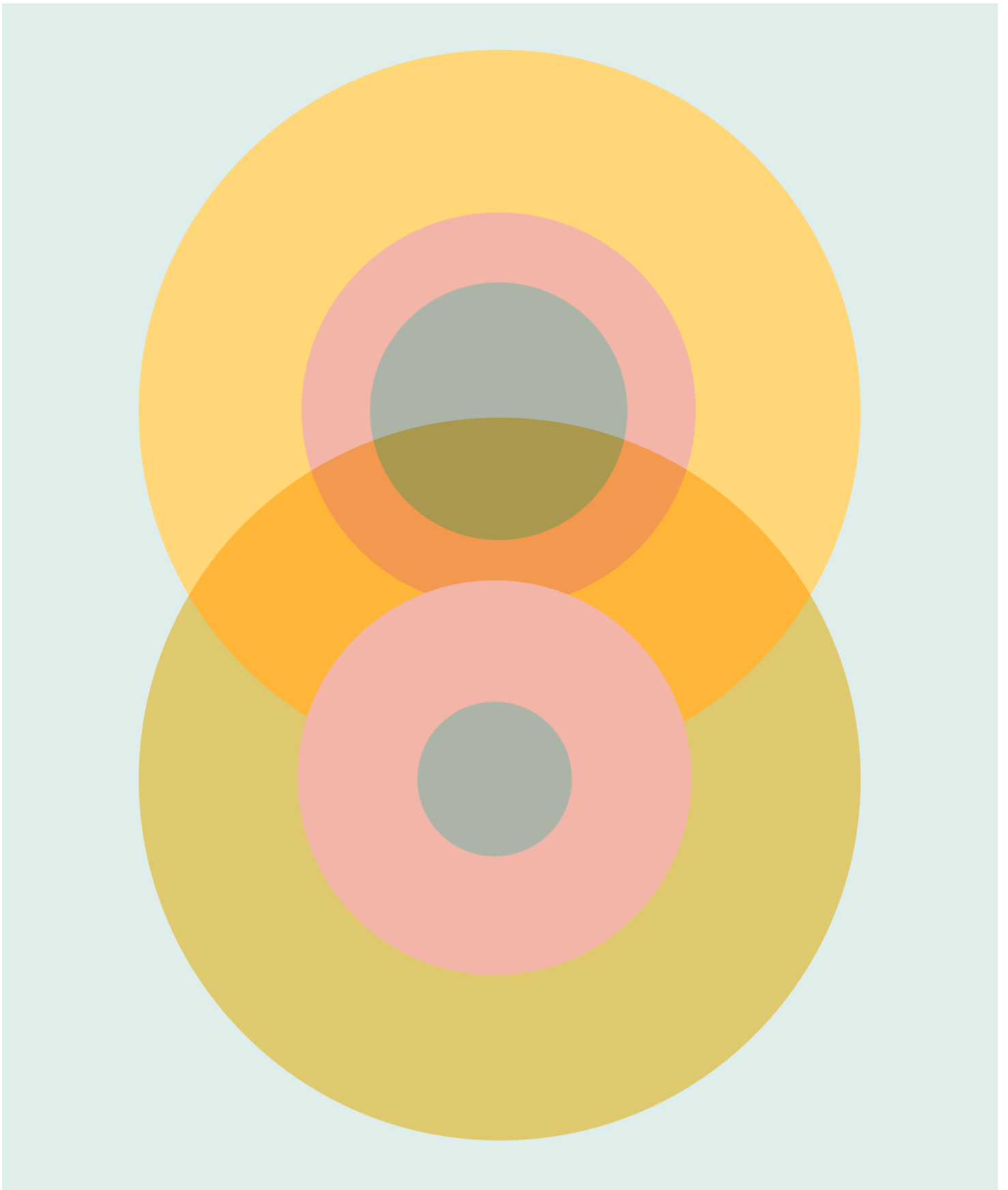




Multimodal machine learning for precision psychiatry

Micah Cearns

“Multimodal machine learning for precision psychiatry” PhD thesis of Micah Cearn, the discipline of Psychiatry in the School of Medicine, the University of Adelaide, Australia.

Cover design and layout: Quim Marin

Copyright ©2021 by Micah Cearn.

All rights reserved. No part of this work may be reproduced or stored in a retrieval system in any form or transmitted by electronic, mechanical, recording, photocopying or other means without the permission of the author.



Multimodal machine learning for precision psychiatry

PhD thesis

To obtain the degree of Doctor of Philosophy (PhD) in Medicine at the University of
Adelaide

By

Micah Cearns

Born on September the 30th 1990

Supervisors

Professor Bernhard T. Baune

A/Professor Scott Clark

February 2021

Adelaide, Australia

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Micah Cearns

Dedication

This work is dedicated to my Fiancé, Aneeta Rawat, my parents and sister, Rosalee, Michael, and Rachel Cearn, and my new family, Netra, Shashi, Ashish, and Mamta Rawat. Thank you all for your support in my life.

Table of Contents

Chapter 1: General introduction.....	6
Chapter 2: Recommendations and future directions for supervised machine learning in psychiatry. <i>Translational psychiatry</i> , 9(1), pp.1-12	25
Chapter 3: Machine learning probability calibration for high-risk clinical decision making. <i>Australian and New Zealand Journal of Psychiatry</i> . (2020): 123-126.....	57
Chapter 4: Systematic misestimation of machine learning performance in neuroimaging studies of depression. <i>Neuropsychopharmacology</i> . (2021), pp. 1-8	
Chapter 5: From multivariate methods to an AI ecosystem <i>Accepted: Molecular Psychiatry</i>	175
Chapter 6: Predicting rehospitalization within 2 years of initial patient admission for a major depressive episode: a multimodal machine learning approach. <i>Translational psychiatry</i> , 9(1), pp.1-9	189
Chapter 7: Stratification of bipolar patients using polygenic scores increases the accuracy of lithium response prediction: A machine learning approach	240
Chapter 8: General discussion and future perspectives.....	284

Chapter 1

General introduction

The problem

In 2017, around 792 million people lived with a mental health disorder, equating to more than 1 in 10 people globally (10.7%)¹. Within these disorders, major depression (MDD) has become the world's leading cause of disability². For patients that receive this diagnosis, approximately 22% will later be diagnosed with bipolar disorder³. After receiving their diagnosis, less than half will be given the care that the World Health Organization deems to be adequate⁴, and when they do, only 30-40% will fully recover after 12-weeks of treatment⁵. For those that do recover, approximately half will follow a relapse trajectory over the following 12 months⁶. Consequently, suicide rates have increased by approximately 25% since 2000, making suicide the 10th most common cause of death world wide^{7,8}. As a result of these trends, one in five adults now has an open prescription for a psychiatric medication in the United States⁹.

These sobering statistics reflect both a need for new treatments and a way to improve and personalise the treatments that we already have. Unfortunately, the pharmaceutical industry has significantly decreased its investment into new treatments for psychiatric disorders⁹, with the field now turning to cost-effective drug repositioning strategies^{10,11}. Some of these endeavours have been successful¹¹, however, it would be unwise to solely rely on this avenue. Alternatively, being able to accurately predict illness trajectories and disorder subtypes for individual patients would provide us with a way to improve and personalise the treatments that we already have^{12,13}. However, the accurate prediction of treatment response and illness trajectory prognostication remains a challenge for modern psychiatry, with contemporary practitioners still relying on a 'wait and see' approach to the treatment of psychiatric disorders¹⁴. Solving this problem would constitute a significant step towards managing this large and growing mental health burden. However, before we can solve this problem, we need to understand why it has arisen.

Firstly, in parallel to other fields (e.g., oncology), a lack of valid, reliable and clinically affordable biomarkers exist that correlate with clinical symptoms; at least with large enough effect to contribute to differential diagnosis, or to facilitate the accurate prediction of illness trajectories¹⁵⁻¹⁸. A multitude of significant univariate associations have been found between biomarkers and psychiatric conditions, yet in their individual form, have lacked clinically utility¹⁹. Given the demonstrated lack of Mendelian inheritance in psychiatric pathology, the power of hindsight yields this conclusion with no great surprise²⁰. Resultantly, the field has

progressed towards a framework of complexity, where no psychiatric trait can be explained by the segregation of a single gene or biomarker²⁰. This complex and polygenic basis helps to explain the heterogeneity of both clinical symptoms and biological markers within and across psychiatric traits²⁰. Furthermore, this observation helps to explain why individual biomarkers tend to yield only small effects when modelled at the group level, and within these groups, show a large degree of variability between individual subjects¹⁹. This framework also makes room for the profound influence of environmental and societal factors on the development and progression of psychiatric illness, and the individual differences often observed between patients²¹⁻²³. Moreover, these observations can be explained mechanistically through the science of epigenetics, where environmental and societal influences can express an individual's genetic predisposition to a psychiatric trait²⁴⁻²⁸. While this biopsychosocial conception of psychiatric phenomenon is near universally accepted^{29,30}, the ways in which psychiatric illness and disease trajectories are modelled are yet to properly account for this structure¹⁴. As the field continues to pursue the panacea of not just statistical but clinical significance in its search for a clinically translatable biopsychosocial basis to pathology, it is important to first consider how we typically model psychiatric phenomenon.

Classical group level statistics

In psychiatric research, we commonly stratify subjects into groups. These groups may be formed on the basis of a diagnosis or whether a subject is a case or control participant in a study. This group-level approach is intuitive as it allows us to understand something about the groups we care about. For example, for individuals with MDD, does one treatment work better than another? These group-level statistics afford us the ability to make an inference about a population through the process of null-hypothesis significance testing¹⁴. Using this process, we can conduct analyses with varying aims, each with their own level of quality and implications for clinical practice. For example, in an exploratory analysis, we might first interpret the magnitude and statistical significance of a variable's association with a disease or illness trajectory (i.e., the groups that we want to learn about). While exploratory findings are far from definitive, these they may lead to further hypothesis generation, later helping to uncover mechanism and establish grounded theory about a group of interest. Further down the spectrum of study design lies the gold standard, the randomised, double-blind, placebo-controlled trial. Finally, we may take the results of these studies and collate them into a meta-analysis or systematic review. This level of analysis confers the highest weight of evidence and is currently the best suited to inform clinical decision making at the group level. Given

the natural variation in patient outcomes, the strength of the placebo effect, and the tendency for a biological system to mean revert over time (i.e., naturally heal itself), these group-level approaches have been integral to the establishment and advancement of modern evidence-based medicine^{31,32}.

While this paradigm shift in how we treat disease has underpinned nearly every medical breakthrough in the last century, the ways in which we treat the psychiatric disorders has hardly changed over the last 50 years. For example, Fluoxetine, first released in 1987, is still used as a first line antidepressant for MDD³³. A recent meta-analysis including newer antidepressants showed that only a few conferred advantage³⁴. Moreover, when a treatment was the novel or experimental drug of comparison, it appeared to be significantly more effective than when that same treatment was the older or control drug of comparison. Adjusting for this novelty effect diminished the differences between newer antidepressants and Fluoxetine³⁴. In the case of BP, Lithium was approved by the FDA in 1970, and to this day, is still one of the most widely used medications for BP treatment³⁵. Although these breakthroughs have been momentous, decades later, they are yet to be superseded. The biopsychosocial complexity of these traits likely contributes to this lack of progress. This begs the question, does understanding the biopsychosocial aetiology of psychiatric disorders for the ‘average’ patient, help us to fully treat and understand every day ‘individual’ patients in the clinic?

Indeed, patients are not modelled as ‘averages’ because they are believed to be the same across all possible covariates and moderators. Instead, there has been a long-standing belief that understanding core disease mechanism at the group-level, afforded by statistical inference, will lead to effective treatment and illness trajectory prognostication. Therefore, clinical trials and group-level studies have been designed to minimise the impact of individual differences through strict inclusion and exclusion criteria. However, given the limited progress of this approach, and the pressing need to advance patient care in the clinic, a new school of thought has arisen where a greater focus is placed on prediction. This shift in focus does not replace the need for explanation, but instead concedes that explanation may not need precede prediction¹⁴. For example, by training models to predict patient outcomes and classify disease states, not only may clinical care be better personalised, but new hypotheses may be generated from the learnt multivariate patterns, leading to further statistical analysis, and a greater understanding of disease. Emerging evidence supports this approach, suggesting that it may predict patient outcomes and treatment better than

traditional DSM and ICD diagnoses¹⁴. It is important to note here that the tools used for prediction can be seen as lying on a continuum with classical statistical methods such as regression, but with the addition of practices designed to reduce overfitting such as parameter regularization and cross validation³⁶. However, to utilise these modified approaches effectively for individual patient predictions, individual differences must not be averaged out, but be seen as fundamental to the facilitation of robust predictions. Indeed, the literature is now making apparent the large and nuanced extent of these differences, and the impact they have on disease status and clinical outcomes¹³.

Heterogeneity

For example, in a group of patients with first-episode psychosis, Cannon et al.³⁷ found those with higher levels of unusual thought content transitioned to schizophrenia sooner than those who did not. In a study of MDD patients, those who experienced abuse as children were significantly more likely to have a better response to cognitive behavioural therapy than to the antidepressant Nefazodone³⁸. In BP, patients who experienced an episodic pattern of manic and depressive episodes and a high age of illness onset were more likely to benefit from prophylactic lithium treatment³⁹. These examples of patient heterogeneity are now commonplace in the literature, representing the spectrum of human individuality and its influence on psychiatric experience.

Indeed, the aggregation of these differences can tell a meaningful clinical story. For example, Peris et al.,⁴⁰ studied whether they could accurately differentiate between participants presenting with nonpsychotic unipolar depression or bipolar depression in an outpatient setting. The majority of variation in symptom and risk factors they noticed was small. Unipolar participants presented with more evident sadness whereas bipolar patients presented with more pessimistic attributes. However, when all of these individual variables were mathematically modelled together, they could distinguish between patients with 87% accuracy. Even though modern-day clinicians are often aware of these differences, it does not imply that they can use them effectively. For example, in an analysis by Chekroud, he assessed the ability of clinicians to use symptom level, demographic, and overall severity data to predict antidepressant response in MDD. Even after considering individual differences between patients, average clinician accuracy was only 49.3%. Given the seemingly infinite and fluctuating combinations of symptoms, environmental influences, and human biology, this poor predictive result comes as no criticism to clinicians. Instead, it

highlights the biopsychosocial complexity of major depression and the cognitive capacity constraints inherent in even the brightest of human minds⁴¹. Interestingly, training a multivariate machine learning (ML) model with the same data led to a model that generalised to the same cohort of unseen patients, attaining accuracies of ~ 60-65% across geographically distinct cohorts. That is, a multivariate predictive model trained to learn about individuals, not averages, yielded a significant improvement in patient care. This model is now used routinely by a US mental healthcare company¹².

This application of machine learning to psychiatric phenomenon provides a new way to consider the individual differences inherent within groups of heterogeneous patients⁴². This tailored modelling approach comes under the broader movement of personalised medicine, a philosophy focused on the distinct biology, behaviour, and experiences of individuals⁴³.

While the idea of a personalised ‘psychiatry’ is seldom new, advances in computing power, the development of high-level programming languages, and increases in the availability of large multimodal datasets are helping this vision become a reality. In particular, these progressions have provided the fundamental building blocks required to train and deploy machine learning models in academic and clinical settings¹³. This represents an important step forwards for personalised psychiatry, as these models are capable of learning and making predictions on a patient-by-patient basis.

Machine learning approaches

Broadly speaking, there are three categories of learning that are commonly used in ML⁴². In ‘supervised’ ML, an algorithm learns a function that maps a set of predictor variables (also known as features in the ML community) to an outcome variable (commonly referred to as a target or label)⁴⁴. This outcome can either be continuous for regression or categorical for classification. In classification problems, the goal is to learn what category an observation belongs to after seeing a set of relevant examples⁴². The second learning category, known as ‘unsupervised’ learning, attempts to discover patterns and structure in data without learning directly from a set of outcome labels⁴⁵. For example, identifying a more homogenous set of major depressive disorder patients (MDD) (i.e., patients with core emotional symptoms) from a broader heterogeneous patient pool (i.e., patients with insomnia, fatigue, and chronic pain). Clustering (e.g., k-means and hierarchical clustering) and dimensionality reduction (e.g., principal components and factor analysis) techniques form the foundation of unsupervised ML^{46,47}. Finally, ‘semi-supervised’ learning combines the mathematical underpinnings of

both ‘supervised’ and ‘unsupervised’ learning. This approach learns from a subset of both labelled and unlabelled data⁴⁸.

In a small handful of well-powered multisite studies, this statistical learning approach to modelling psychiatric disorders has already demonstrated its ability to learn from and make predictions about individual patients and their illness trajectories. However, these studies have predominantly focused on learning from clinical symptom items and overall severity measures, likely due to the scarcity of large biomarker datasets. For example, Chekroud et al¹² successfully predicted response to Citalopram in a study of over 4000 MDD patients using routine clinical data and symptom items whereas Koutsouleris successfully predicted 4- and 52-week treatment outcomes in patients with first-episode psychosis using a combination clinical scale items and demographic variables (4-week balanced accuracy = 75%, 52-week balanced accuracy = 73.8%)⁴⁹.

In the studies that have used biomarkers, these have predominantly focused on neuroimaging predictor modalities and have been composed of only a small number of subjects⁵⁰. An even smaller handful of studies exist using serum, genetic, or electroencephalography (EEG) markers⁵¹⁻⁵⁷. What has not been rigorously studied is the combination of clinical data with imaging, serum, genetic, and EEG biomarkers. Given their biopsychosocial basis, this is a reasonable next step. Indeed, preliminary work focused on psychosis has demonstrated the potential of this multimodal approach, albeit, with a smaller combination of biomarkers. For example, Clark et al⁵⁸ found that using historical (history of drug use), clinical (Positive and Negative Symptoms scale - positive, negative, and general scores, and the global assessment of function), and biomarkers (total omega-3 and nervonic acid) in combination predicted transition to first-episode psychosis better than any individual data modality. In patients at clinical high-risk for psychosis, Koutsouleris et al⁵⁹ found that models trained on a combination of functional and neuroimaging predictors outperformed individual modality models when prognosticating functional outcomes. However, these findings were contested after evidence that routine clinical data was excluded from model training and was instead replaced by eight functional predictors. It was argued that this was done to increase the usability of the model, however, the inclusion of sensitive, invasive, and costly neuroimaging markers was at odds with this rationale⁶⁰.

Methodological inconsistencies

This lack of conclusiveness is not uncommon. Not only because of the large experimenter degrees of freedom that arise in multivariate analysis, as demonstrated above in the Koutsouleris et al⁵⁹ analysis, but due to a lack of methodological consistency and established best practice in the field. For example, in a recent meta-analysis of ML studies using neuroimaging markers for MDD classification⁵⁰, it was noted that data leakage (the leakage of data from outside the training set into model training) was present in at least three studies. This effect arose due to model features being selected outside of cross-validation leading to optimistic bias in model testing⁶¹⁻⁶³. In studies where data transformations have been conducted via cross-validation, reliable schemes have not always been used. For example, the leave-one out cross-validation scheme has been shown to lead to a high degree of variability in performance estimates^{64,65}, however, it still remains a commonly used scheme in psychiatric and neuroimaging studies^{50,65}. Finally, the majority of predictive studies utilising biomarkers have done so on only a handful of subjects. This is not by design but a result of current funding structures and the associated costs of biomarker acquisition and storage. While understandable, the implications of small sample size on model performance are not to be underestimated⁶⁶.

Specifically, ML studies using larger samples have consistently shown weaker performance than studies drawing on smaller ones. For example, Neuhaus & Popescu⁶⁶ studied the relationship between sample size and model accuracy across schizophrenia (total observation $N = 5563$), MDD ($N = 2042$), and attention deficit hyperactivity disorder (ADHD, $N = 8084$) studies, observing an inverse relationship between sample size and balanced accuracy across all disorders (schizophrenia, $r = -0.34$; MDD, $r = -0.32$; and ADHD, $r = -0.43$). These findings are in accordance with a recent review of 200 neuroimaging classification studies that observed a trend towards lower accuracy in larger samples⁶⁷. As model performance is expected to increase with more data, these findings suggest a fundamental problem underlying predictive psychiatric ML studies. Given the size of this performance variability across studies, first overcoming these problems is fundamental before we can address questions about the predictive superiority of one data modality over another, or their combination.

Outline and aims of this thesis

The aim of this thesis is to investigate the ability of ML models to make clinically useful predictions about individual patients. More specifically, it aims to address whether the

combination of clinical predictors with multimodal biomarkers can increase model performance in comparison to models that only use a single data modality (i.e., only clinical or neuroimaging data). However, before these questions can be answered, psychiatric ML practices need to be standardised to ensure that differences between models are attributable to specific data modalities and or their combination instead of natural variation resulting from small samples, data leakage, diverse cross-validation schemes, and experimenter degrees of freedom.

Therefore, the first half of this thesis is devoted to addressing these fundamental methodological issues in the field. First, **Chapter 1** provides a general background on the heterogeneity of psychiatric disorders, an overview of group-level statistics, their limitations for modelling individual differences, the implications this has for clinical translation, and the capability of machine learning methods to assist in overcoming this problem.

Chapter 2 proceeds to lay a foundation for the standardisation of methodological best practice, covering issues of model training and evaluation and clinical application. Specifically, I review literature highlighting the effects of sample size on model performance, and the counterintuitive inverse relationship between decreasing sample size and increasing model performance. In addition, I explore the issue of cross-validation scheme selection and its influence on the variability and bias of model performance estimates. In particular, I discuss commonly used schemes and their specific use cases in psychiatric studies, dependent on factors such as sample size, the presence of multiple data collection sites, and model selection and hyperparameter tuning. More specifically, I considered how CV scheme choice and structure, alongside the correct specification of pipeline architectures within these schemes has implications for data leakage and overfitting, two widespread problems in the field^{50,66,68-70}. Finally, motivated by the observation that many editors and funding bodies lacked a measured way to interpret the maturity and impact of predictive psychiatric studies, I devised a framework and scoring system to measure the maturity of applied machine learning publications.

Next, **Chapter 3** proceeds to build a case for the importance of clinical context in model construction, paying specific attention to the overlooked significance of model calibration and its implications for clinical decision making. Here, I build the case for the importance of model probability calibration and demonstrate the properties of reliability curves for commonly used models in psychiatric machine learning.

Given the counterintuitive inverse relationship between sample size and model performance, **Chapter 4** attempts to quantify where in the cross-validation cycle performance estimates are most susceptible to variation and performance overestimation. Here, we demonstrate that the risk of misestimating model performance predominantly arises from small test set sizes. Moreover, we demonstrate through the use of dummy classifiers that this effect is not specific to any data modality or psychiatric phenotype but is a generalisable statistical effect applicable to all machine learning applications. To overcome this problem, we suggest that future studies require larger testing samples, or conduct further external testing across diverse data collection sites.

Informed by these findings, **Chapter 5** builds a case for a cohesive AI ecosystem, wherein issues of data collection, model construction, evaluation, and clinical translation are considered. In particular, an extensive and open-sourced model evaluation framework is proposed paying particular attention to issues of heterogeneity and sample size. In addition, new methodological techniques are introduced that may help to overcome these issues in future studies.

Once this methodological foundation is laid, the second half of this thesis progresses onto addressing the benefit of introducing multimodal biomarkers into model construction alongside more traditional clinical measures (e.g., symptom items and measures of cognitive function). To this end, **Chapter 6** assesses whether it is possible to predict rehospitalization in an inpatient cohort of major depressive patients using baseline clinical, structural imaging, blood-biomarker, polygenic risk, bioelectrical impedance, and electrocardiography predictors. Here, we find that a multimodal panel composed of structural imaging, blood-biomarker, clinical, medication type, and sleep quality predictors performs best.

Following, **Chapter 7** uses a combination of routine clinical data alongside measures of polygenic risk to predict lithium response in bipolar patients. This paper progresses beyond just using all combined predictors in a supervised ML models feature space, and instead, also uses polygenic risk (PRS) loadings for major depression and schizophrenia to stratify patients prior to model training. This approach affords the parsing of biological heterogeneity before proceeding to supervised model training with routine clinical data. Here, we found that a model stratified with a standardised meta-PRS, combining scores for both major depressive disorder and schizophrenia best predicted continuous and binary lithium response.

Finally, **Chapter 8** concludes the work of this thesis with a general discussion of findings and a set of new considerations for the field. Specifically, we discuss the clinical and biological heterogeneity within psychiatric traits as well as different techniques and modelling approaches that could be used to parse this heterogeneity in future works. Moreover, we address some of the methodological limitations in these studies that have led to failures in replication, and how these problems could be avoided moving forward. We end this thesis by considering the future role of technology in psychiatry, and how future investments into innovative technology may lead to paradigm shifts in the way we diagnose and treat mental health disorders.

References

1. Ritchie H, Roser M. Mental health. Our world in data. In: University of Oxford, Global Change Data Lab; 2018.
2. Friedrich MJ. Depression is the leading cause of disability around the world. *Jama*. 2017;317(15):1517-1517.
3. Ratheesh A, Davey C, Hetrick S, et al. A systematic review and meta-analysis of prospective transition from major depression to bipolar disorder. *Acta Psychiatrica Scandinavica*. 2017;135(4):273-284.
4. Thornicroft G, Chatterji S, Evans-Lacko S, et al. Undertreatment of people with major depressive disorder in 21 countries. *The British Journal of Psychiatry*. 2017;210(2):119-124.
5. Rush AJ, Warden D, Wisniewski SR, et al. STAR* D. *CNS drugs*. 2009;23(8):627-647.
6. Gueorguieva R, Chekroud AM, Krystal JH. Trajectories of relapse in randomised, placebo-controlled trials of treatment discontinuation in major depressive disorder: an individual patient-level data meta-analysis. *The Lancet Psychiatry*. 2017;4(3):230-237.
7. Hedegaard H, Curtin SC, Warner M. *Suicide rates in the United States continue to increase*. US Department of Health and Human Services, Centers for Disease Control and ...; 2018.
8. Ritchie H, Roser M. Our world in data. *Causes of Death*. 2018.
9. Hyman SE. Psychiatric drug development: diagnosing a crisis. Paper presented at: Cerebrum: the Dana forum on brain science 2013.
10. Aan Het Rot M, Zarate Jr CA, Charney DS, Mathew SJ. Ketamine for depression: where do we go from here? *Biological psychiatry*. 2012;72(7):537-547.

11. Rosenblat JD, Carvalho AF, Li M, Lee Y, Subramanieapillai M, McIntyre RS. Oral ketamine for depression: a systematic review. *The Journal of clinical psychiatry*. 2019;80(3):0-0.
12. Chekroud AM, Zotti RJ, Shehzad Z, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry*. 2016;3(3):243-250.
13. Chekroud AM, Lane CE, Ross DA. Computational psychiatry: embracing uncertainty and focusing on individuals, not averages. *Biological psychiatry*. 2017;82(6):e45-e47.
14. Bzdok D, Meyer-Lindenberg A. Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2018;3(3):223-230.
15. Caldieraro MA, Vares EA, Souza LH, et al. Illness severity and biomarkers in depression: Using a unidimensional rating scale to examine BDNF. *Compr Psychiatry*. 2017;75:46-52.
16. Lopresti AL, Maker GL, Hood SD, Drummond PD. A review of peripheral biomarkers in major depression: the potential of inflammatory and oxidative stress biomarkers. *Prog Neuropsychopharmacol Biol Psychiatry*. 2014;48:102-111.
17. Philibert R, Gunter HM, Kolassa IT. The search for peripheral biomarkers for major depression: benefiting from successes in the biology of smoking. *Am J Med Genet B Neuropsychiatr Genet*. 2014;165B(3):230-234.
18. Schneider B, Prvulovic D. Novel biomarkers in major depression. *Current opinion in psychiatry*. 2013;26(1):47-53.
19. Lozupone M, Seripa D, Stella E, et al. Innovative biomarkers in psychiatric disorders: a major clinical challenge in psychiatry. *Expert review of proteomics*. 2017;14(9):809-824.

20. Gelernter J. Genetics of complex traits in psychiatry. *Biological psychiatry*. 2015;77(1):36-42.
21. Cerdá M, Sagdeo A, Johnson J, Galea S. Genetic and environmental influences on psychiatric comorbidity: a systematic review. *J Affect Disorders*. 2010;126(1-2):14-38.
22. Compton MT, Shim RS. The social determinants of mental health. *Focus*. 2015;13(4):419-425.
23. Dunn EC, Masyn KE, Yudron M, Jones SM, Subramanian S. Translating multilevel theory into multilevel research: challenges and opportunities for understanding the social determinants of psychiatric disorders. *Soc Psych Psych Epid*. 2014;49(6):859-872.
24. Kular L, Kular S. Epigenetics applied to psychiatry: Clinical opportunities and future challenges. *Psychiatry Clin Neurosci*. 2018;72(4):195-211.
25. Mahgoub M, Monteggia LM. Epigenetics and psychiatry. *Neurotherapeutics*. 2013;10(4):734-741.
26. El-Sayed AM, Koenen KC, Galea S. Putting the 'epi' into epigenetics research in psychiatry. *J Epidemiol Community Health*. 2013;67(7):610-616.
27. Stahl SM. Epigenetics and methylomics in psychiatry. *J Clin Psychiatry*. 2009;70(9):1204-1205.
28. Kinoshita T. Epigenetics in psychiatry. *Neuropsychobiology*. 2009;60(1):1.
29. Pilgrim D. The biopsychosocial model in Anglo-American psychiatry: Past, present and future? *Journal of Mental Health*. 2002;11(6):585-594.
30. Ghaemi SN. The biopsychosocial model in psychiatry: A critique. *Am J Psychiat*. 2011;121:451-457.
31. Shryock RH. The development of modern medicine. In: LWW; 1948.

32. Shryock RH, Shryock RH. *The development of modern medicine: an interpretation of the social and scientific factors involved*. Knopf New York; 1947.
33. Hillhouse TM, Porter JH. A brief history of the development of antidepressant drugs: from monoamines to glutamate. *Experimental and clinical psychopharmacology*. 2015;23(1):1.
34. Cipriani A, Furukawa TA, Salanti G, et al. Comparative Efficacy and Acceptability of 21 Antidepressant Drugs for the Acute Treatment of Adults With Major Depressive Disorder: A Systematic Review and Network Meta-Analysis. *Focus (Am Psychiatr Publ)*. 2018;16(4):420-429.
35. Machado-Vieira R, Manji HK, Zarate CA, Jr. The role of lithium in the treatment of bipolar disorder: convergent evidence for neurotrophic effects as a unifying hypothesis. *Bipolar Disord*. 2009;11 Suppl 2:92-109.
36. Bennett D, Silverstein SM, Niv Y. The two cultures of computational psychiatry. *JAMA psychiatry*. 2019;76(6):563-564.
37. Cannon TD, Cadenhead K, Cornblatt B, et al. Prediction of psychosis in youth at high clinical risk: a multisite longitudinal study in North America. *Archives of general psychiatry*. 2008;65(1):28-37.
38. Nemeroff CB, Heim CM, Thase ME, et al. Differential responses to psychotherapy versus pharmacotherapy in patients with chronic forms of major depression and childhood trauma. *Proceedings of the National Academy of Sciences*. 2003;100(24):14293-14296.
39. Kleindienst N, Engel R, Greil W. Which clinical factors predict response to prophylactic lithium? A systematic review for bipolar disorders. *Bipolar disorders*. 2005;7(5):404-417.

40. Perlis RH, Brown E, Baker RW, Nierenberg AA. Clinical features of bipolar depression versus major depressive disorder in large multicenter trials. *Am J Psychiatry*. 2006;163(2):225-231.
41. Perlis RH. Abandoning personalization to get to precision in the pharmacotherapy of depression. *World Psychiatry*. 2016;15(3):228-235.
42. Iniesta R, Stahl D, McGuffin P. Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine*. 2016;46(12):2455-2465.
43. Vieta E. Personalised medicine applied to mental health: precision psychiatry. *Revista de psiquiatria y salud mental*. 2015;8(3):117.
44. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Springer series in statistics. In: . Springer; 2001.
45. Ghahramani Z. Unsupervised learning. Paper presented at: Summer School on Machine Learning 2003.
46. Everitt BS. Unresolved problems in cluster analysis. *Biometrics*. 1979;169-181.
47. Lu H, Plataniotis KN, Venetsanopoulos A. *Multilinear subspace learning: dimensionality reduction of multidimensional data*. CRC press; 2013.
48. Zhu X, Goldberg AB. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*. 2009;3(1):1-130.
49. Koutsouleris N, Kahn RS, Chekroud AM, et al. Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *The lancet Psychiatry*. 2016;3(10):935-946.
50. Kambeitz J, Cabral C, Sacchet MD, et al. Detecting Neuroimaging Biomarkers for Depression: A Meta-analysis of Multivariate Pattern Recognition Studies. *Biol Psychiatry*. 2017;82(5):330-338.

51. Khodayari-Rostamabad A, Reilly JP, Hasey GM, de Bruin H, MacCrimmon DJ. A machine learning approach using EEG data to predict response to SSRI treatment for major depressive disorder. *Clin Neurophysiol.* 2013;124(10):1975-1985.
52. Mumtaz W, Ali SSA, Yasin MAM, Malik AS. A machine learning framework involving EEG-based functional connectivity to diagnose major depressive disorder (MDD). *Medical & biological engineering & computing.* 2018;56(2):233-246.
53. Kim EY, Lee MY, Kim SH, Ha K, Kim KP, Ahn YM. Diagnosis of major depressive disorder by combining multimodal information from heart rate dynamics and serum proteomics using machine-learning algorithm. *Progress in Neuro-Psychopharmacology and Biological Psychiatry.* 2017;76:65-71.
54. Maciukiewicz M, Marshe VS, Hauschild A-C, et al. GWAS-based machine learning approach to predict duloxetine response in major depressive disorder. *Journal of psychiatric research.* 2018;99:62-68.
55. Dinga R, Marquand AF, Veltman DJ, et al. Predicting the naturalistic course of depression from a wide range of clinical, psychological, and biological data: a machine learning approach. *Transl Psychiatry.* 2018;8(1):241.
56. Hilbert K, Lueken U, Muehlhan M, Beesdo-Baum K. Separating generalized anxiety disorder from major depression using clinical, hormonal, and structural MRI data: A multimodal machine learning study. *Brain and behavior.* 2017;7(3):e00633.
57. Guilloux J-P, Bassi S, Ding Y, et al. Testing the predictive value of peripheral gene expression for nonremission following citalopram treatment for major depression. *Neuropsychopharmacology.* 2015;40(3):701-710.
58. Clark S, Baune B, Schubert K, et al. Prediction of transition from ultra-high risk to first-episode psychosis using a probabilistic model combining history, clinical assessment and fatty-acid biomarkers. *Translational psychiatry.* 2016;6(9):e897-e897.

