienst N, Engel R, Greil W. Which clinical factors predict response to prophylactic lithium? A systematic review for bipolar disorders. *Bipolar disorders.* 2005;7:404–17.

[235] Kleindienst N, Greil W. Differential efficacy of lithium and carbamazepine in the prophylaxis of bipolar disorder: Results of the MAP study *Neuropsychobiology.* 2000;42:2–10.

[236] Alda M. The phenotypic spectra of bipolar disorder *European Neuropsychopharmacology.* 2004;14.

[237] Kendler KS. From Many to One to Many - The Search for Causes of Psychiatric Illness *JAMA Psychiatry.* 2019;76:1085–1091.

[238] Cruceanu C, Schmouth JF, Torres-Platas SG, et al. Rare susceptibility variants for bipolar disorder suggest a role for G protein-coupled receptors *Molecular Psychiatry.* 2018;23:2050–2056.

[239] Gonzalez-Maeso J, Meana J. Heterotrimeric G Proteins: Insights into the Neurobiology of Mood Disorders *Current Neuropharmacology.* 2006;4:127–138.

[240] Vosahlikova M, Svoboda P. Lithium – Therapeutic tool endowed with multiple beneficiary effects caused by multiple mechanisms *Acta Neurobiologiae Experimentalis.* 2016;76:1–19.

[241] Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, P Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex *Science.* 2001;293:2425–2430.

[242] Sripada C, Rutherford S, Angstadt M, et al. Prediction of neurocognition in youth from resting state fMRI *Molecular Psychiatry.* 2019.

[243] Terao T, Nakano H, Inoue Y, Okamoto T, Nakamura J, Iwata N. Lithium and dementia: A preliminary study *Progress in Neuro-Psychopharmacology and Biological Psychiatry.* 2006;30:1125–1128.

[244] Angst J, Gamma A, Gerber-Werder R, Zarate CA, Manji HK. Does long-term medication with lithium, clozapine or antidepressants prevent or attenuate dementia in bipolar and depressed patients? *International Journal of Psychiatry in Clinical Practice.* 2007;11:2–8.

[245] Nunes PV, Forlenza OV, Gattaz WF. Lithium and risk for Alzheimer's disease in elderly patients with bipolar disorder *British Journal of Psychiatry.* 2007;190:359–360.

[246] Kessing LV, Søndergård L, Forman JL, Andersen PK. Lithium treatment and risk of dementia *Archives of General Psychiatry.* 2008;65:1331–1335.

[247] Kessing LV, Forman JL, Andersen PK. Does lithium protect against dementia? *Bipolar Disorders.* 2010;12:87–94.

[248] Gerhard T, Devanand DP, Huang C, Crystal S, Olfson M. Lithium treatment and risk for dementia in adults with bipolar disorder: Population-based cohort study *British Journal of Psychiatry.* 2015;207:46–51.

[249] Kessing LV, Gerds TA, Knudsen NN, et al. Association of lithium in drinking water with the incidence of dementia *JAMA Psychiatry.* 2017;74:1005–1010.

[250] Fajardo VA, Fajardo VA, Leblanc PJ, Macpherson REK. Examining the Relationship between Trace Lithium in Drinking Water and the Rising Rates of Age-Adjusted Alzheimer's Disease Mortality in Texas *Journal of Alzheimer's Disease.* 2018;61:425–434.

[251] Dunn N, Holmes C, Mullee M. Does lithium therapy protect against the onset of dementia? *Alzheimer Disease and Associated Disorders.* 2005;19:20–22.

[252] Chao A, Chiu CH, Jost L. Phylogenetic diversity measures based on Hill numbers. *Philosophical Transactions of the Royal Society B: Biological Sciences.* 2010;365:3599–3609.

Appendix A

Supplementary Material for *Representational Rényi Heterogeneity*

A.1 Mathematical Appendix

Proposition 5. Rényi heterogeneity (Equation 4.3) obeys the replication principle.

Proof. The Rényi heterogeneity for a single distribution $\mathbf{p}_i = (p_{ij})_{j=1,2,\dots,n_i}$, where $n_i \in \mathbb{N}_+$ is the size of the state space in system i, is

$$\Pi_q(\mathbf{p}_i) = \left(\sum_{j=1}^{n_i} p_{ij}^q\right)^{\frac{1}{1-q}} \tag{A.1}$$

and for the aggregation of N subsystems is

$$\Pi_q(\bar{\mathbf{p}}_i) = \left(\sum_{i=1}^{N}\sum_{j=1}^{n_i} \left(\frac{p_{ij}}{N}\right)^q\right)^{\frac{1}{1-q}}. \tag{A.2}$$

The replication principle asserts that

$$\Pi_q(\bar{\mathbf{p}}) = N\Pi_q(\mathbf{p}_i). \tag{A.3}$$

Let $\lambda_i = \sum_{j=1}^{n_i} p_{ij}^q$ and recall that $\lambda_i = \lambda_k$ for all $(i,k) \in \{1,2,\dots,N\}$. Then,

$$\begin{aligned}
\left(N^{-q}\sum_{i=1}^{N}\sum_{j=1}^{n_i} p_{ij}^q\right)^{\frac{1}{1-q}} &= N\left(\sum_{j=1}^{n_i} p_{ij}^q\right)^{\frac{1}{1-q}} \\
\left(N^{-q}\sum_{i=1}^{N}\lambda_i\right)^{\frac{1}{1-q}} &= N\lambda_i^{\frac{1}{1-q}} \\
\left(N^{1-q}\lambda_i\right)^{\frac{1}{1-q}} &= N\lambda_i^{\frac{1}{1-q}} \\
N\lambda_i^{\frac{1}{1-q}} &= N\lambda_i^{\frac{1}{1-q}}.
\end{aligned} \tag{A.4}$$

Since $\lim_{q\to 1} \lambda_i^{\frac{1}{1-q}}$ exists (it is the perplexity index), the result also holds at $q = 1$. □

Proposition 6. For a system X with probability mass function represented by the vector $\mathbf{p} = (p_i)_{i=1,2,\dots,n}$ on event space $\mathcal{X} = \{1, 2, \dots, n\}$, with distance function $d_X : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ represented by the $n \times n$ matrix $\mathbf{D} = [d_X(i,j)]_{i=1,2,\dots,n}^{j=1,2,\dots,n}$, the functional Hill numbers family of indices

$$F_q\left(\mathbf{D}, \mathbf{p}\right) = \left(\frac{Q_q\left(\mathbf{D}, \mathbf{p}\right)}{Q_1\left(\mathbf{D}, \mathbf{p}\right)}\right)^{\frac{1}{2(1-q)}} \tag{A.5}$$

is insensitive to $d_X(i,j)$ for all $(i,j) \in \mathcal{X}$ when $\mathbf{p}$ is uniform.

Proof. The proof is direct given substitution of $\mathbf{p} = (n^{-1})_{i=1,2,\dots,n}$ into Equation A.5.

$$F_q\left(\mathbf{D}, \mathbf{p}\right) = \left(\frac{Q_q\left(\mathbf{D}, \mathbf{p}\right)}{Q_1\left(\mathbf{D}, \mathbf{p}\right)}\right)^{\frac{1}{2(1-q)}} = \left(\frac{n^{-2q} \sum_{i=1}^{n} \sum_{j=1}^{n} d_X(i,j)}{n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} d_X(i,j)}\right)^{\frac{1}{2(1-q)}} = n \tag{A.6}$$

□

Proposition 7 (Rényi Heterogeneity of a Continuous System)**.** The Rényi heterogeneity of a system X with event space $\mathcal{X} \subseteq \mathbb{R}^n$ and pdf $f \in \mathcal{P}(\mathcal{X})$ is equal to the magnitude of the volume of an n-cube over which there is a uniform probability density with the same Rényi heterogeneity as that given by f.

Proof. Let the basic integral of X be defined as $\int_{\mathcal{X}} f^q(\mathbf{x}) \, \mathrm{d}\mathbf{x}$. Furthermore, let X_* be an idealized reference system with a uniform probability density f_* on $\mathcal{X}$ with lower bounds $\mathbf{0} = (0)_{i=1,\dots,n}$ and upper bounds $\mathbf{u} = (u_*)_{i=1,\dots,n}$ where $u_* \geq 0$ is the side length of an n-cube. We assume that X_* has basic integral $\int_{\mathcal{X}} f_*^q(\mathbf{x}) \, \mathrm{d}\mathbf{x}$ such that

$$\begin{aligned} \int_{\mathcal{X}} f^q(\mathbf{x}) \, \mathrm{d}\mathbf{x} &= \int_{\mathcal{X}} f_*^q(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\ &= \prod_{i=1}^{n} u_*^{1-q} \\ &= u_*^{n(1-q)}. \end{aligned} \tag{A.7}$$

Solving Equation A.7 for u_*^n gives the Rényi heterogeneity of order q. At $q \neq 1$,

$$u_*^n = \left(\int_{\mathcal{X}} f^q(\mathbf{x}) \, \mathrm{d}\mathbf{x}\right)^{\frac{1}{1-q}} \tag{A.8}$$

and in the limit of $q \to 1$, Equation A.8 becomes the exponential of the Shannon (differential) entropy. Thus, Π_q is interpreted as the volume of an n-cube of side length u_*, over which there is a uniform distribution giving the same heterogeneity as X. □

Proposition 8 (Rényi heterogeneity of a multivariate Gaussian)**.** The Rényi heterogeneity of an n-dimensional multivariate Gaussian with probability density function (pdf)

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{n}{2}} \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \tag{A.9}$$

with mean $\boldsymbol{\mu} = (\mu_i)_{i=1,2,\dots,n}$ and covariance matrix $\boldsymbol{\Sigma} = (\Sigma_{ij})_{i=1,2,\dots,n}^{j=1,2,\dots,n}$ is

$$\Pi_q\left(\boldsymbol{\Sigma}\right) = \begin{cases} \text{Undefined} & q = 0 \\ (2\pi e)^{\frac{n}{2}} \sqrt{|\boldsymbol{\Sigma}|} & q = 1 \\ (2\pi)^{\frac{n}{2}} \sqrt{|\boldsymbol{\Sigma}|} & q = \infty \\ (2\pi)^{\frac{n}{2}} q^{\frac{n}{2(q-1)}} \sqrt{|\boldsymbol{\Sigma}|} & \text{Otherwise} \end{cases}. \tag{A.10}$$

Proof. Let $\boldsymbol{\Sigma}^{-1} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{-1}$ be the eigendecomposition of the inverse covariance matrix into an orthonormal matrix of eigenvectors $\mathbf{U}$ and $n \times n$ diagonal matrix $\boldsymbol{\Lambda}$ with eigenvalues $(\lambda_i)_{i=1,2,\dots,n}$ down the leading diagonal. Furthermore, let $\frac{\mathrm{d}x_i}{\mathrm{d}y_j} = U_{ij}$ and use the substitution $\mathbf{y} = \mathbf{U}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}\right)$ to proceed as follows:

$$\begin{aligned} \Pi_q\left(\boldsymbol{\Sigma}\right) &= \left[(2\pi)^{-\frac{qn}{2}} \left|\boldsymbol{\Sigma}\right|^{-\frac{q}{2}} \int e^{-\frac{q}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \,\mathrm{d}\mathbf{x}\right]^{\frac{1}{1-q}} \\ &= \left((2\pi)^{-\frac{qn}{2}} \left|\boldsymbol{\Sigma}\right|^{-\frac{q}{2}} \int e^{-\frac{q}{2}\mathbf{y}^\top \boldsymbol{\Lambda}\mathbf{y}} \,\mathrm{d}\mathbf{y}\right)^{\frac{1}{1-q}} \\ &= \left((2\pi)^{-\frac{qn}{2}} \left|\boldsymbol{\Sigma}\right|^{-\frac{q}{2}} \left(\frac{(2\pi)^n}{q^n \prod_{i=1}^n \lambda_i}\right)^{\frac{1}{2}}\right)^{\frac{1}{1-q}} \\ &= \left((2\pi)^{-\frac{qn}{2}} \left|\boldsymbol{\Sigma}\right|^{-\frac{q}{2}} \left(\frac{(2\pi)^n}{q^n \left|\boldsymbol{\Lambda}\right|}\right)^{\frac{1}{2}}\right)^{\frac{1}{1-q}} \\ &= q^{\frac{n}{2(q-1)}} (2\pi)^{\frac{n}{2}} \sqrt{|\boldsymbol{\Sigma}|} \end{aligned} \tag{A.11}$$

which holds only at $q \notin \{0, 1, \infty\}$. At $q = 1$, we have

$$\begin{aligned} \lim_{q\to1} \log \Pi_q\left(\boldsymbol{\Sigma}\right) &= \lim_{q\to1} \left(\frac{n}{2(q-1)} \log q\right) + \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\left|\boldsymbol{\Sigma}\right| \\ &= \frac{n}{2} + \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\left|\boldsymbol{\Sigma}\right|, \end{aligned} \tag{A.12}$$

and therefore,

$$\Pi_1\left(\mathbf{\Sigma}\right) = (2\pi e)^{\frac{n}{2}}\sqrt{|\mathbf{\Sigma}|}. \tag{A.13}$$

One can then easily show that $\Pi_0(\mathbf{\Sigma})$ is undefined and that as $q \to \infty$,

$$\Pi_\infty\left(\mathbf{\Sigma}\right) = (2\pi)^{\frac{n}{2}}\sqrt{|\mathbf{\Sigma}|}. \tag{A.14}$$

□

A.2 Expected Distance Between two Beta-Distributed Random Variables

To compute the numbers equivalent RQE $\hat{Q}_e$, the functional Hill numbers F_q, and the Leinster-Cobbold index L_q under the beta mixture model, we must derive an analytical expression for the distance matrix. This involves the following integral:

$$d(x,y) = \int_0^1 \int_0^1 |x-y| f(x) g(y)\, \mathrm{d}x\, \mathrm{d}y, \tag{A.15}$$

where $f(x) = \mathrm{Beta}_{\alpha_1,\beta_1}(x)$ and $g(y) = \mathrm{Beta}_{\alpha_2,\beta_2}(y)$. By exploiting the identity

$$|x-y| = x + y - 2\min\{x,y\}, \tag{A.16}$$

and expanding, the integral is greatly simplified and gives the following closed-form solution:

$$d(x,y) = \langle x\rangle - \langle y\rangle + \eta\left(\Phi_a - \alpha_1\Phi_b\right), \tag{A.17}$$

where

$$\eta = \frac{2\Gamma(\alpha_1)\Gamma(\beta_2)\Gamma(\alpha_1+\alpha_2+1)}{B(\alpha_1,\beta_1)B(\alpha_2,\beta_2)}, \tag{A.18}$$

and where $\langle y\rangle = \frac{\alpha_2}{\alpha_2+\beta_2}$, $\langle x\rangle = \frac{\alpha_1}{\alpha_1+\beta_1}$, and the Φ's are regularized hypergeometric functions:

$$\Phi_a = {}_3\tilde{F}_2\left[\begin{array}{c} \alpha_1, \alpha_1+\alpha_2+1, 1-\beta_1 \\ \alpha_1+1, \alpha_1+\alpha_2+\beta_2+1 \end{array}, 1\right] \tag{A.19}$$

$$\Phi_b = {}_3\tilde{F}_2\left[\begin{array}{c} \alpha_1+1, \alpha_1+\alpha_2+1, 1-\beta_1 \\ \alpha_1+2, \alpha_1+\alpha_2+\beta_2+1 \end{array}, 1\right] \tag{A.20}$$

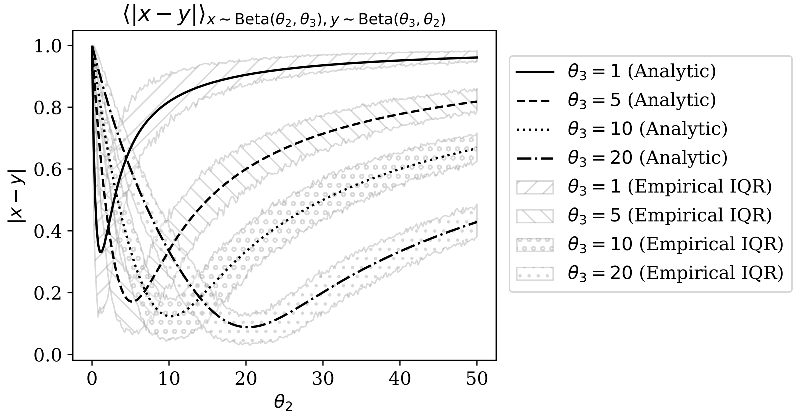


Figure A.1: Numerical verification of the analytical expression for the expected absolute distance between two Beta-distributed random variables. Solid lines are the theoretical predictions. Ribbons show the bounds between 25th-75th percentiles (the interquartile range, IQR) of the simulated values.

Figure A.1 provides numerical verification of this result. One simply uses Equation A.17 to compute the analytic distance matrix

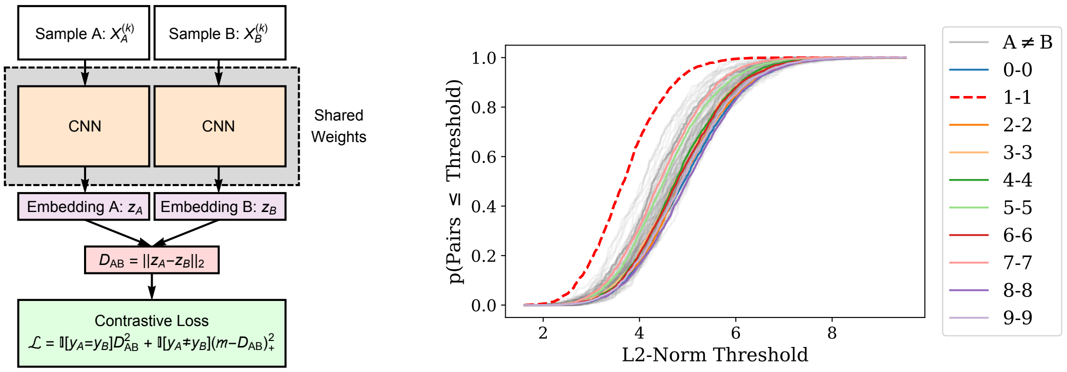
$$\mathbf{D}(\alpha_1, \beta_1, \alpha_2, \beta_2) = \begin{pmatrix} d(x,x) & d(x,y) \\ d(y,x) & d(y,y) \end{pmatrix}, \tag{A.21}$$

which, with the component probabilities (Equation 4.46), can be used to compute $\hat{Q}_e, F_q$, and L_q using the formulas shown in the main body.

A.3 Evidence Supporting Relative Homogeneity of MNIST "Ones"

In our evaluation of non-categorical RRH using the MNIST data, we asserted that the class of handwritten Ones were relatively more homogeneous than other digits. Our initial statement was based simply on visual inspection of samples from the dataset, wherein the Ones ostensibly demonstrate fewer relevant feature variations than other classes. However, to test this hypothesis more objectively, we conducted an empirical evaluation using similarity metric learning.

We implemented a deep neural network architecture known as a "siamese network" [189] to learn a latent distance metric on the MNIST classes. Our siamese network architecture is depicted in Figure A.2a. Training is conducted by sampling batches of 10,000 image pairs from the MNIST test set, where 5000 pairs are drawn from the same class (i.e., a pair of



(a) Depiction of a siamese network architecture. At iteration k, each of two samples, $X_A^{(k)}$ and $X_B^{(k)}$, are passed through a convolutional neural network to yield embeddings z_A and z_B, respectively. The class label for samples A and B are denoted y_A and y_B, respectively. The L2-norm of these embeddings is computed as D_{AB}. The network is optimized on the contrastive loss [190] $\mathcal{L}$. Here $\mathbb{I}[\cdot]$ is an indicator function.

(b) Empirical cumulative distribution functions (CDF) for pairwise distances between images of the listed classes under the siamese network model. The x-axis plots the L2-norm between embedding vectors produced by the siamese network. The y-axis shows the proportion of samples in the respective group (by line color) whose embedded L2 norms were less than the specified threshold on the x-axis. Class groups are denoted by different line colors. For instance, "0-0" refers to pairs where each image is a "zero." We combine all disjoint class pairs, for example "0–8" or "3–4," into a single empirical CDF denoted as "A$\neq$B."

Figure A.2: Depiction of the siamese network architecture and the empirical cumulative distribution function for pairwise distances between digit classes.

Fives or a pair of Threes), and 5000 pairs are drawn from different classes (i.e., the pairs [2,3] or [1,7]). The siamese network is then optimized using gradient-based methods over 100 epochs using the contrastive loss function [190] (Figure A.2a). This analysis may be reproduced in the Supplementary Materials.

After training, we sampled same-class pairs (n=25,000) and different-class pairs (n = 25,000) from the MNIST training set (which contains 60,000 images). Pairwise distances for each sample were computed using the trained siamese network. If the "ones" are indeed the most homogeneous class, they should demonstrate a generally smaller pairwise distance than other digit class pairs. We evaluated this hypothesis by comparing empirical cumulative distribution functions (CDF) on the class-pair distances (Figure A.2b). Our results show that the empirical CDF for "1–1" image pairs dominate that of all other class pairs (where the distance between pairs of "ones" is lower).

Appendix B

Supplementary Material for *Prediction of Lithium Response Using Clinical Data*

B.1 Feature Comparisons Between Responders and Non-responders Across Sites

B.1.1 Cagliari (University)

Table B.1: Demographic descriptive statistics stratified by lithium response for Cagliari (University). *Abbreviations*: N (number or count), "with" (w/) Li(+) (lithium responder), Li(-) (lithium non-responder), BD (bipolar disorder), BD-I (bipolar I disorder), BD-II (bipolar II disorder), NOS (not otherwise specified), MDD (major depressive disorder), SZA (schizoaffective disorder), FDR (first degree relative), SDR (second degree relative) GAF (global assessment of functioning scale), SA (suicide attempt) SI (suicidal ideation), SES (socioeconomic status), UI (unemployment insurance). Normally distributed variables are represented as mean (standard deviation), while non-normally distributed variables are represented as median [interquartile range, IQR]. Categorical variables are represented as count (percentage), with all unique categories listed; where a categorical variable has no subheadings identifying the categories, it is implicitly a binary variable where the count (percentage) refers to the affirmative response of the variable.

Variable	**Li(-)**	**Li(+)**	**p**
n	146	60	
Male (%)	46 (31.5)	19 (31.7)	1
Age	45.48 [19.22, 83.02]	45.20 [18.67, 79.60]	0.04
Diagnosis (%)			0.033
BD I	70 (47.9)	25 (41.7)	
BD II	29 (19.9)	22 (36.7)	
SZA	47 (32.2)	13 (21.7)	
Onset D	26.00 [12.00, 69.00]	26.50 [16.00, 67.00]	0.65

Continued on next page...

Variable	**Li(-)**	**Li(+)**	**p**
Onset M	25.00 [13.00, 66.00]	27.00 [15.00, 70.00]	0.241
Onset m	42.00 [2.00, 66.00]	39.00 [21.00, 67.00]	0.957
Clinical course (%)			0.103
Chronic	6 (4.2)	1 (1.7)	
Chronic deteriorating	7 (4.9)	0 (0.0)	
Completely episodic	62 (43.1)	36 (60.0)	
Episodic + residual	67 (46.5)	23 (38.3)	
Single episode	2 (1.4)	0 (0.0)	
N LT Manias	4.00 [0.00, 28.00]	2.00 [0.00, 26.00]	0.03
N LT Depressions	4.50 [0.00, 41.00]	4.00 [0.00, 36.00]	0.46
N LT Mixed	0.00 [0.00, 5.00]	0.00 [0.00, 1.00]	0.123
N LT Episodes	10.00 [1.00, 66.00]	8.00 [2.00, 72.00]	0.573
LT Psychosis (%)			0.057
Episodic congruent	68 (46.9)	20 (33.3)	
Episodic incongruent	19 (13.1)	6 (10.0)	
Never	54 (37.2)	34 (56.7)	
Outside mood episodes	4 (2.8)	0 (0.0)	
Total ALDA score	3.00 [0.00, 6.00]	7.00 [7.00, 10.00]	<0.001
N Episodes on Li	3.00 [0.00, 25.00]	0.00 [0.00, 5.00]	<0.001
N Episodes pre-Li	3.00 [1.00, 60.00]	6.00 [1.00, 71.00]	<0.001
N SA	0.00 [0.00, 7.00]	0.00 [0.00, 3.00]	0.19
N serious SA	0.00 [0.00, 1.00]	0.00 [0.00, 1.00]	0.977
Age first SA	31.00 [16.00, 57.00]	29.00 [17.00, 47.00]	0.645
N FDR BD1	0.00 [0.00, 2.00]	0.00 [0.00, 1.00]	0.142
N FDR BD2	0.00 [0.00, 3.00]	0.00 [0.00, 2.00]	0.444
N FDR MDD	0.00 [0.00, 5.00]	0.00 [0.00, 2.00]	0.097
N FDR SZA	0.00 [0.00, 2.00]	0.00 [0.00, 2.00]	0.154
N FDR SCZ	0.00 [0.00, 1.00]	0.00 [0.00, 1.00]	0.15
N FDR Anx	0.00 [0.00, 7.00]	0.00 [0.00, 9.00]	0.815
N FDR Unaffected	5.00 [1.00, 14.00]	6.00 [2.00, 13.00]	0.117
N FDR Suicide	0.00 [0.00, 2.00]	0.00 [0.00, 1.00]	0.06

Continued on next page...

Variable	**Li(-)**	**Li(+)**	**p**
N FDR SA	0.00 [0.00, 2.00]	0.00 [0.00, 1.00]	0.934
N SDR Suicide	0.00 [0.00, 2.00]	0.00 [0.00, 1.00]	0.346
N SDR SA	0.00 [0.00, 1.00]	0.00 [0.00, 0.00]	0.265
LT SI	58 (40.6)	16 (27.1)	0.101
No SAD	146 (100.0)	60 (100.0)	NA
No Panic d/o	145 (99.3)	60 (100.0)	1
No GAD	146 (100.0)	60 (100.0)	NA
No OCD	146 (100.0)	60 (100.0)	NA
No addiction	120 (82.2)	56 (93.3)	0.065
Diabetes	4 (11.8)	0 (0.0)	0.609
HTN	12 (34.3)	4 (40.0)	1
SES (%)			0.341
Other	37 (27.6)	17 (28.8)	
Retired	18 (13.4)	11 (18.6)	
Social assist	8 (6.0)	2 (3.4)	
Student	5 (3.7)	2 (3.4)	
Unemployment ins	8 (6.0)	3 (5.1)	
Unknown	5 (3.7)	6 (10.2)	
Work full-time	53 (39.6)	17 (28.8)	
Work part-time	0 (0.0)	1 (1.7)	
Marital status (%)			0.373
Divorced	1 (0.7)	0 (0.0)	
Married	78 (53.4)	27 (45.0)	
Single	57 (39.0)	25 (41.7)	
Widowed	10 (6.8)	8 (13.3)	

B.1.2 Cagliari (Centro Bini)

Table B.2: Demographic descriptive statistics stratified by lithium response for Cagliari (Centro Bini). *Abbreviations*: N (number or count), “with” (w/) Li(+) (lithium responder), Li(-) (lithium non-responder), BD (bipolar disorder), BD-I (bipolar I disorder), BD-II (bipolar II disorder), NOS (not otherwise specified), MDD (major depressive disorder), SZA (schizoaffective disorder), FDR (first degree relative), SDR (second degree relative) GAF (global assessment of functioning scale), SA (suicide attempt) SI (suicidal ideation), SES (socioeconomic status), UI (unemployment insurance). Normally distributed variables are represented as mean (standard deviation), while non-normally distributed variables are represented as median [interquartile range, IQR]. Categorical variables are represented as count (percentage), with all unique categories listed; where a categorical variable has no subheadings identifying the categories, it is implicitly a binary variable where the count (percentage) refers to the affirmative response of the variable.

Variable	**Li(-)**	**Li(+)**	**p**
n	256	68	
Male (%)	104 (40.6)	31 (45.6)	0.549
Age	35.48 [10.65, 75.81]	36.92 [17.39, 71.48]	0.979
Diagnosis (%)			0.893
BD I	173 (67.6)	48 (70.6)	
BD II	79 (30.9)	19 (27.9)	
SZA	4 (1.6)	1 (1.5)	
Onset D	22.71 [4.30, 67.64]	21.16 [6.56, 65.25]	0.512
Onset M	24.39 [8.87, 55.15]	20.23 [13.09, 43.54]	0.113
Onset m	24.34 [12.38, 40.34]	35.74 [31.52, 37.59]	0.051
Clinical course (%)			0.672
Chronic	13 (5.2)	6 (8.8)	
Chronic fluctuating	81 (32.4)	20 (29.4)	
Completely episodic	125 (50.0)	35 (51.5)	
Continuous cycling	31 (12.4)	7 (10.3)	
LT Manias	4.50 [0.00, 120.00]	5.00 [1.00, 50.00]	0.962
LT Depressions	6.00 [0.00, 120.00]	7.00 [0.00, 68.00]	0.7
LT Episodes	12.00 [0.00, 240.00]	12.00 [1.00, 101.00]	0.866
Total ALDA score	2.00 [0.00, 6.00]	8.00 [7.00, 10.00]	<0.001

Continued on next page...

Variable	**Li(-)**	**Li(+)**	**p**
N SA	0.00 [0.00, 9.00]	0.00 [0.00, 3.00]	0.744
N serious SA	0.00 [0.00, 1.00]	0.00 [0.00, 1.00]	0.273
Age first SA	36.20 [16.20, 77.20]	38.75 [18.40, 61.10]	0.844
FDR Mood d/o	137 (63.1)	43 (74.1)	0.158
FDR BD	74 (33.8)	24 (41.4)	0.357
N FDR Suicide	0.00 [0.00, 2.00]	0.00 [0.00, 2.00]	0.331
N FDR SA	0.00 [0.00, 1.00]	0.00 [0.00, 1.00]	0.546
N SDR Suicide	0.00 [0.00, 1.00]	0.00 [0.00, 0.00]	0.201
LT SI	63 (100.0)	23 (100.0)	NA
No SAD	84 (98.8)	37 (100.0)	1
No Panic d/o	45 (53.6)	20 (54.1)	1
No GAD	84 (96.6)	37 (97.4)	1
No OCD	80 (89.9)	37 (100.0)	0.104
No addiction	201 (79.1)	56 (82.4)	0.676
No ADHD	84 (100.0)	37 (100.0)	NA
No Learning d/o	84 (98.8)	37 (100.0)	1
No primary insomnia	84 (100.0)	37 (100.0)	NA
No personality d/o	42 (100.0)	20 (100.0)	NA
Works part-time (%)	19 (18.6)	4 (10.5)	0.371
Marital status (%)			0.896
Divorced	16 (6.2)	4 (5.9)	
Married	107 (41.8)	31 (45.6)	
Single	116 (45.3)	29 (42.6)	
Unknown	3 (1.2)	0 (0.0)	
Widowed	14 (5.5)	4 (5.9)	

B.1.3 IGSLi

Table B.3: Demographic descriptive statistics stratified by lithium response for IGSLi. *Abbreviations*: N (number or count), “with” (w/) Li(+) (lithium responder), Li(-) (lithium non-responder), BD (bipolar disorder), BD-I (bipolar I disorder), BD-II (bipolar II disorder), NOS (not otherwise specified), MDD (major depressive disorder), SZA (schizoaffective disorder), FDR (first degree relative), SDR (second degree relative) GAF (global assessment of functioning scale), SA (suicide attempt) SI (suicidal ideation), SES (socioeconomic status), UI (unemployment insurance). Normally distributed variables are represented as mean (standard deviation), while non-normally distributed variables are represented as median [interquartile range, IQR]. Categorical variables are represented as count (percentage), with all unique categories listed; where a categorical variable has no subheadings identifying the categories, it is implicitly a binary variable where the count (percentage) refers to the affirmative response of the variable.

Variable	**Li(+)**
n	70
Male (%)	28 (40.0)
Age	55.09 [21.66, 77.64]
Diagnosis (%)	
BD I	47 (67.1)
BD II	18 (25.7)
MDD Recurrent	3 (4.3)
MDD Single	1 (1.4)
SZA	1 (1.4)
Age of onset (y)	28.00 [16.00, 63.00]
Onset D	30.00 [16.00, 63.00]
Onset M	32.00 [16.00, 58.00]
Onset m	38.00 [16.00, 63.00]
Polarity first episode (%)	
Biphasic (D-M)	6 (8.7)
Biphasic (M-D)	5 (7.2)
Hypomania	1 (1.4)
Major depression	41 (59.4)
Mania	8 (11.6)

Continued on next page...

Variable	**Li(+)**
Minor depression	8 (11.6)
Continuous cycling course (%)	1 (100.0)
LT Manias	1.00 [0.00, 8.00]
LT Depressions	3.00 [0.00, 15.00]
LT Multiphasic	0.50 [0.00, 13.00]
LT Episodes	6.00 [0.00, 27.00]
No rapid cycling	53 (100.0)
No rapid mood switching	1 (100.0)
GAF Last Ax	95.00 [90.00, 95.00]
Total ALDA score	9.00 [8.00, 10.00]
N SA	0.00 [0.00, 1.00]
FDR Mood d/o	20 (29.4)
FDR BD	2 (2.9)
N FDR BD1	0.00 [0.00, 1.00]
N FDR MDD	0.00 [0.00, 5.00]
N FDR SZA	0.00 [0.00, 1.00]
N FDR SCZ	0.00 [0.00, 1.00]
N FDR Unaffected	0.00 [0.00, 2.00]
Mania at SA	1 (100.0)
LT SI	9 (45.0)
SI related to mood episode	1 (100.0)
No Panic d/o	68 (100.0)
No OCD	68 (100.0)
No addiction	68 (98.6)
Works full-time (%)	1 (100.0)
Single marital (%)	1 (25.0)

B.1.4 Maritimes

Table B.4: Demographic descriptive statistics stratified by lithium response for Maritimes. *Abbreviations*: N (number or count), “with” (w/) Li(+) (lithium responder), Li(-) (lithium non-responder), BD (bipolar disorder), BD-I (bipolar I disorder), BD-II (bipolar II disorder), NOS (not otherwise specified), MDD (major depressive disorder), SZA (schizoaffective disorder), FDR (first degree relative), SDR (second degree relative) GAF (global assessment of functioning scale), SA (suicide attempt) SI (suicidal ideation), SES (socioeconomic status), UI (unemployment insurance). Normally distributed variables are represented as mean (standard deviation), while non-normally distributed variables are represented as median [interquartile range, IQR]. Categorical variables are represented as count (percentage), with all unique categories listed; where a categorical variable has no subheadings identifying the categories, it is implicitly a binary variable where the count (percentage) refers to the affirmative response of the variable.

Variable	**Li(-)**	**Li(+)**	**p**
n	274	69	
Male (%)	98 (35.8)	29 (42.0)	0.41
Age	43.22 [17.00, 82.51]	48.38 [18.73, 78.42]	0.007
Diagnosis (%)			0.905
BD I	193 (70.7)	52 (75.4)	
BD II	69 (25.3)	16 (23.2)	
BD NOS	1 (0.4)	0 (0.0)	
MDD Recurrent	2 (0.7)	0 (0.0)	
MDD Single	1 (0.4)	0 (0.0)	
SZA	7 (2.6)	1 (1.4)	
Age of onset (y)	21.00 [12.00, 60.00]	25.00 [9.00, 56.00]	0.014
Onset D	22.00 [12.00, 60.00]	26.00 [9.00, 60.00]	0.017
Onset M	27.50 [15.00, 61.00]	30.00 [14.00, 66.00]	0.226
Onset m	26.00 [0.00, 60.00]	31.50 [15.00, 56.00]	0.189
Polarity first episode (%)			0.034
Biphasic (D-M)	4 (1.5)	1 (1.5)	
Biphasic (M-D)	10 (3.7)	0 (0.0)	
Hypomania	34 (12.5)	10 (14.9)	
Major depression	167 (61.6)	29 (43.3)	

Continued on next page...

Variable	**Li(-)**	**Li(+)**	**p**
Mania	43 (15.9)	21 (31.3)	
Minor depression	9 (3.3)	5 (7.5)	
Mixed	3 (1.1)	1 (1.5)	
Periodic rapid cycling	1 (0.4)	0 (0.0)	
Clinical course (%)			<0.001
Chronic	24 (8.9)	1 (1.5)	
Chronic deteriorating	4 (1.5)	0 (0.0)	
Chronic fluctuating	97 (35.8)	8 (12.3)	
Completely episodic	56 (20.7)	47 (72.3)	
Continuous cycling	1 (0.4)	0 (0.0)	
Episodic + residual	80 (29.5)	8 (12.3)	
Single episode	9 (3.3)	1 (1.5)	
LT Manias	2.00 [0.00, 99.00]	2.00 [0.00, 34.00]	0.239
LT Depressions	3.00 [0.00, 99.00]	3.00 [0.00, 35.00]	0.051
LT Mixed	0.00 [0.00, 99.00]	0.00 [0.00, 3.00]	0.073
LT Multiphasic	0.00 [0.00, 99.00]	0.00 [0.00, 3.00]	0.001
LT Episodes	6.00 [1.00, 99.00]	5.00 [1.00, 99.00]	0.119
Rapid cycling (%)			0.001
Never	177 (68.1)	58 (90.6)	
Only on Antidepressants	5 (1.9)	0 (0.0)	
Spontaneous	78 (30.0)	6 (9.4)	
Rapid mood switching	15 (28.8)	2 (10.0)	0.169
LT Psychosis (%)			0.089
Episodic congruent	106 (41.6)	36 (56.2)	
Episodic incongruent	41 (16.1)	6 (9.4)	
Never	108 (42.4)	22 (34.4)	
GAF Last Ax	70.00 [35.00, 95.00]	80.00 [40.00, 100.00]	<0.001
Total ALDA score	2.00 [0.00, 6.00]	8.00 [7.00, 10.00]	<0.001
N Episodes on Li	3.00 [0.00, 99.00]	0.00 [0.00, 4.00]	<0.001
N Episodes pre-Li	4.00 [1.00, 99.00]	5.00 [1.00, 70.00]	0.025
N SA	0.00 [0.00, 7.00]	0.00 [0.00, 2.00]	0.071
N serious SA	1.00 [0.00, 7.00]	1.00 [0.00, 1.00]	0.273

Continued on next page...