59. Koutsouleris N, Kambeitz-Ilankovic L, Ruhrmann S, et al. Prediction Models of Functional Outcomes for Individuals in the Clinical High-Risk State for Psychosis or With Recent-Onset Depression: A Multimodal, Multisite Machine Learning Analysis. *JAMA Psychiatry*. 2018;75(11):1156-1172.
60. Koutsouleris N, Upthegrove R, Wood SJ. Importance of Variable Selection in Multimodal Prediction Models in Patients at Clinical High Risk for Psychosis and Recent Onset Depression—Reply. *JAMA psychiatry*. 2019;76(3):339-340.
61. Sacchet MD, Livermore EE, Iglesias JE, Glover GH, Gotlib IH. Subcortical volumes differentiate Major Depressive Disorder, Bipolar Disorder, and remitted Major Depressive Disorder. *J Psychiatr Res*. 2015;68:91-98.
62. Lord A, Horn D, Breakspear M, Walter M. Changes in community structure of resting state functional connectivity in unipolar depression. *PLoS One*. 2012;7(8):e41282.
63. Qin J, Wei M, Liu H, et al. Altered anatomical patterns of depression in relation to antidepressant treatment: Evidence from a pattern recognition analysis on the topological organization of brain networks. *J Affect Disord*. 2015;180:129-137.
64. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Paper presented at: Ijcai1995.
65. Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *Neuroimage*. 2017;145(Pt B):166-179.
66. Neuhaus AH, Popescu FC. Sample Size, Model Robustness, and Classification Accuracy in Diagnostic Multivariate Neuroimaging Analyses. *Biol Psychiatry*. 2018;84(11):e81-e82.

67. Arbabshirani MR, Plis S, Sui J, Calhoun VD. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage*. 2017;145(Pt B):137-165.
68. Kambeitz J, Cabral C, Sacchet MD, et al. Reply to: Sample Size, Model Robustness, and Classification Accuracy in Diagnostic Multivariate Neuroimaging Analyses. *Biol Psychiatry*. 2018.
69. Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage*. 2017;145:166-179.
70. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in Data Mining: Formulation, Detection, and Avoidance. *Acm T Knowl Discov D*. 2012;6(4).

Statement of Authorship

Title of Paper	Recommendations and future directions for supervised machine learning in psychiatry
Publication Status	☒ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. <i>Translational psychiatry</i> , 9(1), 1-12.

Principal Author

Name of Principal Author (Candidate)	Micah Cearns		
Contribution to the Paper	Formulated the paper layout with Tim Hahn. Wrote sections from the 'Introduction' to the end of 'Proof-of-concept studies to clinical application' and 'From ML models to clinician decision support systems' to 'Conclusion'. Created figures 1 and 2. Helped revise and refine the remaining sections.		
Overall percentage (%)	55%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	15/01/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- the candidate's stated contribution to the publication is accurate (as detailed above);
- permission is granted for the candidate to include the publication in the thesis; and
- the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Tim Hahn		
Contribution to the Paper	Wrote sections 'Understanding ML models' through to 'Learning complex models from small samples'. Created figure 3. Helped revise other sections and provided overall feedback.		
Signature		Date	15/01/2021

Name of Co-Author	Bernhard Baune

Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision. As corresponding author, he made communications with journal editors.		
Signature		Date	15/01/2021

Please cut and paste additional co-author panels here as required.

Chapter 2

Recommendations and future directions for supervised machine learning in psychiatry

Micah Cearns, BSc (Hons)^{1,3*}, Tim Hahn, PhD,^{2*} Bernhard T. Baune, MD, PhD^{2,3,4}

¹Discipline of Psychiatry, School of Medicine, University of Adelaide, Australia

²Institute of Translational Psychiatry, University of Münster, Germany

³Department of Psychiatry, University of Münster, Germany

³Department of Psychiatry, Melbourne Medical School, The University of Melbourne, Australia

⁴ The Florey Institute of Neuroscience and Mental Health, The University of Melbourne Parkville, VIC, Australia

* these authors contributed equally to this work and are equal first authors

Translational Psychiatry

Abstract

Machine learning methods hold promise for personalized care in psychiatry, demonstrating the potential to tailor treatment decisions and stratify patients into clinically meaningful taxonomies. Subsequently, publication counts applying machine learning methods have risen, with different data modalities, mathematically distinct models, and samples of varying size being used to train and test models with the promise of clinical translation. Consequently, and in part due to the preliminary nature of such works, many studies have reported largely varying degrees of accuracy, raising concerns over systematic overestimation and methodological inconsistencies. Furthermore, a lack of procedural evaluation guidelines for non-expert medical professionals and funding bodies leaves many in the field with no means to systematically evaluate the claims, maturity, and clinical readiness of a project. Given the potential of machine learning methods to transform patient care, albeit, contingent on the rigour of employed methods and their dissemination, we deem it necessary to provide a review of current methods, recommendations, and future directions for applied machine learning in psychiatry. In this review we will cover issues of best practice for model training and evaluation, sources of systematic error and overestimation, model explainability vs trust, the clinical implementation of AI systems, and finally, future directions for our field.

Introduction

Accurate prediction of intervention response and illness trajectories remains an elusive problem for modern psychiatry, with contemporary practitioners still relying on a ‘wait and see’ approach for the treatment of psychiatric disorders⁽¹⁾. This problem has likely arisen due to an interplay of biopsychosocial factors⁽²⁾ and statistical modelling decisions⁽³⁾. From a biopsychosocial perspective, the high degree of comorbidity between psychiatric conditions^(4, 5), the lack of diagnostic biomarkers to delineate between disorders and illness trajectories^(6, 7), the shared genetic origins of clinically disparate traits⁽⁸⁾, and the imprecision of symptom measures^(9, 10) has likely contributed to the complexity and lack of accuracy in clinical decision making.

Methodologically, psychiatry has commonly focused on statistical inference over prediction^(3, 11). Inferential statistics have afforded the testing of theory driven hypotheses, population inference, and the formulation of grounded theory and mechanism to better understand the aetiology of psychiatric traits^(3, 11). However, both psychiatry and neuroscience have found themselves with significant translation problems. Even in the face of new discoveries and paradigm shifts in the understanding of disorders, the clinical practice of psychiatry and discovery of interventions that outperform placebo has been slow⁽¹²⁾.

Given this complexity, clinicians commonly assume diagnostic homogeneity, where all patients who present with e.g., symptoms of low mood, lack of energy, and negative thoughts are considered to have the same broad diagnosis of major depressive disorder (MDD)⁽¹⁾. However, studies using machine learning methodologies (ML) have begun to identify subtypes of psychiatric disorders with differing symptomology⁽¹³⁾, illness trajectories^(14, 15), and drug response profiles⁽¹⁶⁾. Adding another layer of complexity are the individual differences inherent in these disorders⁽¹⁷⁾. Ignoring this individuality and modelling at the group level fails to represent the heterogeneity of clinical populations⁽¹⁸⁾. Rather than assuming that all patients are accurately represented by measures of central tendency from case control studies, one solution is to utilize modelling techniques that can parse patient heterogeneity⁽¹⁹⁾.

ML models are capable of this task, by learning individual patient characteristics, they can make successive individual (i.e. single subject) predictions. For example, using a ML model trained on multisite data from the STAR*D consortium, Chekroud et al⁽¹⁶⁾, were able to predict remission from MDD after a 12-week course of Citalopram therapy with an accuracy

of 64.6%. The model was then externally validated in the escitalopram and escitalopram-bupropion treatment group of COMED, attaining accuracies 59.6% and 59.7% respectively. Given a report of approximately 49.3% accuracy for clinician prognostication on the same outcome in the STAR*D cohort⁽¹⁶⁾, this is a clinically meaningful increase in prognostic certainty. This study is an exemplar of applied ML in psychiatry. The dataset had a large number of observations allowing for the learning of unique patient characteristics. Model selection was tailored to the available data and rigorously cross-validated across sites using pipeline architecture. It was multisite, affording geographic generalizability and a large clinical scope. Finally, code for the trained model was made available on request, allowing for transparency and dissemination of the studies methods.

Prior to and since this publication, many more ML works have been published using different data modalities, models, sample sizes, and most interestingly, have reported largely varying degrees of accuracy⁽²⁰⁾. Given this variability, concerns have been raised questioning the veracity of findings in our field and hinted at sources of systematic overestimation^(20, 21). Adding further complication, large multisite datasets are the exception, not the rule, commonly have more predictors than observations, are generally imbalanced, and have low signal to noise ratios. Given these circumstances, concerns have been raised that ML may face the same reproducibility crisis as that experienced by group level analyses in recent years⁽²²⁻²⁴⁾. However, when proper methodology has been employed, ML works have been shown to prognosticate significantly better than chance on unseen data⁽¹⁵⁾, generalise across data collection sites⁽¹⁴⁾, and outperform clinician prognostication^(16, 25).

Given the potential of ML models to transform patient care, albeit, contingent on the rigour of employed methods, we deem it necessary to provide a best practice overview for applied ML in psychiatry. In this guide we will cover issues of best practice for model training and evaluation, sources of systematic error and overestimation, model explainability vs trust, the clinical implementation of AI systems, and finally, future directions for our field.

Model training and evaluation

Sample size and systematic overestimation

An array of model training and testing schemes exist in ML. The choice of which depends on the size of a dataset and a practitioner's computational resources. This brings us to our first encounter with the question of sample size in psychiatric ML. How big should a dataset be before a practitioner decides to use an ML strategy? The ignorance of this question and

application of ML to small datasets has given rise to an important concern: Particularly, ML studies using larger samples have commonly shown weaker performance than studies using smaller samples⁽²⁶⁾. This observation has led to questions regarding the validity and reliability of preliminary small N ML studies in psychiatry⁽²¹⁾. To measure the degree of these effects, Neuhaus & Popescu^(20, 26) collated studies across conditions; including schizophrenia (total observation N = 5563), MDD (N = 2042), and attention deficit hyperactivity disorder (ADHD, N = 8084), finding an inverse relationship between sample size and balanced accuracy (schizophrenia, $r = -.34$, $p = 0.012$; MDD, $r = -.32$, $p = 0.053$; and ADHD, $r = -.43$, $p = 0.044$)⁽²⁰⁾. As we would expect model performance to increase with more data, these findings suggest an underlying problem within our field.

One explanation proposed by Schnack & Kahn⁽²¹⁾ is that patient characteristics in smaller samples tend to be more homogenous. In the case of small N, participants may be more likely to be recruited from the same data collection site and of a similar age (for example, in the case of a university recruited convenience sample). In addition, stringent recruitment criterion may be easily met, resulting in a well-defined phenotype that is not truly representative of the parent population of interest. As sample size increases, the geographic, demographic, and phenotypic diversity of a sample will increase also, resulting in decreased model performance, yet increased generalizability and model scope (see from proof-of-concept studies to clinical application below). Theoretically, future works may be able to circumvent this trade-off by subtyping patients into well-defined and clinically meaningful clusters, thus, models could be trained for specific patient subtypes, maintaining phenotypic homogeneity as sample size increases. These considerations underline the importance of further research into patient subtyping in conjunction with supervised ML methods. In addition to the issue of sample homogeneity, the sample size needed to train an ML model is also contingent on the strength of the underlying signal between input features and an outcome of interest, as well as the complexity of the underlying mapping function (the mathematical function used to derive a line or curve between datapoints). As these two factors can vary greatly between research questions and datasets, there can – in principle – be no general rule of thumb to estimate the sample size required for analyses. However, recent investigations show that for a given research question, we can indeed empirically approximate the number of samples needed for training and testing⁽²⁷⁾. Beyond these sample size and heterogeneity constraints, the choice of cross-validation scheme also bears influence on the variability and bias of accuracy estimates.

In the following sections, we will address the methods available for preliminary ML works and their common use cases.

Leave one out cross-validation

Due to sample size constraints, leave one out cross-validation (LOOCV) has become a popular strategy for model performance evaluation in applied psychiatric ML. With this strategy, a model is trained on all available observations minus one ($n - 1$). Following, the trained model is tested on the one held-out observation. This process is repeated until all available observations have been used for testing. Final model performance is then averaged across the held-out samples and an accuracy estimate is derived. However, previous work has demonstrated the variance properties of LOOCV, suggesting that although this method utilises all available data, an appealing property given limited N , it consequently leads to unstable and biased estimates due to the high degree of correlation between samples⁽²⁸⁾. Therefore, repeated random splits are preferred, showing less bias and variability^(28, 29). Notwithstanding, further complications will arise when we conduct multiple transformations on our data and optimize a models hyperparameters. If we were to use LOOCV, yet, conduct transformations within the same cross-validation scheme, we may optimistically bias our model evaluations by using the same test set ($n - 1$) to select parameters and evaluate the model⁽³⁰⁾. In this instance, we need a method that allows us to take advantage of all our data for both model selection and evaluation whilst avoiding any circularity bias. In these situations, each transformation needs to be completed in a nesting procedure.

Nested cross-validation

As the name implies, nested cross-validation allows for the nesting of multiple cross validations. First, an inner cross-validation loop is used to conduct data transformations and/or hyperparameter optimization. This loop is akin to a train/validation partition. Following, this loop is nested inside an outer cross-validation loop that assesses the transformed data and optimized model on different test sets to those used in the inner loop. This outer loop is akin to a validation/test partition and allows for the approximation of the selected models performance.

Once this architecture is defined, the appropriate number of cross-validation iterations need to be set. Previous work by Kohavi has investigated this topic in detail⁽²⁹⁾. For model selection, he found that 10-fold cross-validation best balanced the bias/variance trade-off. Additionally, repeated runs of the 10-fold cycle were recommended to avoid opportune splits

that may lead to overly optimistic estimates. Three to five repeats are commonly used. Finally, given the balance between computational cost and the bias/variance trade-off, it is common to use another 10-fold cycle in the outer cross-validation loop to assess the final model's performance. See figure 1 for visualization of this structure. It is important to note that, no "optimal" number of repeats can be made as this will likely depend on the complexity of the model, the size and amount of signal in a sample, and the resulting stability of the final cross-validation estimates. If, for example, variance is high between cross-validation folds, more repeats should converge on a less variable mean estimate and reduce dependency on spurious data partitions. On the contrary, if the variance of estimates is low between folds, a high number of repeats will be of less use and increase computational cost. However, whether repeated runs actually achieve this intended variance reduction has been contested⁽³¹⁾. Nonetheless, in small samples it may decrease the effects of favourable cross-validation splits without compromising test set sizes in a k-fold scheme.

While this method provides a means to conduct multiple transformations and estimate parameters without overtly double dipping into a dataset, it is possible that a nested scheme may generate test sets that are overly similar to training sets, as well as similar training sets across cross-validation folds^(30, 32). In theory, this could also lead to systematic overestimation. One alternative, in the case of small N is to keep data transformations and hyperparameter optimization to a minimum⁽³³⁾. In the case of many features, only a few subsets could be tested in the inner loop whilst hyperparameters for the model could be left at default to decrease the risk of double dipping. For an example of this method in Python, see the following documentation⁽³⁴⁾.

Train/test/validate

In the case of a sufficiently large sample, the problem of train/validate/test set overlap in a nested scheme can be avoided by using an entirely separate test partition to the train/validate partitions used for model selection. For example, if a balanced sample of $N = 500$ is available, $N = 300$ could be partitioned for training and validation, whilst $N = 200$ could be held out for final testing. In this case, both sample size over/under estimation and partition overlap could be minimized. Commonly, partitions will be stratified by the outcome label for prediction. Transformations and model parameters will be learnt in the train/validation partitions and analysed in the remaining test partition. For deployment of this method in the Python programming language, see the following documentation⁽³⁵⁾.

Leave group out cross-validation

What about when a sample is multisite, and each data collection site has its own unique characteristics? Further, what if outcome distributions vary across these sites? One site, a treatment centre, may have primarily collected a convenience sample of patients, whilst another may have predominantly focused on controls. When we tune a models hyperparameters, how do we avoid tuning them directly to these differences that may proxy for disparities in outcome distributions or processing equipment? Here, the solution is leave-group out cross-validation, also known as Monte Carlo or leave site out cross-validation^(36, 37). In this situation we need to assess whether a model trained on a particular data collection site, will generalize to other sites. To achieve this, we hold out samples according to a third-party array of integer groups that represent each site. These integer codes can then be used to encode site specific cross-validation folds and assess the generalizability of model performance. This approach ensures that all samples in the cross-validation folds come from sites that are not represented at all in the paired training folds. In situations where consortia/multisite data is used, this method is required. For deployment of this method in the Python programming language, see the following documentation⁽³⁸⁾.

Data leakage

While the use of proper cross-validation schemes and minimum sample sizes helps reduce estimate variability and systematic overestimation, another overlooked source of error comes in the form of data leakage. The effects of leakage on estimates are profound yet appear to be rarely discussed in the psychiatric ML literature. Named one of the top 10 data mining mistakes⁽³⁹⁾, leakage refers to the introduction of information about the outcome label (e.g., case/control) that would otherwise not be available to learn from. A trivial example of data leakage would be either selecting features or imputing missing values on an entire dataset before partitioning it into train/validate/test folds. Here, feature distribution information from the validate/test folds would be leaked into the train set, hyperparameters would be tuned to these distributions, and inevitably, test set outcomes would be predicted with a high degree of accuracy. For an in-depth appraisal of this problem and solutions, see Kaufman⁽³⁹⁾.

This problem can be easily avoided through careful selection of features (only selecting features that would truly be available at time of analysis) and the partitioning of cross-validation folds prior to data transformations. However, as sample size is often constrained in psychiatric cohorts, as discussed above, automated k-fold and nested schemes are commonly

used instead of a-priori train/validate/test partitions. As multiple k-fold runs are conducted, it is not as simple as learning transformations on a train partition and then predicting them into validation/test partitions. Here, data leakage risk increases substantially, yet can still be avoided through the use of pipeline architecture.

Pipeline architecture

A machine learning pipeline can be thought of as an object that sequentially chains together a list of transformers and a final estimator into one object. This sequential chaining has three advantages. First, a practitioner only has to call ‘fit’ and ‘predict’ once on a set of data, rather than at each transformation. Secondly, hyperparameters from each estimator can be tuned in unison. Finally, pipelines help avoid the leakage of statistics by guaranteeing that the same samples are used to train the transformers and the final classifier. It is easy to overlook the importance of this final step. As datasets are commonly repositioned for ML analysis, the *a-priori* pre-processing of data sets is common. As demonstrated above, if a transformation as simple as imputation is completed on a dataset in its entirety, this is enough to leak statistics and cause optimistic bias.

To demonstrate the correct use of pipeline architecture in the Python programming language and quantify the magnitude of data leakage effects, we have provided an example Python script in the references⁽⁴⁰⁾. In this example, we randomly generate a balanced large P small N dataset containing 3,000 features and 500 observations. To emulate the low signal to noise ratio commonly inherent in psychiatric cohorts, only 10 of the 3,000 features are related to the binary outcome vector $y = \{1,0\}$. To demonstrate data leakage effects, we first mean centre the dataset and select a subset of features using regularized logistic regression (LASSO)⁽⁴¹⁾ on the full dataset. Following, we train and test a linear SVM with default parameters using 10-fold cross-validation, attaining a test set AUC of 99.89. Next, we conduct the same procedure, now implementing each transformation within a sklearn pipeline to ensure the use of the same cross-validation folds over each transformation. Now, we attain a test AUC of 50.23. Here, we see that when a pipeline is not used, the model overfits to leaked statistics and appears to be a near perfect classifier. However, when we use a pipeline, conduct transformations in the same folds, and only learn from the true signal in the features, predictive accuracy is no better than chance.

Given that at current, the open sourcing of code for peer review is not requested by journals, and the ease of which these mistakes can be made by those newer to the field, it is possible

that beyond just sample size effects, many of the highly optimistic studies ($\geq 90\%$) may be due to data leakage. To counteract this problem, we recommend that journals require code reviews in the peer review process. In addition, the open sourcing of code should be encouraged. Beyond the aforementioned sample size estimation biases, such minor transgressions in code structure may go some way to explain the large degree of variability currently observed in the literature. Finally, permutation tests should be conducted regardless of cross-validation strategy deployed. This way, the null-distribution and statistical significance of a classifier can be obtained⁽⁴²⁾.

From proof-of-concept studies to clinical application

While consortia efforts and open-sourced models are becoming more prevalent, publication numbers of studies at all points along the project maturity continuum continue to rise. Of utmost importance, ensuring methodological rigor and conceptual maturity of these publications is essential. However, beyond best practice recommendations like those provided above, the current lack of practical guidelines to systematically evaluate ML quality and maturity makes it difficult for researchers, stakeholders, journals, and funding agencies to objectively gauge the quality and current clinical utility of an ML model or publication. In addition, the lack of evaluation guidelines risks an overly optimistic or an unduly skeptical perception of findings – both in the scientific community and the public eye. Therefore, we propose a practical set of guidelines for the assessment of clinical utility and maturity for ML models in psychiatry. Based on the conceptual framework of *AI Transparency*⁽³²⁾, we have derived a straightforward “checklist” with which to quantify a project’s maturity ranging from the initial proof-of-concept stage through to the clinical application stage. Specifically, the checklist comprises six categories in which an ML project is evaluated. In each category, scores range from “proof-of-concept stage” (0) to “ready for clinical application” (2).

The first category – generalization – refers to model performance in previously unseen data and has been outlined in detail above. It constitutes the most rudimentary performance measure of an ML model. While employing cross-validation techniques (score 0) avoids data leakage and provides a principled estimate of generalization performance, it is important to note that (nested) cross-validation should be used for initial model evaluation and hyperparameter optimization only. At current, the results of most machine learning studies in psychiatry might well have arisen from small test-sets as is typical for cross-validation. In contrast, using a large, independent test set (score 1) yields a more stable, reliable estimate of

future performance. Finally, using a large, external test set (score 2), i.e. a test set to which the creators of the initial model did not have access at training time, and which was drawn independently from the training set, is optimal. To this end, online model repositories (e.g. www.photon-ai.com/repo) provide valuable infrastructure which greatly simplifies external validation in practice. Here, a research group can make available a trained model, and have it tested on independent data, providing an opportunity to assess the geographic, demographic, and phenotypic generalisability of a published model. In addition, incremental learning can be used to train certain ML models, allowing a research group to further train a pre-trained model with their own data, affording cross institution collaboration without the need for data sharing (for more information, see the partial fit method in sklearn⁽⁴³⁾).

The second category – model scope – refers to the group of individuals about whom the model can make reasonable predictions. If a convenience sample (score 0) is used for training and testing, we cannot expect the model to perform well on different samples. If the sample is representative for a site or local subpopulation (score 1), we can expect it to perform as intended at this site without any guarantees with regard to other sites or subpopulations. Only testing on a representative sample of the target population (score 2) ensures reliable performance estimates that translate into clinical practice. From this point of view, the current use of exclusion criteria appears highly problematic. While perfectly reasonable if we seek to test hypotheses to gain insight into mechanisms or advance theory, excluding patients with e.g. certain comorbidities inevitably entails that our model's utility for this patient group cannot be estimated, thereby severely hampering its applicability in clinical practice.

The third category – incremental utility – refers to the added value a machine learning model confers as compared to current practice⁽⁴⁴⁾. While most studies do not assess incremental utility (score 0), showing higher efficiency or effectiveness with regard to the current state-of-the-art (score 1) is essential. If a model or project cannot show or does not intend to do this, little in the way of clinical translation can be expected. Thus, the essential goal should always be that a project reaches a stage in which it outperforms current state-of-the-art in real-life workflow (score 2). Although experimental approaches such as randomized controlled trials seem optimally suited to show incremental utility, they are rarely used in medical ML research today. This concept also highlights a misalignment of expectations with regards to model performance. A commonly held view is that a model needs to be highly accurate to be of clinical use. However, if a model can outperform its opportunity cost, that is, current state of the art clinical practice free of decision support systems, then patient utility

can be maximized at scale. As previously mentioned, clinician rated accuracy to predict remission in the STAR-D cohort was approximately 49.3%, whilst that of a weak to moderately accurate AI model ranged between 59.6% to 64.6% accuracy. Whilst far from perfect, a clinically meaningful increase in prognostic certainty was attained. Therefore, the absolute accuracy of a classifier should not serve as an indicator of clinical utility, but the relative increase in prognostication compared to current state of the art practice.

Using this simple checklist, we can now evaluate a project or publication with regard to its position on the continuum spanning from “proof-of-concept stage” to “ready for clinical application”. Note that the three categories outlined above build upon each other in the sense that a model with bad generalization cannot have a broad model scope. Likewise, to show incremental utility, a model must generalize well and have reasonable model scope. Thus, insufficient scores in one category cannot be compensated by higher scores in any other category. See Figure 2 for an illustration of this general workflow.

Once an ML model reaches the level we term “ready for clinical application”, other considerations regarding post-deployment evaluation, security, and algorithmic bias come into focus⁽³²⁾. While often crucially important, their impact depends heavily on the context in which the model is used. If, for example, we could be sure that the model draws on measures of causal mechanisms, we can assume that the relationship will not change over time, rendering post-deployment evaluation less important. Also, an ML model deployed in a consumer application for smartphones might require much higher security (e.g. regarding adversarial attacks and data security, i.e., hacking of smartphone input data leading to erroneous model outputs and therapeutic decisions) than is necessary in a closed hardware-based device used in the clinic only. Finally, as ML models are not programmed, but trained, they will mimic systematic biases inherent in their training data. This potential algorithmic bias must be carefully investigated. While it can often be eradicated by retraining without the variables causing the bias, we need to be aware of a bias to be able to remove it. Thus, human involvement is crucial at this point (for an in-depth discussion, see⁽⁴⁵⁾).

Importantly, these guidelines are intended to gauge ML model maturity with regard to clinical application only. It should explicitly not be applied to ML research projects developing methods and proof-of-concept studies. Judging those with the checklist outlined here would inevitably result in low scores and might therefore stifle even the most promising of methodological developments. It is this creativity and ingenuity, however, that allows the

field to move at such breathtaking speed. If researchers claim to develop a clinically useful tool – as is done in the introductions and discussion sections of countless publications and funding proposals, even if nothing but classic statistical group inference or the simplest of ML models are planned – we can judge project planning as well as subsequent results using this approach to delineate projects aiming for ML with high clinical utility from studies seeking insight into mechanisms or proof-of-concept studies. The ability of journals and funding agencies as well as the general public to evaluate the aims and quality of research projects in this manner is crucial for all translational efforts in psychiatry.

Understanding ML Models

In the wake of ever-more powerful machine learning applications emerging across a range of industries, the question of explainability – i.e., understanding which (patterns of) variables lead to which model predictions – has increasingly come into focus. Identifying the (pattern of) variables driving predictions is of obvious scientific interest. Extracting these relevant variables in an ML study would provide theoretic insights similar to classical (usually univariate) statistical inference while retaining predictive performance. In addition, if the relevant variables could be extracted, equally well-performing models could be trained with only a small subset of variables, saving resources on all levels from data acquisition to processing and storage.

With regard to clinical evaluation as outlined in the previous section, quantifying the effect of variables can also be beneficial in at least three ways: First – with regard to model utility – identifying relevant variables might help to detect trivial and erroneous models. For example, Lapuschkin et al.⁽⁴⁶⁾ used Layer-wise Relevance Propagation (LRP)⁽⁴⁷⁾ to show that an ML model trained to detect certain objects in photographs (in this case horses) in fact used nothing but the text of a tag present on all images of horses in the training data. If this tag was inserted into other photographs, the images would be classified as horses independent of their actual content. In addition, identifying relevant variables might enable domain experts to judge whether the associations on which the model relies are likely to remain stable. Second – with regard to model fairness – knowledge of the relevant variables may help to identify algorithmic bias. If, for example, gender and age are identified as the most relevant variables in a model built to identify suitable job applicants, this bias could be explicitly addressed. Third – with regard to model security – identifying the relevant variables provides

information on where the model could be most easily attacked. This, in turn, might help to immunize the model.

Given that (almost all) ML models are by no means “black boxes”, but apply a transparent and deterministic, albeit often rather complex rule to make predictions, a large number of algorithms aiming to help us understand this rule, i.e. to identify relevant variables, have been developed (for a general introduction, see⁽⁴⁸⁾). While it is beyond the scope of this article to review the numerous families of such algorithms, they usually quantify the contribution of each individual variable used in a prediction. This can be done, for example, in a straightforward manner by systematically obscuring certain sets of variables in novel data and analyzing the resulting decrease in performance. Another group of approaches investigates the trained ML model itself. Examples of this include simple weight-maps for Support Vector Machines, decision tree-based importance scores in tree-based models, as well as more complex approaches, such as visualizing the process of layer-wise data transformation in neural networks, including LRP⁽⁴⁷⁾, or more generally applicable approaches e.g. based on Game Theory⁽⁴⁹⁾. While the wealth of research in this area has vastly increased the toolbox available for model interpretation in recent years, it also indicates that no complete solution has been found. In fact, different approaches may lead to different variables being identified.

However, the problems with ML model explainability run much deeper: On the one hand, all algorithms which identify variables driving model prediction require human interpretation. Given the multivariate nature of ML models, this renders such an interpretation extremely difficult for even the simplest of (linear) models. On the other hand, the insight into model utility, fairness, and security we can gain is not only very limited but can usually be accomplished much better by straightforward model evaluation. For example, while quantifying variable relevance might indicate that problematic variables (such as gender or age) are driving predictions, explicitly excluding these variables in the first place is much more effective. However, it is possible that seemingly bias free variables may be confounded even after the exclusion of problematic ones. Further analyses may be required to elucidate such hidden confounds prior to selecting variables for model consideration. Further work in this area is required and ongoing.

While we believe that complete transparency regarding the previously outlined checklist to be essential and facilitated by practices such as variable assessment for bias and the sharing

of code, data, and trained models for scientific reproducibility, we deem the disclosure of “the algorithm” itself (specifically, the mathematical underpinnings of the ML model) to be of no use. In fact, even knowing every single one of the hundreds of millions of parameter values in a given ML model would fail to provide even a spec of practically useful insight into the inner workings of a trained model. While we could re-create every decision, we would have no additional way to investigate its quality – much less assess its real-world impact. The disclosure of detailed information regarding generalization, model scope, and risk profiles, however, ensures utmost transparency.

Future goals

Automatized ML Development and Optimization

The success of ML model development in practice crucially relies on machine learning experts to pre-process data (including e.g. imputation and cleaning), select and construct appropriate features (feature engineering; often with the help of domain experts), select an appropriate model family (e.g. neural networks or random forests etc.), optimize model hyperparameters (e.g. learning rate), and evaluate the model with regard to generalization, model scope, and risk profiles (including incremental utility, fairness, and security as outlined above). The complexity of this generic ML development pipeline as well as the necessity to avoid data leakage and ensure proper evaluation can be challenging even for experienced researchers. Indeed, many studies forego optimizing the ML model development pipeline entirely. Indeed Arbabshirani et al.⁽⁵⁰⁾ showed in a recent review of ML studies in the area of neuroimaging that 73% of studies employed a single machine (namely the Support Vector Machine) and almost no study optimized hyperparameters even for that single machine. This is particularly disconcerting as the *No Free Lunch Theorem*⁽⁵¹⁾ in ML clearly indicates that, without further knowledge of the process generating our data, no ML algorithm will, overall, perform better than any other. Even without any mathematical considerations, it is quite obvious that employing a single machine makes it extremely likely that another ML pipeline, algorithm, or setting thereof might have performed better. However, it may be possible that there are re-occurring data structures for which particular algorithms, pipelines, and hyperparameter values are optimal (e.g., default SVM parameters for variance normalized neuroimaging datasets⁽²⁸⁾).

Against this background, we see increasing efforts to automatize the entire ML development process, from data pre-processing and feature engineering to model selection and

hyperparameter optimization. Examples for this include auto-sklearn⁽⁵²⁾ (a package focused on Bayesian hyperparameter optimization and machine selection), PHOTON (www.photon-ai.com; an ML framework enabling cross-platform ML pipeline construction, optimization, and evaluation), and Auto Keras⁽⁵³⁾ (an open-source package based on TensorFlow for neural network architecture search). We expect this trend to accelerate in the years to come, automatizing most if not all ML model development steps. This, however, will increase the need for proper model evaluation and full *AI Transparency* as outlined above.

Learning Complex Models from Small Samples

While exceptionally successful in many areas, ML model training may require large amounts of data as models comprise at least as many free parameters as there are variables measured⁽⁵⁴⁾. For complex models (e.g. Deep Learning), the number of parameters may easily increase to tens of millions of parameters⁽⁵⁵⁾. Training such large models with “only” hundreds or even thousands of samples may induce so-called overfitting – a situation in which the large number of free parameters allows the model to essentially “memorize” all of the training data, leading to perfect performance on the training set, but extremely bad generalization to new, real-world data. While generally effective, the numerous countermeasures employed lower the complexity of what a model can learn, potentially rendering it unable to capture true associations in the data⁽⁵⁶⁻⁵⁸⁾. Thus, model performance is severely limited by the number of patients available, especially whenever high dimensional data sources such as neuroimaging data are of interest (commonly known as large P small N problems)⁽⁵⁹⁾. Acquiring hundreds of thousands of patient samples, however, is usually not feasible – especially in psychiatry where comprehensive phenotype data is often crucial.

The fundamental problem of model training on small datasets has recently been addressed in other areas with great success: First, using data augmentation – i.e. perturbing existing data to create new samples – as a mathematically principled means to artificially enlarge training set size, enables the generation of an arbitrarily large number of training samples using stochastic and image processing methodology custom-tailored for imaging data^(60, 61) (including e.g. image synthesis, sample pairing techniques etc.; Figure 3a). Second, transfer learning has been used to transform variables into a lower dimensional representation based on what has been learnt from other data sets⁽⁶²⁻⁶⁴⁾. In neuroimaging, for example, we could leverage a cross-domain transfer learning approach by extracting basic visual features of a pre-trained image classification algorithm (i.e. a Convolutional Neural Network, CNN)

trained on 1.2 million natural images (Imagenet⁽⁶⁰⁾, Figure 3b). In the same vein, employing intra-domain transfer learning, we can extract general statistical properties from large datasets of healthy controls using state-of-the-art Generative Adversarial Neural Networks (GAN, Figure 3c). Such GANs can, for example, represent the MRI data on a lower-dimensional manifold and enable the generation of an arbitrary number of MRI images from this distribution. While highly effective, these techniques have thus far not been systematically applied in psychiatry.