Variable	Li(-)	Li(+)	p
Age first SA	26.00 [12.00, 64.00]	23.00 [15.00, 55.00]	0.645
FDR Mood d/o	154 (56.8)	38 (57.6)	1
FDR BD	89 (32.5)	28 (40.6)	0.26
N FDR BD1	0.00 [0.00, 6.00]	0.00 [0.00, 4.00]	0.137
N FDR BD2	1.00 [1.00, 1.00]	0.50 [0.00, 1.00]	0.317
N FDR MDD	0.00 [0.00, 7.00]	1.00 [0.00, 3.00]	0.696
N FDR SZA	0.00 [0.00, 1.00]	0.00 [0.00, 1.00]	0.215
N FDR SCZ	0.00 [0.00, 2.00]	0.00 [0.00, 0.00]	0.031
N FDR Anx	0.00 [0.00, 3.00]	0.00 [0.00, 1.00]	0.443
N FDR Unaffected	0.00 [0.00, 5.00]	0.00 [0.00, 2.00]	0.03
N FDR Suicide	0.00 [0.00, 2.00]	0.00 [0.00, 2.00]	0.51
N FDR SA	0.00 [0.00, 3.00]	0.00 [0.00, 1.00]	0.055
N SDR Suicide	0.00 [0.00, 2.00]	0.00 [0.00, 2.00]	0.066
N SDR SA	0.00 [0.00, 2.00]	0.00 [0.00, 1.00]	0.511
Mood at SA (%)			0.089
Biphasic MD	0 (0.0)	1 (10.0)	
Major depression	61 (88.4)	9 (90.0)	
Mania	4 (5.8)	0 (0.0)	
Mixed	3 (4.3)	0 (0.0)	
Rapid cycling	1 (1.4)	0 (0.0)	
LT SI	98 (44.1)	16 (31.4)	0.131
SI related to mood episode	94 (98.9)	13 (92.9)	0.604
No SAD	208 (77.6)	61 (93.8)	0.005
No Panic d/o	207 (76.7)	55 (83.3)	0.314
No GAD	162 (60.4)	55 (83.3)	0.001
No OCD	238 (87.8)	61 (93.8)	0.241
No addiction	185 (68.5)	52 (78.8)	0.136
No ADHD	258 (97.0)	64 (100.0)	0.341
No Learning d/o	252 (94.7)	64 (100.0)	0.126
No primary insomnia	227 (85.7)	59 (93.7)	0.135
No personality d/o	258 (97.4)	63 (98.4)	0.959
Diabetes	31 (12.0)	3 (5.0)	0.176

Continued on next page...

Variable	Li(-)	Li(+)	p
HTN	36 (14.1)	11 (19.0)	0.466
Menstrual d/o	43 (30.3)	6 (25.0)	0.777
Thyroid d/o	80 (31.6)	17 (28.8)	0.792
TBI	52 (24.2)	10 (23.8)	1
Migraine	40 (16.5)	3 (5.1)	0.041
SES (%)			0.079
Disabled	81 (34.2)	10 (16.9)	
Other	18 (7.6)	7 (11.9)	
Retired	14 (5.9)	7 (11.9)	
Social assist	31 (13.1)	8 (13.6)	
Student	11 (4.6)	1 (1.7)	
Unemployment ins	21 (8.9)	6 (10.2)	
Unknown	4 (1.7)	1 (1.7)	
Work full-time	45 (19.0)	11 (18.6)	
Work part-time	12 (5.1)	8 (13.6)	
Marital status (%)			0.567
Divorced	54 (21.1)	12 (19.4)	
Married	131 (51.2)	38 (61.3)	
Single	60 (23.4)	9 (14.5)	
Unknown	3 (1.2)	1 (1.6)	
Widowed	8 (3.1)	2 (3.2)	

B.1.5 Montreal

Table B.5: Demographic descriptive statistics stratified by lithium response for Montreal. *Abbreviations*: N (number or count), "with" (w/) Li(+) (lithium responder), Li(-) (lithium non-responder), BD (bipolar disorder), BD-I (bipolar I disorder), BD-II (bipolar II disorder), NOS (not otherwise specified), MDD (major depressive disorder), SZA (schizoaffective disorder), FDR (first degree relative), SDR (second degree relative) GAF (global assessment of functioning scale), SA (suicide attempt) SI (suicidal ideation), SES (socioeconomic status), UI (unemployment insurance). Normally distributed variables are represented as mean (standard deviation), while non-normally distributed variables are represented as median [interquartile range, IQR]. Categorical variables are represented as count (percentage), with all unique categories listed; where a categorical variable has no subheadings identifying the categories, it is implicitly a binary variable where the count (percentage) refers to the affirmative response of the variable.

Variable	**Li(-)**	**Li(+)**	**p**
n	80	15	
Male (%)	37 (46.2)	9 (60.0)	0.486
Age	50.09 [22.50, 76.19]	53.06 [41.06, 77.50]	0.069
Diagnosis (%)			0.776
BD I	54 (67.5)	12 (80.0)	
BD II	24 (30.0)	3 (20.0)	
Not primary mood disorder	1 (1.2)	0 (0.0)	
SZA	1 (1.2)	0 (0.0)	
Age of onset (y)	22.50 [7.00, 64.00]	29.00 [16.00, 47.00]	0.212
Onset D	25.00 [10.00, 67.00]	31.00 [16.00, 48.00]	0.419
Onset M	27.00 [16.00, 59.00]	31.00 [18.00, 46.00]	0.581
Onset m	36.50 [13.00, 67.00]	34.00 [16.00, 56.00]	0.81
Polarity first episode (%)			0.065
Biphasic (D-M)	2 (2.5)	0 (0.0)	
Biphasic (M-D)	12 (15.0)	3 (20.0)	
Hypomania	8 (10.0)	6 (40.0)	
Major depression	42 (52.5)	4 (26.7)	
Mania	10 (12.5)	1 (6.7)	
Minor depression	4 (5.0)	0 (0.0)	

Continued on next page...

Variable	Li(-)	Li(+)	p
Periodic rapid cycling	2 (2.5)	1 (6.7)	
Clinical course (%)			0.003
Chronic fluctuating	15 (18.8)	1 (6.7)	
Completely episodic	23 (28.7)	12 (80.0)	
Episodic + residual	40 (50.0)	2 (13.3)	
Single episode	2 (2.5)	0 (0.0)	
LT Manias	2.00 [0.00, 15.00]	2.00 [0.00, 6.00]	0.926
LT Depressions	5.00 [0.00, 99.00]	5.00 [0.00, 20.00]	0.3
LT Episodes	11.00 [1.00, 80.00]	9.50 [3.00, 99.00]	0.374
Rapid cycling (%)			0.05
Never	48 (60.8)	14 (93.3)	
Only on Antidepressants	4 (5.1)	0 (0.0)	
Spontaneous	27 (34.2)	1 (6.7)	
Rapid mood switching	49 (62.0)	5 (41.7)	0.307
LT Psychosis (%)			0.929
Episodic congruent	36 (45.0)	8 (53.3)	
Episodic incongruent	11 (13.8)	2 (13.3)	
Never	28 (35.0)	4 (26.7)	
Outside of mood episodes	5 (6.2)	1 (6.7)	
GAF Last Ax	75.00 [52.00, 95.00]	90.00 [0.00, 99.00]	0.001
Total ALDA score	3.00 [0.00, 6.00]	7.00 [7.00, 8.00]	<0.001
N Episodes on Li	3.00 [0.00, 99.00]	1.00 [0.00, 5.00]	<0.001
N Episodes pre-Li	5.00 [1.00, 99.00]	6.50 [3.00, 99.00]	0.539
N SA	0.00 [0.00, 6.00]	0.00 [0.00, 2.00]	0.168
N serious SA	1.00 [0.00, 6.00]	1.00 [0.00, 1.00]	0.291
Age first SA	31.00 [10.00, 56.00]	32.00 [29.00, 40.00]	0.628
FDR Mood d/o	22 (59.5)	5 (83.3)	0.505
FDR BD	15 (18.8)	6 (40.0)	0.139
N FDR BD1	0.00 [0.00, 2.00]	0.00 [0.00, 1.00]	0.085
N FDR MDD	1.00 [0.00, 3.00]	1.00 [0.00, 5.00]	0.234
N FDR SCZ	0.00 [0.00, 2.00]	0.00 [0.00, 0.00]	0.514
N FDR Anx	0.00 [0.00, 2.00]	0.00 [0.00, 3.00]	0.068

Continued on next page...

Variable	**Li(-)**	**Li(+)**	**p**
N FDR Unaffected	1.00 [0.00, 4.00]	1.00 [0.00, 2.00]	0.388
N FDR Suicide	0.00 [0.00, 1.00]	0.00 [0.00, 0.00]	0.292
N FDR SA	0.00 [0.00, 2.00]	0.00 [0.00, 2.00]	0.084
N SDR Suicide	0.00 [0.00, 1.00]	0.00 [0.00, 1.00]	0.971
N SDR SA	0.00 [0.00, 1.00]	0.00 [0.00, 0.00]	0.674
Mood at SA (%)			0.926
Major depression	26 (92.9)	2 (100.0)	
Mania	1 (3.6)	0 (0.0)	
Minor depression	1 (3.6)	0 (0.0)	
LT SI	44 (55.7)	8 (53.3)	1
SI related to mood episode	37 (86.0)	7 (100.0)	0.67
No SAD	64 (81.0)	10 (71.4)	0.645
No Panic d/o	63 (78.8)	13 (86.7)	0.725
No GAD	66 (84.6)	14 (100.0)	0.253
No OCD	74 (93.7)	14 (93.3)	1
No addiction	50 (62.5)	8 (53.3)	0.704
No ADHD	68 (85.0)	14 (93.3)	0.651
No Learning d/o	73 (92.4)	14 (93.3)	1
No primary insomnia	74 (92.5)	14 (93.3)	1
No personality d/o	39 (48.8)	11 (73.3)	0.142
Diabetes	5 (6.3)	3 (20.0)	0.217
HTN	12 (15.2)	2 (13.3)	1
Menstrual d/o	15 (34.9)	3 (50.0)	0.789
Thyroid d/o	26 (32.5)	6 (40.0)	0.79
TBI	24 (30.8)	4 (26.7)	0.992
Migraine	29 (37.2)	2 (13.3)	0.135
SES (%)			0.847
Disabled	8 (10.0)	1 (6.7)	
Other	4 (5.0)	1 (6.7)	
Retired	15 (18.8)	5 (33.3)	
Social assist	10 (12.5)	1 (6.7)	
Student	6 (7.5)	0 (0.0)	

Continued on next page...

Variable	Li(-)	Li(+)	p
Unemployment ins	9 (11.2)	1 (6.7)	
Work full-time	20 (25.0)	4 (26.7)	
Work part-time	8 (10.0)	2 (13.3)	
Marital status (%)			0.547
Divorced	24 (30.4)	2 (13.3)	
Married	28 (35.4)	7 (46.7)	
Single	26 (32.9)	6 (40.0)	
Widowed	1 (1.3)	0 (0.0)	

B.1.6 Ontario

Table B.6: Demographic descriptive statistics stratified by lithium response for Ontario. *Abbreviations*: N (number or count), “with” (w/) Li(+) (lithium responder), Li(-) (lithium non-responder), BD (bipolar disorder), BD-I (bipolar I disorder), BD-II (bipolar II disorder), NOS (not otherwise specified), MDD (major depressive disorder), SZA (schizoaffective disorder), FDR (first degree relative), SDR (second degree relative) GAF (global assessment of functioning scale), SA (suicide attempt) SI (suicidal ideation), SES (socioeconomic status), UI (unemployment insurance). Normally distributed variables are represented as mean (standard deviation), while non-normally distributed variables are represented as median [interquartile range, IQR]. Categorical variables are represented as count (percentage), with all unique categories listed; where a categorical variable has no subheadings identifying the categories, it is implicitly a binary variable where the count (percentage) refers to the affirmative response of the variable.

Variable	Li(-)	Li(+)	p
n	19	98	
Male (%)	8 (42.1)	47 (48.0)	0.828
Age	42.53 [28.11, 66.66]	47.74 [21.85, 80.16]	0.275
Diagnosis (%)			0.018
BD I	9 (47.4)	64 (65.3)	
BD II	6 (31.6)	29 (29.6)	
MDD Recurrent	2 (10.5)	5 (5.1)	

Continued on next page...

Variable	Li(-)	Li(+)	p
MDD Single	1 (5.3)	0 (0.0)	
SZA	1 (5.3)	0 (0.0)	
Age of onset (y)	26.00 [16.00, 53.00]	24.00 [12.00, 64.00]	0.902
Onset D	26.50 [18.00, 53.00]	25.00 [12.00, 46.00]	0.775
Onset M	27.50 [18.00, 38.00]	27.00 [13.00, 60.00]	1
Onset m	24.00 [18.00, 46.00]	32.00 [19.00, 57.00]	0.364
Polarity first episode (%)			0.361
Biphasic (D-M)	1 (6.7)	4 (4.8)	
Biphasic (M-D)	1 (6.7)	4 (4.8)	
Hypomania	1 (6.7)	3 (3.6)	
Major depression	11 (73.3)	42 (50.0)	
Mania	1 (6.7)	22 (26.2)	
Minor depression	0 (0.0)	9 (10.7)	
Completely episodic course (%)	3 (100.0)	8 (100.0)	NA
LT Manias	1.00 [0.00, 2.00]	2.00 [0.00, 11.00]	0.059
LT Depressions	2.00 [1.00, 8.00]	2.50 [0.00, 15.00]	0.935
LT Multiphasic	0.00 [0.00, 2.00]	0.00 [0.00, 13.00]	0.93
LT Episodes	4.50 [1.00, 8.00]	6.00 [1.00, 26.00]	0.027
Spontaneous rapid cyc. (%)	0 (0.0)	1 (2.4)	1
Rapid mood switching	0 (0.0)	1 (25.0)	1
LT Psychosis (%)			0.605
Episodic congruent	1 (12.5)	4 (13.8)	
Episodic incongruent	1 (12.5)	1 (3.4)	
Never	6 (75.0)	24 (82.8)	
GAF Last Ax	NA [Inf, -Inf]	90.00 [90.00, 95.00]	NA
Total ALDA score	4.00 [0.00, 6.00]	8.50 [7.00, 10.00]	<0.001
N SA	0.00 [0.00, 0.00]	0.00 [0.00, 3.00]	0.167
FDR Mood d/o	8 (44.4)	47 (48.0)	0.986
FDR BD	8 (44.4)	43 (43.9)	1
N FDR BD1	0.00 [0.00, 5.00]	0.00 [0.00, 5.00]	0.415
N FDR MDD	0.00 [0.00, 1.00]	0.00 [0.00, 4.00]	0.382

Continued on next page...

Variable	**Li(-)**	**Li(+)**	**p**
N FDR SZA	0.00 [0.00, 1.00]	0.00 [0.00, 1.00]	0.124
N FDR Anx	0.00 [0.00, 2.00]	0.00 [0.00, 1.00]	0.745
N FDR Unaffected	0.00 [0.00, 0.00]	0.00 [0.00, 2.00]	0.543
N FDR Suicide	2.00 [0.00, 2.00]	0.00 [0.00, 2.00]	0.192
N FDR SA	2.00 [0.00, 2.00]	0.00 [0.00, 2.00]	0.378
N SDR Suicide	0.00 [0.00, 0.00]	0.00 [0.00, 1.00]	0.513
LT SI	3 (42.9)	15 (42.9)	1
SI related to mood episode	3 (100.0)	10 (100.0)	NA
No SAD	0 (NaN)	3 (100.0)	NA
No Panic d/o	3 (75.0)	26 (100.0)	0.273
No GAD	1 (100.0)	3 (60.0)	1
No OCD	3 (100.0)	26 (100.0)	NA
No addiction	6 (85.7)	26 (78.8)	1
No ADHD	1 (100.0)	2 (100.0)	NA
No Learning d/o	1 (100.0)	2 (100.0)	NA
No primary insomnia	1 (100.0)	2 (40.0)	1
No personality d/o	1 (100.0)	2 (100.0)	NA
No Diabetes	1 (100.0)	0 (NaN)	NA
No HTN	1 (100.0)	0 (NaN)	NA
Menstrual d/o	0 (0.0)	1 (100.0)	1
No Thyroid d/o	1 (100.0)	0 (NaN)	NA
No TBI	1 (100.0)	0 (NaN)	NA
Migraine	0 (0.0)	1 (100.0)	1
Works full-time (%)	1 (100.0)	3 (100.0)	NA
Single marital status (%)	1 (50.0)	0 (0.0)	0.699

B.1.7 Poznan

Table B.7: Demographic descriptive statistics stratified by lithium response for Poznan. *Abbreviations*: N (number or count), "with" (w/) Li(+) (lithium responder), Li(-) (lithium non-responder), BD (bipolar disorder), BD-I (bipolar I disorder), BD-II (bipolar II disorder), NOS (not otherwise specified), MDD (major depressive disorder), SZA (schizoaffective disorder), FDR (first degree relative), SDR (second degree relative) GAF (global assessment of functioning scale), SA (suicide attempt) SI (suicidal ideation), SES (socioeconomic status), UI (unemployment insurance). Normally distributed variables are represented as mean (standard deviation), while non-normally distributed variables are represented as median [interquartile range, IQR]. Categorical variables are represented as count (percentage), with all unique categories listed; where a categorical variable has no subheadings identifying the categories, it is implicitly a binary variable where the count (percentage) refers to the affirmative response of the variable.

Variable	**Li(-)**	**Li(+)**	**p**
n	52	59	
Male (%)	20 (38.5)	13 (22.0)	0.093
Age	64.08 [38.72, 86.22]	65.59 [37.81, 108.69]	0.091
Dx = BD II (%)	11 (21.2)	17 (28.8)	0.479
Onset D	31.00 [19.00, 59.00]	32.00 [18.00, 57.00]	0.976
Onset M	32.00 [20.00, 58.00]	33.00 [16.00, 52.00]	0.827
Onset m	36.00 [20.00, 53.00]	44.00 [19.00, 57.00]	0.785
Clinical course (%)			0.154
Chronic	19 (36.5)	29 (50.0)	
Chronic deteriorating	5 (9.6)	2 (3.4)	
Chronic fluctuating	22 (42.3)	25 (43.1)	
Episodic + residual	6 (11.5)	2 (3.4)	
LT Manias	2.00 [0.00, 16.00]	2.00 [0.00, 6.00]	0.172
LT Depressions	3.00 [1.00, 18.00]	2.00 [1.00, 13.00]	0.044
LT Mixed	0.00 [0.00, 4.00]	0.00 [0.00, 2.00]	0.017
LT Multiphasic	0.00 [0.00, 6.00]	0.00 [0.00, 0.00]	0.164
LT Episodes	7.00 [0.00, 30.00]	5.00 [0.00, 16.00]	0.004
Rapid cycling (%)			0.001
Never	39 (75.0)	57 (98.3)	

Continued on next page...

Variable	Li(-)	Li(+)	p
Only on Antidepressants	8 (15.4)	1 (1.7)	
Spontaneous	5 (9.6)	0 (0.0)	
No rapid mood switching	51 (100.0)	57 (100.0)	NA
No LT Psychosis (%)	44 (84.6)	53 (91.4)	0.423
Total ALDA score	5.00 [0.00, 6.00]	9.00 [7.00, 10.00]	<0.001
N Episodes on Li	2.00 [0.00, 12.00]	1.00 [0.00, 6.00]	<0.001
N Episodes pre-Li	4.50 [2.00, 18.00]	4.00 [1.00, 12.00]	0.11
N SA	0.00 [0.00, 7.00]	0.00 [0.00, 9.00]	0.329
FDR BD	29 (55.8)	25 (42.4)	0.223
N FDR BD1	1.00 [0.00, 1.00]	0.00 [0.00, 1.00]	0.161
No SAD	25 (48.1)	30 (50.8)	0.919
No Panic d/o	3 (5.8)	16 (27.1)	0.006
No GAD	13 (25.5)	19 (32.2)	0.574
No OCD	45 (86.5)	49 (83.1)	0.806
No addiction	21 (40.4)	28 (47.5)	0.577
No ADHD	3 (5.8)	9 (15.3)	0.194
No Learning d/o	2 (3.8)	15 (25.4)	0.004
No primary insomnia	17 (32.7)	50 (84.7)	<0.001
No personality d/o	23 (44.2)	33 (55.9)	0.298
Diabetes	3 (5.8)	4 (6.8)	1
HTN	30 (58.8)	34 (57.6)	1
Menstrual d/o	1 (3.0)	0 (0.0)	0.913
Thyroid d/o	7 (13.5)	5 (8.5)	0.591
TBI	49 (94.2)	25 (42.4)	<0.001
Migraine	28 (53.8)	10 (16.9)	<0.001
SES (%)			0.29
Retired	24 (46.2)	25 (42.4)	
Social assist	2 (3.8)	0 (0.0)	
Work full-time	18 (34.6)	19 (32.2)	
Work part-time	8 (15.4)	15 (25.4)	
Marital status (%)			0.75
Divorced	6 (11.5)	6 (10.2)	

Continued on next page...