From ML models to clinical decision support systems

At current, applied ML in psychiatry shows promise yet is still in its early days. Once best practice is attained and proof-of-concept studies have been conducted, forward testing will be necessary to demonstrate prognostic stability, incremental utility, and real-world estimates of performance in the same context as they will be clinically deployed. In such a forward test, a clinician could make their prediction and clinical decision (e.g., will a patient enter remission if they prescribe a certain drug? If so, prescribe drug). In parallel, a trained model could also make its prediction, with both the model and clinician assessed at a 12-week endpoint. However, given the interpersonal nature of psychiatric care, it is unlikely that even if ML models prove to outperform clinician prognostication, they will ever solely drive the clinical decision-making process. Therefore, the testing of AI decision support systems alongside clinicians will likely provide a more realistic approximation of what to expect in terms of socially accepted clinical use. In this third study arm, a clinician could make a prediction and decision not only based off of their own clinical expertise, but the binary prediction of an ML model as well as its predicted probabilities⁽⁶⁵⁾. Here, the synergy between a clinician and an AI decision support system could be measured. In this case, not only the binary prediction but the probabilistic estimates are of importance. Therefore, model calibration should also be carefully considered (see Niculescu-Mizil & R. Caruana⁽⁶⁶⁾ and the sklearn documentation⁽⁶⁷⁾). If the collaboration of the clinician and the AI system significantly outperform clinician prognostication alone, the model could then move towards clinical deployment.

Psychiatry and beyond

Whilst the current overview has focused on the application of ML to psychiatric phenotypes, these recommendations will generalize to most areas of applied ML in medicine. However, as psychiatric disorders are defined based on deviations of phenotypic characteristics – not

causal biological models – a disorder can be highly heterogeneous with regard to its biological underpinnings. Therefore, diagnostic labels in psychiatry likely contain noise not seen in other medical disciplines. This presents a set of unique challenges that are distinct to psychiatry, with implications for model selection, accuracy, reliability, reproducibility, and ceiling effects on model performance. In areas that do not face these challenges, for example, oncology, specific problems discussed, e.g., sample sized based systematic overestimation and overfitting risks may be less prevalent. Therefore, domain application should always be considered. Moving forward, only by parsing this heterogeneity in psychiatric phenotypes will we be able to lay the groundwork for more targeted models and improved patient care^(68, 69).

Conclusion

In this overview, we have suggested and discussed best-practice guidelines with the intention that they may help stakeholders, journals, and funding agencies to obtain a more realistic view of 1) what can be expected of a planned research project in terms of clinical utility, 2) how much closer a particular finding has brought us to clinical application, and 3) what remains to be done before we can expect improvements in daily practice. In addition, understanding how to develop, train, and evaluate ML models and publications might help researchers new to the field of ML in psychiatry to better plan and monitor their ML projects, creating robust best-practice procedures in the mid-term. In the long run, we hope that these guidelines can help to channel funding as well as media attention towards the most promising developments regarding improved patient care, circumventing the evident dangers of the current hype around machine learning and artificial intelligence.

Acknowledgments

This work was supported by the Australian Government Research Training Program Scholarship.

Conflicts of interest

None

References

1. Bzdok D, Meyer-Lindenberg A. Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2018 Mar;3(3):223-30. DOI: 10.1016/j.bpsc.2017.11.007.
2. Engel GL. The Clinical-Application of the Biopsychosocial Model. *J Med Philos*. 1981;6(2):101-23. DOI: DOI 10.1093/jmp/6.2.101.
3. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. *Nat Methods*. 2018 Apr;15(4):233-4. DOI: 10.1038/nmeth.4642.
4. AL-Asadi AM, Klein B, Meyer D. Multiple comorbidities of 21 psychological disorders and relationships with psychosocial variables: a study of the online assessment and diagnostic system within a web-based population. *Journal of Medical Internet Research*. 2015;17(3).
5. Anker E, Bendiksen B, Heir T. Comorbid psychiatric disorders in a clinical sample of adults with ADHD, and associations with education, work and social characteristics: a cross-sectional study. *BMJ Open*. 2018 Mar 2;8(3):e019700. DOI: 10.1136/bmjopen-2017-019700.
6. Strawbridge R, Young AH, Cleare AJ. Biomarkers for depression: recent insights, current challenges and future prospects. *Neuropsychiatr Dis Treat*. 2017;13:1245-62. DOI: 10.2147/NDT.S114542.
7. Yahata N, Kasai K, Kawato M. Computational neuroscience approach to biomarkers and treatments for mental disorders. *Psychiatry Clin Neurosci*. 2017 Apr;71(4):215-37. DOI: 10.1111/pcn.12502.
8. Cross-Disorder Group of the Psychiatric Genomics C, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*. 2013 08/11/online;45:984. DOI: 10.1038/ng.2711
<https://www.nature.com/articles/ng.2711#supplementary-information>.
9. Del Boca FK, Noll JA. Truth or consequences: the validity of self-report data in health services research on addictions. *Addiction*. 2000 Nov;95 Suppl 3:S347-60.

10. Nguyen TMU, Caze AL, Cottrell N. What are validated self-report adherence scales really measuring?: a systematic review. *British journal of clinical pharmacology*. 2014;77(3):427-45.
11. Yarkoni T, Westfall J. Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspect Psychol Sci*. 2017 Nov;12(6):1100-22. DOI: 10.1177/1745691617693393.
12. Kirsch I, Sapirstein G. Listening to Prozac but hearing placebo: A meta-analysis of antidepressant medication. *Prevention & Treatment*. 1998;1(2):2a.
13. Chekroud AM, et al. Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. *JAMA psychiatry*. 2017;74(4):370-8.
14. Koutsouleris N, et al. Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *Lancet Psychiatry*. 2016 Oct;3(10):935-46. DOI: 10.1016/S2215-0366(16)30171-7.
15. Redlich R, et al. Prediction of Individual Response to Electroconvulsive Therapy via Machine Learning on Structural Magnetic Resonance Imaging Data. *JAMA Psychiatry*. 2016 Jun 1;73(6):557-64. DOI: 10.1001/jamapsychiatry.2016.0316.
16. Chekroud AM, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry*. 2016 Mar;3(3):243-50. DOI: 10.1016/S2215-0366(15)00471-X.
17. Sackett PR, Lievens F, Van Iddekinge CH, Kuncel NR. Individual differences and their measurement: A review of 100 years of research. *Journal of Applied Psychology*. 2017;102(3):254.
18. Speelman CP, McGann M. Editorial: Challenges to Mean-Based Analysis in Psychology: The Contrast Between Individual People and General Science. *Frontiers in psychology*. 2016;7:1234-. DOI: 10.3389/fpsyg.2016.01234. Cited in: PubMed.
19. Chekroud AM, Lane CE, Ross DA. Computational Psychiatry: Embracing Uncertainty and Focusing on Individuals, Not Averages. *Biol Psychiatry*. 2017 Sep 15;82(6):e45-e7. DOI: 10.1016/j.biopsych.2017.07.011.

20. Neuhaus AH, Popescu FC. Sample Size, Model Robustness, and Classification Accuracy in Diagnostic Multivariate Neuroimaging Analyses. *Biol Psychiatry*. 2018 Dec 1;84(11):e81-e2. DOI: 10.1016/j.biopsych.2017.09.032.
21. Schnack HG, Kahn RS. Detecting Neuroimaging Biomarkers for Psychiatric Disorders: Sample Size Matters. *Front Psychiatry*. 2016;7:50. DOI: 10.3389/fpsyt.2016.00050.
22. Hutson M. Artificial intelligence faces reproducibility crisis. *Science*. 2018 Feb 16;359(6377):725-6. DOI: 10.1126/science.359.6377.725.
23. Open Science C. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science*. 2015 Aug 28;349(6251):aac4716. DOI: 10.1126/science.aac4716.
24. Baker M. Over half of psychology studies fail reproducibility test. *Nature News*. 2015.
25. Koutsouleris N, et al. Prediction Models of Functional Outcomes for Individuals in the Clinical High-Risk State for Psychosis or With Recent-Onset Depression: A Multimodal, Multisite Machine Learning Analysis. *JAMA Psychiatry*. 2018 Nov 1;75(11):1156-72. DOI: 10.1001/jamapsychiatry.2018.2165.
26. Kambeitz J, et al. Reply to: Sample Size, Model Robustness, and Classification Accuracy in Diagnostic Multivariate Neuroimaging Analyses. *Biol Psychiatry*. 2018 Dec 1;84(11):e83-e4. DOI: 10.1016/j.biopsych.2018.01.023.
27. Kaehler CC, M. Opel, N. Redlich, Mehler, D. Emden, D. Winter, N. Leenings, R. Eickoff, S. Kircher, T. Krug, A. Nenadic, I. Arolt, V. Clark, S. Baune, B. Jiang, X. Dannlowskir, U. Hahn, T. Systematic Overestimation of Machine Learning Accuracies in Structural Neuroimaging Studies of Depression. 2019.
28. Varoquaux G, et al. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage*. 2017;145:166-79.
29. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Ijcai*. 1995: Montreal, Canada; p. 1137-45.
30. Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res*. 2010;11(Jul):2079-107.

31. Vanwinckelen G, Blockeel H. On estimating model accuracy with repeated cross-validation. In: Proceedings of the 21st Belgian-Dutch Conference on Machine Learning. 2012 p. 39-44.
32. Hahn T, Ebner-Priemer U, Meyer-Lindenberg A. Transparent Artificial Intelligence—A Conceptual Framework for Evaluating AI-based Clinical Decision Support Systems. Available at SSRN 3303123. 2018.
33. Varoquaux G, et al. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. Neuroimage. 2017 Jan 15;145(Pt B):166-79. DOI: 10.1016/j.neuroimage.2016.10.038.
34. Pedregosa F, et al. Nested versus non-nested cross-validation [Webpage]. Scikit learn documentation: Scikit learn; 2019 [cited 2019 03/07/2019]. Explanation and code for nested cross-validation. Available from: https://scikit-learn.org/stable/auto_examples/model_selection/plot_nested_cross_validation_iris.html
35. Pedregosa F, et al. Train, test, split. 2019 [cited 2019 03/07/2019]. Train/test/split cross-validation documentation. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
36. Roberts DR, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017;40(8):913-29.
37. Xu Q-S, Liang Y-Z. Monte Carlo cross validation. Chemometrics and Intelligent Laboratory Systems. 2001;56(1):1-11.
38. Pedregosa F, et al. Leave one group out cross-validation. 2019 [cited 2019 03/07/2019]. Leave one group out cv python code example in scikit learn. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.LeaveOneGroupOut.html#sklearn.model_selection.LeaveOneGroupOut
39. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in Data Mining: Formulation, Detection, and Avoidance. Acm T Knowl Discov D. 2012 Dec;6(4). DOI: Artn 15 10.1145/2382577.2382579.

40. Cearns M. Code based data leakage gist for Translational Psychiatry. 2019 [cited 2019 03/07/2019]. Available from: <https://gist.github.com/Micah0808/6d9e4d0919c9f43dcb3e53d21f405c97>
41. Tang J, Alelyani S, Liu H. Feature selection for classification: A review. *Data classification: algorithms and applications*. 2014:37.
42. Ojala M, Garriga GC. Permutation Tests for Studying Classifier Performance. *J Mach Learn Res*. 2010 Jun;11:1833-63.
43. Pedregosa F, et al. Strategies to scale computationally: bigger data. 2019. Available from: https://scikit-learn.org/0.15/modules/scaling_strategies.html
44. Gabrieli JD, Ghosh SS, Whitfield-Gabrieli S. Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron*. 2015;85(1):11-26.
45. FAT M. Fairness, Accountability, and Transparency in Machine Learning. Retrieved December. 2018;24:2018.
46. Lapuschkin S, et al. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications*. 2019;10(1):1096.
47. Bach S, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*. 2015;10(7):e0130140.
48. Molnar C. Interpretable machine learning: A guide for making black box models explainable. Christoph Molnar, Leanpub. 2018.
49. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*. 2017 p. 4765-74.
50. Arbabshirani MR, Plis S, Sui J, Calhoun VD. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage*. 2017 Jan 15;145(Pt B):137-65. DOI: 10.1016/j.neuroimage.2016.02.079.
51. Wolpert DH, Macready WG. No free lunch theorems for search: Technical Report SFI-TR-95-02-010, Santa Fe Institute; 1995.
52. Feurer M, et al. Efficient and robust automated machine learning. In: *Advances in neural information processing systems*. 2015 p. 2962-70.

53. Jin H, Song Q, Hu X. Efficient neural architecture search with network morphism. arXiv preprint arXiv:1806.10282. 2018.
54. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Springer series in statistics. :: Springer; 2001.
55. Goodfellow I, Bengio Y, Courville A. Deep learning: MIT press; 2016.
56. Hahn T, et al. Integrating neurobiological markers of depression. Arch Gen Psychiatry. 2011 Apr;68(4):361-8. DOI: 10.1001/archgenpsychiatry.2010.178.
57. Rondina JM, et al. SCoRS-A Method Based on Stability for Feature Selection and Mapping in Neuroimaging (vol 33, pg 85, 2014). Ieee T Med Imaging. 2014 Mar;33(3):794-. DOI: 10.1109/Tmi.2014.2307811.
58. Hahn T, et al. Predicting Treatment Response to Cognitive Behavioral Therapy in Panic Disorder With Agoraphobia by Integrating Local Neural Information. Jama Psychiatry. 2015 Jan;72(1):68-74. DOI: 10.1001/jamapsychiatry.2014.1741.
59. Arbabshirani MR, Plis S, Sui J, Calhoun VD. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. Neuroimage. 2017 Jan 15;145:137-65. DOI: 10.1016/j.neuroimage.2016.02.079.
60. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Communications of the Acm. 2017 Jun;60(6):84-90. DOI: 10.1145/3065386.
61. Inoue H. Data augmentation by pairing samples for images classification. arXiv preprint arXiv:1801.02929. 2018.
62. Esteva A, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017 Feb 2;542(7639):115-8. DOI: 10.1038/nature21056.
63. Cheng B, et al. Multimodal manifold-regularized transfer learning for MCI conversion prediction. Brain Imaging Behav. 2015 Dec;9(4):913-26. DOI: 10.1007/s11682-015-9356-x.
64. Donahue J, Krähenbühl P, Darrell T. Adversarial feature learning. arXiv preprint arXiv:1605.09782. 2016.

65. Hahn T, et al. A novel approach to probabilistic biomarker-based classification using functional near-infrared spectroscopy. *Human Brain Mapping*. 2013 May;34(5):1102-14. DOI: 10.1002/hbm.21497.
66. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: *Proceedings of the 22nd international conference on Machine learning*. 2005: ACM; p. 625-32.
67. Pedregosa F, et al. Probability calibration. 2019. Available from: <https://scikit-learn.org/stable/modules/calibration.html>
68. Marquand AF, Wolfers T, Mennes M, Buitelaar J, Beckmann CF. Beyond Lumping and Splitting: A Review of Computational Approaches for Stratifying Psychiatric Disorders. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2016 Sep;1(5):433-47. DOI: 10.1016/j.bpsc.2016.04.002.
69. Marquand AF, Rezek I, Buitelaar J, Beckmann CF. Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biological psychiatry*. 2016;80(7):552-61.

Figure 1. Visualization of a nested cross-validation scheme. All steps from 2a – 2c should be conducted inside a pipeline, inside the inner cross-validation loop.

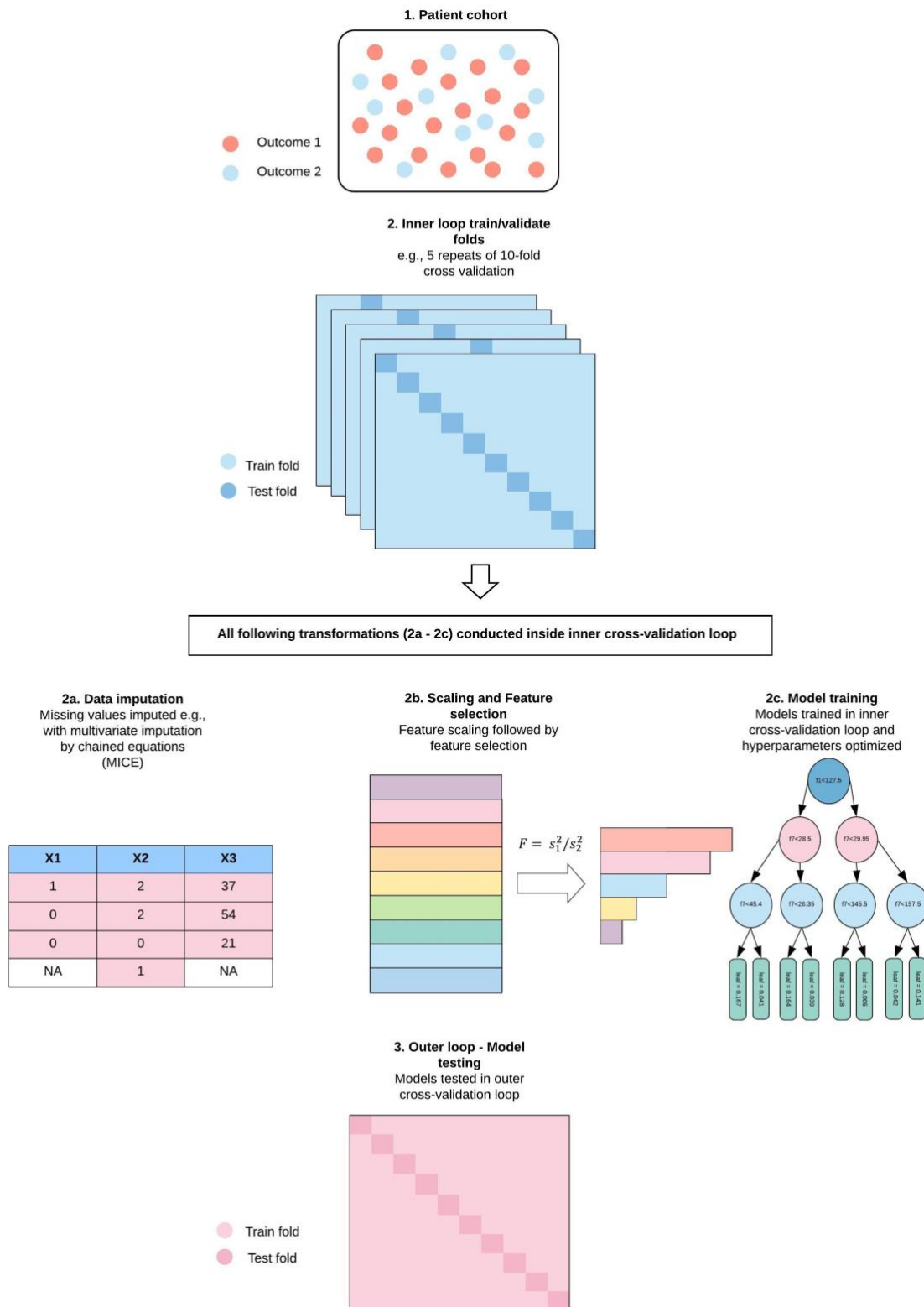


Figure 2. Illustration of the full best practice workflow from pipeline construction through to project maturity assessment.

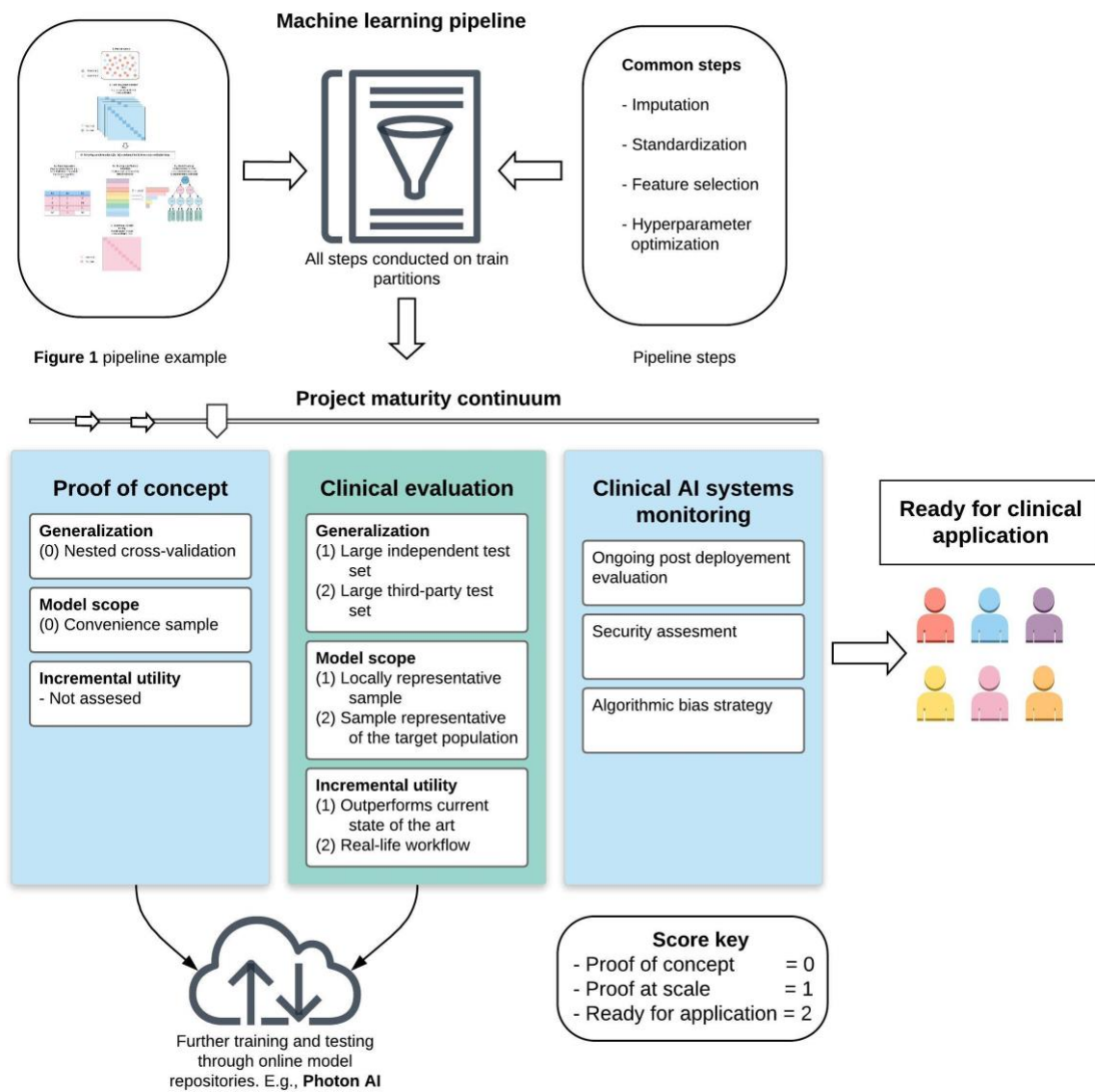
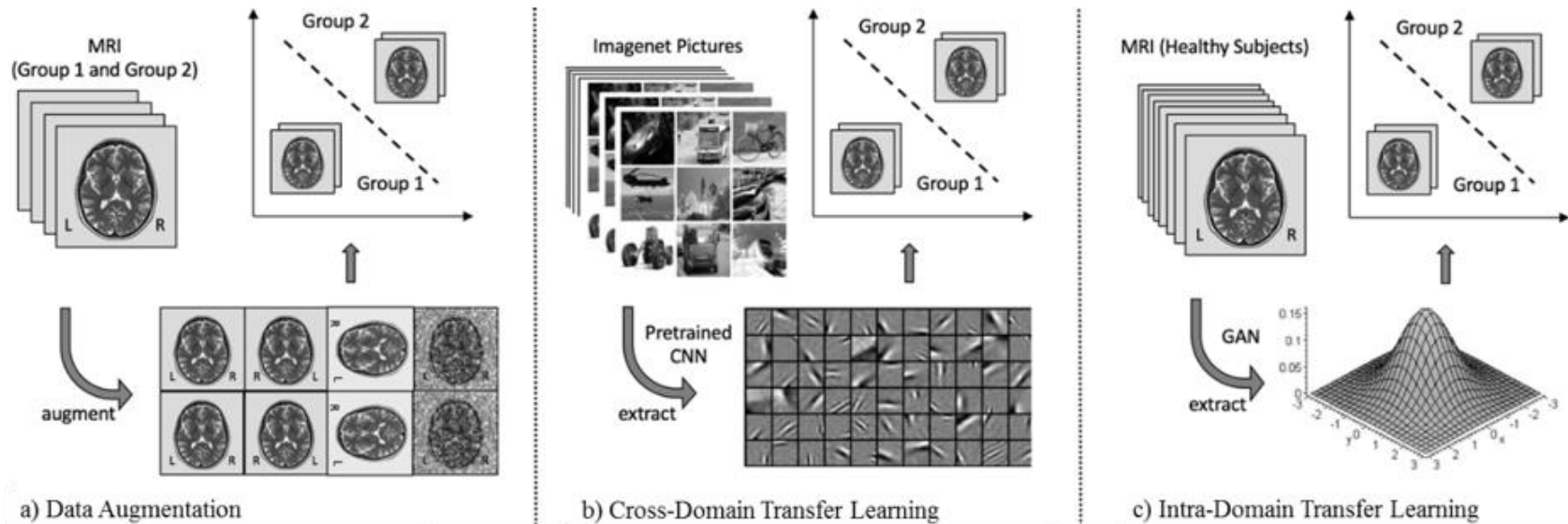


Figure 3. Illustration of workflows for the different techniques exemplified using Magnetic Resonance Imaging (MRI) data. (a) Data augmentation approach using stochastic and image processing methodology. (b) Cross-domain Transfer Learning applying low-level filters learnt by a Convolutional Neural Network (CNN) from the Imagenet database. (c) Intra-domain Transfer Learning deriving a statistical embedding from a large database of MRI images employing a Generative Adversarial Network (GAN).



Statement of Authorship

Title of Paper	Machine learning probability calibration for high-risk clinical decision-making
Publication Status	☒ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Cearns, M., Hahn, T., Clark, S., & Baune, B. T. (2020). Machine learning probability calibration for high-risk clinical decision-making.

Principal Author

Name of Principal Author (Candidate)	Micah Cearns		
Contribution to the Paper	Formulated the idea for the paper, wrote the manuscript, ran the analysis, and created the figure.		
Overall percentage (%)	95%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	15/01/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Tim Hahn		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision.		
Signature		Date	15/01/2021
Name of Co-Author	Scott Clark		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision.		

Signature		Date	15/01/2021

Name of Co-Author	Bernhard Baune		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision. As corresponding author, he made communications with journal editors.		
Signature		Date	15/01/2021

Please cut and paste additional co-author panels here as required.

Chapter 3

Machine learning probability calibration for high-risk clinical decision making

Micah Cearns, BPsy (Hons)¹, Tim Hahn, PhD², Scott Clark, PhD¹, and Bernhard T. Baune, PhD^{1,2,3,4}

¹ Discipline of Psychiatry, School of Medicine, University of Adelaide, Australia

² Department of Psychiatry, University of Münster, Germany

³ Department of Psychiatry, Melbourne Medical School, The University of Melbourne, Melbourne, VIC Australia

⁴ The Florey Institute of Mental Health, The University of Melbourne, Melbourne, VIC Australia

Australia and New Zealand Journal of Psychiatry

A clinician is seeing a patient named James for the first time in years. James suffers from bipolar type two, yet due to side effects from his first medication, sodium valproate, he has chosen to self-manage his symptoms with the help of a psychologist. However, James's partner has recently fallen ill, and their first child is due in two months. He has been sleeping less and working more, whilst the dynamic range of his highs and lows has continued to widen. For the first time since diagnosis, the only form of respite presents itself in ego dystonic thoughts of harm. Given his first and only experience with psychiatric medication, the hope of reprieve from a new class of drug seems like an unlikely but necessary panacea. Whilst apprehensive, James decides to reach back out to his psychiatrist and discuss trying a new medication.

Scenarios such as these are all too common in clinical care. High risk situations where a patient is in need of fast and effective treatment. Unfortunately, this need often goes unmet, with patient adherence to first treatments often wavering. If attrition rates in the second and third treatment levels of the sequenced treatment alternatives to relieve depression (STAR*D) study are of any indication, patient hopes for successful treatment are quick to fade (Rush et al., 2006). In the face of this dilemma, two solutions are commonly proposed. First, the discovery of new pharmacological, biological or psychological interventions with novel mechanisms of action. Second, personalizing treatment selection for the interventions that we already have through the use of machine learning (ML) methodologies (Chekroud et al., 2016). In recent years, the utility of this approach has been demonstrated and a greater dearth of prognostic outcomes from suicide prediction to rehospitalization have been explored.

One area of psychiatry that has been quick to adopt ML methodologies is the field of neuroimaging (Kambeitz et al., 2017). Of interest, support vector machines (SVM) have been chosen in approximately 60 - 73% of current neuroimaging works of major depressive disorder, (MDD) with theoretical or clinical justifications for model selection lacking (Kambeitz et al., 2017). This is a strange phenomenon as the purpose of moving to a statistical learning framework is to build translational models for clinical use whilst demonstrating the potential prognostic value of certain predictor modalities.

This observation begs the question, if ML models ever make it to clinical care, how might we actually use them? And based on this use, how should this inform the models we choose and how we build them in our current research efforts? Given the interpersonal nature of clinical practice, we deem it highly unlikely that ML applications alone will supersede the need for human practitioners in the foreseeable future. What is more likely is that ML models will

exist in the form of clinical decision support systems that will be used to complement the professional judgement of clinicians. Therefore, we see the clinician/ML collaboration as a synergistic relationship, with clinicians calling on ML systems in high-risk scenarios when a second opinion is required. In this sense, we can think of this human/ML collaboration as forming its own meta-classifier, where a clinician's original probabilistic prediction is combined with that of an ML model, subsequently updated, and used to make a binary decision derived from this new meta-probability estimate (treat with drug A? Yes/No). Therefore, the need for robust probability calibration, defined as the degree of convergence between a models predicted probability of an event and its actual probability is essential for this collaborative workflow(Niculescu-Mizil and Caruana, 2005).

An everyday example of probability calibration can be found when we check the weather. Here, we may see a statement about how likely it is that it will rain. If our weather app tells us that there is a 70% chance of rain, and on 70 out of 100 days that it makes this statement, it does in fact rain, we would say that our weather app is well calibrated. Whilst predicting the weather and prognosticating patient outcomes are fundamentally different problems, they both share the same underlying properties of uncertainty. Therefore, a clinician should not just know how *accurately* an ML model will classify a prognostic or diagnostic endpoint, but the underlying *probability* of the respective endpoint itself. This *probability* can then provide a measure of *confidence* in a model's binary prediction in the face of this uncertainty.

To further reiterate this point, we can think about two levels of accuracy in an ML model (levels **A** and **B**). A model that is 70% accurate at the **group level (A)** means that 70% of the predictions it makes are correct. Whilst at the **individual level (B)**, a probability of 70% for “yes” means that we are 70% sure that this individual prediction for **group level membership (A)** is correct. If a model is poorly calibrated, the *confidence* we can have in this probabilistic estimation will be low (it may give a probability of 90% “yes” although the true probability of “yes” is only 60%). While it might still be true that the patient belongs to the “yes” class, the model will be considerably overconfident in this **group level (A)** classification. At current, many works in psychiatry only test whether **(A)** is true. Case in point, the “go to” SVM (both linear and non-linear) chosen in the majority of applied neuroimaging ML works has some of the worst calibrated probabilities of any classical ML model, showing a sigmoid curve plotted across probability estimates on the X axis of the reliability curve (Figure 1, Linear: orange curve, Non-linear: purple curve). This statistical property is commonly observed in

maximum-margin methods which focus on hard samples that are close to a model's decision boundary (Niculescu-Mizil and Caruana, 2005).

To make clear the clinical ramifications of this behaviour, imagine consulting James from the introductory example. A decision support system trained with a linear SVM predicts that he will not respond to Lithium, the clinicians next choice for treatment. In addition, the underlying probability behind this binary classification is 10%, suggesting that the model is highly confident that James will not respond (in comparison to giving James a 49% probability of response, for example). The clinician is starting to second guess their decision, especially given the promising accuracy of the model demonstrated in clinical trials. Imagine now that we calibrate the model, the model still classifies James as someone that will not respond to Lithium, however, it is only 47% confident in this group level classification (**A**). As the model is uncertain, the clinician now relies solely on their clinical expertise and prescribes Lithium, leading to a successful treatment response for James.

Given this dissonance between clinical need and commonly selected model's probabilistic behaviour, what can be done to mitigate this discrepancy? Firstly, as logistic regression (LR) models directly optimize for log-loss, their probabilities are well calibrated by virtue of their underlying mathematics (Figure 1). In addition, the adding of L1 and L2 regularization terms to the objective function (known as LASSO and Ridge) builds in embedded feature selection to the model fitting procedure (in the case of L1, or in the combination of both norms, known as the Elastic Net). Given that modern psychiatric datasets commonly contain more features than observations (known as large P small N problems), such a constraint is optimal for both feature selection and subsequent model fitting. In addition, as both LR and linear SVMs attempt to fit a linear decision boundary between data points, with the former fitting a line and the latter a hyperplane, accuracy measurement variation between the two models should theoretically be small. Therefore, we might start with the simpler well-calibrated LR model first and only then move to a linear or radial basis function SVM (or other non-linear models) if a major improvement is observed in classification accuracy.

In situations where a clinically meaningful increase in accuracy is achieved compared to LR, attaining well calibrated probabilities is still possible. First, a practitioner can directly optimize for negative log-loss rather than metrics such as area under the receiver operator characteristic curve (AUC) or balanced accuracy. Secondly, post-processing probability calibration methods are available and have demonstrated success when samples are large enough to include this procedure (Nixon et al., 2019). Commonly available procedures include

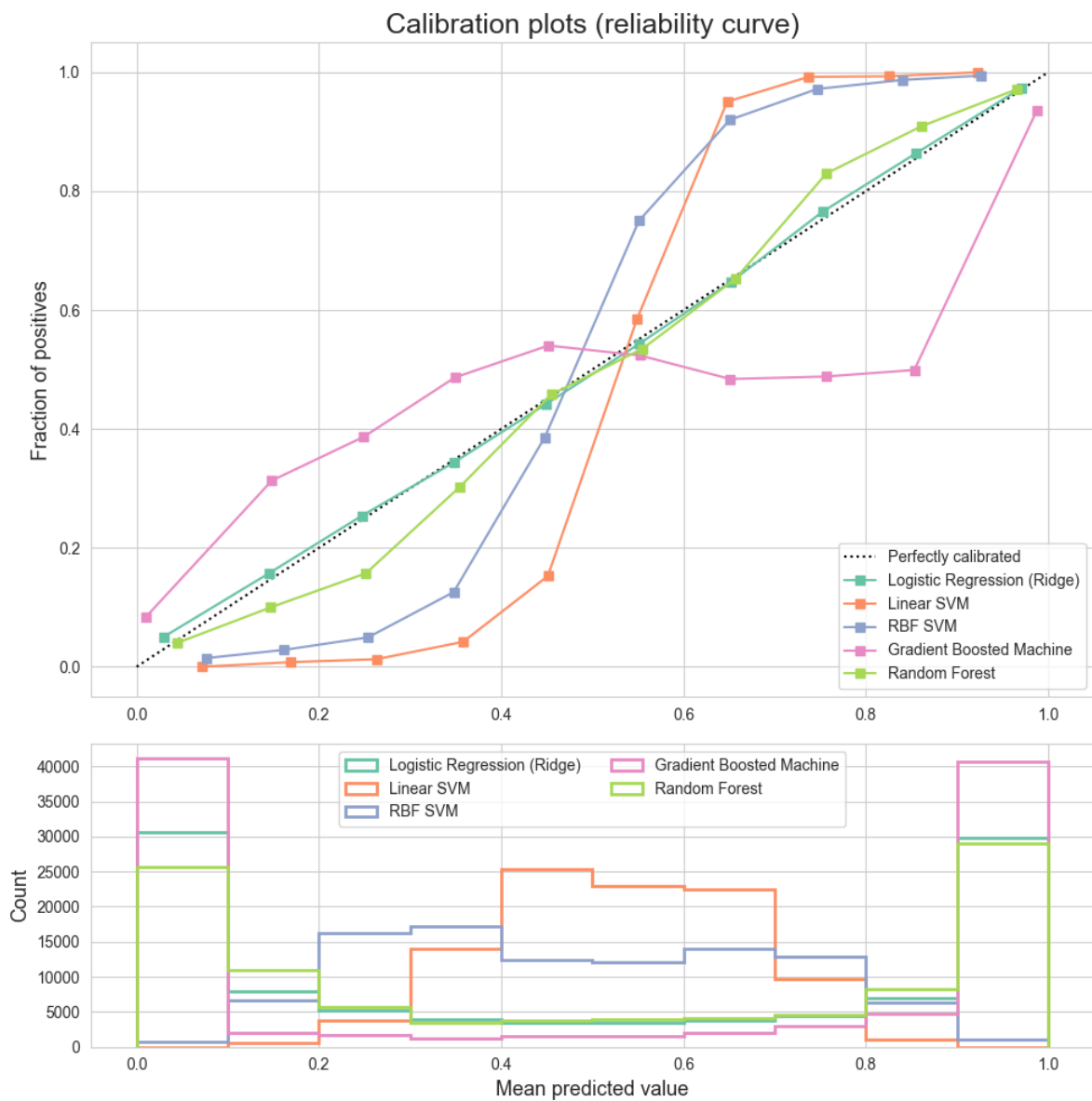
sigmoid and isotonic regression for post-processing calibration. However, samples are needed that are separate to those used in model training, imposing further sample size requirements on samples that may already be limited. Alternatively, post-processing calibration can be nested, however, optimistic bias may theoretically arise due to similarity between train and test set samples used in a nesting procedure. The balancing of these constraints needs to be considered when calibrating an ML models probabilities post-hoc. Given these restrictions, a clinically meaningful increase may need to be seen in model accuracy to invoke post-processing calibration. One solution to this problem would be to start with a regularized LR model first, forming a basis to quantify changes in subsequent calibration/classification accuracy trade-offs (in the case of multiclass classification, a one-vs-rest or multinomial LR can be used). In addition, the statistical significance of univariate coefficients forming the underlying multivariate pattern can be disseminated, providing insight to clinicians hoping to complement their own clinical judgement with a nuance that extends beyond continuous probability estimates and binary outputs.

Finally, as a clinician hoping to form an understanding of a model's calibration, what metrics should be of concern? The first commonly used scoring-rule is log-loss, which compares a model's predicted probabilities to its true binary class labels. Subsequently, a score is calculated that penalizes each probability estimate based on the distance from its expected class label(Niculescu-Mizil and Caruana, 2005). As the penalty is logarithmic, it offers a small score for small differences and a large score for large differences. These scores are averaged across samples, providing an aggregate score for model calibration. Another common scoring rule is known as the Brier score. This score measures the mean squared difference between the predicted probability assigned to a possible outcome and its true class labels(Niculescu-Mizil and Caruana, 2005). Therefore, the Brier score is akin to a mean squared error measurement of a model's probability prediction. Consequently, for both log-loss and the Brier score, lower scores equate to better calibrated probabilities. Finally, plotting a reliability curve (Figure 1) provides an intuitive visualization of a model's calibration, allowing a clinician or ML practitioner to derive a visual understanding of discrepancies between actual and predicted probability values from a model.

In conclusion, we suggest that ML model selection needs to be driven and justified by future best practice clinical workflows. As we deem it plausible that these workflows will entail human/ML collaboration with the purpose of engaging in high-risk clinical decision making, we recommend that a greater emphasis be placed on model calibration in future works. As a

trade-off between calibration and rudimentary measures of accuracy may ensue, active thought surrounding clinical context should always form the basis of model selection. This context should be made explicit, drive the model selection process, and be informed by clinical expertise. To form a basis to capture calibration/accuracy trade-offs, we suggest starting with basic penalized LR models that are well calibrated by of their mathematics. If marginal trade-offs in calibration/accuracy are observed that are justifiable from the required clinical context, the progression to higher complexity models with poorer default calibration could be made. Hyperparameters could then be directly optimized for negative log-loss or calibrated with post-processing methods if required. Finally, whilst the go to linear SVM has established itself as a useful model in psychiatric classification studies, it is important to be aware of its poor default calibration and the implications this has for clinical care. We urge ML practitioners to apply thought to clinical context when selecting models in future works.

Figure 1. Reliability curves for commonly used classifiers in psychiatric machine learning studies. The closer a curve is to the line of equality (the straight dashed line), the better calibrated the ML models probabilities. All classifiers were trained and tested with default parameters on a simulated binary dataset of 100,000 observations. Without post-processing calibration, logistic regression returns well calibrated probabilities by default as it is optimized for log-loss, whilst other commonly used classifiers show varying non-linear reliability curves. Abbreviations: SVM (support vector machine), RBF (radial basis function).



References

- Chekroud AM, Zotti RJ, Shehzad Z, et al. (2016) Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry* 3(3): 243-250.
- Kambeitz J, Cabral C, Sacchet MD, et al. (2017) Detecting Neuroimaging Biomarkers for Depression: A Meta-analysis of Multivariate Pattern Recognition Studies. *Biol Psychiatry* 82(5): 330-338.
- Niculescu-Mizil A and Caruana R. (2005) Predicting good probabilities with supervised learning. *Proceedings of the 22nd international conference on Machine learning*. ACM, 625-632.
- Nixon J, Dusenberry M, Zhang L, et al. (2019) Measuring Calibration in Deep Learning. *arXiv preprint arXiv:1904.01685*.
- Rush AJ, Trivedi MH, Wisniewski SR, et al. (2006) Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR* D report. *American Journal of Psychiatry* 163(11): 1905-1917.

Statement of Authorship

Title of Paper	Systematic Misestimation of Machine Learning Performance in Neuroimaging Studies of Depression
Publication Status	☐ Published ☐ Accepted for Publication ☒ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Flint, C., Cearns, M., Opel, N., Redlich, R., Mehler, D., Emden, D., ... & Hahn, T. (2019). Systematic Overestimation of Machine Learning Performance in Neuroimaging Studies of Depression. <i>arXiv preprint arXiv:1912.06686</i> .

Principal Author

Name of Principal Author (Candidate)	Micah Cearns		
Contribution to the Paper	Formulated the idea for the paper alongside Claas Flint, wrote a large proportion of the manuscript and conducted revisions, created figure 1.		
Overall percentage (%)	50%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	15/01/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Claas Flint		
Contribution to the Paper	Contributed to study design and writing. Wrote the code for the analyses and figures and formatted all results.		
Signature		Date	15/01/2021
Name of Co-Author	Nils Opel		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision.		

Signature		Date	15/01/2021
Name of Co-Author	Ronny Redlich		
Contribution to the Paper	Critical revision of the manuscript, analysis and supervision.		
Signature		Date	15/01/2021
Name of Co-Author	David Mehler		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision.		
Signature		Date	15/01/2021
Name of Co-Author	Daniel Emden		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision.		
Signature		Date	15/01/2021
Name of Co-Author	Ramona Leenings		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision.		
Signature		Date	15/01/2021
Name of Co-Author	Simon Eickoff		
Contribution to the Paper	Feedback on the analytical pipelines, discussion of the results, critical editing of the manuscript.		
Signature		Date	15/01/2021
Name of Co-Author	Tilo Kircher		
Contribution to the Paper	Funding and organisation of data collection. Critical revision of the manuscript, overall guidance and supervision.		

Signature		Date	15/01/2021
Name of Co-Author	Axel Krug		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision.		
Signature		Date	15/01/2021
Name of Co-Author	Igor Nenadic		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision.		
Signature		Date	15/01/2021
Name of Co-Author	Volker Arolt		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision.		
Signature		Date	15/01/2021
Name of Co-Author	Nils Winter		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision.		
Signature		Date	15/01/2021
Name of Co-Author	Scott Clark		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision.		
Signature		Date	15/01/2021
Name of Co-Author	Bernhard Baune		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision.		
Signature		Date	15/01/2021

Name of Co-Author	Xiaoyi Jiang		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision.		
Signature		Date	15/01/2021
Name of Co-Author	Udo Dannlowski		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision. Supervision of imaging analyses.		
Signature		Date	15/01/2021
Name of Co-Author	Tim Hahn		
Contribution to the Paper	Senior and corresponding author. Helped plan analyses and study design. Assisted with writing and general revisions.		
Signature		Date	15/01/2021

Please cut and paste additional co-author panels here as required.

Chapter 4

Systematic Misestimation of Machine Learning Performance in Neuroimaging Studies of Depression

Claas Flint^{1, 2 *}, Micah Cearns^{3, 5 *}, Nils Opel¹, Ronny Redlich¹, David M. A. Mehler¹, Daniel Emden¹, Nils R. Winter¹, Ramona Leenings¹, Simon B. Eickhoff^{4, 8}, Tilo Kircher⁶, Axel Krug⁶, Igor Nenadic⁶, Volker Arolt¹, Scott Clark³, Bernhard T. Baune^{1, 5, 7}, Xiaoyi Jiang², Udo Dannlowski^{1**}, Tim Hahn^{1**, ♦}

¹ Department of Psychiatry, University of Münster, Germany

² Faculty of Mathematics and Computer Science, University of Münster, Germany

³ Discipline of Psychiatry, School of Medicine, University of Adelaide, Australia

⁴ Institute of Neuroscience and Medicine (INM-7) Research Center Jülich

⁵ Department of Psychiatry, Melbourne Medical School, The University of Melbourne, Parkville, Australia

⁶ Department of Psychiatry and Psychotherapy, University of Marburg, Germany

⁷ The Florey Institute of Neuroscience and Mental Health, The University of Melbourne, Parkville, Australia

⁸ Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

* these authors contributed equally to this work and are equal first authors

Neuropsychopharmacology

Abstract

We currently observe a disconcerting phenomenon in machine learning studies in psychiatry: While we would expect larger samples to yield better results due to the availability of more data, larger machine learning studies consistently show much weaker performance than the numerous small-scale studies. Here, we systematically investigated this effect focusing on one of the most heavily studied questions in the field, namely the classification of patients suffering from Major Depressive Disorder (MDD) and healthy controls based on neuroimaging data. Drawing upon structural MRI data from a balanced sample of $N=1,868$ MDD patients and healthy controls from our recent international Predictive Analytics Competition (PAC), we first trained and tested a classification model on the full dataset which yielded an accuracy of 61%. Next, we mimicked the process by which researchers would draw samples of various sizes ($N=4$ to $N=150$) from the population and showed a strong risk of misestimation. Specifically, for small sample sizes ($N=20$), we observe accuracies of up to 95%. For medium sample sizes ($N=100$) accuracies up to 75% were found. Importantly, further investigation showed that sufficiently large test sets effectively protect against performance misestimation whereas larger datasets per se do not. While these results question the validity of a substantial part of the current literature, we outline the relatively low-cost remedy of larger test sets, which is readily available in most cases.

Introduction

In psychiatry, we are witnessing an explosion of interest in machine learning (ML) and artificial intelligence for prediction and biomarker discovery, paralleling similar developments in personalized medicine [1–4]. In contrast to the majority of investigations employing classic group-level statistical inference, ML approaches aim to build models which allow for individual (i.e. single subject) predictions, thus enabling direct assessment of individual differences and clinical utility [5]. While this constitutes a major advancement for clinical translation, recent results of large-scale investigations have given rise to a fundamental concern in the field: Specifically, machine learning studies including larger samples did not yield stronger performance, but consistently showed weaker results than studies drawing on small samples, calling into question the validity and generalizability of a large number of highly published proof-of-concept studies.

The magnitude of this issue was impressively illustrated by the results of the Predictive Analytics Competition (PAC 2018; <http://photon-ai.com/pac>) in which participants developed machine learning models to classify healthy controls (HC) and depressive patients (MDD) based on structural MRI data from $N=2,240$ participants. Despite the best efforts of ~ 170 machine learners in 49 teams from around the world, performance ranged between 60% and 65% accuracy in a large, independent test set. This is in strong contrast to the numerous smaller studies showing accuracies of 80% or more [6–8].

Further empirical studies focusing on other disorders support this observed effect of performance deterioration with increasing sample size: In a large-scale investigation, Neuhaus & Popescu [9] aggregated original studies across disorder categories, including schizophrenia (total observation $N=5,563$), MDD ($N=2,042$), and attention deficit hyperactivity disorder (ADHD, $N=8,084$), finding an inverse relationship between sample size and balanced accuracy (schizophrenia, $r=-.34$; MDD, $r=-.32$; and ADHD, $r=-.43$). Similar results were observed in a recent review of 200 neuroimaging classification studies of brain disorders, finding a general trend towards lower reported accuracy scores in larger samples [10]. Given that model performance would be expected to increase with more data, these results hint at a fundamental issue hampering current predictive biomarker studies in psychiatry.

From a methodological point of view, it has been known since the early 90's that training samples should be large when there is a high number of features (i.e. measured variables) and a complex classification rule being fit to a dataset [11]. Recent works have further reiterated

this point [12]. Moreover, these effects may have been further amplified by certain cross-validation schemes. For example, Kambeitz et al. [13] observed higher accuracy estimates in studies using hold-out cross-validation strategies compared to 10-fold and leave-one-out (LOOCV), whilst Varoquaux et al. observed that LOOCV leads to unstable and biased estimates, concluding that repeated random splits should be preferred [14].

Although these findings have sparked in-depth conceptual considerations [15], empirical investigations of this problem have been limited to specific cross-validation schemes and small test set sizes [16]. Here, we aim to systematically investigate the effects of both train and test set sample sizes on machine learning model performance in neuroimaging based MDD classification. In addition, as it is possible that effects of systematic misestimation have arisen due to suboptimal pipeline configurations (i.e., the disproportionate use of LOOCV and linear Support Vector Machines on samples containing more predictors than observations), we also test a further 48 different pipeline configurations to quantify the influence of these additional factors. To demonstrate that this effect was not dependent on the data or any pipeline configurations used in our analyses, we repeat the process using a dummy classifier. To quantify the magnitude of these effects in each configuration, we drew samples of various sizes from the PAC dataset – mimicking the process by which researchers would draw samples from the population of ML studies reported in the literature. The resulting probability distributions are investigated.

Materials and Methods

To investigate the effects of both train and test set sample sizes on machine learning model performance in neuroimaging based MDD classification, we repeatedly drew samples of different sizes from the PAC dataset to imitate the procedure reported in the literature. Subsequently, the resulting probability distributions are investigated.

Dataset description

The PAC dataset comprised anonymized, pre-processed MRI data of N=2,240 individuals obtained from two large, independent, ongoing studies - the Münster Neuroimaging Cohort [17,18] (N=724 MDD; N=285 HC) and the FOR2107-study [19] (N=582 MDD, N=649 HC). Case/control status was diagnosed with the SCID-IV [20] interview employed by trained clinical raters in both studies. In both cohorts, exclusion criteria were any neurological or severe medical condition, MRI contraindications, substance-related disorders, Benzodiazepine

treatment and head injuries. For healthy controls, any current or previous psychiatric disorder or use of any psychotropic substances. The Münster Neuroimaging Cohort was scanned at one single MRI-scanner with the same sequence, while the FOR2107-study was scanned at two independent sites[21], yielding 3 different scanner types/sequences. The structural T1-weighted magnetic resonance imaging (MRI) scans were pre-processed with the CAT12 toolbox (<http://www.neuro.uni-jena.de/cat>, r1184) using default parameters to obtain modulated, normalized grey matter segments (resolution $1.5 \times 1.5 \times 1.5 \text{ mm}^3$) which were used for the present analysis. Furthermore, age, gender, scanner type, and total intracranial volume (TIV) were provided.

Machine learning pipeline

To ensure the unbiased approximation of the model's performance in previously unseen patients (i.e., model generalization), we trained and tested all models in a pipeline to prevent information leaking between patients used for training and validation. First, we used random under-sampling to derive a balanced sample of 934 MDD cases and 934 healthy controls (HC) (see Supplementary appendix A for summary statistics). To reduce the computational effort, all images have been scaled down to a voxel size of $3 \times 3 \times 3 \text{ mm}^3$. Following, every image was converted to a vector, where every voxel served as a feature. For standardization, the features were scaled to have zero mean and unit variance. Finally, a linear support vector machine (SVM) with default parameters (sklearn [22], v0.20) was trained and model performance was calculated and analysed in the subsequent analyses. The effect of varying scanner-distributions across samples was found to be negligible (see Supplementary appendix C).

Overall sample size effects

To first examine the effects of overall-set size, we randomly sampled the PAC dataset in steps of 1 from $N=4$ to 150. For each N value, we created 1,000 balanced samples. This yielded 147,000 random samples with equal numbers of patients and healthy controls. Applying the SVM pipeline described above to each of these samples independently allowed us to obtain a distribution of accuracy scores for each N value, respectively. On each sample, one SVM was trained with default parameters using LOOCV. Thus, we trained a total of 11,315,000 SVMs. Since the computational effort increases quadratically with increasing sample size, and in addition, the most recent neuroimaging ML studies using LOOCV rarely exceed $N=150$, we

decided to stop at this value. Following, we evaluated the distribution of accuracy scores estimated for each sample size (N=4 to 150).

Training set size effects

Second, to examine the effects of training set size, we varied the size of each training set and then tested performance on a fixed hold-out set. Specifically, we randomly sampled the PAC dataset in steps of 1 from N=4 to 150 as was done in the previous analysis. For each N value, we created 1,000 balanced samples and used each of them to train different models. To then test the effects of training set size on test set performance, we tested each trained model on a balanced test set of N=300. From the resulting distributions, we quantified the probability of overestimating accuracy as a function of training set size.

Test set size effects

To assess the effects of test set size on classification accuracy, we randomly sampled a balanced group of 300 subjects from our full sample of 934 MDD cases and 934 HC's. From this sample, we randomly sub-sampled *test* sets of N=4 to 150 in steps of 1. For each value of N, we took 1,000 samples, resulting in the creation of 147,000 random *test* samples. From the remaining 784 MDD cases and 784 HC's, we took out another 20% sample (MDD=157 and HC=157). The prediction of this sample provides a reliable basis for overall performance estimation and allows for a comparison with the results obtained from smaller samples. We then trained a single SVM on the remaining sample (MDD=627, HC=627). The trained SVM was then tested on each test set sample (N=4 to 150). From the resulting *test* set accuracy distributions, we derived the probability of obtaining accuracy scores between 50% and 90% accuracy by chance. See Figure 1 for an overview of all analyses.

Generalizability of statistical effect

Alternative pipeline configurations

As we have attempted to hold other components of the modelling process constant so any observed effects of systematic misestimation can be attributed to sample size alone, it is also possible that our results may be dependent on the basic configuration of our machine learning pipeline (e.g., the use of a linear SVM, default hyperparameters, and LOOCV). Therefore, we trained a further 48 pipeline configurations, including the use of both linear and radial basis

function SVMs and a Random Forest classifier, all of which have demonstrated their efficacy in neuroimaging classification studies [13]. Within these configurations, we conducted dimensionality reduction using principal components analysis as well as f-test based feature selection. This allowed us to assess whether our findings were being confounded by the large number of predictors used in the main analysis. For further information see Supplement Appendix E and for all results see supplementary results E.1, supplementary figures E.4 – E.51 and supplementary tables E.9 – E.56.

Dummy classifier

In order to show that the observed effects were not specific to the PAC dataset or any of the alternative pipeline configurations in these analyses, we repeated the procedures described above with a dummy classifier. Our dummy classifier assumed a prior probability for MDD vs control classification based on the percentage proportion of each class in the training data (prevalence). As the dataset was balanced with random under-sampling, the prior and subsequent ground truth of the model was equal to exactly 50%. This approach allowed us to compare our distribution of dummy performance estimates derived from our subsample analysis to this ground truth value. Importantly, this approach allows for the quantification of accuracy misestimation as a function of sample size completely independent of any unique characteristics that may be specific to the PAC dataset or our pipeline configurations. Therefore, we can then be sure that any subsequent changes in classifier performance are attributable to sample size alone.

Results

Overall sample size effects

In the overall sample size analysis using LOOCV, we were able to show that the risk of overestimating the classifier performance increases with decreasing overall sample size (Figure 2a, Supplements Table B.2a). Specifically, accuracies of 70% or higher are observed with a probability of 13% on sample sizes of $N=20$ whereas this probability is reduced to 2% for sample sizes of $N=100$. In addition, the sample size has a profound impact on the variability of accuracy estimates: For samples of size $N=20$, accuracies ranged from 10% to 95% (standard deviation=15%) while for samples of $N=100$, accuracies ranged between 35% and 81% (standard deviation=6%) (Figure 2b, Supplements Table B.2b). Note that this effect is symmetrical and also applies to the underestimation of performance (Figure 2b). Additionally,

the results of the dummy classifier (Figure 2c, d, Supplements Table B.3) show that the observed overestimation effect is a general effect of sample size as previously pointed out by Varoquaux [16]. As the regularization of the SVM is sensitive to the total number of outliers, which may increase in parallel with sample size, we conducted an additional analysis with adjusted C parameters, with the observed effect remaining constant across these analyses (see Supplements Figure D.1 and D.2).

Training set size effects

When examining the effects of training set size (N=4 to 150) using a large test set for evaluation (N=300), we did not observe any systematic misestimation (Figure 3a, Supplements Table B.4a). In fact, models trained on virtually any training set size from N=4 to 150 were sufficient to obtain maximum model performance. However, increasing training sample size decreased the probability of obtaining very low performance estimate. For training sets of size N=20, accuracies ranged from 32% to 69% (standard deviation=7.1%) while for training sets of N=100, accuracies ranged between 51% and 70% (standard deviation=3.0%) (Figure 3b, Supplements Table B.4b). In accordance with the overall sample size analysis, the results of the dummy classifier (Figure 3c, d, Supplements Table B.5) showed that this observed effect was general in nature.

Test set size effects

For our analysis varying test set size, we found a similar pattern of systematic misestimation as that in our first overall sample size analysis using LOOCV. With a sample size of N=20, we obtain results of 70% accuracy or higher with a probability of 30%, whilst the mean accuracy on the full dataset (N=268) was only 61%. This probability dropped to 13% when the test sample size was N=100. For test sets of size N=20, accuracies ranged from 35% to 95% (standard deviation=10%) whilst for test sets of size N=100, accuracies ranged from 51% and 79% (standard deviation=4%) (Figure 4b, Supplements Table B.6b). Running the analysis again using our dummy classifier, we were able to show that the general pattern of systematic misestimation was independent of the specific dataset used (Figure 4c, d, Supplements Table B.7). To show the independent and generalizable character of the observed effect, we repeated the analysis on the 48 unique pipeline configurations discussed above (see Supplementary appendix F). Specifically, the results are comparable to the originally used configuration, i.e.,

an SVM with a linear kernel and no preprocessing. Finally, our analysis of scanner sites revealed no effects on model performance (see Supplementary appendix C).

Discussion

Sparked by the observation that machine learning studies drawing on larger neuroimaging samples consistently showed weaker results than studies drawing on smaller ones, we drew samples of various sizes from the PAC dataset, thereby mimicking the process by which researchers would draw samples from the population of ML studies reported in the literature. When applying a linear SVM with LOOCV, as is the most common approach in the neuroimaging literature [10], we observed a higher risk of misestimations, which may lead to artificially high performance estimates in smaller samples. Importantly, our analyses revealed that this is primarily due to a small test set, not training set size. Generally, this shows that a small test sample size may explain many of the highly optimistic results published in recent years. When considering the well-established effect of publication bias, even if underestimated results are equally as likely, they will have a significantly lower chance of being published.

Our results are the first to disentangle the effects of training and test set size effects which are typically inseparable in common cross-validation frameworks such as LOOCV. This delineation of effects enabled two important insights for biomarker discovery and outcome prediction. First, researchers need to validate their models on large, independent test sets. Our results indicate that in the PAC dataset, a test set size of $N=100$ was already sufficient to lower the probability of obtaining artificially good performance (i.e., 70% or higher) to 13%. With a median N of less than 100 in many published studies [10], this may seem challenging. However, online infrastructure for independent machine learning model evaluation is available (e.g., www.photon-ai.com/repo). If researchers open-source their models, anyone - independent of technical knowledge or machine learning expertise - can evaluate them. This way, large independent test datasets can be obtained in a short time without the need for data sharing. This is not restricted to neuroimaging data, but any machine learning model. In addition, efforts from consortia will also help to mitigate this problem and should be considered by machine learning practitioners.

Second, the size of the training set alone cannot serve as an indicator of later model performance. Larger training sets are more likely to generalize to new data and broaden model scope (i.e., about which groups within a population a given model can make reasonable predictions), however, in the current analysis, the linear rule learned by an SVM on high-

dimensional neuroimaging data could be approximated with only a handful of samples. From a training set size of 30 onward, we no longer observed any increase in model performance. This somewhat counterintuitive effect arises whenever a simple rule is approximated. For higher complexity models (i.e., models capable of learning more complex rules), we, of course, expect performance increases as training sample size increases. However, considering the results of the PAC competition (<https://photon-ai.com/pac>), high complexity models such as Deep Learning approaches did not yield higher performance when trained with $\sim 1,000$ samples. Thus, we conclude that simple models are competitive for sample sizes of up to $N=1,000$ for this particular classification problem. Whether more complex rules can be discovered beyond this point or whether other fundamental issues hamper biomarker discovery in psychiatry (cf. e.g. biotype identification [19] and normative modelling approaches [23]) remains an open question.

An intuitive criticism of our main analyses would be that it has merely replicated methods similar to those of previous low-quality works (for example, studies using only linear SVMs with default parameters, tested in LOOCV schemes, on samples with many more predictors than observations). Whilst this pipeline configuration was used in the current analysis to a) hold constant properties, that if varied, may have been indistinguishably responsible for changes in accuracy misestimation, and b) replicate the most commonly used ML pipeline configuration in our field, it was important to conduct complementary analyses to rule out these confounds. Therefore, we tested a further 48 ML pipeline configurations (see Supplementary appendix F) using both linear (a linear SVM) and non-linear (a radial basis function SVM and a Random Forest) classifiers. In addition, we conducted PCA based dimensionality reduction as well as f-test based feature selection within these classifiers to delineate whether our findings were confounded by the large size of the predictor space relative to our number of observations. Importantly, all pipeline configurations demonstrated the same pattern of systematic misestimation as that in our main analyses. The second potential criticism of the current work is that these findings may merely be modality-specific, limiting the generalizability of these findings across domains. However, the use of a dummy classifier that completely ignored the input predictor space (the voxels), and instead, classified samples based only on their prevalence in training (MDD=50%, Control=50%), showed the same pattern of sample size based systematic misestimation across all pipeline configurations, thus, demonstrating a generalizable statistical effect regardless of the data modality used.

Given the profound effect of the test set size on systematic misestimation, it is important to consider why an effect of overestimation may arise. Previous work by Schnack & Kahn [24] suggests that patient characteristics in smaller samples tend to be more homogenous. In the case of small N, participants may be more likely to be recruited from the same data collection site and of a similar age (for example, in the case of a university recruited convenience sample). In addition, stringent recruitment criterion may be easily met, resulting in a well-defined phenotype that is not truly representative of the parent population of interest. Whilst this explanation makes sense for samples collected in this manner, it fails to explain why we observed this phenomenon in our random sampling procedure, and more importantly, with our dummy classifiers that paid no attention to participants, their characteristics, or the inputted predictor variables. This observation suggests a mechanism for systematic misestimation that is not just sample/patient-specific or contingent on sample homogeneity, but instead, inherent in the natural variation that arises in small test samples.

In addition to sample size, other issues such as data leakage [13], are likely contributing to the systematic overestimation seen in the literature. Dedicated cross-platform software to help avoid data leakage is freely available (e.g. PHOTON, www.photon-ai.com or Sci-kit learn [22]). Finally, code should be made available on request or provided in an online repository (e.g., GitHub or GitLab) upon submission for review. In addition, a more elaborate evaluation framework including the analysis of model scope assessment as well as incremental utility and risk analysis is needed to move the field beyond proof-of-concept studies. The success of current translational efforts in neuroimaging and psychiatry will crucially depend on the timely adoption of guidelines and rules for medical machine learning models (for an in-depth introduction, see [15]).

In summary, our results indicate that - while many of the most highly published results might strongly overestimate true performance - evaluation on large test sets constitutes a straightforward remedy. Given that simple, low-complexity models such as linear SVMs did not gain from larger training set size, researchers should not discard their models due to low training N but seek evaluation on a large test set for any model showing good performance.

Funding and Disclosure

This work was funded by the German Research Foundation (DFG, grant FOR2107 DA1151/5-1 and DA1151/5-2 to UD; SFB-TRR58, Projects C09 and Z02 to UD) and the Interdisciplinary Center for Clinical Research (IZKF) of the medical faculty of Münster (grant Dan3/012/17 to UD). TH was supported by the German Research Foundation (DFG grants HA7070/2-2, HA7070/3, HA7070/4).

This work is part of the German multicenter consortium “Neurobiology of Affective Disorders. A translational perspective on brain structure and function”, funded by the German Research Foundation (Deutsche Forschungsgemeinschaft DFG; Forschungsgruppe/Research Unit FOR2107).

Principal investigators (PIs) with respective areas of responsibility in the FOR2107 consortium are:

Work Package WP1, FOR2107/MACS cohort and brain imaging: Tilo Kircher (speaker FOR2107; DFG grant numbers KI 588/14-1, KI 588/14-2), Udo Dannlowski (co-speaker FOR2107; DA 1151/5-1, DA 1151/5-2), Axel Krug (KR 3822/5-1, KR 3822/7-2), Igor Nenadic (NE 2254/1-2), Carsten Konrad (KO 4291/3-1). WP2, animal phenotyping: Markus Wöhr (WO 1732/4-1, WO 1732/4-2), Rainer Schwarting (SCHW 559/14-1, SCHW 559/14-2). WP3, miRNA: Gerhard Schratt (SCHR 1136/3-1, 1136/3-2). WP4, immunology, mitochondriae: Judith Alferink (AL 1145/5-2), Carsten Culmsee (CU 43/9-1, CU 43/9-2), Holger Garn (GA 545/5-1, GA 545/7-2). WP5, genetics: Marcella Rietschel (RI 908/11-1, RI 908/11-2), Markus Nöthen (NO 246/10-1, NO 246/10-2), Stephanie Witt (WI 3439/3-1, WI 3439/3-2). WP6, multi-method data analytics: Andreas Jansen (JA 1890/7-1, JA 1890/7-2), Tim Hahn (HA 7070/2-2), Bertram Müller-Myhsok (MU1315/8-2), Astrid Dempfle (DE 1614/3-1, DE 1614/3-2). CP1, biobank: Petra Pfefferle (PF 784/1-1, PF 784/1-2), Harald Renz (RE 737/20-1, 737/20-2). CP2, administration. Tilo Kircher (KI 588/15-1, KI 588/17-1), Udo Dannlowski (DA 1151/6-1), Carsten Konrad (KO 4291/4-1).

Data access and responsibility: All PIs take responsibility for the integrity of the respective study data and their components. All authors and coauthors had full access to all study data.

Acknowledgements and members by Work Package (WP):

WP1: Henrike Bröhl, Katharina Brosch, Bruno Dietsche, Rozbeh Elahi, Jennifer Engelen, Sabine Fischer, Jessica Heinen, Svenja Klingel, Felicitas Meier, Tina Meller, Torsten Sauder, Simon Schmitt, Frederike Stein, Annette Tittmar, Dilara Yüksel (Dept. of Psychiatry, Marburg University). Mechthild Wallnig, Rita Werner (Core-Facility Brainimaging, Marburg University). Carmen Schade-Brittinger, Maik Hahmann (Coordinating Centre for Clinical Trials, Marburg). Michael Putzke (Psychiatric Hospital, Friedberg). Rolf Speier, Lutz Lenhard (Psychiatric Hospital, Haina). Birgit Köhnlein (Psychiatric Practice, Marburg). Peter Wulf, Jürgen Kleebach, Achim Becker (Psychiatric Hospital Hephata, Schwalmstadt-Treysa). Ruth Bär (Care facility Bischoff, Neunkirchen). Matthias Müller, Michael Franz, Siegfried Scharmann, Anja Haag, Kristina Spenner, Ulrich Ohlenschläger (Psychiatric Hospital Vitos, Marburg). Matthias Müller, Michael Franz, Bernd Kundermann (Psychiatric Hospital Vitos, Gießen). Christian Bürger, Fanni Dzvonyar, Verena Enneking, Stella Fingas, Janik Goltermann, Hannah Lemke, Susanne Meinert, Jonathan Repple, Kordula Vorspohl, Bettina Walden, Dario Zaremba (Dept. of Psychiatry, University of Münster). Harald Kugel, Jochen Bauer, Walter Heindel, Birgit Vahrenkamp (Dept. of Clinical Radiology, University of Münster). Gereon Heuft, Gudrun Schneider (Dept. of Psychosomatics and Psychotherapy, University of Münster). Thomas Reker (LWL-Hospital Münster). Gisela Bartling (IPP Münster). Ulrike Buhlmann (Dept. of Clinical Psychology, University of Münster).

WP2: Marco Bartz, Miriam Becker, Christine Blöcher, Annuska Berz, Moria Braun, Ingmar Conell, Debora dalla Vecchia, Darius Dietrich, Ezgi Esen, Sophia Estel, Jens Hensen, Ruhkshona Kayumova, Theresa Kisko, Rebekka Obermeier, Anika Pützer, Nivethini Sangarapillai, Özge Sungur, Clara Raithel, Tobias Redecker, Vanessa Sandermann, Finnja Schramm, Linda Tempel, Natalie Vermehren, Jakob Vörckel, Stephan Weingarten, Maria Willadsen, Cüneyt Yildiz (Faculty of Psychology, Marburg University).

WP4: Jana Freff, Silke Jörgens, Kathrin Schwarte (Dept. of Psychiatry, University of Münster). Susanne Michels, Goutham Ganjam, Katharina Elsässer (Faculty of Pharmacy, Marburg University). Felix Ruben Picard, Nicole Löwer, Thomas Ruppertsberg (Institute of Laboratory Medicine and Pathobiochemistry, Marburg University).