Variable	Li(-)	Li(+)	p
Married	37 (71.2)	38 (64.4)	
Single	1 (1.9)	1 (1.7)	
Widowed	8 (15.4)	14 (23.7)	

B.2 Sensitivity Analyses with Hyperparameter Optimization and Different Model Architectures

The aggregate analysis was repeated using hyperparameter optimization. However, here a nesting procedure must be used in order to prevent "information leak" from validation to training data. The procedure is outlined in Algorithm 1. Here, we used 10 fold cross validation, and 100 samples from the uniform distribution over missing data. Results are shown in Table B.8 in comparison to the unoptimized run, demonstrating that both analyses are virtually the same with respect to performance.

Results of tests across different model architectures are shown in Figure B.1.

B.3 Were Prediction Errors Related to Missingness?

We investigated the relationship between proportion of missing data and model prediction error for the random forest classifier using the following linear mixed effects model (in the syntax of the lme4 package for the R statistical programming language):

```
Error ~ P_Missing + (P_Missing + 1|Centre)
```

This model assumes that error is a function of the proportion of missing data for each subject, but that the effect of missingness on prediction error (i.e. random slope), and prediction error itself (i.e. random intercept), varies across sites. Error was first converted from the [0, 1] interval to the [-1, 1] interval using the (bijective) transform $2x - 1$. Results are shown in Table B.9, showing that the proportion of data missing were not related to prediction error of the random forest classifier to a statistically significant degree.

B.4 Analysis of Predicting Lithium Non-response

The entire analysis was repeated exactly as described in the Methods section of the main paper, with only one difference. Here, the "positive" class was defined as an Alda score <4.

Algorithm 1: Pseudocode for procedure outlining hyperparameter optimization on aggregate data.

```
Input:
    • Data $D$
    • Number of outer ($K$) and inner ($L$) cross-validation folds (outer) $K$
    • nsamples_perfold: Number of uniform samples over missing values (outer loop)
    • nsamples_hpopt: Number of uniform samples over missing values (inner loop)
results = $\{\}$
foldwise_indices = StratifiedKFoldPartitionFunction($\mathbf{D}$, $K$)
for $k \in \{1, 2, \ldots, K\}$ do
    train_indices, test_indices = foldwise_indices[k]
    Xtrain, ytrain = $\mathbf{D}$[train_indices]
    Xtest, ytest = $\mathbf{D}$[test_indices]
    $\mathcal{M}$ = RandomForestClassifier
    param_bounds = $\{$"NEstimators": (10, 1000)$\}$
    target_metric = MatthewsCorrelationCoefficient
    nfolds = L
    W = OptimizeHyperparameters($D$[train_indices], $\mathcal{M}$, param_bounds, target_metric, nfolds,
      nsamples_hpopt)
    result_samples = $\{\}$
    for $i \in \{1, 2, \ldots,$ nsamples_per_fold $\}$ do
        Xtrain = SampleUniformlyOverMissingData(Xtrain)
        Xtest = SampleUniformlyOverMissingData(Xtest)
        Xtrain, ytrain = SmoteTomekRebalancing(Xtrain, ytrain)
        $\mathcal{M}$ = RandomForestClassifier(NEstimators=W)
        $\mathcal{M}^*$ = TrainModel($\mathcal{M}$, Xtrain, ytrain)
        $\hat{\mathbf{y}}$ = PredictClasses($\mathcal{M}^*$, Xtest)
        $\hat{\mathbf{p}}$ = PredictClassProbabilities($\mathcal{M}^*$, Xtest)
        test_statistics = Performance(ytest, $\hat{\mathbf{y}}$, $\hat{\mathbf{p}}$)
        result_samples = AppendToList(result_samples, test_statistics)
    end
    results = AppendToList(results, Expectation(result_samples))
end
```

Table B.8: Comparison of aggregate analysis results with and without hyperparameter optimization. The run without hyperparameter optimization was done with a random forest classifier with 100 trees set a priori. *Abbreviations*: positive and negative predictive values (PPV, NPV), area under the receiver operating characteristic curve (AUC), Cohen's kappa (Kappa).

	Optimized	
Statistic	**Yes**	**No**
Accuracy	0.78 (0.77-0.8)	0.78 (0.76, 0.80)
Sensitivity	0.54 (0.49, 0.59)	0.54 (0.50, 0.57)
Specificity	0.91 (0.9, 0.92)	0.91 (0.89, 0.93)
PPV	0.77 (0.75, 0.79)	0.76 (0.72, 0.80)
NPV	0.79 (0.77, 0.8)	0.79 (0.77, 0.80)
AUC	0.81 (0.8, 0.83)	0.81 (0.79, 0.83)
Kappa	0.48 (0.45, 0.52)	0.48 (0.44, 0.52)

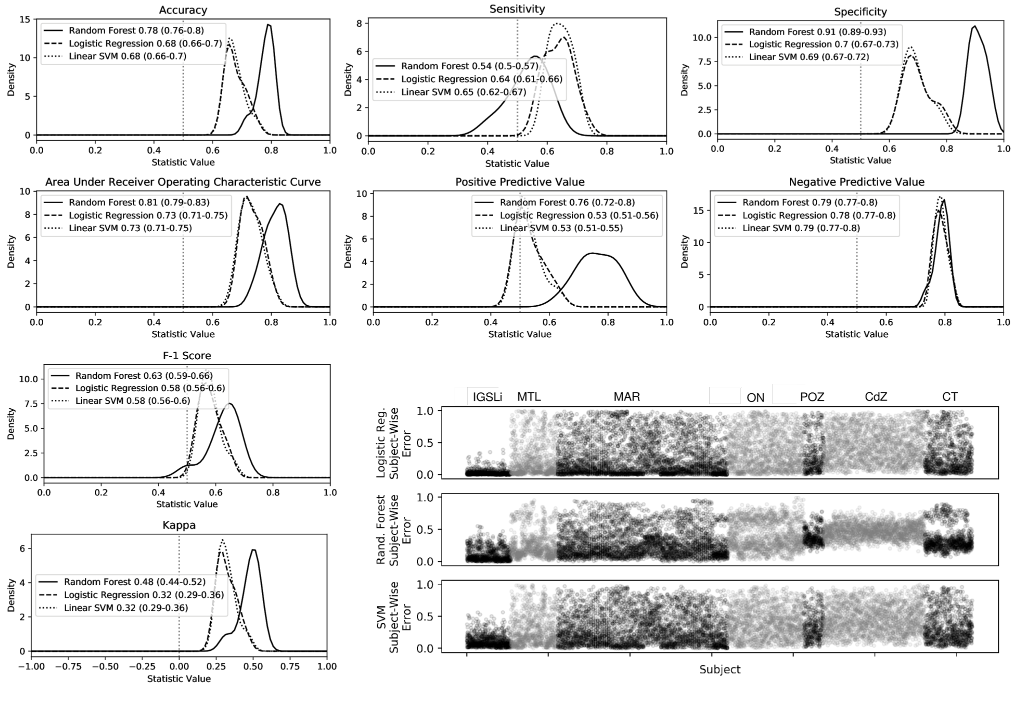


Figure B.1: Post-hoc comparison of Random Forest, Logistic Regression, and Linear SVM models. The models were trained on the exact same data, under the exact same conditions as shown in Figure 1. The plots on the bottom right hand side show the prediction error for each subject in the dataset (across samples under the missing data marginalization), where subjects are grouped according to their sites of origin. *Abbreviations*: International Group for the study of Lithium (IGSLi), Montreal (MTL), maritimes (MAR), Ontario (ON), Poznan (POZ), Cagliari (Centro Bini; CT), Cagliari (University; CdZ).

Table B.9: Results of linear mixed effects regression model of random forest prediction error against proportion of missing variables, where both the intercept and slope were taken to be random effects across centres.

	Dependent Variable $(2 * Error - 1)$
PMissing	0.223 (0.230)
Intercept	-0.448 (0.151)***
Observations	1266
Log-Likelihood	-537.914
Akaike Inf. Criteria	1087.828
Bayesian Inf. Criteria	1118.690
* p<0.1; ** p<0.05; *** p<0.01;	

In this section we present the results of the aggregate and site-level analyses, and the feature importances for the Aggregate, Aggregate (without Maritimes), Maritimes, and Poznan data in order to compare the informative features between analyses.

B.5 Predictive Capacity of a Model Including only Clinical Course and Rapid Cycling Variables

We have repeated the aggregate analysis with the same model architecture (random forest classifier [RFC] with 100 estimators) in an identical fashion to that reported in the main text, albeit with restrictions to the included features. Specifically, we tested classification performance under three conditions:

1. Including clinical course and rapid cycling variables (CC+RC)
2. Including clinical course only (CC)
3. Including rapid cycling only (RC)

We note that this cannot substitute for a more complete study of variable importance. Rather, we intend this analysis primarily as a supplement to further qualify our statements in the discussion regarding the potential importance of clinical course and rapid cycling. Results are shown in Table B.12, with comparison to the original aggregate analysis (labeled ALL). Clinical course achieved a better classification performance than RC alone (kappa 0.31, 95% CI [0.26, 0.36] vs. 0.14 [0.11, 0.17]), although both were inferior to the aggregate model with all variables (kappa 0.46 [0.4, 0.51]). Combining both CC and RC together improves only slightly (kappa 0.35 [0.28, 0.38]; and with questionable significance) over the CC model alone. Using CC alone appears to have a better specificity (0.89 [0.62, 0.94]) than RC alone (0.50 [0.48, 0.52]), albeit with a potentially lower sensitivity (0.42 [0.37, 0.67] vs. 0.65 [0.64, 0.68]). Using only CC and RC (whether alone or combined) results in a lower positive predictive value than the complete model (PPV for all variables was 0.74 [0.69, 0.79]), although the 95% CI for the PPV using only CC is wide (0.63 [0.51, 0.76]).

We then applied the same procedure to the site-level analysis (Table B.13), the leave-one-site-out (LOSO) analysis (Table B.14), and the predict-one-site-out (POSO) analysis (Table B.13). The CC+RC variables together were most informative within the Maritimes (kappa 0.40 [0.15, 0.61]) and Montreal (0.27 [0.23, 0.32]) sites. Clinical course alone was also most informative in those sites (MAR kappa 0.47 [0.32, 0.57]; MTL 0.35 [0.31, 0.36]), with rapid cycling showing a similar, but attenuated pattern (MAR kappa 0.12 [0.05, 0.15]; MTL 0.14 [0.11, 0.17]). These results would suggest that the performance of the aggregate models reported in Table B.12 were driven entirely by the Maritime and Montreal samples. However, the LOSO analysis results contribute some nuance to the interpretation.

If the Maritimes and Montreal sites were the only samples in which CC and RC were important for classification, then leaving those sites out from the aggregate analysis should be the only scenarios in which overall classification performance declines. However, Table B.14 shows that in the CC+RC condition, exclusion of any site (with the exception of Centro Bini) impairs out of sample classification performance. Under the CC condition, exclusion of any site reduced the aggregate classification performance substantially, while under the RC condition, performance improves with exclusion of the Centro Bini sample (kappa improves from 0.14 [0.11, 0.17] to 0.34 [0.30, 0.43]). These results suggest that each site contributed important information about the relationship between clinical course and lithium response, but that information about rapid cycling was contributed in a more heterogeneous pattern.

Finally, the most important result from the POSO analysis (reported in Table B.13 alongside the site-level results) is the inability of a classifier to predict the Maritimes' sample given information from all other samples. Recall that in the main analysis (with all variables), only the Maritimes sample was classifiable to a non-trivial extent (kappa 0.16 [0.12, 0.19]). However, using CC+RC only, the ability to predict the Maritimes data was substantially lower (kappa 0.02 [0.02, 0.02]), and this was also reflected in the sole use of

CC or RC variables.

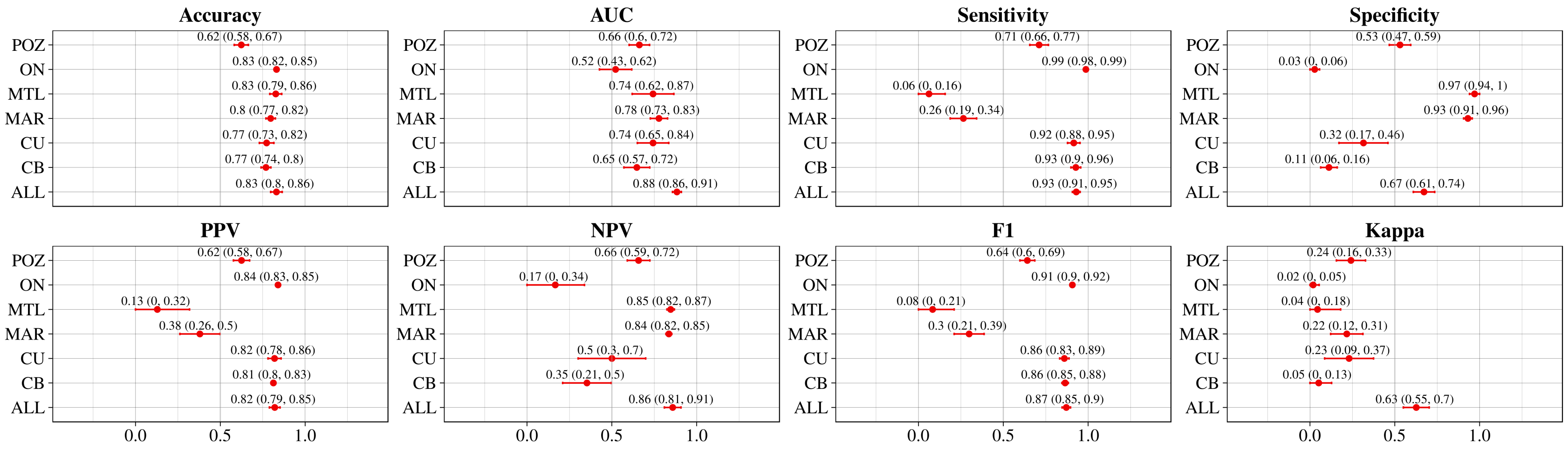


Figure B.2: Performance of random forest classifier on data pooled between sites (ALL), and within each contributing site for the analysis of predicting non-response (Alda score <4). Sites are indicated along the y-axes. *Abbreviations*: Poznan (POZ), Ottawa/Hamilton (ON; ”Ontario”), Montreal (MTL), Maritimes (MAR), Cagliari (Centro Bini; CB), Cagliari (University; CU). Statistic abbreviations: Area under receiver operating characteristic curve (AUC), positive predictive value (PPV), negative predictive value (NPV), F1-score (F1), and Cohen’s Kappa (Kappa).

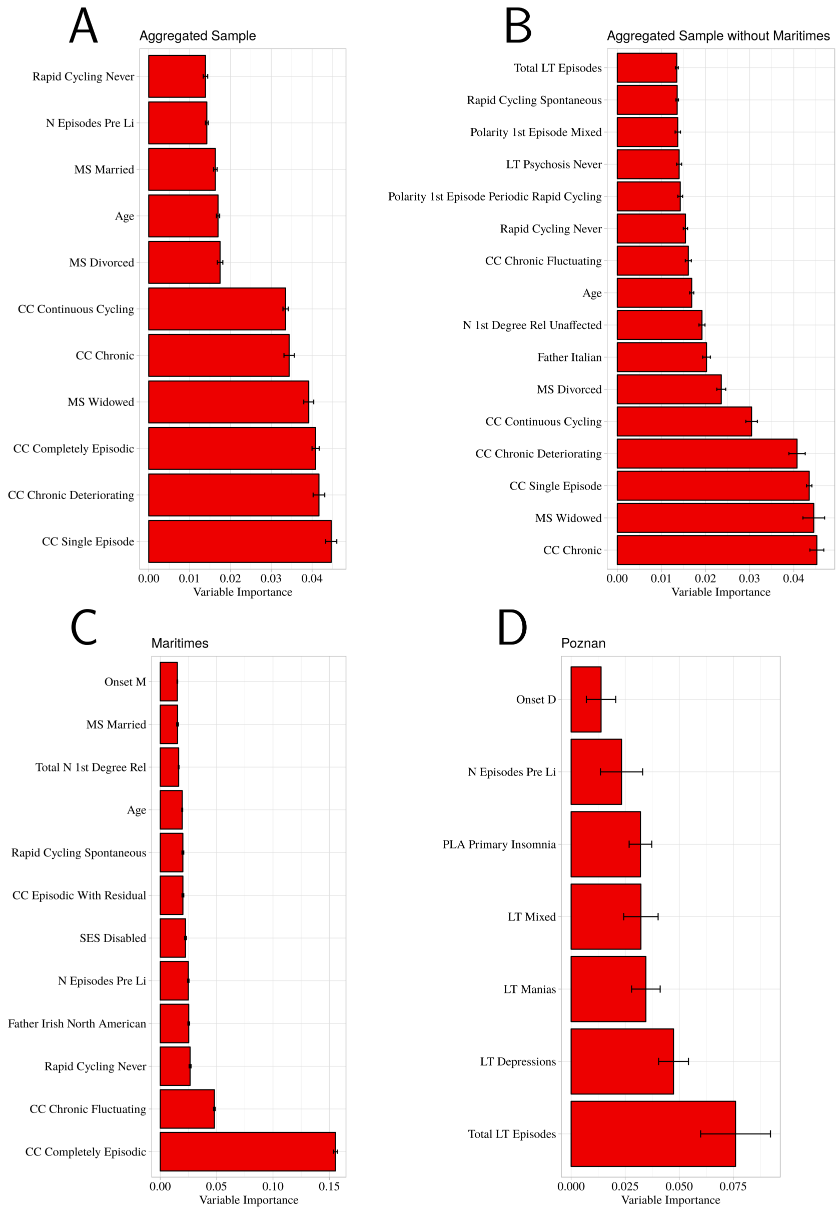


Figure B.3: Variable importance across (A) Aggregate dataset, (B) Aggregate dataset excluding the Maritimes data, (C) Maritimes site-level data, and (D) Poznan site level data. These were obtained from the analyses of predicting lithium non-response (Alda score <4). Due to space constraints, only those variables with coefficients above the overall mean were included in these plots. Notwithstanding, only bars that strongly deviate from the height of others should be considered "important." Bars are variable importance means over the 10 folds, and error bars are standard errors. *Abbreviations*: lifetime (LT), clinical course (CC), global assessment of functioning (GAF), marital status (MS), proportion of life affected (PLA), schizophrenia (SCZ).

Table B.10: Results of predict one site out analysis on the non-responder (Alda score <4) prediction. *Abbreviations*: Cagliari (Centro Bini; CB), Cagliari (University; CU), International Group for Study of Lithium (IGSLi), Maritimes (MAR), Montreal (MTL), Ontario (ON), Poznan (POZ), positive and negative predictive values (PPV, NPV), area under the receiver operating characteristic curve (AUC), Cohen's kappa (Kappa).