WP5: Helene Dukal, Christine Hohmeyer, Lennard Stütz, Viola Schwerdt, Fabian Streit, Josef Frank, Lea Sirignano (Dept. of Genetic Epidemiology, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University).

WP6: Anastasia Benedyk, Miriam Bopp, Roman Keßler, Maximilian Lückel, Verena Schuster, Christoph Vogelbacher (Dept. of Psychiatry, Marburg University). Jens Sommer,

Olaf Steinsträter (Core-Facility Brainimaging, Marburg University). Thomas W.D. Möbius (Institute of Medical Informatics and Statistics, Kiel University).

CP1: Julian Glandorf, Fabian Kormann, Arif Alkan, Fatana Wedi, Lea Henning, Alena Renker, Karina Schneider, Elisabeth Folwarczny, Dana Stenzel, Kai Wenk, Felix Picard, Alexandra Fischer, Sandra Blumenau, Beate Kleb, Doris Finholdt, Elisabeth Kinder, Tamara Wüst, Elvira Przypadlo, Corinna Brehm (Comprehensive Biomaterial Bank Marburg, Marburg University).

The FOR2107 cohort project (WP1) was approved by the Ethics Committees of the Medical Faculties, University of Marburg (AZ: 07/14) and University of Münster (AZ: 2014-422-b-S).

Conflict of Interest

Biomedical financial interests or potential conflicts of interest: Tilo Kircher received unrestricted educational grants from Servier, Janssen, Recordati, Aristo, Otsuka, neuraxpharm.

The other authors (Claas Flint, Micah Cearns, Nils Opel, Ronny Redlich, David M. A. Mehler, Daniel Emden, Nils R. Winter, Ramona Leenings, Simon B. Eickhoff, Axel Krug, Igor Nenadic, Volker Arolt, Scott Clark, Bernhard T. Baune, Xiaoyi Jiang, Udo Dannlowski, Tim Hahn) declare no conflicts of interest.

Author contributions

Claas Flint^{1, 2, 3, 4}, Micah Cearns^{1,2,3,4}, Nils Opel^{1,3}, Ronny Redlich^{1,3}, David M. A. Mehler^{1,3}, Daniel Emden^{1,3}, Nils R. Winter^{1,3}, Ramona Leenings^{1,3}, Simon B. Eickhoff^{1,3}, Tilo Kircher^{1,3}, Axel Krug^{1,3}, Igor Nenadic^{1,3}, Volker Arolt^{1,3}, Scott Clark^{1,3}, Bernhard T. Baune^{1,3}, Xiaoyi Jiang^{1,3}, Udo Dannlowski^{1,2,3,4}, Tim Hahn^{1,2,3,4}

¹ these authors contributed substantially to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work.

² these authors drafting the work or revising it critically for important intellectual content.

³ these authors did the final approval of the version to be published.

⁴ these authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Darcy AM, Louie AK, Roberts LW. Machine learning and the profession of medicine. *JAMA - J Am Med Assoc.* 2016;315:551–552.
2. Eyre HA, Singh AB, Reynolds C. Tech giants enter mental health. *World Psychiatry.* 2016;15:21–22.
3. Gabrieli JDE, Ghosh SS, Whitfield-Gabrieli S. Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron.* 2015;85:11–26.
4. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science.* 2015;349:255–260.
5. Hahn T, Nierenberg AA, Whitfield-Gabrieli S. Predictive analytics in mental health: Applications, guidelines, challenges and perspectives. *Mol Psychiatry.* 2017;22:37–43.
6. Johnston BA, Steele JD, Tolomeo S, Christmas D, Matthews K. Structural MRI-based predictions in patients with treatment-refractory depression (TRD). *PLoS One.* 2015;10:1–16.
7. Mwangi B, Ebmeier KP, Matthews K, Douglas Steele J. Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder. *Brain.* 2012;135:1508–1521.
8. Patel MJ, Andreescu C, Price JC, Edelman KL, Reynolds CF, Aizenstein HJ. Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. *Int J Geriatr Psychiatry.* 2015;30:1056–1067.
9. Neuhaus AH, Popescu FC. Sample Size, Model Robustness, and Classification Accuracy in Diagnostic Multivariate Neuroimaging Analyses. *Biol Psychiatry.* 2018;84:e81–e82.
10. Arbabshirani MR, Plis S, Sui J, Calhoun VD. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage.* 2017;145:137–165.
11. Raudys S, Jain A. Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners. *IEEE Trans Pattern Anal Mach Intell.* 1991;13:252–264.
12. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data

- hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol.* 2014;14:137.
13. Kambeitz J, Cabral C, Sacchet MD, Gotlib IH, Zahn R, Serpa MH, et al. Detecting Neuroimaging Biomarkers for Depression: A Meta-analysis of Multivariate Pattern Recognition Studies. *Biol Psychiatry.* 2017;82:330–338.
 14. Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *Neuroimage.* 2017;145:166–179.
 15. Hahn T, Ebner-Priemer U, Meyer-Lindenberg A. Transparent Artificial Intelligence – A Conceptual Framework for Evaluating AI-based Clinical Decision Support Systems. *OSF Prepr.* 2019. 2019. <https://doi.org/10.31219/OSF.IO/UZEHJ>.
 16. Varoquaux G. Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage.* 2018;180:68–77.
 17. Dannlowski U, Kugel H, Grotegerd D, Redlich R, Suchy J, Opel N, et al. NCAN Cross-Disorder Risk Variant Is Associated with Limbic Gray Matter Deficits in Healthy Subjects and Major Depression. *Neuropsychopharmacology.* 2015;40:2510–2516.
 18. Dannlowski U, Grabe HJ, Wittfeld K, Klaus J, Konrad C, Grotegerd D, et al. Multimodal imaging of a tescalcin (TESC)-regulating polymorphism (rs7294919)-specific effects on hippocampal gray matter structure. *Mol Psychiatry.* 2015;20:398–404.
 19. Kircher T, Wöhr M, Nenadic I, Schwarting R, Schratt G, Alferink J, et al. Neurobiology of the major psychoses: a translational perspective on brain structure and function—the FOR2107 consortium. *Eur Arch Psychiatry Clin Neurosci.* 2018;0:0.
 20. Wittchen H-U, Wunderlich U, Gruschwitz S, Zaudig M. SKID I. Strukturiertes Klinisches Interview für DSM-IV. Achse I: Psychische Störungen. Interviewheft und Beurteilungsheft. Eine deutschsprachige, erweiterte Bearb. d. amerikanischen Originalversion des SKID I. Göttingen: Hogrefe; 1997.
 21. Vogelbacher C, Möbius TWD, Sommer J, Schuster V, Dannlowski U, Kircher T, et al. The Marburg-Münster Affective Disorders Cohort Study (MACS): A quality assurance

- protocol for MR neuroimaging data. *Neuroimage*. 2018;172:450–460.
22. Scikit-learn. Pipeline: chaining estimators. 2018. <https://scikit-learn.org/stable/modules/compose.html#pipeline>. Accessed 29 January 2019.
 23. Marquand AF, Rezek I, Buitelaar J, Beckmann CF. Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. *Biol Psychiatry*. 2016;80:552–561.
 24. Schnack HG, Kahn RS. Detecting neuroimaging biomarkers for psychiatric disorders: Sample size matters. *Front Psychiatry*. 2016;7.

Figure 1. Workflow for overall, train, and test set size systematic overestimation analyses.

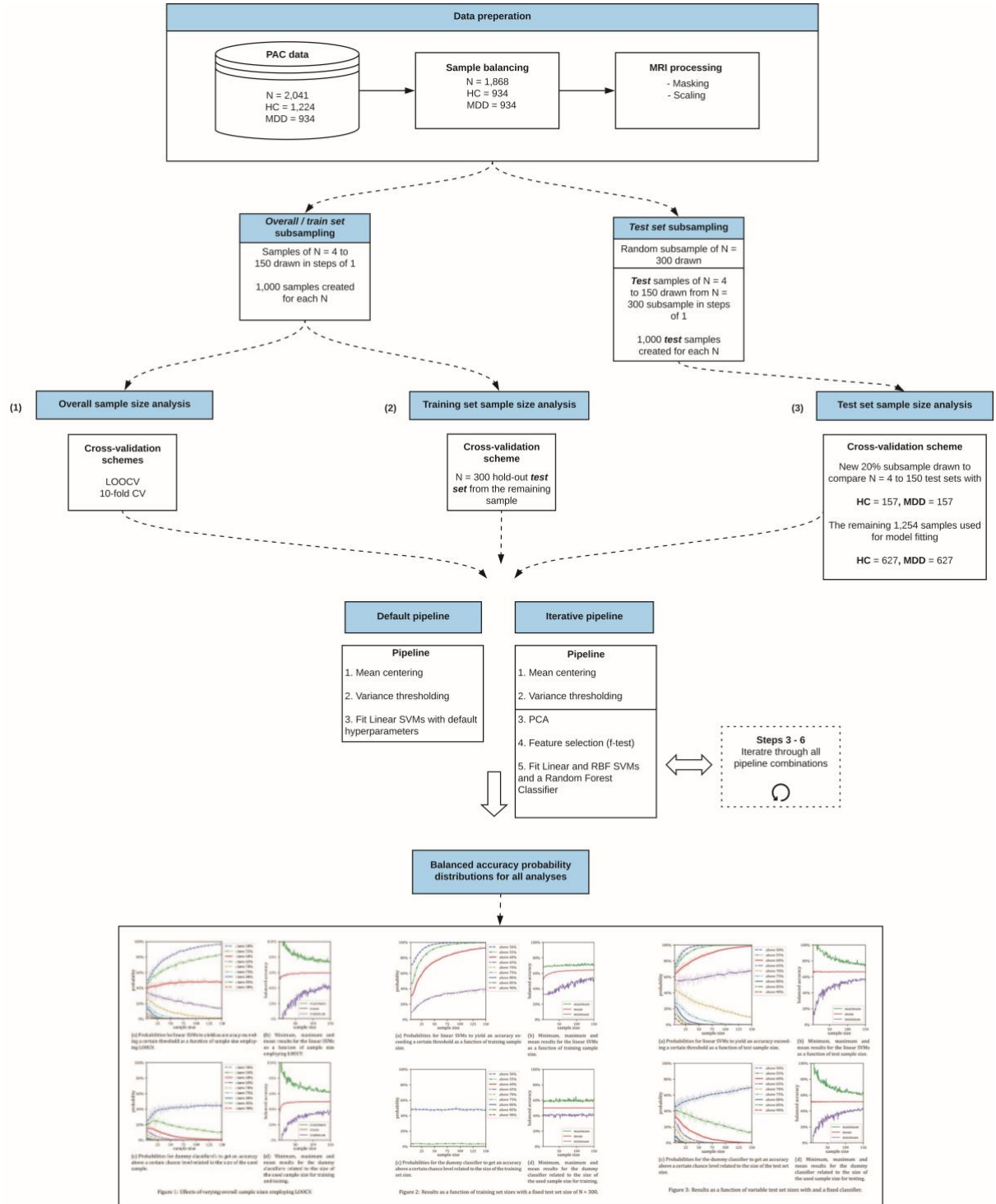


Figure 2. Effects of varying overall sample sizes employing LOOCV. (a) Probabilities for linear SVMs to yield an accuracy exceeding a certain threshold as a function of sample size employing LOOCV. (b) Minimum, maximum and mean results for the linear SVMs as a function of sample size employing LOOCV. (c) Probabilities for dummy classifiers to get an accuracy above a certain chance level related to the size of the used sample. (d) Minimum, maximum and mean results for the dummy classifiers related to the size of the used sample size for training and testing.

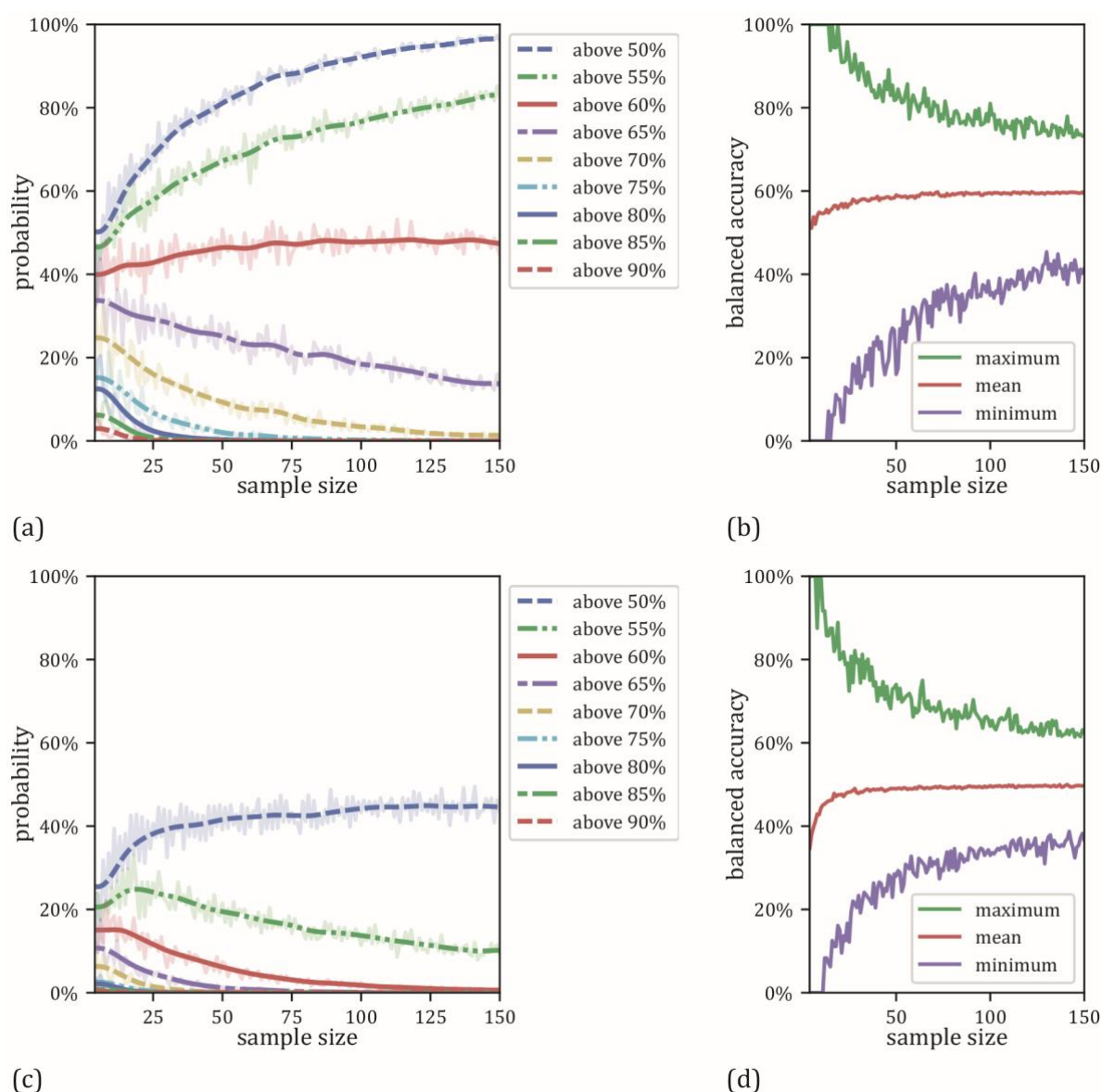


Figure 3. Results as a function of training set sizes with a fixed test set size of $N = 300$. (a) Probabilities for linear SVMs to yield an accuracy exceeding a certain threshold as a function of training sample size. (b) Minimum, maximum and mean results for the linear SVMs as a function of training sample size. (c) Probabilities for the dummy classifier to get an accuracy above a certain chance level related to the size of the training set size. (d) Minimum, maximum and mean results for the dummy classifier related to the size of the used sample size for training.

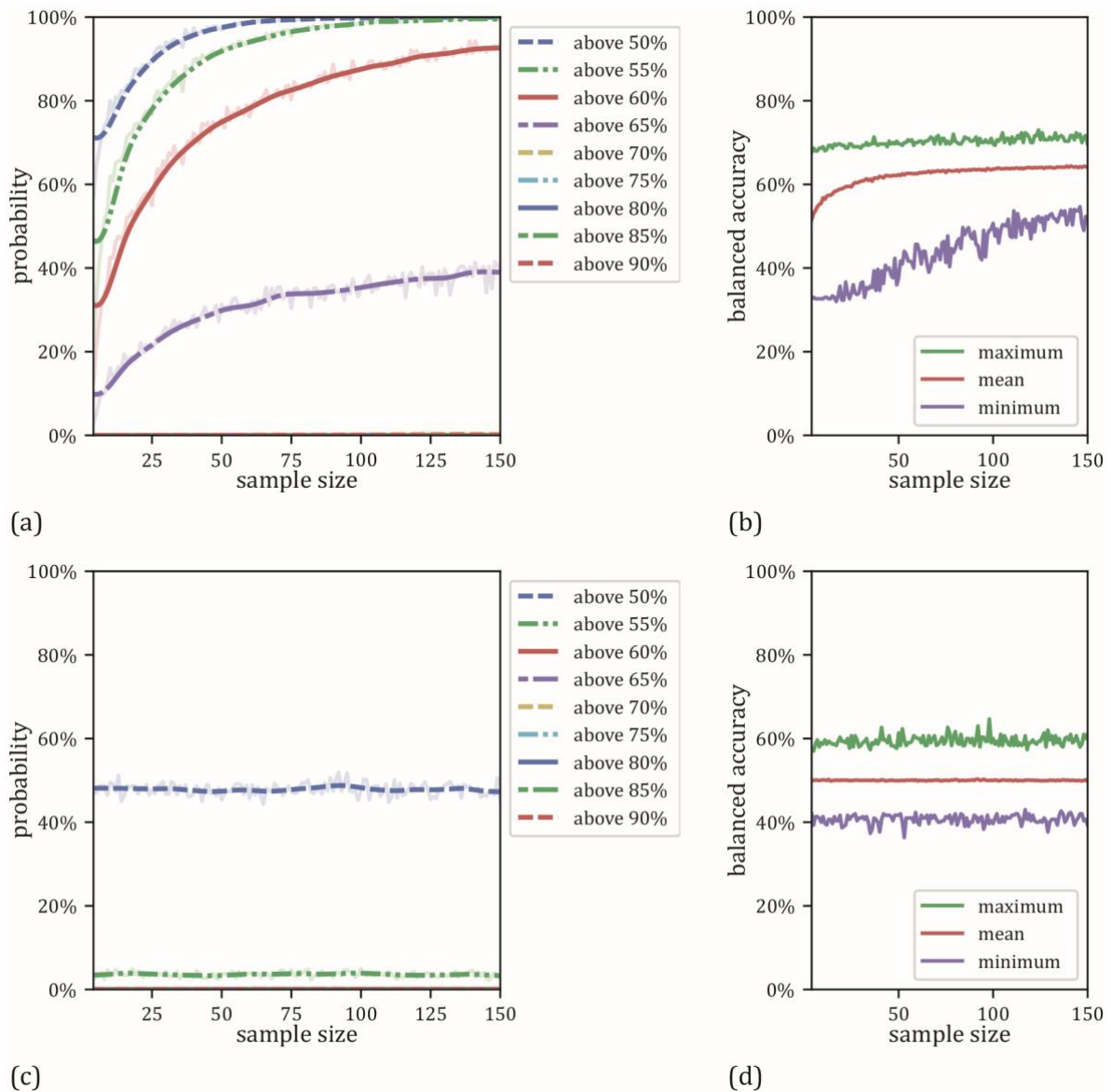
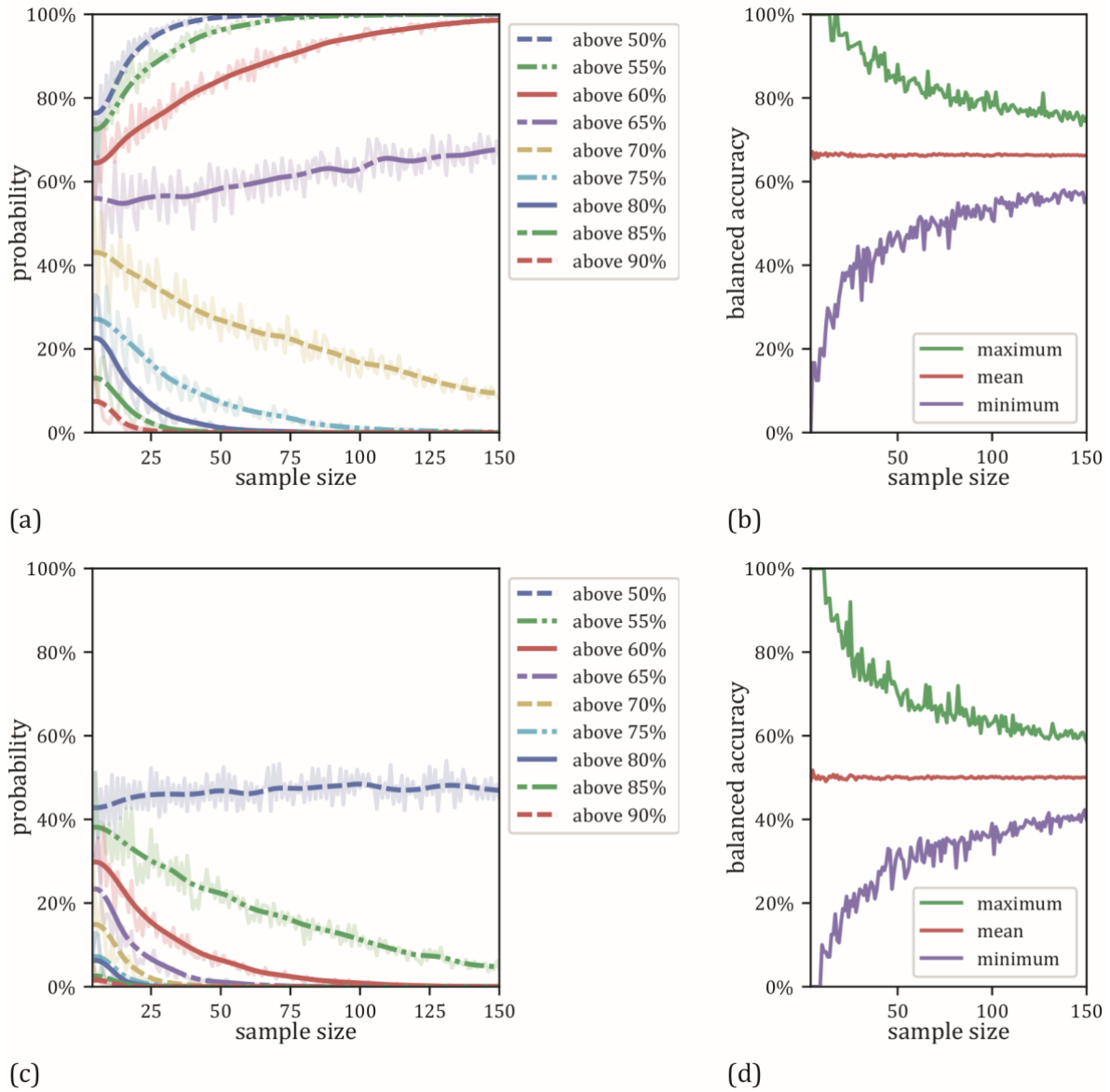


Figure 4. Results as a function of variable test set sizes with and a fixed classifier. (a) Probabilities for linear SVMs to yield an accuracy exceeding a certain threshold as a function of test sample size. (b) Minimum, maximum and mean results for the linear SVMs as a function of test sample size. (c) Probabilities for the dummy classifier to get an accuracy above a certain chance level related to the size of the test set size. (d) Minimum, maximum and mean results for the dummy classifier related to the size of the used sample size for testing.



SUPPLEMENTAL MATERIALS

Systematic Overestimation of Machine Learning Performance in Neuroimaging Studies of Depression

A Summary statistics for the final study sample

		MDD	HC
Sex	n	934	934
	female	556	548
	male	378	368
Scanner	1	285	512
	2	395	236
	3	254	186

(a) Summary of sample size (n), sex and scanner site.

		Mean	SD	Min	Max
MDD	Age ($n = 934$)	37.76	12.94	16	65
	TIV ($n = 933$)	1575.94	152.97	1115.00	2189.00
HC	Age ($n = 934$)	34.15	12.42	17	65
	TIV ($n = 933$)	1575.61	156.38	1146.23	2706.76

(b) Summage of Age (in years) and TIV (total intracranial volume).

Table A.1: Summary statistics for the final study sample. MDD = Major Depressive Disorder; HC = Healthy Controls.

B Additional tabluar presentations

This section contains the tabular presentation of the graphical visualized data in the paper.

		≥ accuracy (%)								
overall sample size	n	50	55	60	65	70	75	80	85	90
	10	0.54	0.54	0.43	0.33	0.15	0.15	0.05	0.05	0.01
	20	0.64	0.52	0.40	0.24	0.13	0.06	0.01	0.01	0.00
	30	0.70	0.61	0.39	0.29	0.11	0.06	0.01	0.00	0.00
	40	0.78	0.63	0.47	0.23	0.09	0.02	0.00	0.00	0.00
	50	0.82	0.70	0.51	0.27	0.08	0.02	0.00	0.00	0.00
	60	0.81	0.62	0.41	0.19	0.05	0.01	0.00	0.00	0.00
	70	0.88	0.72	0.42	0.20	0.07	0.01	0.00	0.00	0.00
	80	0.90	0.73	0.48	0.18	0.04	0.00	0.00	0.00	0.00
	90	0.89	0.76	0.45	0.20	0.03	0.00	0.00	0.00	0.00
	100	0.93	0.75	0.47	0.17	0.02	0.00	0.00	0.00	0.00
	125	0.94	0.81	0.48	0.14	0.01	0.00	0.00	0.00	0.00
150	0.96	0.83	0.45	0.12	0.01	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the overall sample size.

	n	mean	median	min	max	std
overall sample size	10	0.55	0.60	0.00	1.00	0.20
	20	0.56	0.60	0.10	0.95	0.15
	30	0.58	0.58	0.20	0.87	0.12
	40	0.59	0.60	0.25	0.82	0.10
	50	0.59	0.60	0.16	0.82	0.09
	60	0.58	0.58	0.32	0.82	0.08
	70	0.59	0.60	0.36	0.81	0.07
	80	0.59	0.60	0.36	0.76	0.07
	90	0.59	0.60	0.34	0.77	0.07
	100	0.60	0.60	0.35	0.81	0.06
	125	0.59	0.60	0.41	0.74	0.05
	150	0.59	0.59	0.41	0.73	0.05

(b) Summary of the accuracy value distributions for certain overall sample size.

Table B.1: Effects of varying overall sample sizes for training and testing a linear SVM employing LOOCV.

		≥ accuracy (%)								
overall sample size	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.25	0.25	0.17	0.10	0.03	0.03	0.01	0.01	0.00
	20	0.32	0.18	0.10	0.04	0.02	0.01	0.00	0.00	0.00
	30	0.33	0.22	0.06	0.03	0.01	0.00	0.00	0.00	0.00
	40	0.37	0.18	0.09	0.02	0.00	0.00	0.00	0.00	0.00
	50	0.39	0.21	0.07	0.01	0.00	0.00	0.00	0.00	0.00
	60	0.40	0.16	0.04	0.01	0.00	0.00	0.00	0.00	0.00
	70	0.40	0.16	0.02	0.00	0.00	0.00	0.00	0.00	0.00
	80	0.41	0.13	0.03	0.00	0.00	0.00	0.00	0.00	0.00
	90	0.41	0.14	0.02	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.43	0.11	0.02	0.00	0.00	0.00	0.00	0.00	0.00
	125	0.46	0.13	0.01	0.00	0.00	0.00	0.00	0.00	0.00
150	0.44	0.09	0.01	0.00	0.00	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the overall sample size.

	<i>n</i>	mean	median	min	max	std
overall sample size	10	0.44	0.40	0.00	1.00	0.17
	20	0.47	0.45	0.15	0.80	0.11
	30	0.48	0.47	0.23	0.80	0.09
	40	0.49	0.50	0.25	0.70	0.08
	50	0.49	0.48	0.28	0.74	0.07
	60	0.49	0.50	0.28	0.68	0.06
	70	0.49	0.50	0.31	0.69	0.06
	80	0.49	0.50	0.30	0.66	0.06
	90	0.49	0.49	0.34	0.68	0.05
	100	0.50	0.50	0.33	0.64	0.05
	125	0.50	0.50	0.33	0.63	0.05
	150	0.50	0.49	0.37	0.63	0.04

(b) Summary of the accuracy value distributions for certain overall sample size.

Table B.2: Effects of varying overall sample sizes for training and testing a Dummy Classifier employing LOOCV.

train sample size	$\geq$ accuracy (%)									
	n	50	55	60	65	70	75	80	85	90
	10	0.80	0.62	0.42	0.16	0.00	0.00	0.00	0.00	0.00
	20	0.87	0.75	0.55	0.20	0.00	0.00	0.00	0.00	0.00
	30	0.94	0.83	0.65	0.26	0.00	0.00	0.00	0.00	0.00
	40	0.97	0.90	0.72	0.29	0.00	0.00	0.00	0.00	0.00
	50	0.97	0.93	0.76	0.31	0.00	0.00	0.00	0.00	0.00
	60	0.99	0.95	0.76	0.30	0.00	0.00	0.00	0.00	0.00
	70	0.99	0.96	0.83	0.35	0.00	0.00	0.00	0.00	0.00
	80	0.99	0.97	0.83	0.34	0.00	0.00	0.00	0.00	0.00
90	1.00	0.98	0.89	0.36	0.00	0.00	0.00	0.00	0.00	
100	1.00	0.98	0.87	0.35	0.00	0.00	0.00	0.00	0.00	
125	1.00	0.99	0.91	0.39	0.00	0.00	0.00	0.00	0.00	
150	1.00	0.99	0.94	0.39	0.00	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the train sample size.

train sample size	<i>n</i>	mean	median	min	max	std
	10	0.57	0.58	0.33	0.69	0.08
	20	0.59	0.61	0.32	0.69	0.07
	30	0.61	0.62	0.36	0.69	0.06
	40	0.62	0.63	0.35	0.70	0.05
	50	0.62	0.63	0.38	0.70	0.05
	60	0.63	0.64	0.46	0.70	0.04
	70	0.63	0.64	0.45	0.71	0.04
	80	0.63	0.64	0.45	0.70	0.03
	90	0.64	0.64	0.49	0.71	0.03
	100	0.64	0.64	0.51	0.70	0.03
	125	0.64	0.64	0.51	0.71	0.03
	150	0.64	0.65	0.51	0.70	0.03

(b) Summary of the accuracy value distributions for certain train sample size.

Table B.3: Results of a fixed validation set of $n = 300$ and varying train sample sizes on a linear SVM using default parameter. ($C = 1$)

train sample size	≥ accuracy (%)									
	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.49	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	20	0.47	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	30	0.47	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	40	0.47	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	50	0.45	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	60	0.46	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	70	0.48	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	80	0.49	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	90	0.47	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.48	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
125	0.45	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
150	0.45	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the train sample size.

train sample size	<i>n</i>	mean	median	min	max	std
	10	0.50	0.50	0.41	0.59	0.03
20		0.50	0.50	0.39	0.59	0.03
30		0.50	0.50	0.41	0.58	0.03
40		0.50	0.50	0.38	0.60	0.03
50		0.50	0.50	0.42	0.61	0.03
60		0.50	0.50	0.40	0.60	0.03
70		0.50	0.50	0.41	0.58	0.03
80		0.50	0.50	0.42	0.62	0.03
90		0.50	0.50	0.42	0.58	0.03
100		0.50	0.50	0.40	0.60	0.03
125		0.50	0.50	0.40	0.59	0.03
150		0.50	0.50	0.39	0.61	0.03

(b) Summary of the accuracy value distributions for certain train sample size.

Table B.4: Results of a fixed validation set of $n = 300$ and varying train sample sizes on a Dummy Classifier.

		$\geq$ accuracy (%)								
test sample size	n	50	55	60	65	70	75	80	85	90
	10	0.78	0.78	0.66	0.54	0.29	0.29	0.10	0.10	0.02
	20	0.91	0.81	0.69	0.48	0.30	0.17	0.06	0.06	0.00
	30	0.96	0.90	0.73	0.57	0.27	0.15	0.02	0.01	0.00
	40	0.98	0.91	0.80	0.49	0.25	0.09	0.02	0.01	0.00
	50	0.99	0.97	0.85	0.58	0.22	0.07	0.00	0.00	0.00
	60	1.00	0.97	0.84	0.56	0.23	0.05	0.00	0.00	0.00
	70	1.00	0.99	0.88	0.61	0.21	0.04	0.00	0.00	0.00
	80	1.00	0.99	0.90	0.56	0.17	0.02	0.00	0.00	0.00
	90	1.00	1.00	0.94	0.63	0.17	0.02	0.00	0.00	0.00
	100	1.00	0.99	0.94	0.58	0.13	0.01	0.00	0.00	0.00
	125	1.00	1.00	0.97	0.63	0.13	0.00	0.00	0.00	0.00
150	1.00	1.00	0.98	0.66	0.08	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	<i>n</i>	mean	median	min	max	std
test sample size	10	0.66	0.70	0.20	1.00	0.15
	20	0.67	0.65	0.35	0.95	0.11
	30	0.66	0.67	0.40	0.93	0.08
	40	0.66	0.65	0.42	0.88	0.07
	50	0.66	0.66	0.48	0.88	0.06
	60	0.66	0.67	0.47	0.83	0.06
	70	0.66	0.67	0.50	0.83	0.05
	80	0.66	0.66	0.49	0.80	0.05
	90	0.66	0.67	0.53	0.78	0.04
	100	0.66	0.66	0.51	0.79	0.04
	125	0.66	0.66	0.54	0.78	0.03
	150	0.66	0.66	0.56	0.75	0.03

(b) Summary of the accuracy value distributions for certain test sample size.

Table B.5: Effects of varying test sample sizes on a fixed pretrained linear SVM.

test sample size	≥ accuracy (%)									
	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.39	0.39	0.29	0.17	0.05	0.05	0.01	0.01	0.00
	20	0.42	0.24	0.13	0.05	0.02	0.01	0.00	0.00	0.00
	30	0.42	0.28	0.10	0.05	0.01	0.00	0.00	0.00	0.00
	40	0.44	0.19	0.08	0.01	0.00	0.00	0.00	0.00	0.00
	50	0.44	0.20	0.06	0.01	0.00	0.00	0.00	0.00	0.00
	60	0.45	0.15	0.03	0.00	0.00	0.00	0.00	0.00	0.00
	70	0.45	0.17	0.02	0.00	0.00	0.00	0.00	0.00	0.00
	80	0.44	0.10	0.01	0.00	0.00	0.00	0.00	0.00	0.00
90	0.43	0.12	0.01	0.00	0.00	0.00	0.00	0.00	0.00	
100	0.47	0.10	0.01	0.00	0.00	0.00	0.00	0.00	0.00	
125	0.53	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
150	0.45	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

test sample size	n	mean	median	min	max	std
	10	0.50	0.50	0.10	1.00	0.16
	20	0.50	0.50	0.15	0.85	0.11
	30	0.50	0.50	0.23	0.73	0.09
	40	0.50	0.50	0.28	0.72	0.07
	50	0.50	0.50	0.32	0.70	0.06
	60	0.50	0.50	0.30	0.67	0.06
	70	0.50	0.50	0.34	0.67	0.05
	80	0.50	0.50	0.35	0.64	0.05
	90	0.50	0.50	0.36	0.67	0.04
	100	0.50	0.50	0.39	0.63	0.04
	125	0.50	0.50	0.39	0.62	0.04
	150	0.50	0.50	0.40	0.59	0.03

(b) Summary of the accuracy value distributions for certain test sample size.

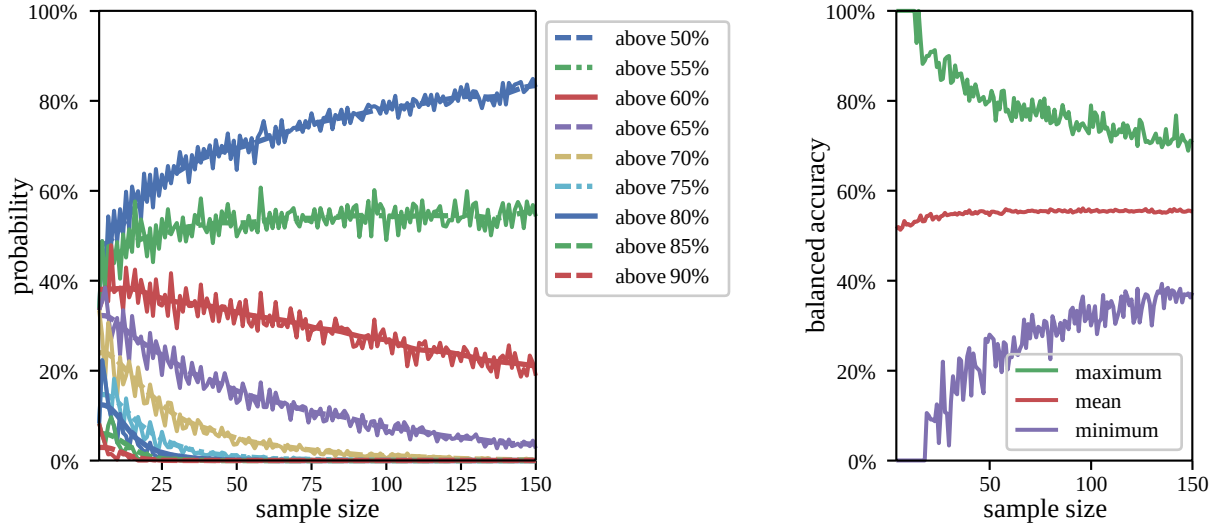
Table B.6: Effects of varying test sample sizes on a fixed pretrained Dummy Classifier.

C Influence of the scanner distribution

Scanner site distributions were not well balanced in the PAC sample. This imbalance was even stronger in our randomly drawn sub-samples. In order to determine the influence of the different scanner-sites on model accuracy, we took the same methodical approach that we used for the “train sample size effect analysis” and determined the scanner distribution for each set. Following, we calculated Spearman’s rho between the scanner-distribution and the accuracy of the hold-out test set ($n = 300$). The scanner-distribution for each set was approximated using Gini’s index. Here, we show that the scanner distribution has a statistically significant influence on test set accuracy but explains only 0.79% of the variance.

D Adjustment of SVM regularization based on sample size

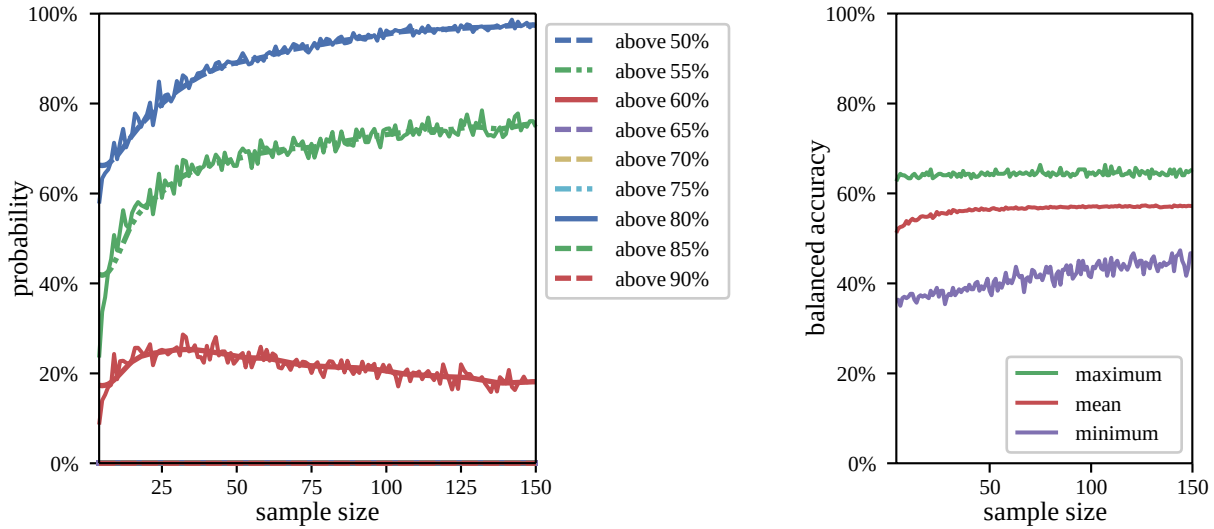
The regularization of a linear SVM depends on the absolute number of outliers in a sample. To exclude the possibility that the use of default hyperparameters (a constant value of $C = 1$) caused the effect that we observed, we have adjusted the SVM regularization (our C hyperparameter) based on the size of each analyses sub-sample. In this adjustment, a constant value k is divided by the sample size. In order to approximate the default parameter $C = 1$ on average, we have set $k = 75$. For example, C would be $C = 100/75 = 1.33$ for a sample size of $n = 100$. In this case, the adjusted regularization did not deliver results as good as what we observed with $C = 1$ (see Figure D.1, D.2). Therefore, we conclude that the default value of $C = 1$ across changing N values did not increase the probability of overestimating accuracy as sample size decreased.



(a) Probabilities for linear SVMs to yield an accuracy exceeding a certain threshold as a function of sample size employing LOOCV.

(b) Minimum, maximum and mean results for the linear SVMs as a function of sample size employing LOOCV.

Figure D.1: Effects of varying overall sample sizes employing LOOCV with an adjusted C parameter with $C = \frac{75}{\text{sample size}}$.



(a) Probabilities for linear SVMs to yield an accuracy exceeding a certain threshold as a function of training sample size.

(b) Minimum, maximum and mean results for the linear SVMs as a function of training sample size.

Figure D.2: Results as a function of training set sizes with a fixed test set size of $N = 300$ and an adjusted C parameter with $C = \frac{75}{\text{sample size}}$.

E Alternative Machine Configurations

In order to exclude the possibility that the observed effects can only be traced back to our specific configuration, we have tested other usual configurations. The configurations consist of a preprocessing and a classification. In the preprocessing step, we used a method for reduction of feature space and a method for features selection. The reduction of the feature space was achieved with a Principal Component Analysis (PCA), where only a certain number the first components were used. Afterwards, for feature selection, an ANOVA was calculated and a specific number of feature beginning with the highest F-value were taken. The actual configuration used for preprocessing is listed in Table E.1. For the classification, we chose three specific machines:

1. An SVM with a linear kernel and default parameter. ($C = 1.0$)
2. An SVM with a rbf kernel and default parameter ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)
3. A Random Forrest and default parameter ($n_{\text{estimators}} = 100$)

In combination with the preprocessing, this results in 48 configurations. Since the found effect is limited to the test set size, we have only repeated our analyses for the test set for each of these configurations.

		PCA $n_{\text{components}}$				
		no	all	500	50	10
ANOVA k_{best}	no	×	×	×	×	×
	10	×	×	×	×	
	100	×	×	×		
	1000	×	×			
	10 000	×				
	50 000	×				

Table E.1: The $\times$ marks the used combinations of configurations for the preprocessing. The *no*-label means that this preprocessing step were left out.

E.1 Results

The results are comparable to the original used configuration, an SVM with a linear kernel and no preprocessing (see Figure E.1 to E.48 and Table E.2 to E.49). The results thus underline a generally valid character of the findings. An

outlier in the results can be found for a specific configuration, a PCA with 10 components and the SVM with rbf kernel (see Figure E.20 and Table E.21). This result can be explained due overfitting, the machine constantly returns one constant class as prediction.

E.1.1 Graphical results representation

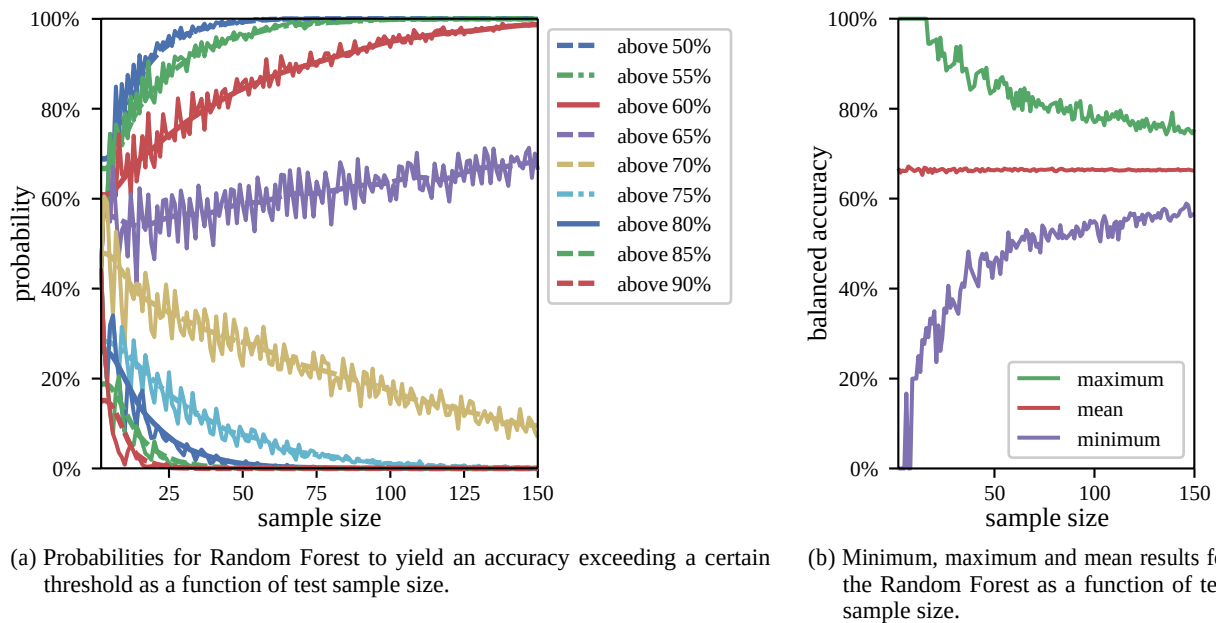


Figure E.1: Results as a function of variable test set sizes with a fixed classifier. A **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)

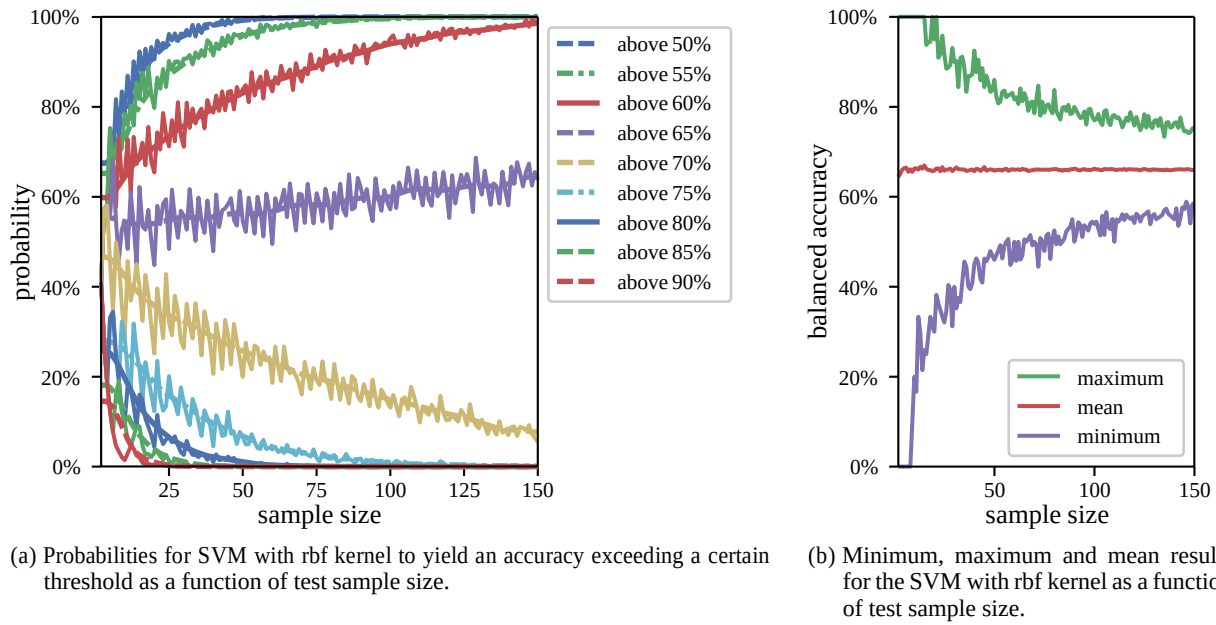


Figure E.2: Results as a function of variable test set sizes with a fixed classifier. An **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)

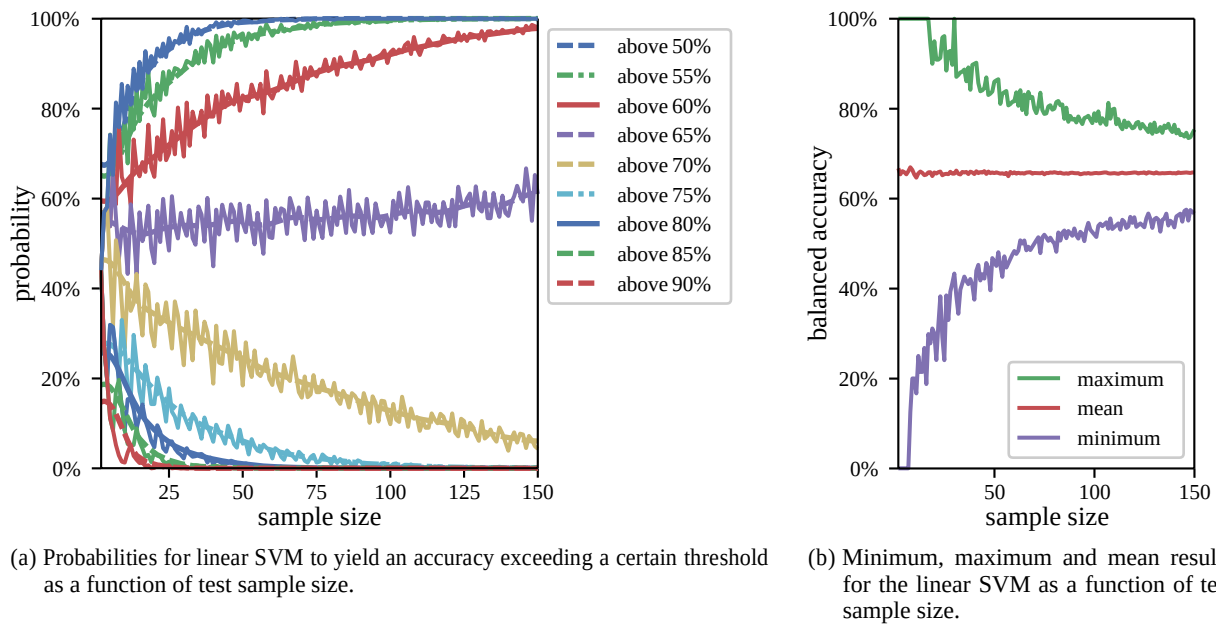
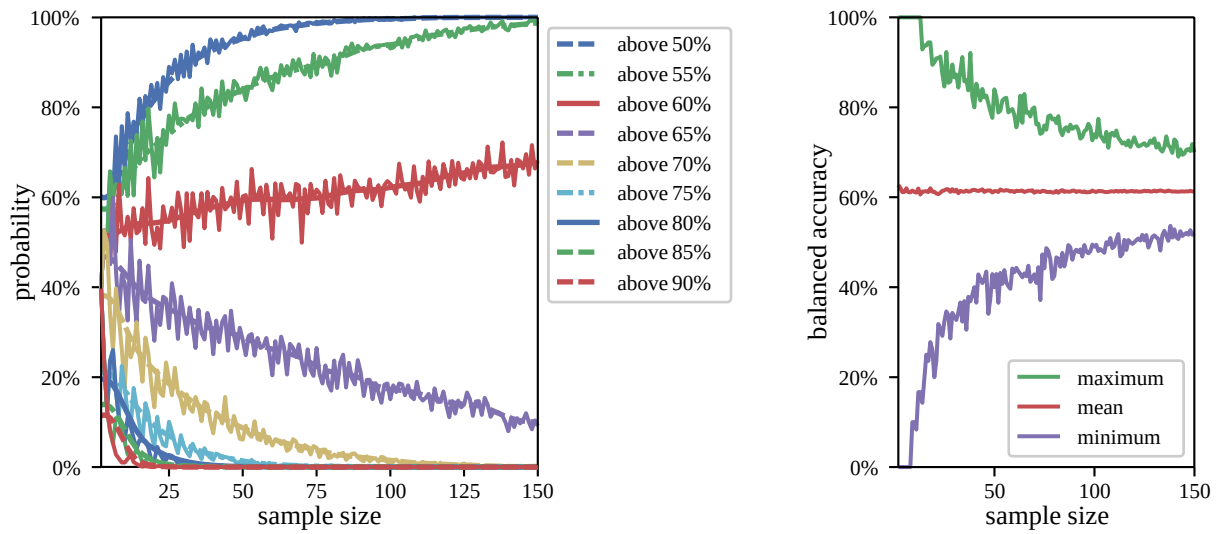


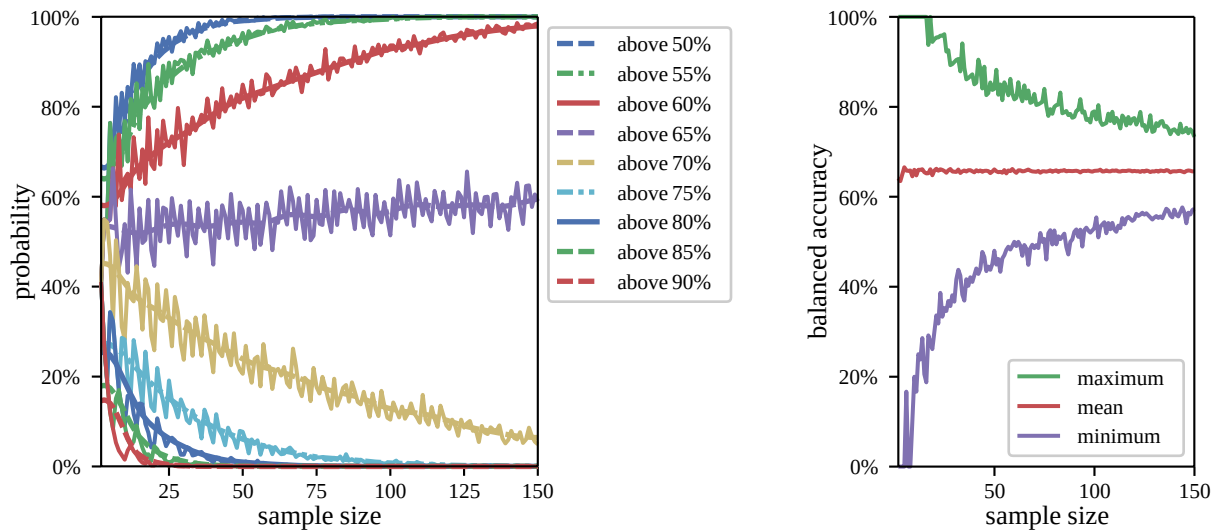
Figure E.3: Results as a function of variable test set sizes with a fixed classifier. An **linear SVM** was trained with default parameters. ($C = 1.0$)



(a) Probabilities for Random Forest to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the Random Forest as a function of test sample size.

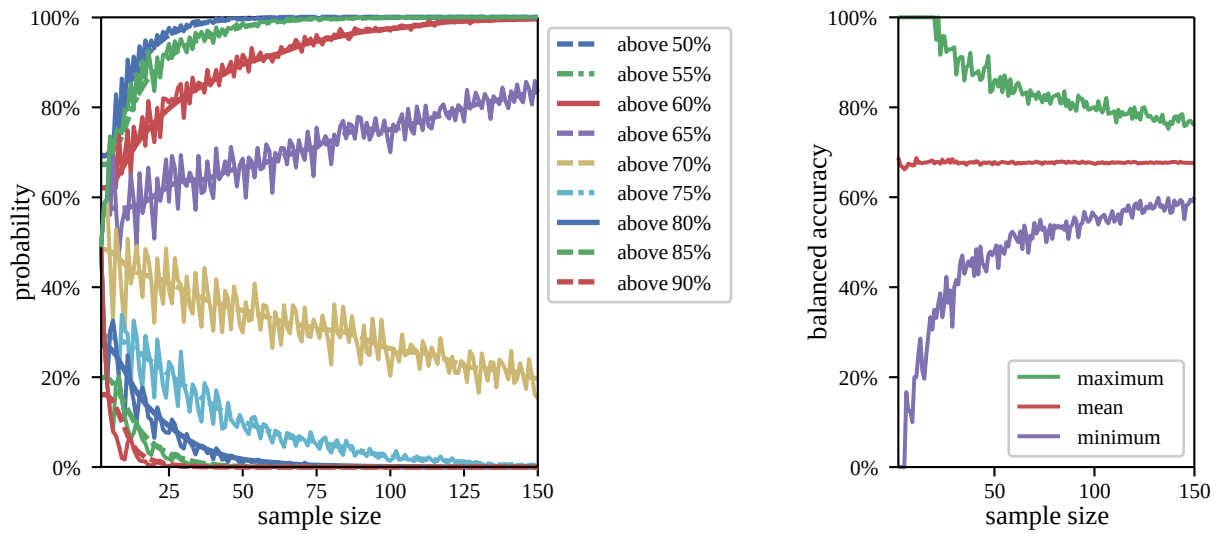
Figure E.4: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)



(a) Probabilities for SVM with rbf kernel to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the SVM with rbf kernel as a function of test sample size.

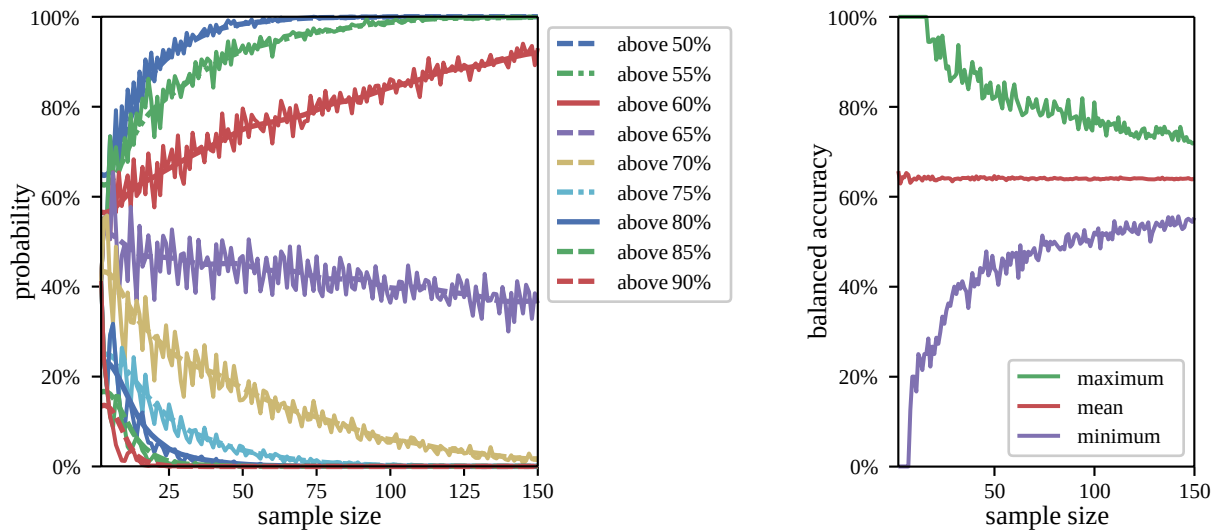
Figure E.5: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)



(a) Probabilities for linear SVM to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the linear SVM as a function of test sample size.

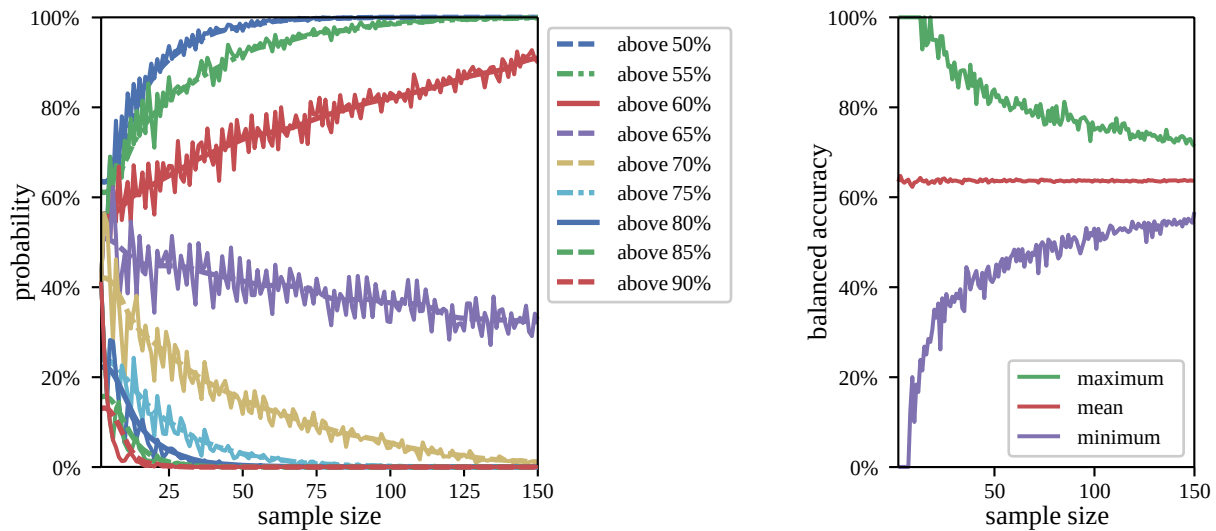
Figure E.6: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)



(a) Probabilities for Random Forest to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the Random Forest as a function of test sample size.

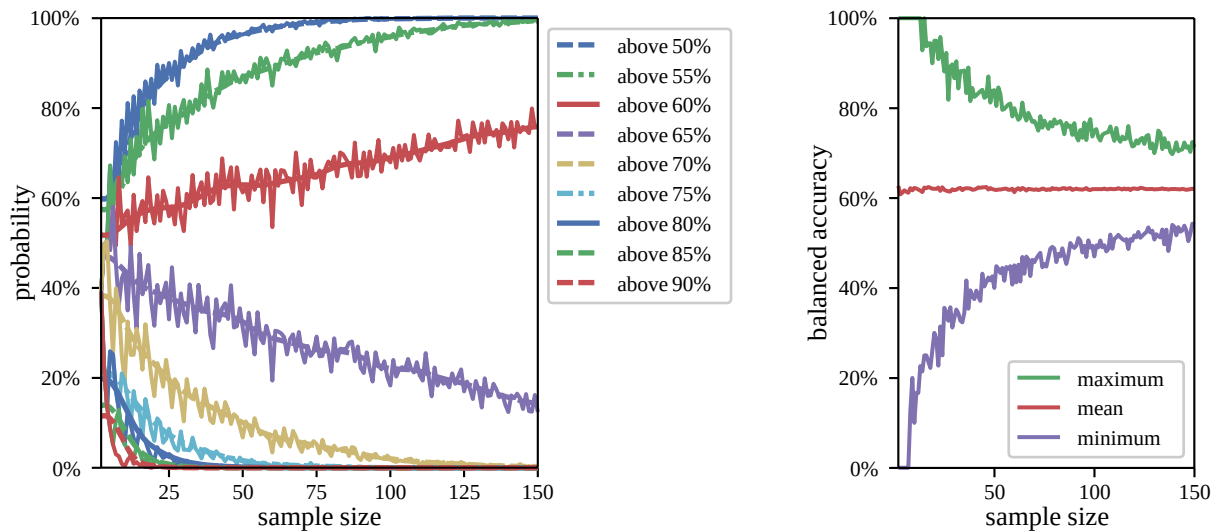
Figure E.7: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the pipeline and the top 100 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)



(a) Probabilities for SVM with rbf kernel to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the SVM with rbf kernel as a function of test sample size.

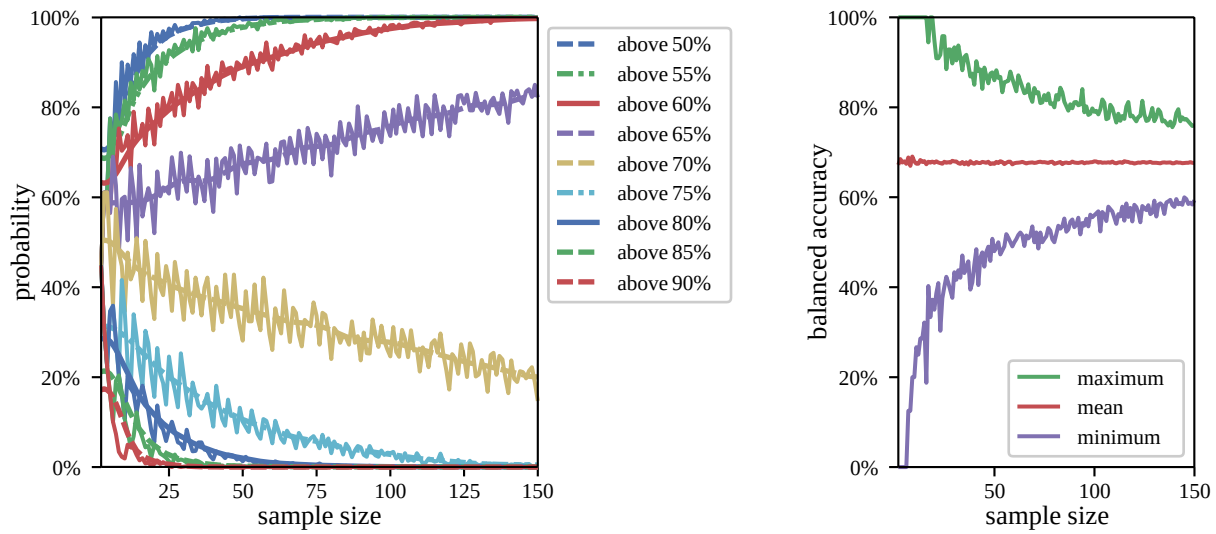
Figure E.8: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the pipeline and the top 100 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)



(a) Probabilities for linear SVM to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the linear SVM as a function of test sample size.

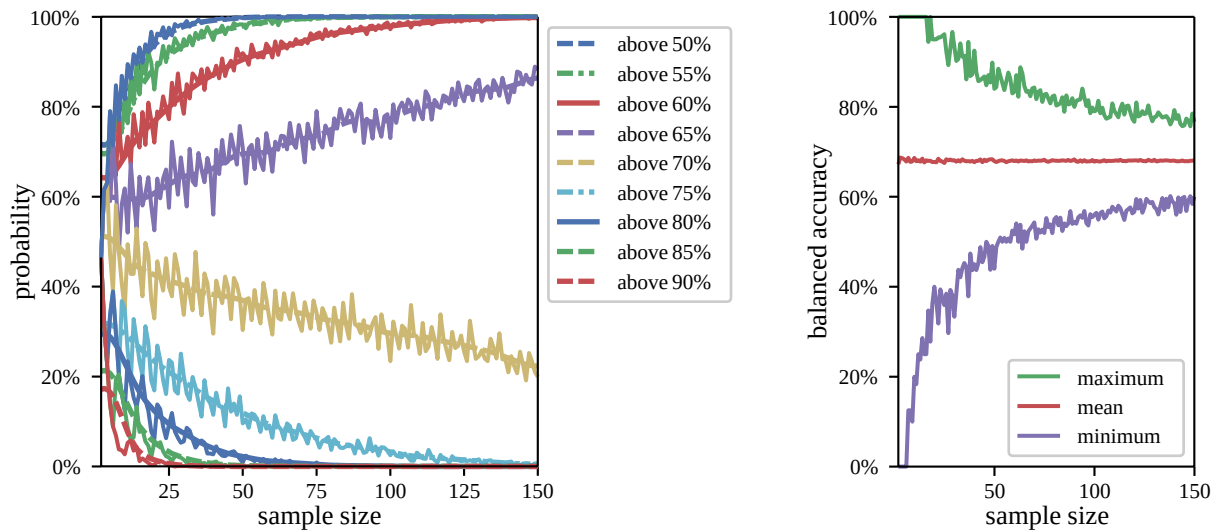
Figure E.9: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the pipeline and the top 100 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)



(a) Probabilities for Random Forest to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the Random Forest as a function of test sample size.

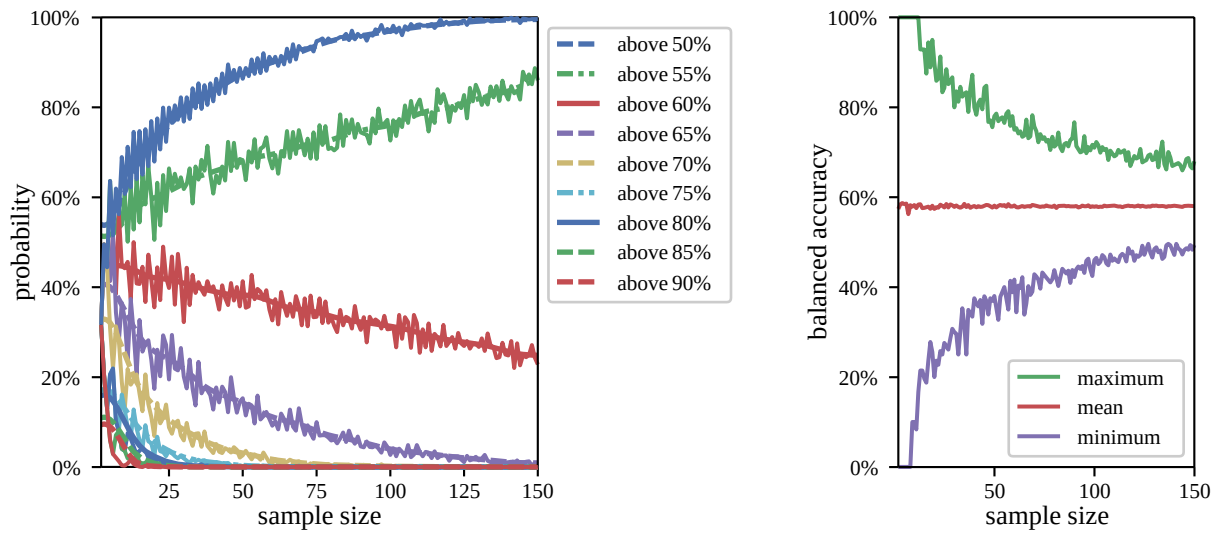
Figure E.10: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 1000 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)



(a) Probabilities for SVM with rbf kernel to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the SVM with rbf kernel as a function of test sample size.