Site	**Accuracy**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**AUC**	**Kappa**
CU	0.76 (0.76, 0.77)	0.99 (0.99, 1.0)	0.01 (0.0, 0.02)	0.77 (0.76, 0.77)	-	0.67 (0.65, 0.69)	0.01 (-0.0, 0.02)
CB	0.59 (0.56, 0.61)	0.68 (0.64, 0.72)	0.2 (0.17, 0.22)	0.78 (0.77, 0.78)	0.13 (0.12, 0.14)	0.39 (0.37, 0.41)	-0.1 (-0.12, -0.08)
IGSLi	0.98 (0.98, 0.99)	0.0 (0.0, 0.0)	0.98 (0.98, 0.99)	0.0 (0.0, 0.0)	1.0 (1.0, 1.0)	-	0.0 (0.0, 0.0)
MAR	0.76 (0.76, 0.77)	0.88 (0.87, 0.89)	0.3 (0.29, 0.31)	0.84 (0.83, 0.84)	0.38 (0.36, 0.41)	0.65 (0.64, 0.66)	0.2 (0.18, 0.21)
MTL	0.96 (0.95, 0.97)	0.99 (0.99, 1.0)	0.2 (0.05, 0.35)	0.97 (0.96, 0.97)	-	0.97 (0.94, 0.99)	0.24 (0.05, 0.43)
ON	0.88 (0.87, 0.89)	0.14 (0.1, 0.18)	0.95 (0.94, 0.96)	0.18 (0.13, 0.23)	0.93 (0.92, 0.93)	0.64 (0.6, 0.69)	0.09 (0.05, 0.14)
POZ	0.22 (0.21, 0.23)	1.0 (1.0, 1.0)	0.03 (0.03, 0.04)	0.2 (0.2, 0.2)	1.0 (1.0, 1.0)	0.5 (0.44, 0.56)	0.01 (0.01, 0.02)

Table B.11: Results of leave one site out analysis on the non-responder (Alda score <4) prediction. *Abbreviations*: Cagliari (Centro Bini; CB), Cagliari (University; CU), International Group for Study of Lithium (IGSLi), Maritimes (MAR), Montreal (MTL), Ontario (ON), Poznan (POZ), positive and negative predictive values (PPV, NPV), area under the receiver operating characteristic curve (AUC), Cohen's kappa (Kappa)

Site	**Accuracy**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**AUC**	**Kappa**
CU	0.84 (0.82, 0.86)	0.93 (0.9, 0.95)	0.72 (0.66, 0.77)	0.83 (0.8, 0.86)	0.88 (0.84, 0.91)	0.89 (0.87, 0.92)	0.66 (0.61, 0.71)
CB	0.85 (0.82, 0.88)	0.92 (0.9, 0.94)	0.77 (0.71, 0.83)	0.83 (0.8, 0.87)	0.89 (0.86, 0.92)	0.92 (0.9, 0.94)	0.7 (0.64, 0.75)
IGSLi	0.81 (0.79, 0.83)	0.91 (0.89, 0.93)	0.61 (0.54, 0.67)	0.83 (0.8, 0.85)	0.77 (0.73, 0.81)	0.85 (0.83, 0.88)	0.54 (0.49, 0.59)
MAR	0.84 (0.82, 0.86)	0.93 (0.9, 0.96)	0.73 (0.68, 0.79)	0.81 (0.79, 0.84)	0.91 (0.87, 0.94)	0.88 (0.86, 0.9)	0.68 (0.64, 0.71)
MTL	0.83 (0.81, 0.84)	0.93 (0.91, 0.96)	0.68 (0.63, 0.73)	0.81 (0.79, 0.83)	0.88 (0.84, 0.91)	0.88 (0.85, 0.9)	0.63 (0.6, 0.67)
ON	0.82 (0.8, 0.84)	0.93 (0.92, 0.94)	0.57 (0.5, 0.65)	0.83 (0.81, 0.85)	0.79 (0.76, 0.83)	0.85 (0.83, 0.88)	0.54 (0.48, 0.61)
POZ	0.84 (0.82, 0.85)	0.96 (0.95, 0.97)	0.62 (0.58, 0.66)	0.82 (0.8, 0.84)	0.9 (0.88, 0.92)	0.89 (0.86, 0.91)	0.62 (0.58, 0.66)

Table B.12: Results of classification on the pooled sample using models with only clinical course (CC), rapid cycling (RC), or clinical course and rapid cycling variables (CC+RC). Statistics are presented as means and 95% confidence intervals over the cross validation folds. Performance statistics include area under the receiver operating characteristic curve (AUC), accuracy (Acc), sensitivity (Sn), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV), and Cohen's kappa.

Variables	**AUC**	**Acc**	**Sn**	**Sp**	**PPV**	**NPV**	**Kappa**
ALL	0.8 [0.78, 0.82]	0.77 [0.75, 0.79]	0.53 [0.48, 0.57]	0.9 [0.88 0.92]	0.74 [0.69, 0.79]	0.78 [0.77, 0.80]	0.46 [0.4, 0.51]
CC+RC	0.74 [0.73,0.77]	0.68 [0.66,0.71]	0.68 [0.67,0.69]	0.70 [0.64,0.74]	0.54 [0.51,0.57]	0.80 [0.79,0.82]	0.35 [0.28,0.38]
CC	0.75 [0.72,0.77]	0.69 [0.66,0.74]	0.42 [0.37,0.67]	0.89 [0.62,0.94]	0.63 [0.51,0.76]	0.74 [0.73,0.78]	0.31 [0.26,0.36]
RC	0.58 [0.57,0.60]	0.55 [0.54,0.57]	0.65 [0.64,0.68]	0.50 [0.48,0.52]	0.41 [0.40,0.42]	0.74 [0.72,0.75]	0.14 [0.11,0.17]

Table B.13: Results of classification on each site's sample using models with only clinical course (CC), rapid cycling (RC), or clinical course and rapid cycling variables (CC+RC). We report the results within each site's sample, individually (denoted "SITE"), as well as in the predict-one-site-out analysis (denoted "POSO"). Statistics are presented as means and 95% confidence intervals over the cross validation folds. Sites include Centro Bini (CB; Cagliari), the University of Cagliari (CU), the Canadian Maritimes (MAR), Montreal (MTL), Ottawa & Hamilton Ontario (ON), and Poznan (POZ). Performance statistics include area under the receiver operating characteristic curve (AUC), accuracy (Acc), sensitivity (Sn), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV), and Cohen's kappa. Note that in the POSO results, the 95% confidence intervals are taken over samples from the uninformative distribution over missing data, and not over folds (since there is only one fold, effectively).

		AUC	**Acc**	**Sn**	**Sp**	**PPV**	**NPV**	**Kappa**
Clinical Course + Rapid Cycling								
CB	SITE	0.49 [0.44,0.53]	0.62 [0.59,0.64]	0.28 [0.20,0.33]	0.71 [0.67,0.75]	0.19 [0.14,0.21]	0.78 [0.77,0.80]	-0.02 [-0.09,0.03]
	POSO	0.43 [0.43,0.43]	0.44 [0.44,0.44]	0.43 [0.43,0.43]	0.44 [0.44,0.44]	0.17 [0.17,0.17]	0.75 [0.75,0.75]	-0.08 [-0.08,-0.08]
CU	SITE	0.55 [0.51,0.59]	0.60 [0.58,0.64]	0.38 [0.30,0.44]	0.69 [0.68,0.74]	0.33 [0.26,0.38]	0.74 [0.72,0.75]	0.07 [-0.01,0.13]
	POSO	0.46 [0.46,0.46]	0.68 [0.68,0.68]	0. [0.00,0.]	0.95 [0.95,0.95]	0.01 [0.01,0.01]	0.70 [0.70,0.70]	-0.06 [-0.06,-0.06]
IGSLi	SITE	-	-	-	-	-	-	-
	POSO	-	0.48 [0.48,0.48]	0.48 [0.48,0.48]	-	1. [1.00,1.]	0. [0.00,0.]	0. [0.00,0.]
MAR	SITE	0.74 [0.70,0.88]	0.74 [0.69,0.83]	0.72 [0.50,1.]	0.81 [0.69,0.87]	0.40 [0.29,0.57]	0.91 [0.85,1.]	0.40 [0.15,0.61]
	POSO	0.52 [0.52,0.52]	0.74 [0.74,0.74]	0.11 [0.11,0.11]	0.90 [0.90,0.90]	0.23 [0.23,0.23]	0.80 [0.80,0.80]	0.02 [0.02,0.02]
MTL	SITE	0.75 [0.71,0.75]	0.75 [0.68,0.75]	0.80 [0.60,0.90]	0.74 [0.64,0.78]	0.29 [0.29,0.33]	0.95 [0.92,0.98]	0.27 [0.23,0.32]
	POSO	0.50 [0.50,0.50]	0.65 [0.65,0.65]	0.13 [0.13,0.13]	0.75 [0.75,0.75]	0.09 [0.09,0.09]	0.82 [0.82,0.82]	-0.10 [-0.10,-0.10]
ON	SITE	0.52 [0.51,0.53]	0.72 [0.71,0.73]	0.83 [0.82,0.85]	0.13 [0.10,0.18]	0.84 [0.83,0.85]	0.17 [0.14,0.20]	-0. [-0.03,0.02]
	POSO	0.51 [0.51,0.51]	0.49 [0.49,0.49]	0.47 [0.47,0.47]	0.56 [0.56,0.56]	0.85 [0.85,0.85]	0.17 [0.17,0.17]	0.02 [0.02,0.02]

Continued on next page...

		AUC	Acc	Sn	Sp	PPV	NPV	Kappa
POZ	SITE	0.58 [0.45,0.72]	0.67 [0.56,0.75]	1. [0.80,1.]	0.50 [0.25,0.50]	0.62 [0.57,0.67]	0.97 [0.50,1.]	0.27 [0.05,0.50]
	POSO	0.46 [0.46,0.46]	0.47 [0.47,0.47]	0. [0.00,0.]	1. [1.00,1.]	0. [0.00,0.]	0.47 [0.47,0.47]	0. [0.00,0.]
Clinical Course Only								
CB	SITE	0.48 [0.35,0.63]	0.42 [0.41,0.56]	0.30 [0.25,0.75]	0.47 [0.40,0.53]	0.14 [0.09,0.21]	0.75 [0.70,0.79]	-0.06 [-0.16,0.01]
	POSO	0.49 [0.49,0.49]	0.41 [0.41,0.41]	0.63 [0.63,0.63]	0.35 [0.35,0.35]	0.20 [0.20,0.20]	0.79 [0.79,0.79]	-0.01 [-0.01,-0.01]
CU	SITE	0.64 [0.50,0.67]	0.57 [0.48,0.64]	0.75 [0.38,0.75]	0.56 [0.50,0.60]	0.38 [0.29,0.43]	0.83 [0.67,0.86]	0.19 [-0.05,0.29]
	POSO	0.51 [0.51,0.51]	0.49 [0.49,0.49]	0.42 [0.42,0.42]	0.52 [0.52,0.52]	0.26 [0.26,0.26]	0.68 [0.68,0.68]	-0.05 [-0.05,-0.05]
IGSLi	SITE	-	-	-	-	-	-	-
	POSO	-	0.97 [0.97,0.97]	0.97 [0.97,0.97]	-	1. [1.00,1.]	0. [0.00,0.]	0. [0.00,0.]
MAR	SITE	0.77 [0.64,0.88]	0.80 [0.70,0.85]	0.75 [0.58,0.97]	0.81 [0.75,0.88]	0.50 [0.38,0.60]	0.93 [0.88,0.99]	0.47 [0.32,0.57]
	POSO	0.49 [0.49,0.49]	0.22 [0.22,0.23]	0.99 [0.99,0.99]	0.03 [0.03,0.04]	0.20 [0.20,0.20]	0.90 [0.89,0.91]	0.01 [0.01,0.01]
MTL	SITE	0.70 [0.67,0.76]	0.74 [0.71,0.75]	0.80 [0.70,0.90]	0.74 [0.69,0.75]	0.33 [0.33,0.35]	0.95 [0.93,0.98]	0.35 [0.31,0.36]
	POSO	0.46 [0.46,0.46]	0.18 [0.18,0.18]	1. [1.00,1.]	0.03 [0.03,0.03]	0.16 [0.16,0.16]	1. [1.00,1.]	0.01 [0.01,0.01]
ON	SITE	0.46 [0.43,0.47]	0.72 [0.71,0.72]	0.83 [0.82,0.83]	0.12 [0.09,0.15]	0.83 [0.82,0.84]	0.11 [0.08,0.13]	-0.06 [-0.09,-0.03]
	POSO	0.50 [0.50,0.50]	0.85 [0.85,0.85]	1. [1.00,1.]	0.06 [0.06,0.06]	0.85 [0.85,0.85]	1. [1.00,1.]	0.09 [0.09,0.09]
POZ	SITE	0.48 [0.35,0.59]	0.54 [0.46,0.56]	0.70 [0.60,0.85]	0.25 [0.03,0.50]	0.52 [0.50,0.56]	0.50 [0.33,0.66]	0. [-0.13,0.10]
	POSO	0.47 [0.47,0.47]	0.53 [0.53,0.53]	1. [1.00,1.]	0. [0.00,0.]	0.53 [0.53,0.53]	-	0. [0.00,0.]
Rapid Cycling Only								
CB	SITE	0.51 [0.50,0.55]	0.64 [0.62,0.66]	0.30 [0.28,0.40]	0.73 [0.69,0.75]	0.24 [0.21,0.27]	0.80 [0.79,0.81]	0.03 [0.00,0.09]
	POSO	0.42 [0.41,0.42]	0.42 [0.42,0.43]	0.41 [0.40,0.42]	0.43 [0.42,0.43]	0.16 [0.16,0.16]	0.73 [0.73,0.73]	-0.10 [-0.11,-0.10]
CU	SITE	0.53 [0.48,0.54]	0.57 [0.54,0.60]	0.40 [0.33,0.44]	0.67 [0.59,0.71]	0.31 [0.28,0.36]	0.72 [0.70,0.74]	0.04 [-0.01,0.09]
	POSO	0.45 [0.44,0.45]	0.67 [0.67,0.67]	0.01 [0.00,0.02]	0.95 [0.94,0.95]	0.05 [0.00,0.11]	0.70 [0.70,0.70]	-0.06 [-0.07,-0.05]
IGSLi	SITE	-	-	-	-	-	-	-

Continued on next page...

		AUC	**Acc**	**Sn**	**Sp**	**PPV**	**NPV**	**Kappa**
	POSO	-	0.46 [0.39,0.47]	0.46 [0.39,0.47]	-	1. [1.00,1.]	0. [0.00,0.]	0. [0.00,0.]
MAR	SITE	0.60 [0.55,0.64]	0.43 [0.40,0.47]	0.85 [0.75,1.]	0.31 [0.25,0.38]	0.25 [0.21,0.27]	0.91 [0.86,1.]	0.12 [0.05,0.15]
	POSO	0.52 [0.51,0.52]	0.74 [0.73,0.74]	0.12 [0.10,0.13]	0.90 [0.89,0.90]	0.21 [0.19,0.24]	0.80 [0.80,0.80]	0.01 [-0.01,0.04]
MTL	SITE	0.65 [0.63,0.69]	0.47 [0.45,0.50]	1. [0.90,1.]	0.41 [0.37,0.43]	0.22 [0.21,0.24]	1. [0.96,1.]	0.14 [0.11,0.17]
	POSO	0.56 [0.56,0.56]	0.84 [0.84,0.84]	0. [0.00,0.]	1. [1.00,1.]	0. [0.00,0.]	0.84 [0.84,0.84]	0. [0.00,0.]
ON	SITE	0.52 [0.50,0.54]	0.48 [0.46,0.49]	0.46 [0.42,0.48]	0.57 [0.54,0.66]	0.84 [0.83,0.87]	0.17 [0.17,0.18]	0.02 [0.01,0.04]
	POSO	0.47 [0.44,0.50]	0.45 [0.44,0.48]	0.43 [0.41,0.47]	0.58 [0.54,0.62]	0.84 [0.83,0.86]	0.16 [0.16,0.18]	-0. [-0.01,0.04]
POZ	SITE	0.62 [0.50,0.75]	0.62 [0.56,0.75]	1. [1.00,1.]	0.25 [0.00,0.50]	0.57 [0.56,0.67]	1. [1.00,1.]	0.25 [0.00,0.50]
	POSO	0.50 [0.50,0.50]	0.47 [0.47,0.47]	0. [0.00,0.]	1. [1.00,1.]	0. [0.00,0.]	0.47 [0.47,0.47]	0. [0.00,0.]

Table B.14: Results of the leave-one-site-out (LOSO) analysis using models with only clinical course (CC), rapid cycling (RC), or clinical course and rapid cycling variables (CC+RC). Statistics are presented as means and 95% confidence intervals over the cross validation folds. Sites include Centro Bini (CB; Cagliari), the University of Cagliari (CU), the Canadian Maritimes (MAR), Montreal (MTL), Ottawa & Hamilton Ontario (ON), and Poznan (POZ). Performance statistics include area under the receiver operating characteristic curve (AUC), accuracy (Acc), sensitivity (Sn), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV), and Cohen's kappa.

	AUC	**Acc**	**Sn**	**Sp**	**PPV**	**NPV**	**Kappa**
Clinical Course + Rapid Cycling							
CC+RC (ALL SITES)	0.74 [0.73,0.77]	0.68 [0.66,0.71]	0.68 [0.67,0.69]	0.70 [0.64,0.74]	0.54 [0.51,0.57]	0.80 [0.79,0.82]	0.35 [0.28,0.38]
CB	0.71 [0.70,0.73]	0.72 [0.68,0.75]	0.47 [0.40,0.51]	0.89 [0.85,0.92]	0.74 [0.65,0.77]	0.72 [0.68,0.74]	0.37 [0.26,0.44]
CU	0.58 [0.53,0.59]	0.63 [0.61,0.66]	0.33 [0.32,0.39]	0.81 [0.75,0.83]	0.48 [0.44,0.54]	0.69 [0.67,0.70]	0.15 [0.09,0.19]
IGSLi	0.54 [0.52,0.55]	0.59 [0.58,0.65]	0.32 [0.19,0.38]	0.72 [0.68,0.83]	0.34 [0.32,0.36]	0.70 [0.70,0.71]	0.03 [0.02,0.06]
MAR	0.58 [0.56,0.61]	0.61 [0.59,0.65]	0.38 [0.34,0.41]	0.79 [0.76,0.81]	0.51 [0.49,0.59]	0.64 [0.64,0.67]	0.14 [0.12,0.23]
MTL	0.56 [0.53,0.59]	0.62 [0.56,0.63]	0.32 [0.27,0.41]	0.78 [0.67,0.83]	0.47 [0.39,0.48]	0.67 [0.66,0.67]	0.12 [0.05,0.13]
ON	0.53 [0.49,0.54]	0.64 [0.60,0.65]	0.26 [0.19,0.30]	0.79 [0.73,0.83]	0.34 [0.28,0.36]	0.71 [0.70,0.72]	0.04 [-0.02,0.06]
POZ	0.59 [0.56,0.60]	0.66 [0.65,0.67]	0.37 [0.33,0.42]	0.80 [0.78,0.81]	0.47 [0.44,0.49]	0.72 [0.71,0.74]	0.17 [0.13,0.21]
Clinical Course Only							
CC (ALL SITES)	0.75 [0.72,0.77]	0.69 [0.66,0.74]	0.42 [0.37,0.67]	0.89 [0.62,0.94]	0.63 [0.51,0.76]	0.74 [0.73,0.78]	0.31 [0.26,0.36]
CB	0.55 [0.51,0.57]	0.46 [0.44,0.52]	0.84 [0.46,0.94]	0.21 [0.10,0.54]	0.41 [0.40,0.41]	0.68 [0.64,0.77]	0.05 [0.01,0.05]
CU	0.52 [0.50,0.55]	0.49 [0.43,0.53]	0.66 [0.53,0.68]	0.40 [0.27,0.50]	0.34 [0.32,0.37]	0.66 [0.64,0.69]	0.01 [0.00,0.04]
IGSLi	0.51 [0.50,0.53]	0.43 [0.38,0.53]	0.61 [0.36,0.88]	0.33 [0.18,0.60]	0.30 [0.29,0.31]	0.68 [0.62,0.72]	-0.02 [-0.03,0.01]
MAR	0.53 [0.52,0.55]	0.45 [0.44,0.46]	0.96 [0.95,0.97]	0.11 [0.09,0.14]	0.42 [0.41,0.42]	0.75 [0.68,0.83]	0.05 [0.02,0.05]
MTL	0.53 [0.50,0.56]	0.43 [0.40,0.44]	0.92 [0.82,0.97]	0.14 [0.08,0.20]	0.38 [0.37,0.38]	0.78 [0.66,0.86]	0.04 [0.02,0.06]

Continued on next page...