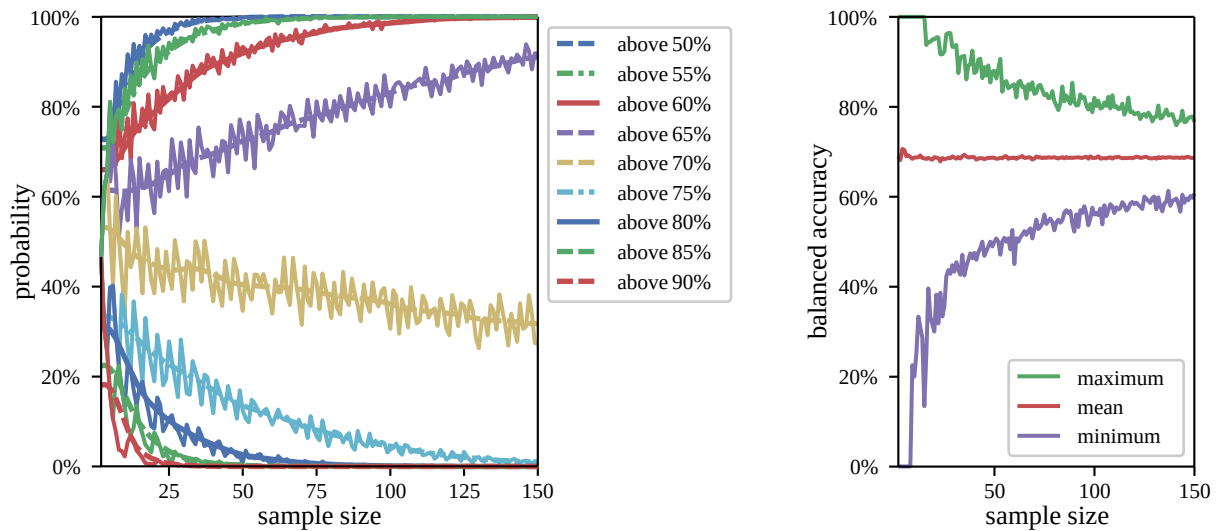
Figure E.11: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 1000 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)



(a) Probabilities for linear SVM to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the linear SVM as a function of test sample size.

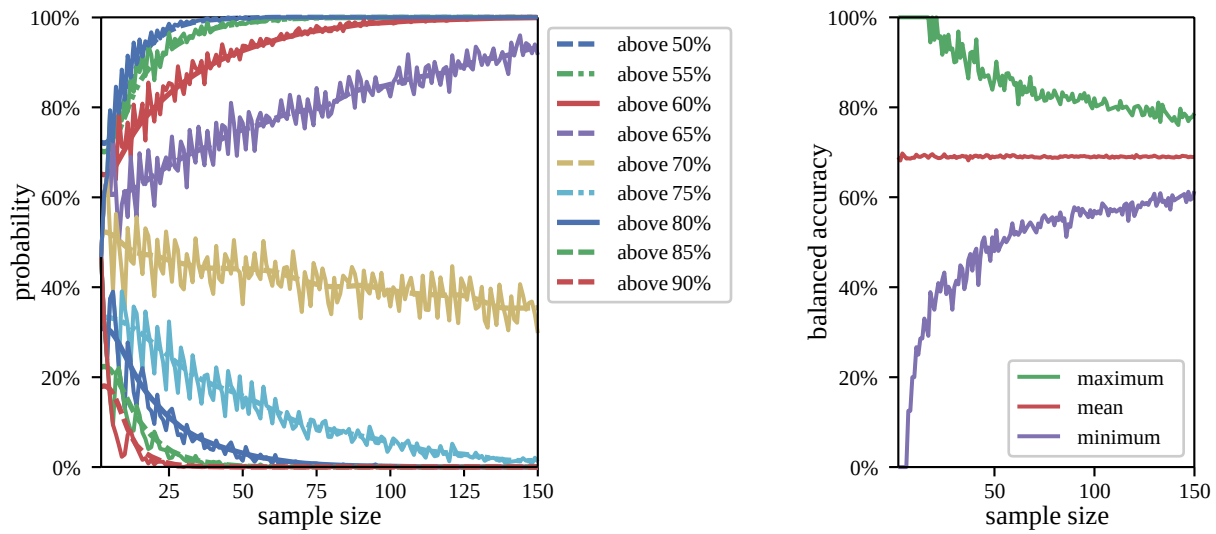
Figure E.12: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 1000 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)



(a) Probabilities for Random Forest to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the Random Forest as a function of test sample size.

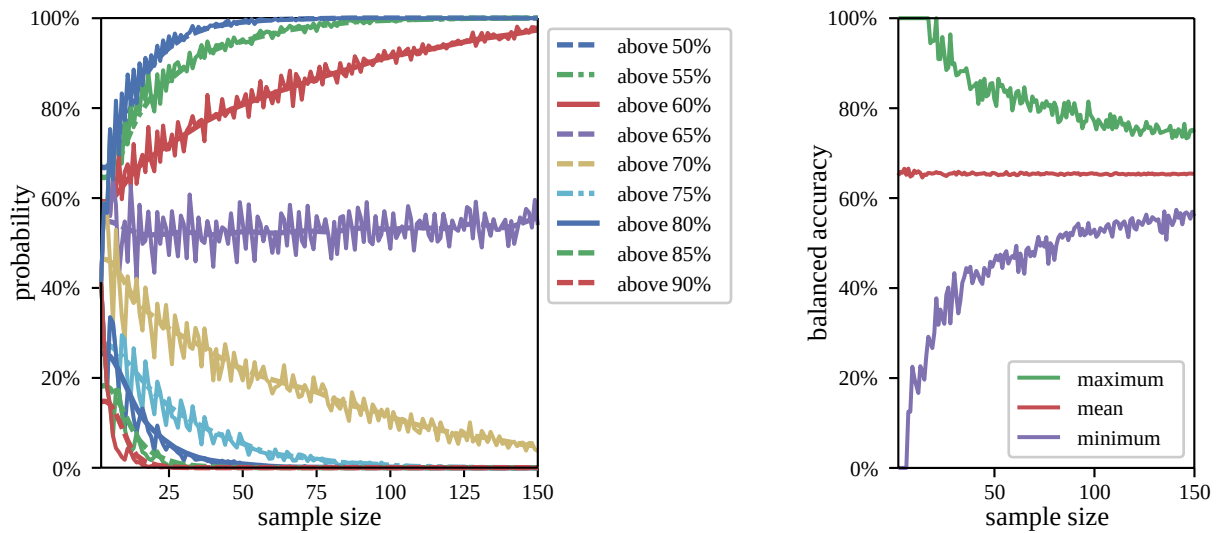
Figure E.13: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 000 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)



(a) Probabilities for SVM with rbf kernel to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the SVM with rbf kernel as a function of test sample size.

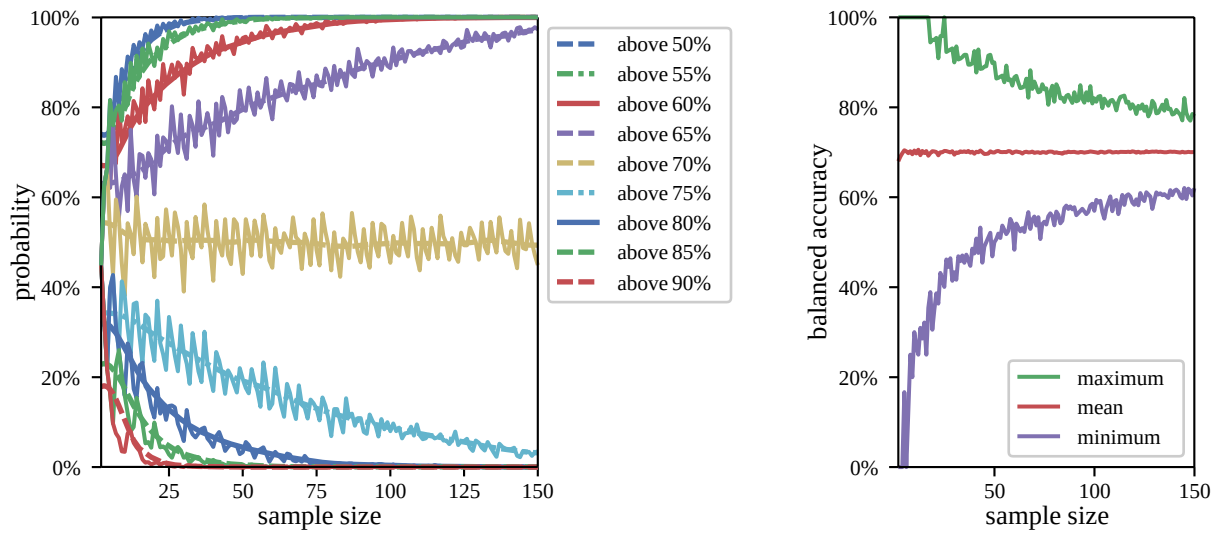
Figure E.14: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 000 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)



(a) Probabilities for linear SVM to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the linear SVM as a function of test sample size.

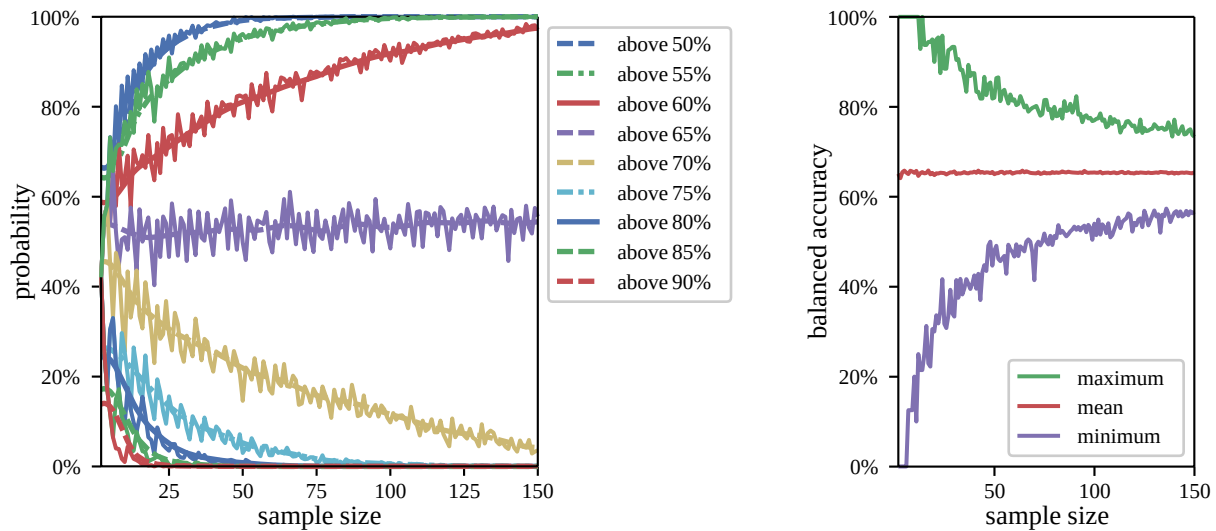
Figure E.15: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 000 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)



(a) Probabilities for Random Forest to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the Random Forest as a function of test sample size.

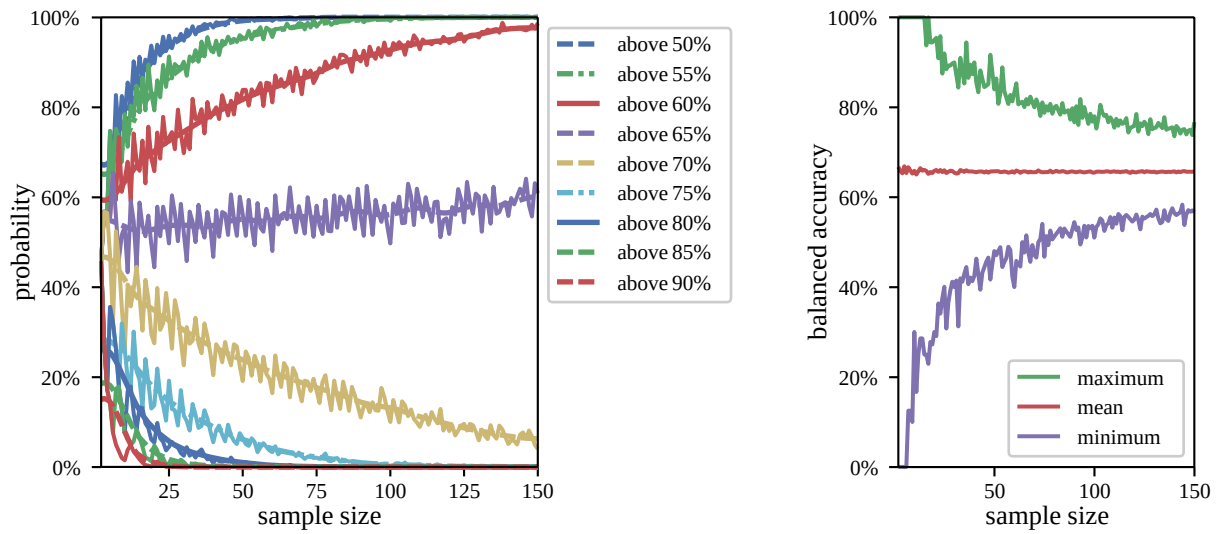
Figure E.16: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 50 000 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)



(a) Probabilities for SVM with rbf kernel to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the SVM with rbf kernel as a function of test sample size.

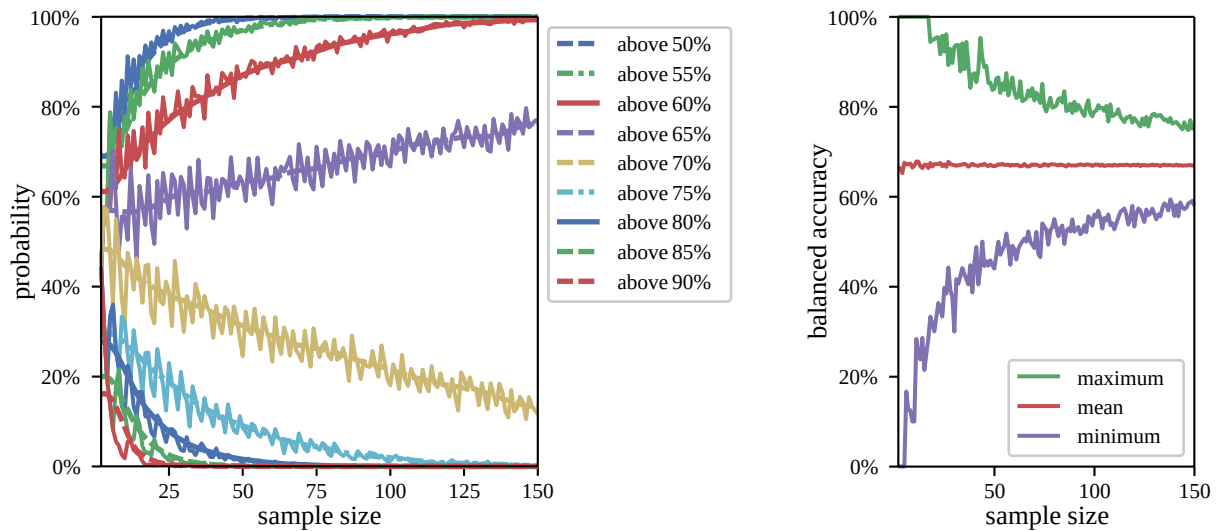
Figure E.17: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 50 000 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)



(a) Probabilities for linear SVM to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the linear SVM as a function of test sample size.

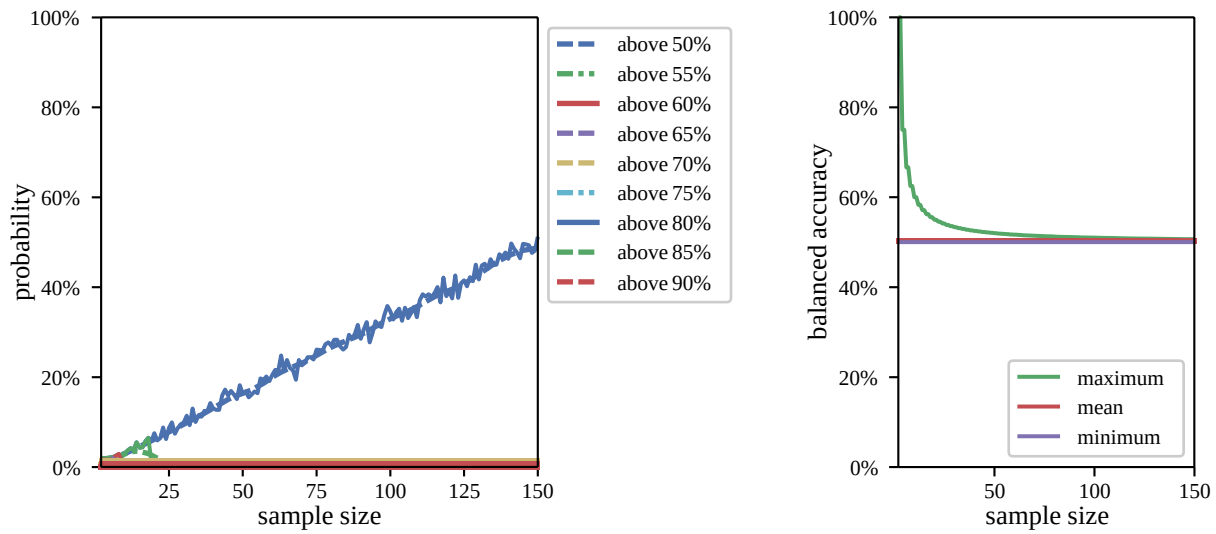
Figure E.18: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 50 000 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)



(a) Probabilities for Random Forest to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the Random Forest as a function of test sample size.

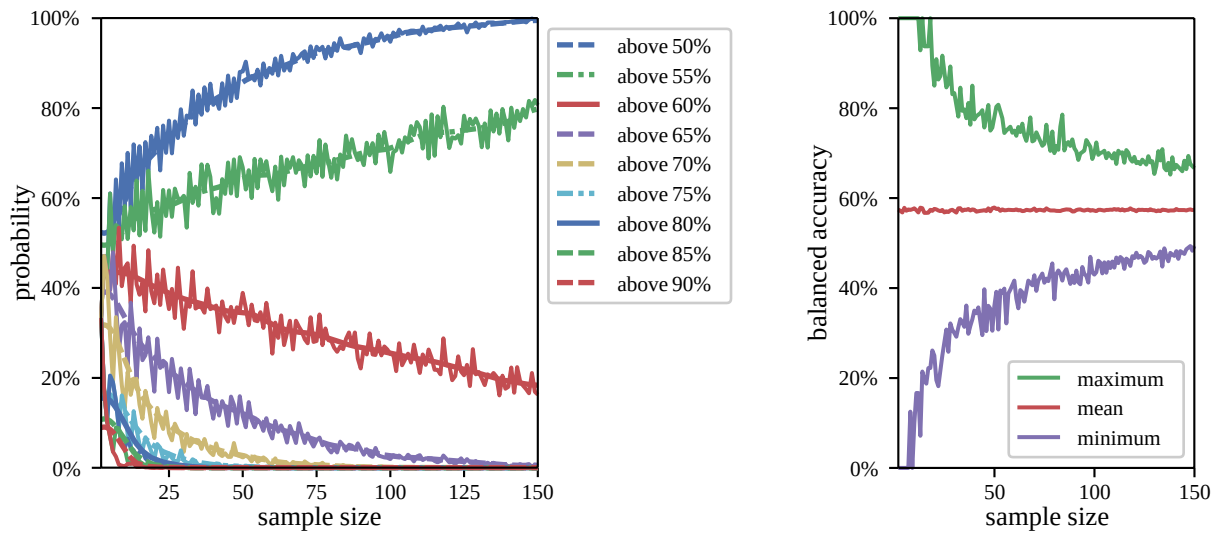
Figure E.19: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 10 **components** were retained. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)



(a) Probabilities for SVM with rbf kernel to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the SVM with rbf kernel as a function of test sample size.

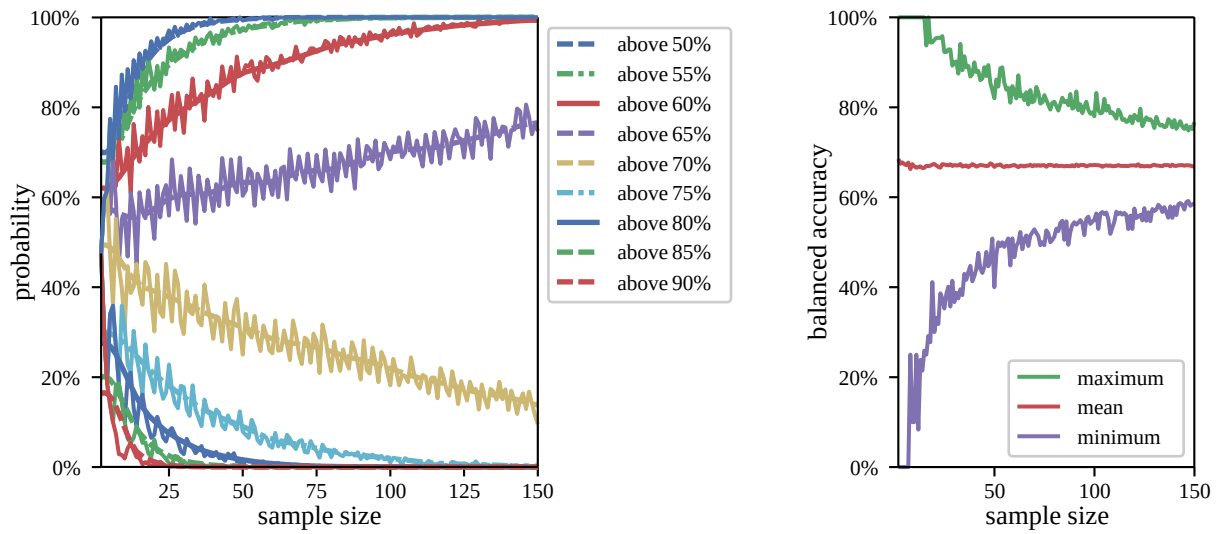
Figure E.20: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 10 **components** were retained. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)



(a) Probabilities for linear SVM to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the linear SVM as a function of test sample size.

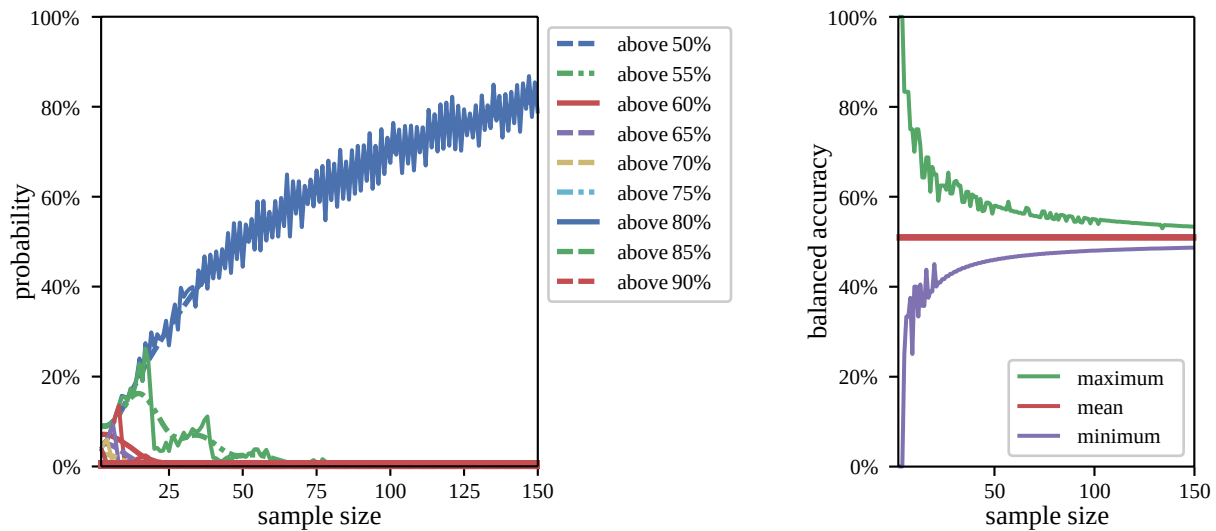
Figure E.21: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 10 **components** were retained. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)



(a) Probabilities for Random Forest to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the Random Forest as a function of test sample size.

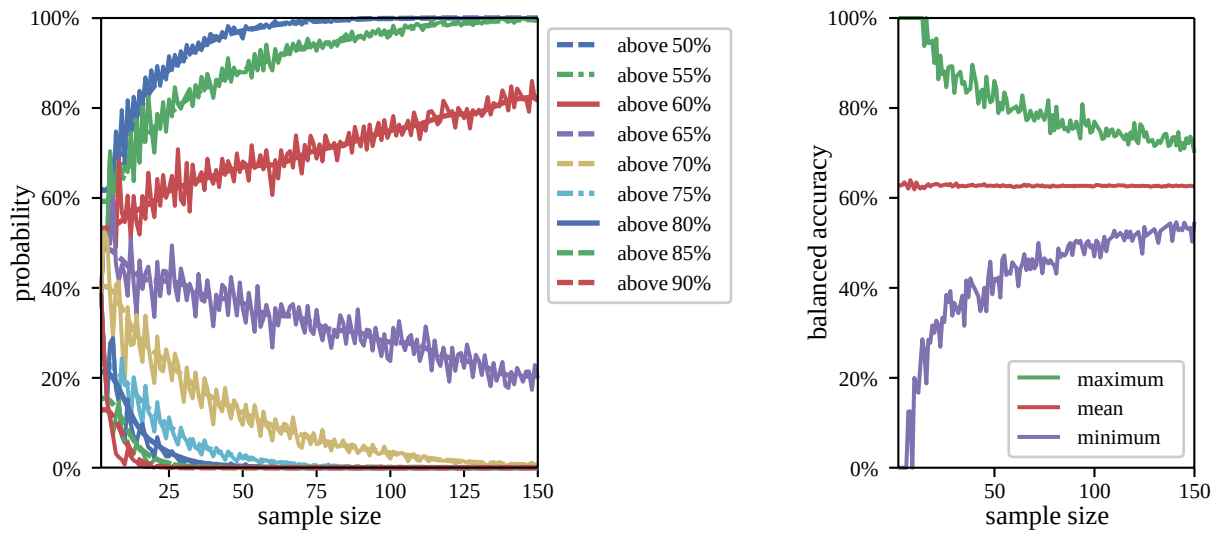
Figure E.22: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 50 **components** were retained. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)



(a) Probabilities for SVM with rbf kernel to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the SVM with rbf kernel as a function of test sample size.

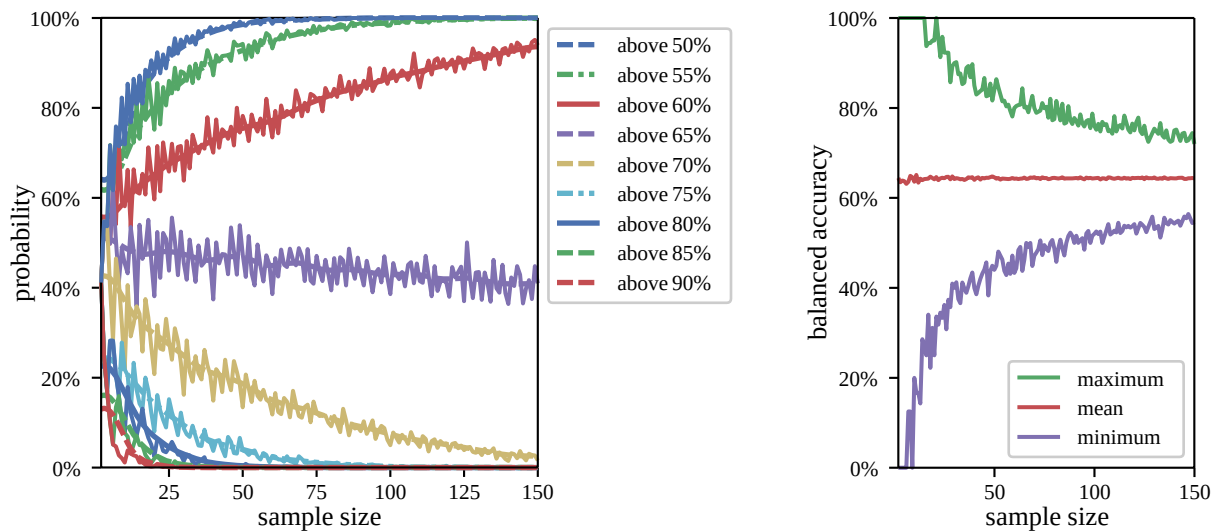
Figure E.23: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 50 **components** were retained. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)



(a) Probabilities for linear SVM to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the linear SVM as a function of test sample size.

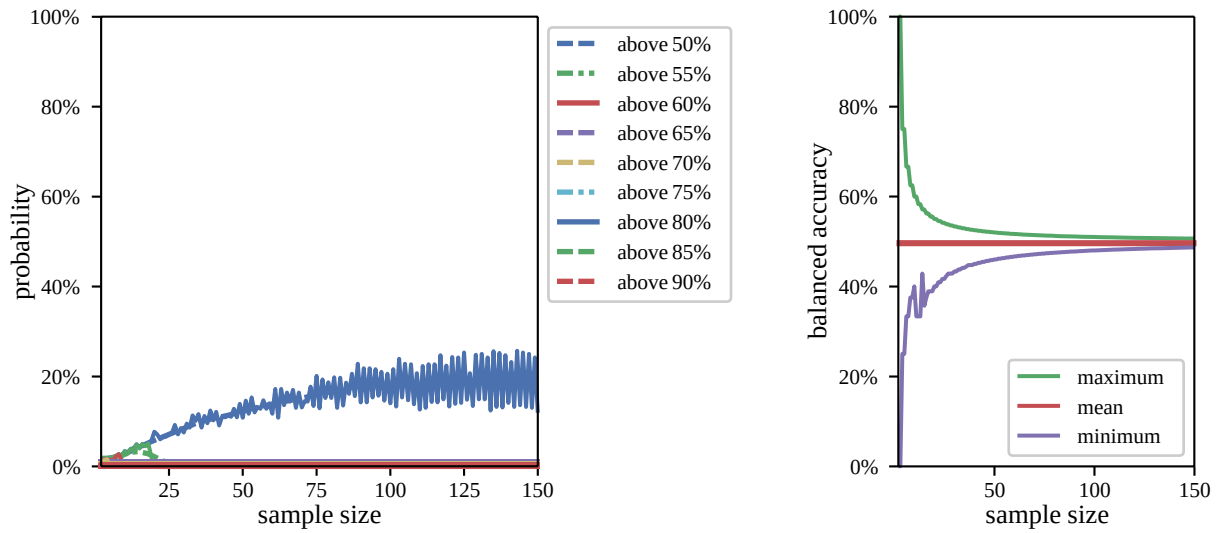
Figure E.24: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 50 **components** were retained. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)



(a) Probabilities for Random Forest to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the Random Forest as a function of test sample size.

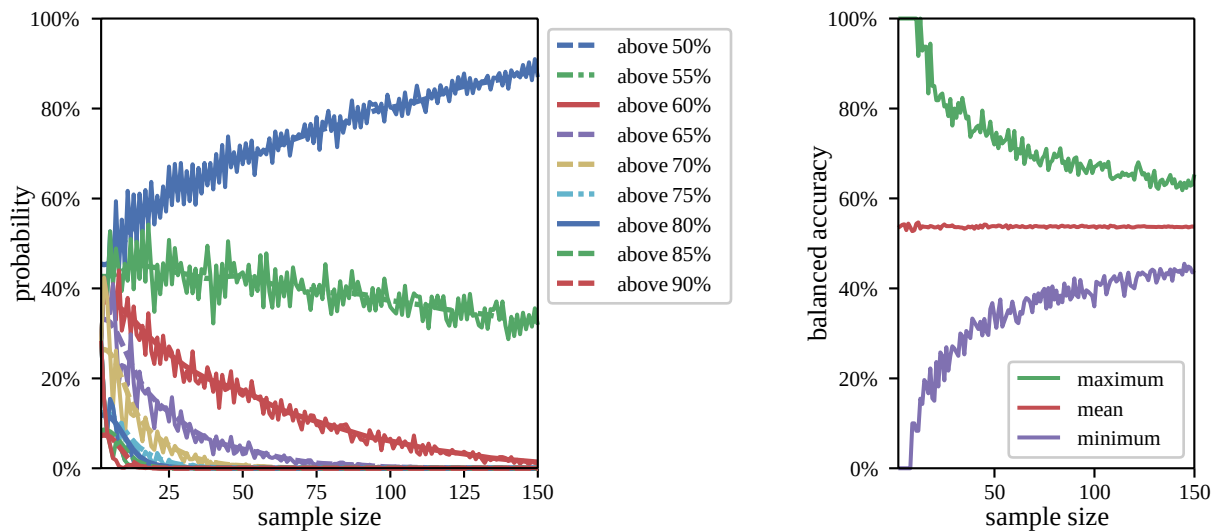
Figure E.25: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 50 **components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)



(a) Probabilities for SVM with rbf kernel to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the SVM with rbf kernel as a function of test sample size.