	AUC	Acc	Sn	Sp	PPV	NPV	Kappa
ON	0.50 [0.49,0.51]	0.42 [0.36,0.44]	0.69 [0.62,0.84]	0.31 [0.14,0.36]	0.29 [0.29,0.30]	0.70 [0.65,0.76]	-0.01 [-0.02,0.01]
POZ	0.52 [0.48,0.53]	0.49 [0.39,0.54]	0.56 [0.36,0.79]	0.46 [0.23,0.66]	0.34 [0.30,0.35]	0.68 [0.64,0.71]	0.01 [-0.05,0.03]
Rapid Cycling Only							
RC (ALL SITES)	0.58 [0.57,0.60]	0.55 [0.54,0.57]	0.65 [0.64,0.68]	0.50 [0.48,0.52]	0.41 [0.40,0.42]	0.74 [0.72,0.75]	0.14 [0.11,0.17]
CB	0.69 [0.66,0.72]	0.71 [0.69,0.74]	0.47 [0.38,0.50]	0.88 [0.86,0.90]	0.70 [0.66,0.76]	0.71 [0.69,0.74]	0.34 [0.30,0.43]
CU	0.59 [0.57,0.60]	0.65 [0.63,0.65]	0.37 [0.33,0.39]	0.79 [0.78,0.83]	0.50 [0.47,0.52]	0.69 [0.68,0.70]	0.18 [0.13,0.19]
IGSLi	0.52 [0.50,0.54]	0.65 [0.64,0.67]	0.20 [0.19,0.20]	0.85 [0.83,0.88]	0.37 [0.35,0.42]	0.71 [0.70,0.71]	0.05 [0.04,0.08]
MAR	0.56 [0.55,0.58]	0.59 [0.59,0.62]	0.34 [0.33,0.35]	0.77 [0.75,0.78]	0.50 [0.48,0.55]	0.63 [0.63,0.65]	0.11 [0.10,0.16]
MTL	0.53 [0.51,0.55]	0.61 [0.60,0.64]	0.31 [0.29,0.33]	0.77 [0.76,0.83]	0.45 [0.44,0.51]	0.67 [0.66,0.68]	0.10 [0.08,0.14]
ON	0.51 [0.51,0.53]	0.66 [0.66,0.67]	0.17 [0.15,0.18]	0.87 [0.86,0.89]	0.35 [0.33,0.37]	0.71 [0.71,0.72]	0.04 [0.02,0.06]
POZ	0.58 [0.58,0.60]	0.66 [0.64,0.68]	0.37 [0.36,0.40]	0.80 [0.77,0.82]	0.47 [0.44,0.52]	0.72 [0.72,0.73]	0.17 [0.16,0.22]

In sum, the results of this supplementary analysis further highlight the between-site heterogeneity in feature importance. More specifically, while the performance of this restricted variable set (CC+RC, mainly) was relatively comparable to the aggregate analysis with all variables, the LOSO performance using only CC+RC (or one of them) was less robust and showed greater relative variability across sites (mean kappa range of 0.03-0.37). Conversely, in the main analysis with all variables the LOSO performance was more robust to leaving any one site out (mean kappa range 0.36-0.51). The POSO analyses are slightly less contributory, since only one site in the main analysis could be predicted to a non-trivial degree when left out (Maritimes); using CC/RC, however, the classification of the held out Maritimes data was trivial. This suggests that information about variables other than CC and RC are important for classification of the Maritimes data when it is held out of the aggregate sample.

B.6 Co-occurrence Tables

Tables B.15, B.16, and B.17 highlight the co-occurrence of clinical course, rapid cycling, and family history variables.

Table B.15: Co-occurrence of clinical course and family history variables. Values are presented as either count (percentage) or median [min, max]. *Abbreviations*: Chronic (C), chronic deteriorating (CD), chronic fluctuating (CF), completely episodic (E), continuous cycling (CC), episodic with residual symptoms (ER), single episode (S), first degree relative (1DR), second degree relative (2DR), bipolar disorder (BD), unipolar depression (MDD), schizoaffective disorder (SZA), schizophrenia (SCZ). Hypothesis test done with the Kruskal-Wallis test.

	C	**CD**	**CF**	**E**	**CC**	**ER**	**S**	**p**
n	99	18	269	407	40	228	14	
1DR Mood D/O (%)	28 (66.7)	2 (50.0)	117 (58.8)	164 (62.8)	19 (59.4)	64 (59.3)	4 (40.0)	0.74
1DR Bipolar (%)	42 (46.7)	4 (36.4)	82 (32.2)	108 (37.9)	11 (33.3)	39 (28.3)	4 (33.3)	0.12
N 1DR BD1	0 [0,3]	0 [0,1]	0 [0,3]	0 [0,6]	0 [0,1]	0 [0,4]	0 [0,3]	0.002
N 1DR BD2	0 [0,1]	0 [0,0]	0 [0,0]	0 [0,2]	0 [0,0]	0 [0,3]	0.50 [0,1]	<0.001
N 1DR MDD	0 [0,5]	0 [0,1]	0 [0,5]	0 [0,7]	0 [0,0]	0 [0,6]	0 [0,4]	0.033
N 1DR SZA	0 [0,0]	0 [0,0]	0 [0,1]	0 [0,1]	0 [0,0]	0 [0,2]	0 [0,0]	0.072
N 1DR SCZ	0 [0,1]	0 [0,1]	0 [0,2]	0 [0,1]	0 [0,0]	0 [0,2]	0 [0,1]	0.681
N 1DR Anxiety	0 [0,1]	0 [0,0]	0 [0,3]	0 [0,7]	0 [0,0]	0 [0,9]	0 [0,0]	0.552
N 1DR Unaffected	0 [0,11]	3 [0,11]	0 [0,4]	1 [0,13]	0 [0,0]	1.50 [0,14]	0 [0,5]	<0.001
N 1DR Completed suicide	0 [0,1]	0 [0,0]	0 [0,1]	0 [0,2]	0 [0,2]	0 [0,2]	0 [0,1]	0.539
N 1DR Attempted suicide	0 [0,3]	0 [0,0]	0 [0,3]	0 [0,2]	0 [0,1]	0 [0,2]	0 [0,1]	0.001
N 2DR Completed suicide	0 [0,1]	0 [0,1]	0 [0,1]	0 [0,2]	0 [0,0]	0 [0,2]	0 [0,2]	0.045
N 2DR Attempted suicide	0 [0,1]	0 [0,2]	0 [0,1]	0 [0,1]	0 [0,0]	0 [0,2]	0 [0,0]	0.517
Total N 1DR	0 [0,20]	3 [0,6]	0 [0,17]	1 [0,20]	0 [0,4]	5 [0,14]	5 [0,15]	<0.001

Table B.16: Co-occurrence of rapid cycling variable and clinical course. Only complete cases of the co-occurrences were analyzed here. Chi2 test used for hypothesis testing. *Abbreviations*: antidepressants (AD), chronic (C), chronic deteriorating (CD), chronic fluctuating (CF), completely episodic (E), continuous cycling (CC), episodic with residual symptoms (ER), single episode (S)

	Rapid Cycling			
	Never	**Only on AD**	**Spontaneous**	**p**
N	495	18	118	
Clinical Course (%)				<0.001
C	62 (15.6)	4 (22.2)	6 (5.1)	
CD	10 (2.5)	0 (0.0)	1 (0.9)	
CF	80 (20.1)	8 (44.4)	71 (60.7)	
E	135 (33.9)	0 (0.0)	9 (7.7)	
CC	1 (0.3)	0 (0.0)	1 (0.9)	
ER	98 (24.6)	6 (33.3)	29 (24.8)	
S	12 (3.0)	0 (0.0)	0 (0.0)	

Table B.17: Co-occurrence of clinical course and rapid cycling. *Abbreviations*: Chronic (C), chronic deteriorating (CD), chronic fluctuating (CF), completely episodic (E), continuous cycling (CC), episodic with residual symptoms (ER), single episode (S). Hypothesis testing done with Chi2 test.

	Clinical Course							
	C	**CD**	**CF**	**E**	**CC**	**ER**	**S**	**p**
n	99	18	269	407	40	228	14	
Rapid Cycling (%)								<0.001
Never	62 (86.1)	10 (90.9)	80 (50.3)	135 (93.8)	1 (50.0)	98 (73.7)	12 (100.0)	
Only on Antidepressants	4 (5.6)	0 (0.0)	8 (5.0)	0 (0.0)	0 (0.0)	6 (4.5)	0 (0.0)	
Spontaneous	6 (8.3)	1 (9.1)	71 (44.7)	9 (6.2)	1 (50.0)	29 (21.8)	0 (0.0)	

Appendix C

Supplementary Material for *Exemplar Scoring Identifies Genetically Separable Phenotypes of Lithium Responsive Bipolar Disorder*

C.1 Gene Set Analysis

At each fold of cross-validation (under all settings of q), the logistic regression coefficients were saved. The SNPs whose logistic regression coefficients were of the same sign (i.e. positive or negative) across all folds were ranked in terms of their absolute median coefficient values and linked to gene identifiers using the NCBI gene database. Each gene was assigned the maximal absolute value of the logistic regression coefficients for all SNPs tagged by that gene; the remainder (duplicates) were deleted, such that each included gene had only one numerical value associated with it. We then applied the statistical enrichment test in the PANTHER classification system v. 14.1 [232]. We repeated the statistical enrichment test for the following annotation sets: PANTHER pathways, GO molecular function (complete), GO biological processes (complete), GO cellular components (complete). To further evaluate the degree to which the enrichment analyses speak specifically to findings among the best exemplars, we repeated the same procedures outlined here using the logistic regression coefficients for the poor exemplars.

C.2 Population Stratification

To evaluate for the presence of population stratification in our genomic sample, we plot the first several principal components of the subjects' genotypes in Figure C.1. For comparison, Figure C.2 demonstrates the first several principal components from 14 sites of the full Consortium on Lithium Genetics (ConLiGen) genomic sample.

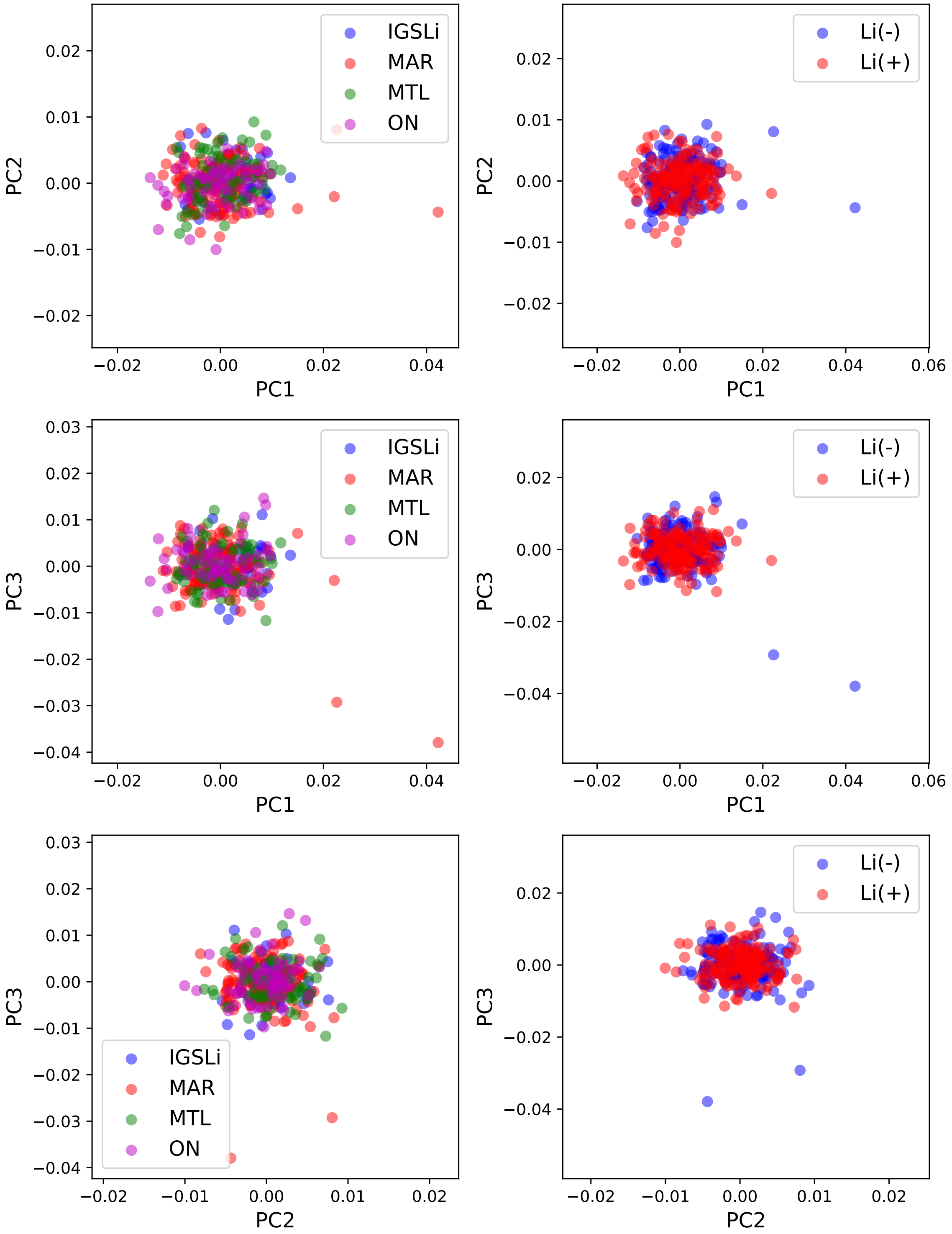


Figure C.1: Principal components analysis of the genomic dataset from Halifax (as coded in the ConLiGen studies [30]). The left column is coloured by the site of origin, whereas the right column of plots is coloured by lithium responsiveness. *Abbreviations*: International Group for the Study of Lithium (IGSLi), Maritimes (MAR), Montreal (MTL), Ontario (ON; also known as Ottawa/Hamilton).

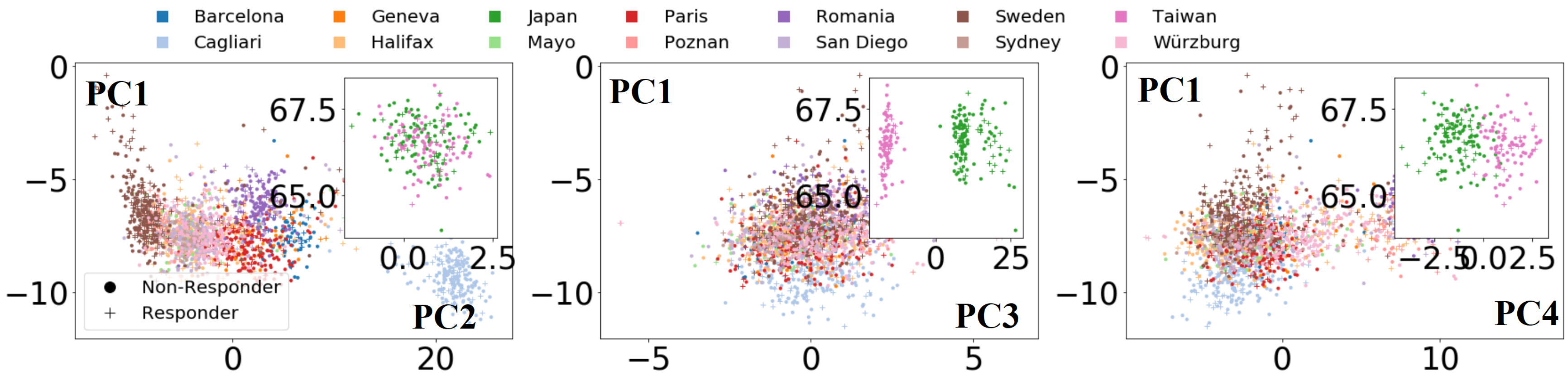


Figure C.2: Principal components analysis of the genomic dataset from the Consortium on Lithium Genetics sample [30] (Figure used with permission from Stone et al., submitted manuscript)

C.3 Supplementary Tables

Clinical demographic comparisons between the best exemplars, poor exemplars, and the aggregated sample of genotyped patients is presented in Table C.1, with stratification by lithium response. The results of gene enrichment analysis are presented in Table C.3, with specific genes enriched in the best exemplar group (related to glutamate receptors and signalling processes) shown in Tables C.4 and C.5.

Table C.1: Demographic comparisons for subjects whose genomic data (from the Consortium for Lithium Genetics; ConLiGen) overlapped with our clinical dataset. Comparisons were done in between lithium responders (LR(+)) and non-responders (LR(-)) for the total group (“ALL), the best exemplars (“Best; exemplar score $\geq$ 75th percentile), and the poorest exemplars (“Poor; exemplar score $\leq$ 25th percentile).

	ALL			**Poor**			**Best**		
	LR(-)	**LR(+)**	**p**	**LR(-)**	**LR(+)**	**p**	**LR(-)**	**LR(+)**	**p**
n	162	159		41	40		40	39	
Centre (%)			<0.001			<0.001			<0.001
IGSLi	0 (0.0)	56 (35.2)		0 (0.0)	8 (20.0)		0 (0.0)	33 (84.6)	
Maritimes	92 (56.8)	37 (23.3)		22 (53.7)	10 (25.0)		23 (57.5)	0 (0.0)	
Montreal	62 (38.3)	12 (7.5)		14 (34.1)	3 (7.5)		17 (42.5)	0 (0.0)	
Ontario	8 (4.9)	54 (34.0)		5 (12.2)	19 (47.5)		0 (0.0)	6 (15.4)	
GWAS Wave 2 (%)	93 (57.4)	20 (12.6)	<0.001	22 (53.7)	5 (12.5)	<0.001	27 (67.5)	0 (0.0)	<0.001
Male (%)	66 (40.7)	70 (44.0)	0.629	19 (46.3)	18 (45.0)	1	16 (40.0)	16 (41.0)	1
Age (y)	48.53 [21.59, 82.51]	50.94 [21.66, 80.16]	0.009	49.43 (14.47)	52.63 (13.83)	0.312	42.50 (12.81)	59.10 (11.07)	<0.001
Diagnosis (%)			0.154			0.015			0.11
BD I	108 (66.7)	112 (70.4)		34 (82.9)	20 (50.0)		24 (60.0)	26 (66.7)	
BD II	52 (32.1)	41 (25.8)		7 (17.1)	18 (45.0)		16 (40.0)	9 (23.1)	
MDD Recurrent	0 (0.0)	4 (2.5)		0 (0.0)	1 (2.5)		0 (0.0)	3 (7.7)	
MDD Single	0 (0.0)	1 (0.6)		0 (0.0)	1 (2.5)		0 (0.0)	0 (0.0)	
SZA	2 (1.2)	1 (0.6)		0 (0.0)	0 (0.0)		0 (0.0)	1 (2.6)	
Age of Onset (y)	22. [7., 64.]	26. [13., 63.]	<0.001	25. [12., 54.]	25. [14., 48.]	0.429	17. [7., 30.]	28. [17., 63.]	<0.001
Onset Dep. (y)	24. [12., 67.]	29. [14., 63.]	<0.001	30.29 (10.08)	27.62 (8.60)	0.243	20. [12., 35.]	32. [19., 63.]	<0.001
Onset M (y)	29. [15., 59.]	30. [13., 66.]	0.292	30. [17., 56.]	32. [18., 66.]	0.193	27.50 [15., 47.]	32. [17., 52.]	0.039
Onset HypoM (y)	30. [0., 67.]	36.50 [16., 63.]	0.254	31. (13.39)	35.74 (11.36)	0.253	24. [12., 62.]	38. [20., 63.]	0.054
Polarity 1st Ep. (%)			0.006			0.118			0.073
Biphasic (D-M)	2 (1.2)	9 (5.9)		1 (2.6)	3 (8.3)		1 (2.5)	3 (7.9)	
Biphasic (M-D)	10 (6.2)	8 (5.3)		3 (7.7)	2 (5.6)		3 (7.5)	1 (2.6)	
Hypomania	18 (11.2)	13 (8.6)		6 (15.4)	2 (5.6)		5 (12.5)	2 (5.3)	

Continued on next page...

	ALL			Poor			Best		
	LR(-)	**LR(+)**	**p**	**LR(-)**	**LR(+)**	**p**	**LR(-)**	**LR(+)**	**p**
Major dep.	99 (61.9)	68 (44.7)		14 (35.9)	21 (58.3)		28 (70.0)	21 (55.3)	
Mania	20 (12.5)	36 (23.7)		12 (30.8)	4 (11.1)		1 (2.5)	7 (18.4)	
Minor dep.	7 (4.4)	16 (10.5)		2 (5.1)	4 (11.1)		1 (2.5)	4 (10.5)	
Mixed	2 (1.2)	1 (0.7)		1 (2.6)	0 (0.0)		1 (2.5)	0 (0.0)	
Periodic rapid cyc.	2 (1.2)	1 (0.7)		0 (0.0)	0 (0.0)		0 (0.0)	0 (0.0)	
Clinical Course (%)			<0.001			0.01			NaN
Chronic	7 (4.5)	0 (0.0)					1 (2.5)	0 (0.0)	
Chronic fluctuating	49 (31.6)	6 (12.2)		3 (8.1)	4 (30.8)		19 (47.5)	0 (0.0)	
Completely episodic	41 (26.5)	35 (71.4)		29 (78.4)	4 (30.8)		0 (0.0)	0 (0.0)	
Episodic + Residual	56 (36.1)	7 (14.3)		4 (10.8)	5 (38.5)		20 (50.0)	0 (0.0)	
Single episode	2 (1.3)	1 (2.0)		1 (2.7)	0 (0.0)		0 (0.0)	0 (0.0)	
N LT Manias	2. [0., 99.]	2. [0., 34.]	0.052	2. [0., 11.]	1. [0., 25.]	0.041	2. [0., 99.]	1. [0., 8.]	0.103
N LT Dep.	4. [0., 99.]	3. [0., 35.]	0.001	3. [0., 27.]	3. [0., 25.]	0.277	7. [0., 99.]	3. [0., 15.]	<0.001
N LT Mixed	0. [0., 99.]	0. [0., 3.]	<0.001	0. [0., 1.]	0. [0., 2.]	0.403	0. [0., 99.]	0. [0., 0.]	<0.001
N LT Multiphasic	0. [0., 99.]	0. [0., 13.]	0.454	0. [0., 1.]	0. [0., 9.]	0.109	1. [0., 99.]	0. [0., 13.]	0.203
Total LT Episodes	8. [1., 99.]	6. [0., 99.]	0.007	6. [1., 33.]	6. [0., 50.]	0.899	11.50 [3., 99.]	6.50 [2., 27.]	0.002
Rapid Cycling (%)			<0.001			0.798			<0.001
Never	92 (60.1)	104 (94.5)		0 (0.0)	0 (0.0)		12 (30.0)	25 (100.0)	
Only on AD	6 (3.9)	0 (0.0)		0 (0.0)	0 (0.0)		2 (5.0)	0 (0.0)	
Spontaneous	55 (35.9)	6 (5.5)		1 (2.8)	2 (7.4)		26 (65.0)	0 (0.0)	
Rapid mood switch	46 (54.1)	4 (21.1)	0.019	6 (30.0)	1 (20.0)	1	14 (63.6)	0 (0.0)	-
LT Psychosis (%)			0.888			0.469			0.659
Episodic congruent	57 (37.7)	27 (42.9)		14 (37.8)	9 (50.0)		18 (45.0)	0 (0.0)	
Episodic incong.	21 (13.9)	7 (11.1)		6 (16.2)	1 (5.6)		6 (15.0)	0 (0.0)	
Never	70 (46.4)	28 (44.4)		17 (45.9)	8 (44.4)		15 (37.5)	1 (100.0)	
Out of episodes	3 (2.0)	1 (1.6)					1 (2.5)	0 (0.0)	
GAF last Ax	70. [35., 95.]	90. [0., 100.]	<0.001	80. [50., 95.]	90. [40., 95.]	0.006	70. [40., 90.]	90. [90., 95.]	<0.001
Total ALDA Score	2. [0., 6.]	8. [7., 10.]	<0.001	4. [0., 6.]	8. [7., 10.]	<0.001	2. [0., 6.]	8. [8., 10.]	<0.001
N Episodes on Li	3. [0., 99.]	0. [0., 5.]	<0.001	2.50 [0., 99.]	0. [0., 2.]	0.002	3.50 [1., 99.]	-	-
N Episodes Pre-Li	4. [1., 99.]	4.50 [2., 99.]	0.373	3. [1., 99.]	5. [2., 99.]	0.078	8.50 [1., 99.]	-	-

Continued on next page...

	ALL			Poor			Best		
	LR(-)	LR(+)	p	LR(-)	LR(+)	p	LR(-)	LR(+)	p
N SA	0. [0., 6.]	0. [0., 3.]	0.119	0. [0., 2.]	0. [0., 2.]	0.235	1. [0., 6.]	0. [0., 3.]	0.022
N serious SA	1. [0., 6.]	0.50 [0., 2.]	0.044	1. (0.82)	1. (1.00)	1	1. [0., 3.]	0. [0., 0.]	0.177
Age First SA (y)	27.50 [12., 64.]	30.50 [16., 55.]	0.308	39.75 (12.69)	34. (10.39)	0.552	24.30 (9.60)	-	-
N FDR Mood d/o	75 (57.7)	61 (41.2)	0.009	18 (60.0)	11 (30.6)	0.031	24 (70.6)	16 (42.1)	0.028
FDR BD (%)	56 (34.6)	41 (25.9)	0.12	20 (48.8)	14 (35.9)	0.348	10 (25.0)	0 (0.0)	0.003
N FDR BD-I	0. [0., 4.]	0. [0., 5.]	0.13	0. [0., 4.]	0. [0., 5.]	0.3	0. [0., 2.]	0. [0., 0.]	0.001
N FDR MDD	1. [0., 7.]	0. [0., 5.]	0.01	0. [0., 7.]	0. [0., 3.]	0.158	1. [0., 3.]	0. [0., 5.]	0.019
N FDR SZA	0. [0., 1.]	0. [0., 1.]	0.678	0. [0., 1.]	0. [0., 0.]	0.165	0. [0., 1.]	0. [0., 1.]	0.986
N FDR SCZ	0. [0., 2.]	0. [0., 1.]	0.01	0. [0., 2.]	0. [0., 1.]	0.216	0. [0., 2.]	0. [0., 0.]	0.127
N FDR Ans d/o	0. [0., 3.]	0. [0., 3.]	0.044	0. [0., 2.]	0. [0., 1.]	0.129	0. [0., 2.]	0. [0., 1.]	0.006
N FDR Unaff.	0. [0., 5.]	0. [0., 2.]	<0.001	0. [0., 3.]	0. [0., 1.]	0.014	0. [0., 4.]	0. [0., 1.]	<0.001
N FDR Suicide	0. [0., 2.]	0. [0., 2.]	0.801	0. [0., 1.]	0. [0., 2.]	0.384	0. [0., 1.]	0. [0., 0.]	0.779
N FDR SA	0. [0., 2.]	0. [0., 2.]	0.193	0. [0., 2.]	0. [0., 2.]	0.183	0. [0., 2.]	0. [0., 0.]	0.743
N SDR Suicide	0. [0., 1.]	0. [0., 2.]	0.738	0. [0., 1.]	0. [0., 0.]	0.232	0. [0., 1.]	0. [0., 0.]	0.819
N SDR SA	0. [0., 1.]	0. [0., 1.]	0.387	0. [0., 1.]	0. [0., 0.]	0.499	0. [0., 1.]	0. [0., 0.]	0.819
LT Hx SI	73 (54.5)	27 (38.0)	0.036	8 (29.6)	6 (37.5)	0.845	32 (82.1)	7 (43.8)	0.012
SI episode related (%)			0.284			-			1
No	1 (1.4)	1 (6.7)		0 (0.0)	0 (0.0)		0 (0.0)	0 (0.0)	
Sometimes	5 (7.1)	0 (0.0)		0 (0.0)	0 (0.0)		0 (0.0)	0 (0.0)	
Yes	64 (91.4)	14 (93.3)		8 (100.0)	5 (100.0)		26 (89.7)	1 (100.0)	
Social Anx. d/o (%)	28 (18.3)	6 (12.8)	0.508	2 (5.7)	3 (25.0)	0.184	8 (20.0)	0 (0.0)	1
Panic d/o (%)	32 (20.6)	5 (4.3)	<0.001	1 (2.8)	2 (7.7)	0.772	12 (30.0)	0 (0.0)	0.001
GAD (%)	37 (24.2)	3 (6.4)	0.014	4 (11.4)	1 (8.3)	1	14 (35.0)	0 (0.0)	1
OCD (%)	13 (8.4)	1 (0.8)	0.011	1 (2.8)	1 (3.8)	1	6 (15.0)	0 (0.0)	0.039
SUD (%)	43 (27.7)	20 (16.4)	0.036	6 (16.7)	7 (25.0)	0.611	14 (35.0)	1 (2.6)	0.001
ADHD (%)	12 (7.8)	1 (2.2)	0.31	5 (13.9)	1 (7.7)	0.928	3 (7.5)	0 (0.0)	-
LD (%)	7 (4.6)	1 (2.2)	0.765	2 (5.7)	0 (0.0)	0.946	2 (5.0)	0 (0.0)	-
Insom (%)	18 (11.7)	3 (6.7)	0.491	2 (5.6)	0 (0.0)	0.96	6 (15.0)	0 (0.0)	-
PD (%)	34 (22.2)	4 (8.7)	0.067	2 (5.6)	0 (0.0)	0.96	13 (33.3)	0 (0.0)	-
Diabetes (%)	22 (14.7)	4 (9.3)	0.512	4 (11.4)	0 (0.0)	0.575	6 (15.4)	0 (0.0)	-

Continued on next page...

	ALL			Poor			Best		
	LR(-)	**LR(+)**	**p**	**LR(-)**	**LR(+)**	**p**	**LR(-)**	**LR(+)**	**p**
HTN (%)	25 (16.9)	6 (14.3)	0.867	5 (14.7)	2 (20.0)	1	5 (13.2)	0 (0.0)	-
Menstrual abn (%)	22 (28.9)	8 (42.1)	0.408	1 (7.1)	3 (75.0)	0.028	4 (17.4)	0 (0.0)	-
Thyroid d/o (%)	51 (34.5)	16 (37.2)	0.88	13 (37.1)	4 (36.4)	1	13 (33.3)	0 (0.0)	-
TBI (%)	27 (20.9)	7 (20.0)	1	6 (21.4)	4 (44.4)	0.357	8 (22.2)	0 (0.0)	-
Migraine (%)	41 (29.1)	4 (9.1)	0.013	8 (25.0)	1 (8.3)	0.423	8 (21.1)	0 (0.0)	-
SES (%)			0.181			0.113			-
Work full-time	27 (19.3)	11 (23.4)		14 (42.4)	3 (23.1)		5 (12.8)	0 (0.0)	
Work part-time	12 (8.6)	7 (14.9)		1 (3.0)	2 (15.4)		4 (10.3)	0 (0.0)	
Unemployment ins	20 (14.3)	4 (8.5)		6 (18.2)	1 (7.7)		4 (10.3)	0 (0.0)	
Social assist.	19 (13.6)	6 (12.8)		0 (0.0)	2 (15.4)		9 (23.1)	0 (0.0)	
Disabled	34 (24.3)	7 (14.9)		5 (15.2)	2 (15.4)		9 (23.1)	0 (0.0)	
Other	3 (2.1)	4 (8.5)		0 (0.0)	0 (0.0)		1 (2.6)	0 (0.0)	
Retired	19 (13.6)	8 (17.0)		7 (21.2)	3 (23.1)		2 (5.1)	0 (0.0)	
Student	6 (4.3)	0 (0.0)		0 (0.0)	0 (0.0)		5 (12.8)	0 (0.0)	
Marital status (%)			0.547			0.444			-
Single	34 (23.3)	12 (23.1)		3 (9.4)	4 (25.0)		17 (42.5)	0 (0.0)	
Married	76 (52.1)	32 (61.5)		19 (59.4)	9 (56.2)		13 (32.5)	0 (0.0)	
Divorced	32 (21.9)	7 (13.5)		9 (28.1)	3 (18.8)		9 (22.5)	0 (0.0)	
Widowed	4 (2.7)	1 (1.9)		1 (3.1)	0 (0.0)		1 (2.5)	0 (0.0)	

Table C.2: Results of classifying lithium response based on the genomic data of all subjects (ALL; n=321), the poor exemplars (¡25th percentile of exemplar score; n=81), and the best exemplars (¿75th percentile of exemplar score; n=79). Each panel shows the results for a different classification performance metric. Classification was done using logistic regression with an L2 penalty (regularization weight set to C=1 a priori) with stratification done over each value of the resolution parameter q=1 and q=2. *Abbreviations*: accuracy (Acc), area under the receiver operating characteristic curve (AUC), sensitivity (Sens), specificity (Spec), Cohen's kappa (Kappa), Matthews' correlation coefficient (MCC), positive predictive value (PPV), negative predictive value (NPV). Results are presented as means and 95% confidence intervals.

q	**Group**	**Acc**	**AUC**	**Sens**	**Spec**	**PPV**	**NPV**	**F1**	**Kappa**	**MCC**
1	Best	0.75 [0.66,0.87]	0.88 [0.83,0.98]	0.75 [0.50,0.94]	0.88 [0.75,1.]	0.90 [0.75,1.]	0.71 [0.67,0.95]	0.71 [0.67,0.86]	0.50 [0.31,0.74]	0.58 [0.41,0.77]
	Poor	0.65 [0.53,0.75]	0.66 [0.61,0.80]	0.50 [0.31,0.75]	0.75 [0.75,0.79]	0.67 [0.53,0.75]	0.67 [0.53,0.73]	0.62 [0.39,0.73]	0.28 [0.06,0.50]	0.29 [0.06,0.50]
2	Best	0.75 [0.65,0.75]	0.81 [0.66,0.86]	0.75 [0.54,0.75]	0.75 [0.56,0.94]	0.75 [0.67,0.95]	0.75 [0.67,0.79]	0.71 [0.67,0.79]	0.50 [0.29,0.50]	0.50 [0.39,0.58]
	Poor	0.50 [0.50,0.72]	0.53 [0.45,0.72]	0.50 [0.06,0.50]	0.75 [0.50,1.]	0.50 [0.12,0.90]	0.50 [0.50,0.67]	0.50 [0.08,0.67]	0. [0.00,0.44]	0. [0.00,0.50]
	ALL	0.66 [0.60,0.70]	0.70 [0.62,0.75]	0.59 [0.48,0.62]	0.70 [0.59,0.83]	0.67 [0.59,0.78]	0.65 [0.61,0.67]	0.64 [0.58,0.67]	0.31 [0.20,0.39]	0.32 [0.20,0.44]

Table C.3: Results of gene enrichment analysis using the PANTHER gene ontology system. Analyses are presented for (A) pathways, (B) gene ontology cellular components, and (C) gene ontology biological processes. *Abbreviations*: false discovery rate (FDR).

	Best				Poor			
	N	**+/-**	**P value**	**FDR**	**N**	**+/-**	**P value**	**FDR**
Pathways								
Muscarinic acetylcholine receptor 1 and 3 signaling pathway	27	+	0.00011	0.017				
Alzheimer disease-amyloid secretase pathway	30	+	0.00045	0.034				
Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha mediated pathway	53	+	0.00081	0.041				
Histamine H1 receptor mediated signaling pathway	27	+	0.00103	0.039				
Cellular component								
glutamatergic synapse (GO:0098978)	159	+	6.02E-08	3.05E-05				
synapse (GO:0045202)	468	+	3.90E-09	5.93E-06				
neuron projection (GO:0043005)	489	+	6.68E-05	7.25E-03				
neuron part (GO:0097458)	657	+	7.12E-07	1.08E-04				
cation channel complex (GO:0034703)					109	+	6.01E-05	7.10E-03
ion channel complex (GO:0034702)					139	+	1.09E-04	9.83E-03
transmembrane transporter complex (GO:1902495)					145	+	2.54E-04	1.70E-02
transporter complex (GO:1990351)					146	+	3.88E-04	2.48E-02
ionotropic glutamate receptor complex (GO:0008328)	28	+	1.20E-04	1.07E-02	27	+	1.32E-04	1.07E-02
plasma membrane part (GO:0044459)	1041	+	7.13E-04	4.52E-02				
neurotransmitter receptor complex (GO:0098878)	28	+	1.20E-04	1.14E-02	27	+	1.32E-04	1.13E-02
integral component of postsynaptic density membrane (GO:0099061)	35	+	3.16E-04	2.40E-02	33	+	6.62E-05	7.26E-03
intrinsic component of postsynaptic density membrane (GO:0099146)	36	+	1.95E-04	1.64E-02	34	+	1.37E-04	1.05E-02
intrinsic component of postsynaptic specialization membrane (GO:0098948)					39	+	5.09E-05	6.51E-03
intrinsic component of postsynaptic membrane (GO:0098936)	66	+	4.85E-04	3.35E-02	62	+	3.72E-05	5.71E-03
intrinsic component of synaptic membrane (GO:0099240)	91	+	6.09E-05	7.72E-03	85	+	1.24E-05	3.18E-03
synapse part (GO:0044456)	368	+	5.20E-07	9.87E-05	374	+	3.33E-05	5.68E-03
synapse (GO:0045202)					469	+	7.61E-07	3.89E-04

Continued on next page...

	Best				Poor			
	N	+/-	P value	FDR	N	+/-	P value	FDR
synaptic membrane (GO:0097060)	202	+	2.23E-08	1.70E-05	203	+	3.30E-07	2.53E-04
postsynaptic membrane (GO:0045211)	150	+	7.48E-08	2.84E-05	154	+	2.33E-07	3.58E-04
postsynapse (GO:0098794)	243	+	1.50E-07	4.56E-05	250	+	4.97E-04	2.93E-02
postsynaptic specialization membrane (GO:0099634)	54	+	4.22E-04	3.05E-02	50	+	3.99E-05	5.57E-03
postsynaptic specialization (GO:0099572)	146	+	1.59E-06	2.20E-04				
neuron part (GO:0097458)					662	+	1.38E-06	5.28E-04
postsynaptic density membrane (GO:0098839)	45	+	6.17E-05	7.22E-03	44	+	1.05E-04	1.00E-02
postsynaptic density (GO:0014069)	140	+	5.32E-07	8.99E-05				
asymmetric synapse (GO:0032279)	142	+	1.77E-07	4.49E-05				
neuron to neuron synapse (GO:0098984)	149	+	2.67E-07	5.81E-05				
integral component of postsynaptic specialization membrane (GO:0099060)					38	+	2.49E-05	4.78E-03
integral component of postsynaptic membrane (GO:0099055)	63	+	2.76E-04	2.21E-02	59	+	1.32E-05	2.89E-03
integral component of synaptic membrane (GO:0099699)	84	+	1.05E-04	1.07E-02	78	+	1.01E-05	3.10E-03
cell junction (GO:0030054)	478	+	6.31E-04	4.17E-02	478	+	9.19E-05	9.40E-03
neuron projection (GO:0043005)					501	+	1.73E-04	1.26E-02
presynaptic membrane (GO:0042734)					80	+	3.97E-04	2.44E-02
presynapse (GO:0098793)					196	+	2.46E-04	1.72E-02
Biological process								
regulation of cell morphogenesis involved in differentiation (GO:0010769)	116	+	1.09E-05	2.57E-02				
regulation of cell morphogenesis (GO:0022604)	189	+	2.31E-05	2.28E-02				
synapse organization (GO:0050808)	114	+	1.70E-05	2.87E-02				
axon guidance (GO:0007411)	121	+	1.75E-05	2.58E-02				
cell development (GO:0048468)	595	+	2.96E-05	2.50E-02				
cell differentiation (GO:0030154)	1174	+	8.13E-05	4.80E-02				
developmental process (GO:0032502)	1819	+	6.38E-05	4.19E-02				
anatomical structure development (GO:0048856)	1742	+	4.46E-05	3.51E-02				
generation of neurons (GO:0048699)	551	+	5.13E-05	3.79E-02				
neurogenesis (GO:0022008)	583	+	1.86E-05	2.19E-02				

Continued on next page...

	Best				Poor			
	N	**+/-**	**P value**	**FDR**	**N**	**+/-**	**P value**	**FDR**
nervous system development (GO:0007399)	819	+	6.40E-06	3.78E-02				
system development (GO:0048731)	1462	+	3.69E-06	4.35E-02				
multicellular organism development (GO:0007275)	1631	+	6.43E-05	3.99E-02				
neuron projection guidance (GO:0097485)	123	+	1.93E-05	2.07E-02				
regulation of neuron projection development (GO:0010975)	190	+	1.79E-05	2.35E-02				
regulation of neuron differentiation (GO:0045664)	243	+	9.89E-06	3.89E-02				
regulation of plasma membrane bounded cell projection organization (GO:0120035)	249	+	1.02E-05	3.00E-02				
regulation of cell projection organization (GO:0031344)	250	+	1.40E-05	2.76E-02				
glutamate receptor signaling pathway (GO:0007215)	30	+	2.49E-05	2.26E-02				
circulatory system development (GO:0072359)	314	+	5.58E-05	3.88E-02				
modulation of chemical synaptic transmission (GO:0050804)					182	+	4.51E-06	5.32E-02
regulation of trans-synaptic signaling (GO:0099177)					182	+	4.51E-06	2.66E-02
cell-cell adhesion via plasma-membrane adhesion molecules (GO:0098742)					99	+	9.19E-06	3.62E-02

Table C.4: Genes enriched in the best exemplars group related to glutamatergic synapses (gene ontology "cellular component" category).

Gene	Gene Symbol	Protein Class
ABR	Active breakpoint cluster region-related protein	guanyl-nucleotide exchange factor (PC00113)
ACAN	Aggrecan core protein	extracellular matrix glycoprotein (PC00100)
ACTN1, ACTN2	Alpha-actinin-1 & 2	
ADAM22, ADAM23	Disintegrin and metalloproteinase domain-containing protein 22 & 23	metalloprotease (PC00153)
ADCY1, ADCY8	Adenylate cyclase type 1 & 8	
ADGRL3	Adhesion G protein-coupled receptor L3	G-protein coupled receptor (PC00021), antibacterial response protein (PC00051), protease (PC00190)
ADORA2B	Adenosine receptor A2b	G-protein coupled receptor (PC00021)
ADRA1A	Alpha-1A adrenergic receptor	G-protein coupled receptor (PC00021)
APBA1	Amyloid-beta A4 precursor protein-binding family A member 1	membrane trafficking regulatory protein (PC00151)
ARHGAP22, ARHGAP39, ARHGAP44	Rho GTPase-activating protein 22	
ATP2B2, ATP2B4	Plasma membrane calcium-transporting ATPase 2 & 4	cation transporter (PC00068), hydrolase (PC00121), ion channel (PC00133)

Continued on next page...

Gene	Gene Symbol	Protein Class
BAIAP2	Brain-specific angiogenesis inhibitor 1-associated protein 2	receptor (PC00197)
BCR	Breakpoint cluster region protein	guanyl-nucleotide exchange factor (PC00113)
CACNA1A	Voltage-dependent P/Q-type calcium channel subunit alpha-1A	
CACNG2, CACNG3, CACNG4	Voltage-dependent calcium channel gamma-2 subunit	voltage-gated calcium channel (PC00240)
CADPS, CADPS2	Calcium-dependent secretion activator 1 & 2	calcium-binding protein (PC00060)
CAMK4	Calcium/calmodulin-dependent protein kinase type IV	non-motor microtubule binding protein (PC00166), non-receptor serine/threonine protein kinase (PC00167)
CDH8, CDH10, CDH11	Cadherin-8,10,11	
CHMP2B	Charged multivesicular body protein 2b	
CHRM2, CHRM3	Muscarinic acetylcholine receptor M2 & M3	G-protein coupled receptor (PC00021)
CLSTN1, CLSTN2	Calsyntenin-1 & 2	calcium-binding protein (PC00060), cell adhesion molecule (PC00069)
CNR1	Cannabinoid receptor 1	G-protein coupled receptor (PC00021)
CPLX2	Complexin-2	
CTBP2	C-terminal-binding protein 2	transcription cofactor (PC00217)

Continued on next page...

Gene	Gene Symbol	Protein Class
CTTNBP2	Cortactin-binding protein 2	
DGKB	Diacylglycerol kinase beta	kinase (PC00137)
DGKI	Diacylglycerol kinase iota	kinase (PC00137)
DLG2	Disks large homolog 2	transmembrane receptor regulatory/adaptor protein (PC00226)
DLGAP4	Disks large-associated protein 4	transmembrane receptor regulatory/adaptor protein (PC00226)
DNM2, DNM3	Dynamin-2 & 3	hydrolase (PC00121), microtubule family cytoskeletal protein (PC00157), small GTPase (PC00208)
DRD2, DRD3	D(2) & D(3) dopamine receptors	G-protein coupled receptor (PC00021)
EFNB2	Ephrin-B2	membrane-bound signaling molecule (PC00152)
EPHA4, EPHA7	Ephrin type-A receptors 4 & 7	
EPHB1, EPHB2	Ephrin type-B receptors 1 & 2	
ERBB4	Receptor tyrosine-protein kinase erbB-4	
ERC2	ERC protein 2	G-protein modulator (PC00022), membrane traffic protein (PC00150)
FARP1	FERM, ARHGEF and pleckstrin domain-containing protein 1	
FYN	Tyrosine-protein kinase Fyn	

Continued on next page...

Gene	Gene Symbol	Protein Class
FZD3	Frizzled-9	G-protein coupled receptor (PC00021), protease inhibitor (PC00191), signaling molecule (PC00207)
GABRR1	Gamma-aminobutyric acid receptor subunit rho-1	GABA receptor (PC00023), acetylcholine receptor (PC00037)
GPC6	Glypican-6	
GPM6A	Neuronal membrane glycoprotein M6-a	myelin protein (PC00161)
GRIA1	Glutamate receptor 1	
GRID1, GRID2	Glutamate receptor ionotropic, delta-1 & 2	
GRIK2, GRIK5	Glutamate receptor ionotropic, kainate 2 & 5	
GRIN2A, GRIN3A	Glutamate receptor ionotropic, NMDA 2A & 3A	
GRIP1, GRIP2	Glutamate receptor-interacting protein 1 & 2	
GRM1, GRM3	Metabotropic glutamate receptor 1 & 3	G-protein coupled receptor (PC00021)
GSG1L	Germ cell-specific gene 1-like protein	cytoskeletal protein (PC00085)
GSK3B	Glycogen synthase kinase-3 beta	non-receptor serine/threonine protein kinase (PC00167)
HIP1	Huntingtin-interacting protein 1	non-motor actin binding protein (PC00165)
HOMER1, HOMER2	Homer protein homolog 1 & 2	

Continued on next page...

Gene	Gene Symbol	Protein Class
HTR2A	5-hydroxytryptamine receptor 2A	G-protein coupled receptor (PC00021)
IL1RAP	Interleukin-1 receptor accessory protein	type I cytokine receptor (PC00231)
ITGB1, ITGB3	Integrin beta-1 & 3	cell adhesion molecule (PC00069), receptor (PC00197)
ITSN1	Intersectin-1	G-protein modulator (PC00022); calcium-binding protein (PC00060); membrane traffic protein (PC00150)
KCND2	Potassium voltage-gated channel subfamily D member 2	
LGI1	Leucine-rich glioma-inactivated protein 1	
LRFN5	Leucine-rich repeat and fibronectin type-III domain-containing protein 5	
LRRC4C	Leucine-rich repeat-containing protein 4C	
LRRK2	Leucine-rich repeat serine/threonine-protein kinase 2	
LRRN2	Leucine-rich repeat transmembrane neuronal protein 2	
LRRTM4	Leucine-rich repeat transmembrane neuronal protein 4	extracellular matrix protein (PC00102), receptor (PC00197)

Continued on next page...

Gene	Gene Symbol	Protein Class
LYN	Tyrosine-protein kinase Lyn	
MAPK10, MAPK14	Mitogen-activated protein kinase 10 & 14	non-receptor serine/threonine protein kinase (PC00167)
MTOR	Serine/threonine-protein kinase mTOR	non-receptor serine/threonine protein kinase (PC00167); nucleic acid binding (PC00171); nucleotide kinase (PC00172)
NAPB	Beta-soluble NSF attachment protein	membrane traffic protein (PC00150)
NDRG1	Protein NDRG1	serine protease (PC00203)
NETO1	Neuropilin and tolloid-like protein 1	
NLGN1	Neuroligin-1	
NOS1AP	Carboxyl-terminal PDZ ligand of neuronal nitric oxide synthase protein	signaling molecule (PC00207)
NRCAM	Neuronal cell adhesion molecule	
NRG1, NRG3	Pro-neuregulin-1 & 3, membrane-bound isoform	growth factor (PC00112)
NRP1, NRP2	Neuropilin-1 & 2	
NRXN1	Neurexin-1	
NTNG1, NTNG2	Netrin-G1 & G2	extracellular matrix linker protein (PC00101), protease inhibitor (PC00191), receptor (PC00197)
NTRK3	NT-3 growth factor receptor	
OLFM2	Noelin-2	receptor (PC00197); structural protein (PC00211)
P2RY1	P2Y purinoceptor 1	

Continued on next page...

Gene	Gene Symbol	Protein Class
PAK2	Serine/threonine-protein kinase PAK 2	
PLCB1, PLCB4	1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase beta-1 & 4	calcium-binding protein (PC00060), guanyl-nucleotide exchange factor (PC00113), phospholipase (PC00186), signaling molecule (PC00207)
PLEKHA5	Pleckstrin homology domain-containing family A member 5	
PLPPR4	Phospholipid phosphatase-related protein type 4	phosphatase (PC00181); pyrophosphatase (PC00196)
PPFIA2	Liprin-alpha-2 & 3	
PPFIA3		
PPM1H	Protein phosphatase 1H	kinase inhibitor (PC00139), protein phosphatase (PC00195)
PPP1R9A	Neurabin-1	
PPP3CA	Serine/threonine-protein phosphatase 2B catalytic subunit alpha isoform	
PRKAR1A	cAMP-dependent protein kinase type I-alpha regulatory subunit	
PSD2	PH and SEC7 domain-containing protein 2	
PTK2B	Protein-tyrosine kinase 2-beta	
PTPRD	Receptor-type tyrosine-protein phosphatase delta	protein phosphatase (PC00195); receptor (PC00197)

Continued on next page...

Gene	Gene Symbol	Protein Class
PTPRO, PTPRS, PTPRT	Receptor-type tyrosine-protein phosphatase O, S, & T	protein phosphatase (PC00195)
RAC1	Ras-related C3 botulinum toxin substrate 1	small GTPase (PC00208)
RAP1A	Ras-related protein Rap-1A	small GTPase (PC00208)
RGS7BP	Regulator of G-protein signaling 7-binding protein	
RNF216	E3 ubiquitin-protein ligase RNF216	
SCN2A	Sodium channel protein types 2 & 10 10 subunit alpha	voltage-gated calcium channel (PC00240)
SCN10A		voltage-gated sodium channel (PC00243)
SH3GL1, SHGL2, SHGL3	Endophilin-A2,A1, & A3	
SHANK2	SH3 and multiple ankyrin repeat domains protein 2	
SHISA6, SHISA9	Protein shisa-6 & 9	
SLC1A2, SLC1A6	Excitatory amino acid transporter 2	cation transporter (PC00068)
SLC6A17	Sodium-dependent neutral amino acid transporter SLC6A17	cation transporter (PC00068)
SNAP25	Synaptosomal-associated protein 25	SNARE protein (PC00034)

Continued on next page...

Gene	Gene Symbol	Protein Class
SORCS3	VPS10 domain-containing receptor SorCS3	receptor (PC00197), transporter (PC00227)
SPARC, SPARCL1	SPARC & SPARC-like protein 1	cell adhesion molecule (PC00069), extracellular matrix glycoprotein (PC00100), growth factor (PC00112)
SPTBN1	Spectrin beta chain, non-erythrocytic 1	
SRC	Proto-oncogene tyrosine-protein kinase Src	
STX3	Syntaxin-3	SNARE protein (PC00034)
SV2A	Synaptic vesicle glycoprotein 2A	
SYN3	Synapsin-3	membrane trafficking regulatory protein (PC00151); non-motor actin binding protein (PC00165)
SYNPO	Synaptopodin	non-motor actin binding protein (PC00165)
SYT1, SYT6	Synaptotagmin-1 & 6	membrane trafficking regulatory protein (PC00151)
TANC2	Protein TANC2	
TIAM1	T-lymphoma invasion and metastasis-inducing protein 1	
TNIK	TRAF2 and NCK-interacting protein kinase	
TNR	Tenascin-R	signaling molecule (PC00207)
UNC13A	Protein unc-13 homolog A	
WASF3	Wiskott-Aldrich syndrome protein family member 3	non-motor actin binding protein (PC00165)

Continued on next page...

Gene	Gene Symbol	Protein Class
WNT7A	Protein Wnt-7a	signaling molecule (PC00207)
YWHAZ	14-3-3 protein zeta/delta	chaperone (PC00072)

Table C.5: Genes enriched among the best exemplars in the gene ontology "biological process" category of the glutamate receptor signaling pathway.

Gene	Gene Symbol	Protein Class
APP	Amyloid-beta A4 protein	protease inhibitor (PC00191)
GNAQ	Guanine nucleotide-binding protein G(q) subunit alpha	heterotrimeric G-protein (PC00117)
GRIA1, GRIA4	Glutamate receptor 1 & 4	
GRID1, GRID2	Glutamate receptor ionotropic, delta-1, 2	
GRIK1, GRIK2, GRIK4, GRIK5	Glutamate receptor ionotropic, kainate 1,2,4,5	
GRIN2A, GRIN2B, GRIN2D, GRIN3A	Glutamate receptor ionotropic, NMDA 2A, 2B, 2D, 3A	
GRM1, GRM3, GRM4, GRM5, GRM6, GRM7, GRM8	Metabotropic glutamate receptor 1,3,4,5,6,7,8	G-protein coupled receptor (PC00021)
HOMER1, HOMER2	Homer protein homolog 1 & 2	

Continued on next page...

Gene	Gene Symbol	Protein Class
KCNB1	Potassium voltage-gated channel subfamily B member 1	
PLCB1	1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase beta-1	calcium-binding protein (PC00060), guanyl-nucleotide exchange factor (PC00113), phospholipase (PC00186), signaling molecule (PC00207)
PTK2B	Protein-tyrosine kinase 2-beta	
SSR1	Somatostatin receptor type 1	G-protein coupled receptor (PC00021)
TIAM1	T-lymphoma invasion and metastasis-inducing protein 1	
TRPM1, TRPM3	Transient receptor potential cation channel subfamily M member 1 & 3	ion channel (PC00133), receptor (PC00197)

Appendix D

Description of Supplementary Files

Supplementary files are available at Dalspace and in repositories for the respective Chapters' publications (see each Chapter for a link).

D.1 Supplementary Files for Chapter 3

`MMH_Supplementary_Code_1.nb` **Main analyses.** Mathematica notebook containing the primary analyses accompanying *The Meaning and Measure of Heterogeneity*.

D.2 Supplementary Files for Chapter 4

`RRH_Supplementary_Code_1.ipynb` **Main analyses of existing heterogeneity indices.** Jupyter notebook containing the primary analyses of the existing non-categorical heterogeneity indices including evaluation under the beta mixture model and representational Rényi heterogeneity in the convolutional variational autoencoder (CVAE) model.

`RRH_Supplementary_Code_2.ipynb` **Evaluation of MNIST homogeneity using siamese networks.** Jupyter notebook evaluating the hypothesis that the MNIST "Ones" are the most homogeneous class.

D.3 Supplementary Files for Chapter 6

`ASYM_Supplementary_Data_1.csv` **Total Alda score ratings.** Inter-rater reliability data for the total Alda score.

`ASYM_Supplementary_Data_2.csv` **Alda A-score ratings.** Inter-rater reliability data for the Alda A-score.

`ASYM_Supplementary_Code_1.nb` **Alda score analysis code.** Mathematica notebook containing the empirical evaluation of the Alda Score of Lithium response. This notebook also contains additional analysis of the A-score alone.

`ASYM_Supplementary_Code_2.nb` **Theoretical analysis code.** Mathematica notebook containing the theoretical analyses of discrete vs. continuous mutual information in asymmetrically reliable data.

`ASYM_Supplementary_Code_3.ipynb` **Code for statistical power tests.** Jupyter notebook containing the theoretical analyses of the statistical power of classical associative tests under asymmetrically reliable data.

Appendix E

Article Permissions

Permissions for the papers included in Chapters 2 and 5 are included in the following pages. Chapters 4 and 6 were published under the CC BY licence.

5 January 2020

Journal of Psychiatry & Neuroscience

To the editor:

I am preparing my doctoral thesis for submission to the Faculty of Graduate Studies at Dalhousie University, Halifax, Nova Scotia, Canada. I am seeking your permission to include a manuscript version of the following paper(s) as a chapter in the thesis:

> Nunes A, Trappenberg T, and Alda M.We Need an Operational Framework for Heterogeneity in Psychiatric Research. *Journal of Psychiatry & Neuroscience.* 2020;45(1):3-6

Canadian graduate theses are reproduced by the Library and Archives of Canada (formerly National Library of Canada) through a non-exclusive, world-wide license to reproduce, loan, distribute, or sell theses. I am also seeking your permission for the material described above to be reproduced and distributed by the LAC(NLC). Further details about the LAC(NLC) thesis program are available on the LAC(NLC) website (www.nlc-bnc.ca).

Full publication details and a copy of this permission letter will be included in the thesis.

Yours sincerely,

Abraham Nunes

Permission is granted for:

a) the inclusion of the material described above in your thesis.

b) for the material described above to be included in the copy of your thesis that is sent to the Library and Archives of Canada (formerly National Library of Canada) for reproduction and distribution.

Name: [redacted] Title: Publisher, CMAJ Group

Signature: Date: Jan 8, 2020

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS

Jan 30, 2020

This Agreement between Dalhousie University -- Abraham Nunes ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	4758801348991
License date	Jan 30, 2020
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Acta Psychiatrica Scandinavica
Licensed Content Title	Prediction of lithium response using clinical data
Licensed Content Author	A. Nunes, R. Ardau, A. Berghöfer, et al
Licensed Content Date	Nov 22, 2019
Licensed Content Volume	141
Licensed Content Issue	2
Licensed Content Pages	11
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article

Format	Print and electronic
Portion	Full article
Will you be translating?	No
Title of your thesis / dissertation	Measurement of Heterogeneity in Computational Psychiatry
Expected completion date	Jun 2020
Expected size (number of pages)	257
Requestor Location	Dalhousie University 60 Bayview Rd. Halifax, NS B3H 2E2 Canada Attn: Dalhousie University
Publisher Tax ID	EU826007151
Total	0.00 CAD
Terms and Conditions	

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.

- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, **and any CONTENT (PDF or image file) purchased as part of your order,** is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.

- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.**For STM Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts,** You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.

- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.

- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-

Creative Commons Attribution Non-Commercial License

The Creative Commons Attribution Non-Commercial (CC-BY-NC)License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License

The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by commercial "for-profit" organizations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee.

Further details can be found on Wiley Online Library
http://olabout.wiley.com/WileyCDA/Section/id-410895.html

Other Terms and Conditions:

v1.10 Last updated September 2015

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.