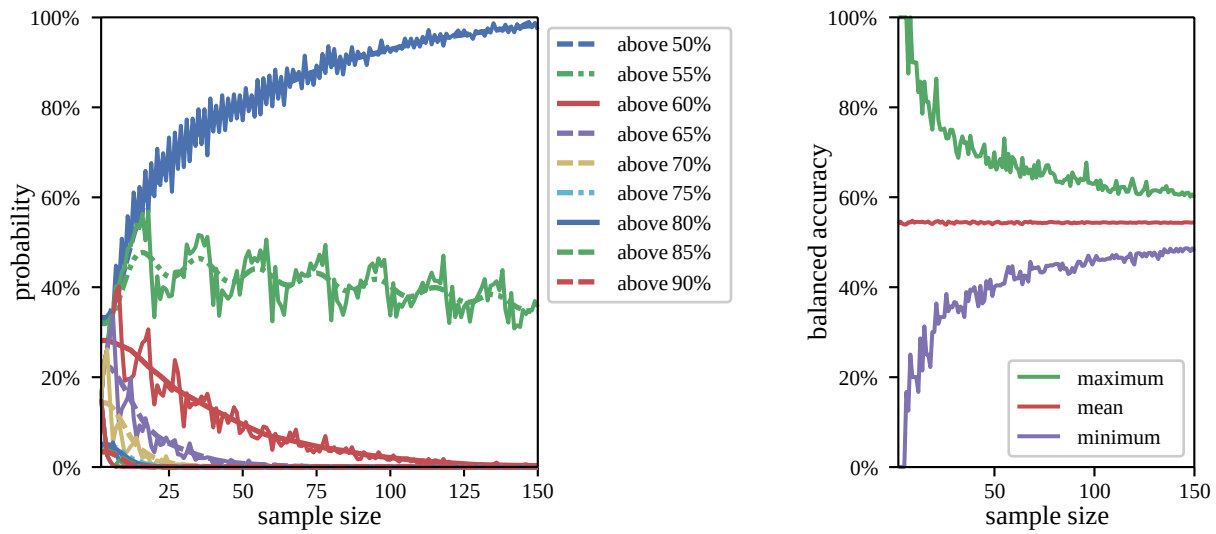
Figure E.26: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 50 **components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)



(a) Probabilities for linear SVM to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the linear SVM as a function of test sample size.

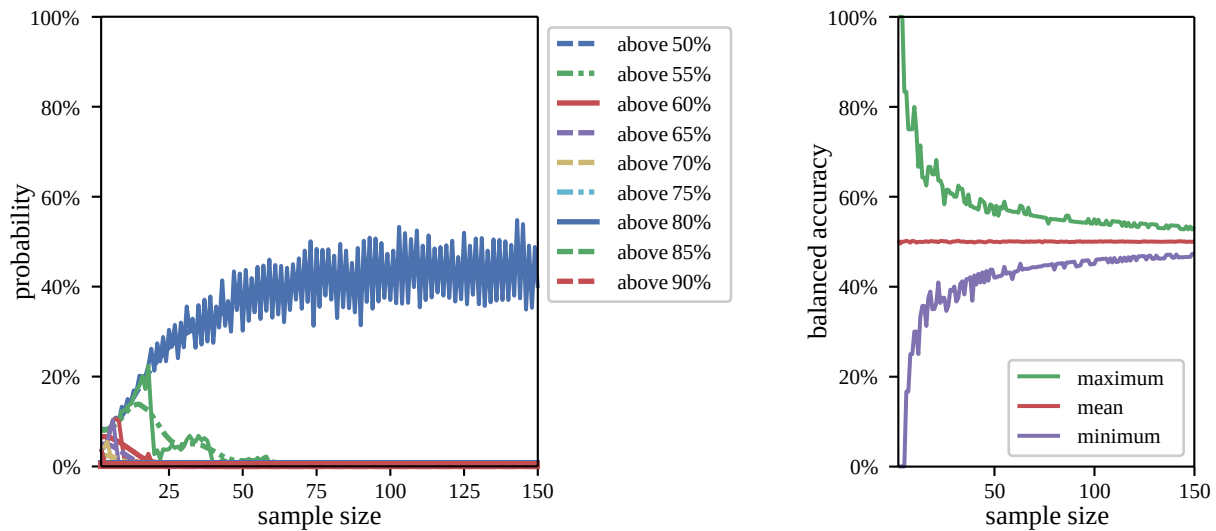
Figure E.27: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 50 **components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)



(a) Probabilities for Random Forest to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the Random Forest as a function of test sample size.

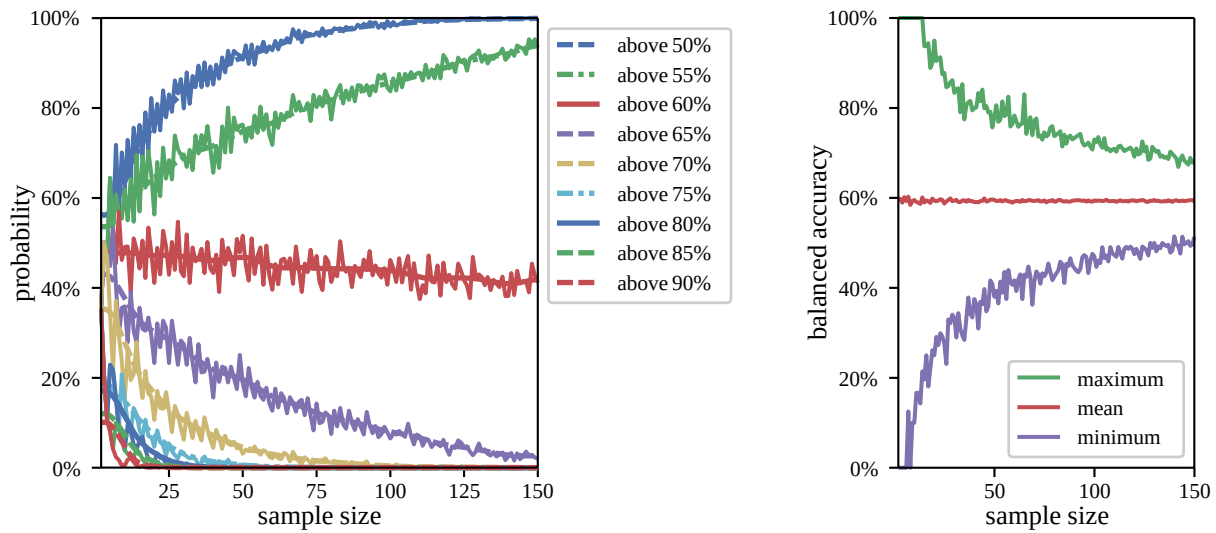
Figure E.28: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 500 **components** were retained. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)



(a) Probabilities for SVM with rbf kernel to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the SVM with rbf kernel as a function of test sample size.

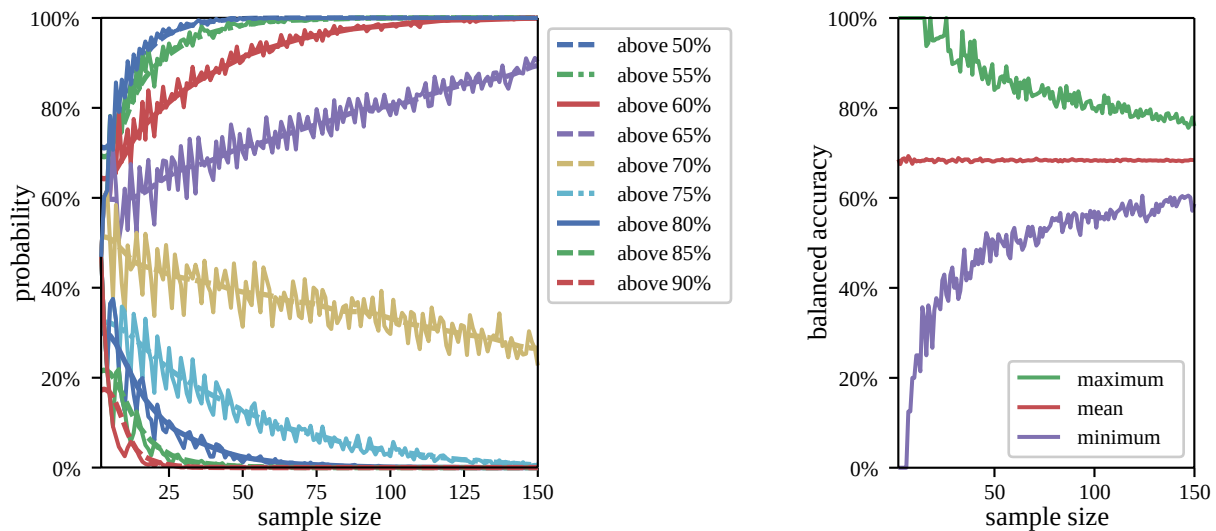
Figure E.29: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 500 **components** were retained. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)



(a) Probabilities for linear SVM to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the linear SVM as a function of test sample size.

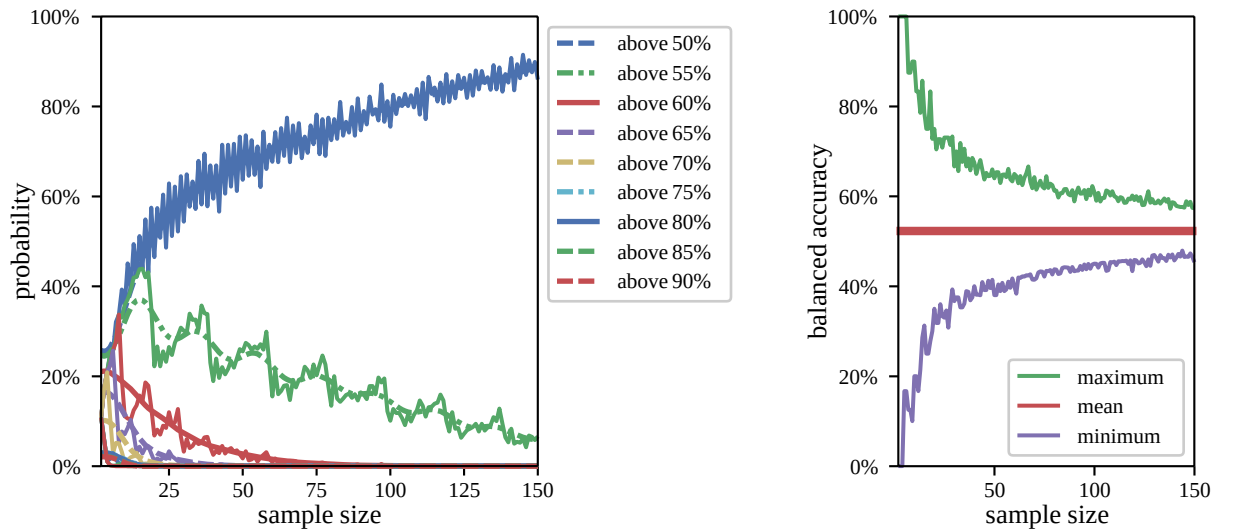
Figure E.30: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 500 **components** were retained. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)



(a) Probabilities for Random Forest to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the Random Forest as a function of test sample size.

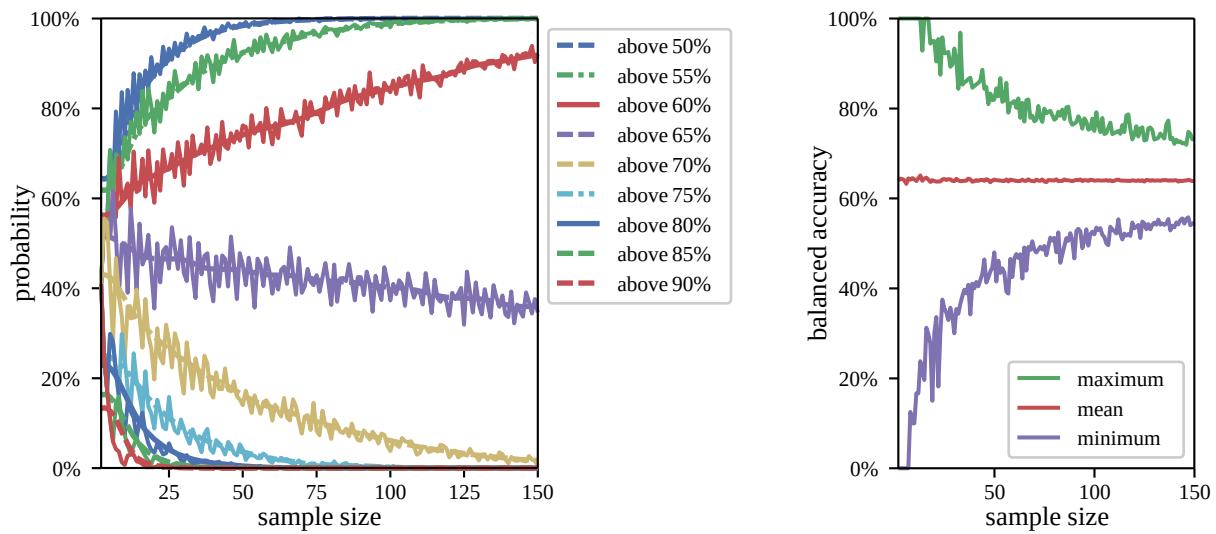
Figure E.31: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 500 **components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)



(a) Probabilities for SVM with rbf kernel to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the SVM with rbf kernel as a function of test sample size.

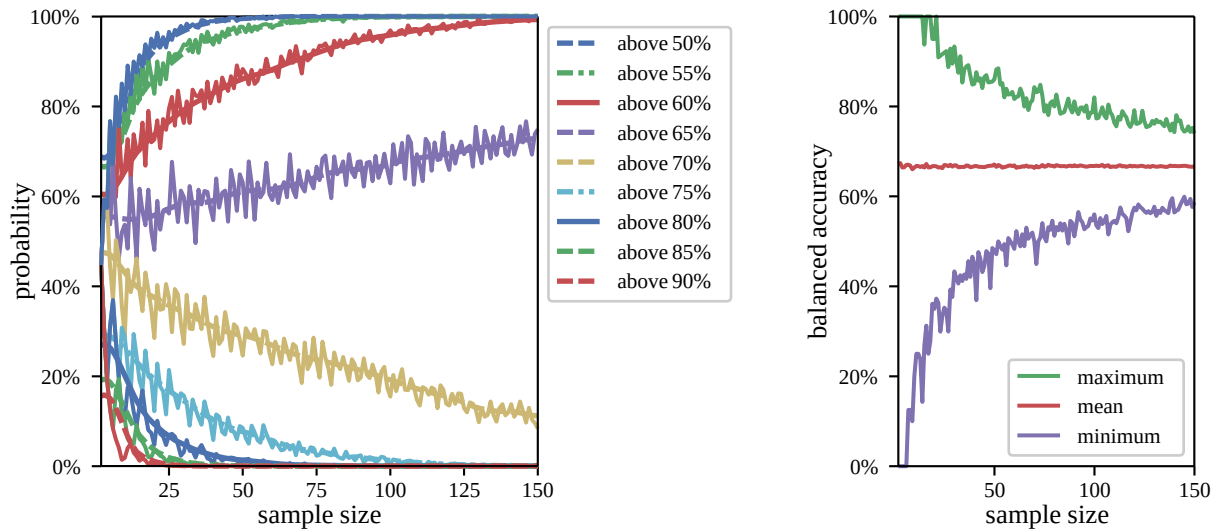
Figure E.32: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 500 **components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)



(a) Probabilities for linear SVM to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the linear SVM as a function of test sample size.

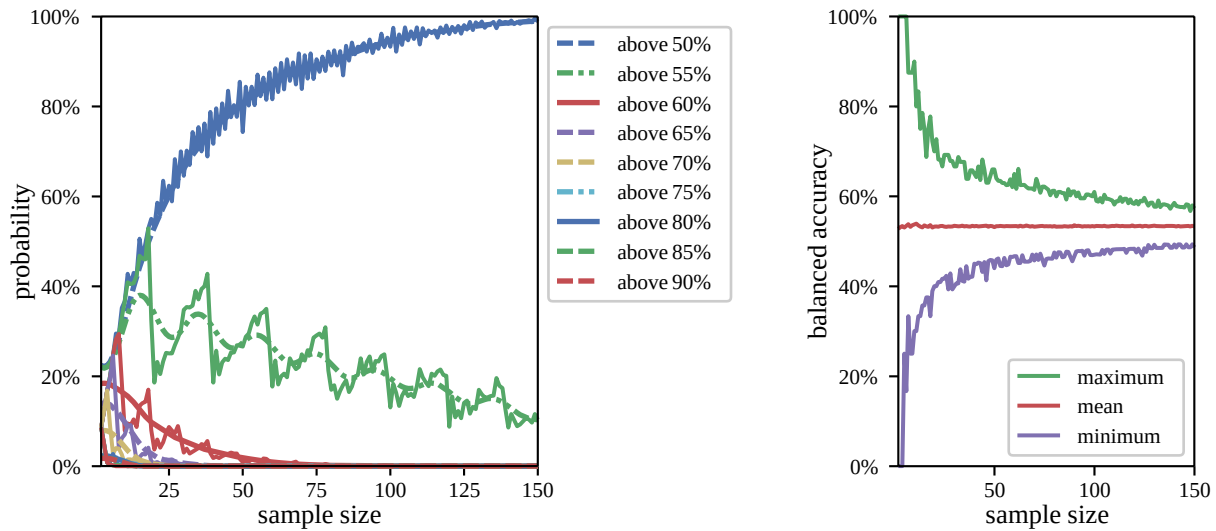
Figure E.33: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 500 **components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)



(a) Probabilities for Random Forest to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the Random Forest as a function of test sample size.

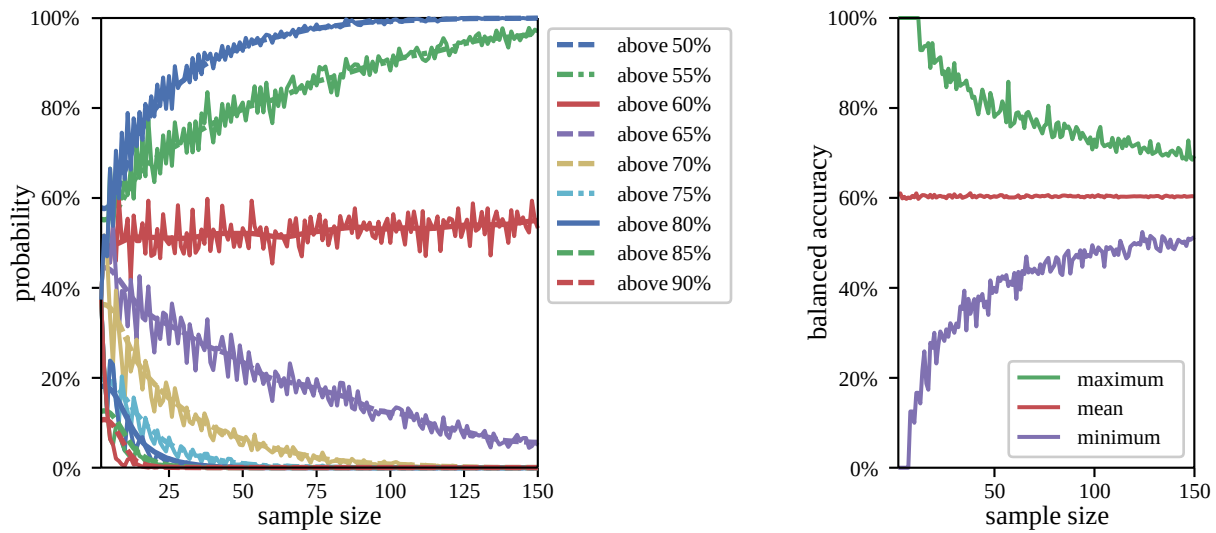
Figure E.34: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 500 **components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 100 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)



(a) Probabilities for SVM with rbf kernel to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the SVM with rbf kernel as a function of test sample size.

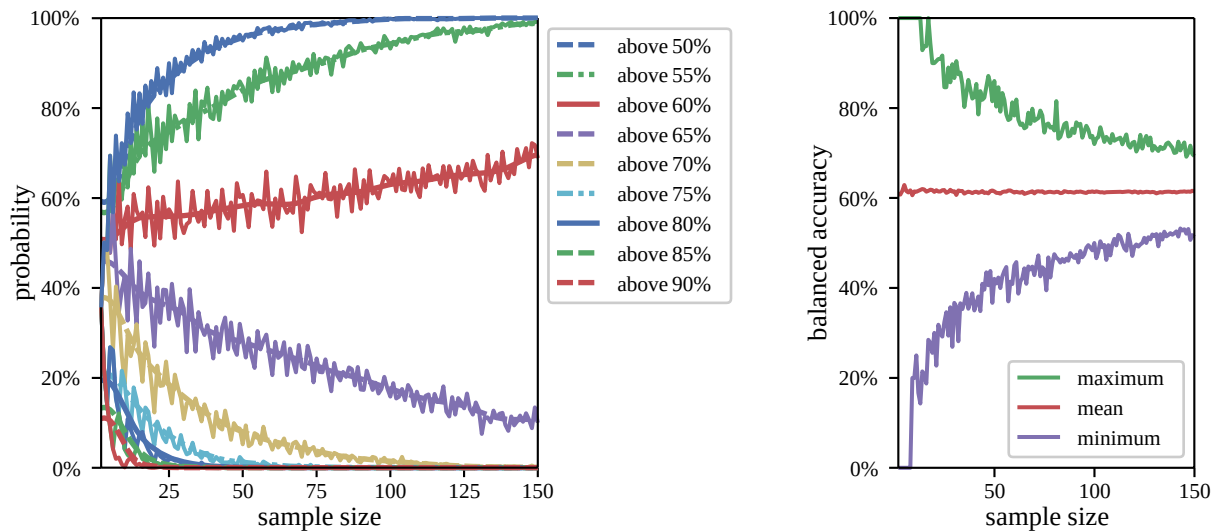
Figure E.35: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 500 **components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 100 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)



(a) Probabilities for linear SVM to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the linear SVM as a function of test sample size.

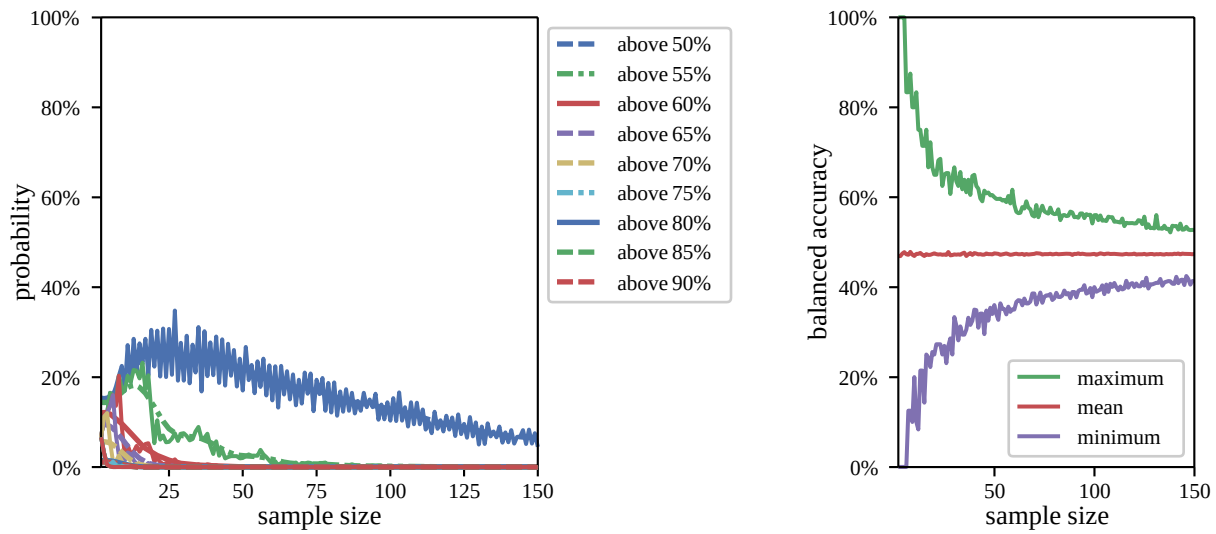
Figure E.36: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 500 **components** were retained. For **feature selection** an ANOVA was computed inside the pipeline and the top 100 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)



(a) Probabilities for Random Forest to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the Random Forest as a function of test sample size.

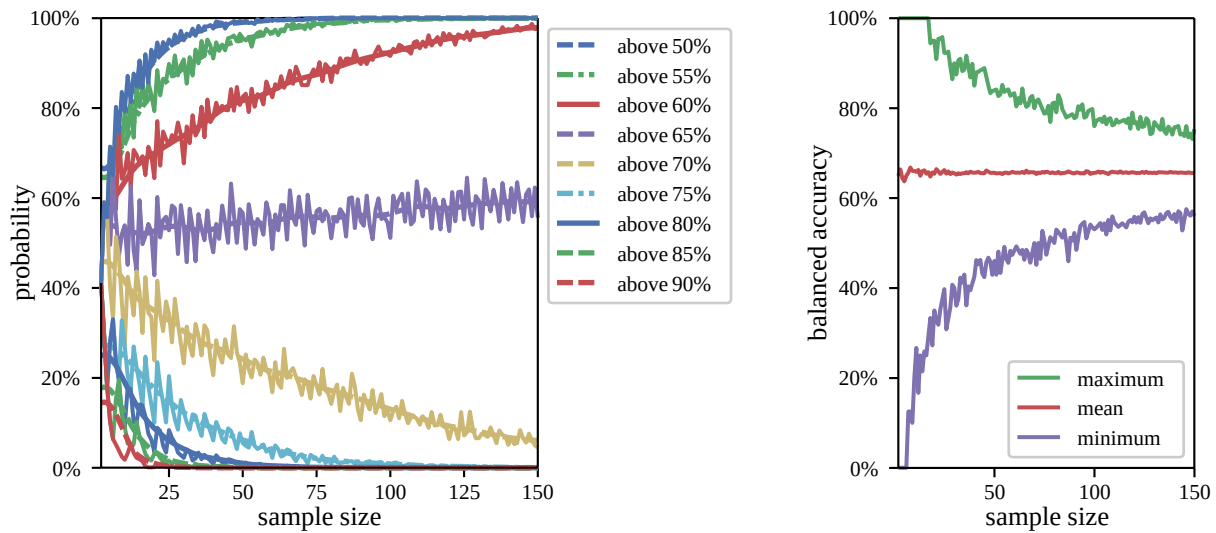
Figure E.37: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)



(a) Probabilities for SVM with rbf kernel to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the SVM with rbf kernel as a function of test sample size.

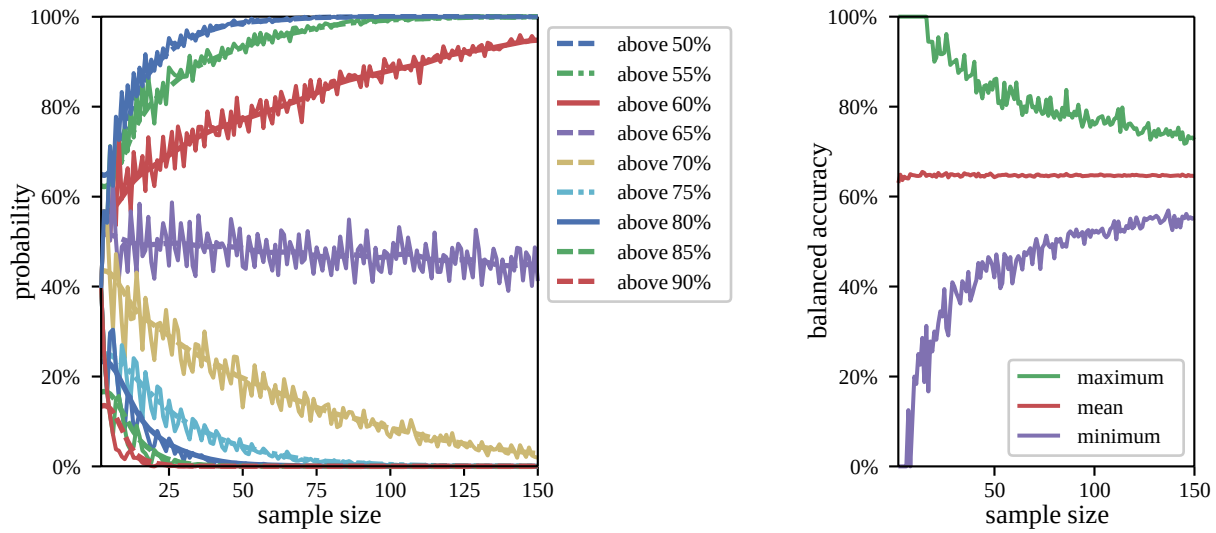
Figure E.38: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)



(a) Probabilities for linear SVM to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the linear SVM as a function of test sample size.

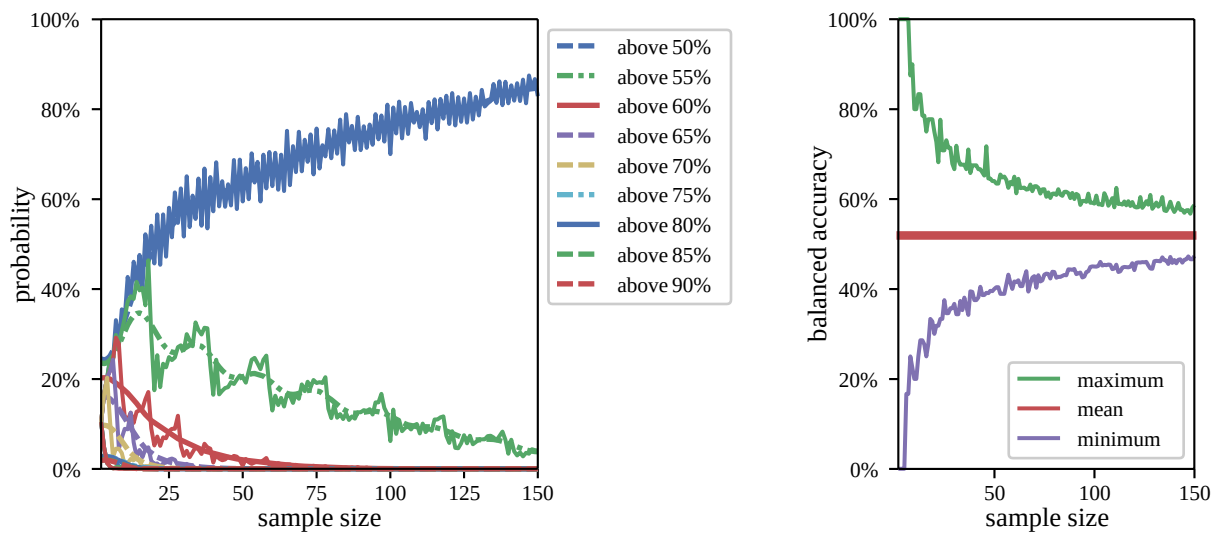
Figure E.39: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)



(a) Probabilities for Random Forest to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the Random Forest as a function of test sample size.

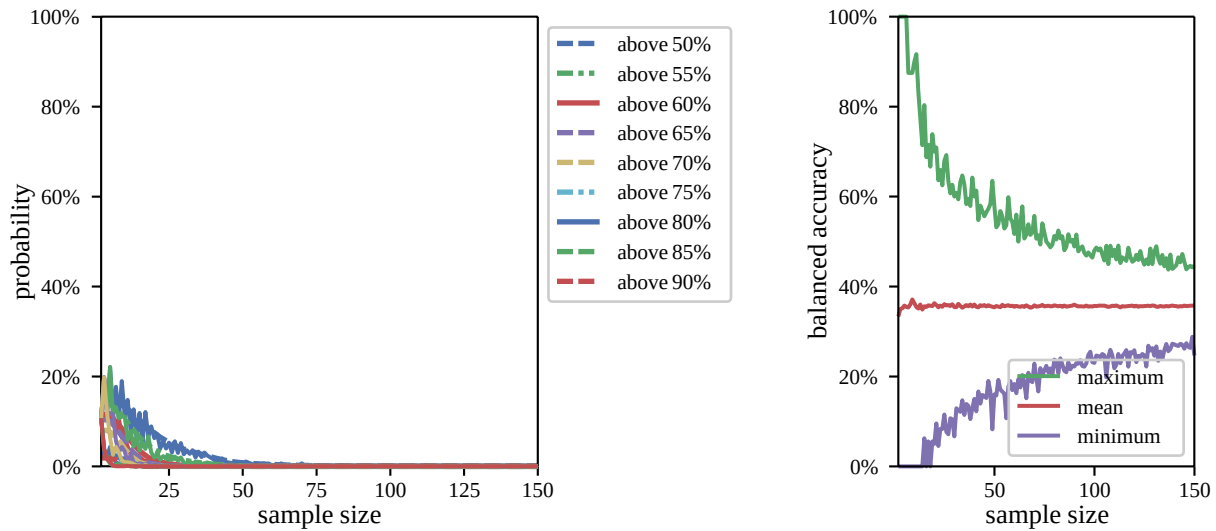
Figure E.40: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)



(a) Probabilities for SVM with rbf kernel to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the SVM with rbf kernel as a function of test sample size.

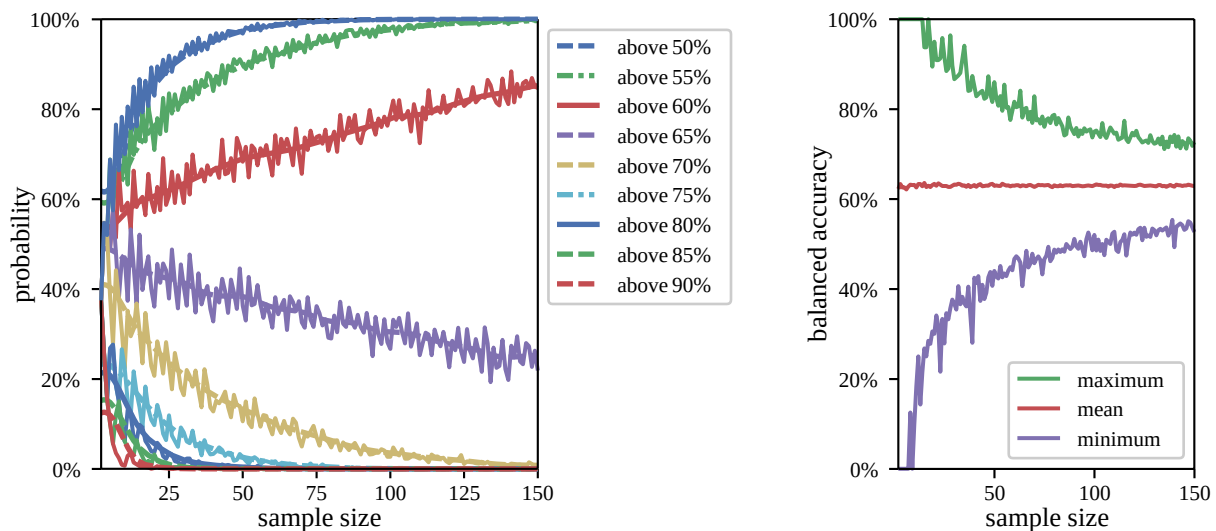
Figure E.41: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)



(a) Probabilities for linear SVM to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the linear SVM as a function of test sample size.

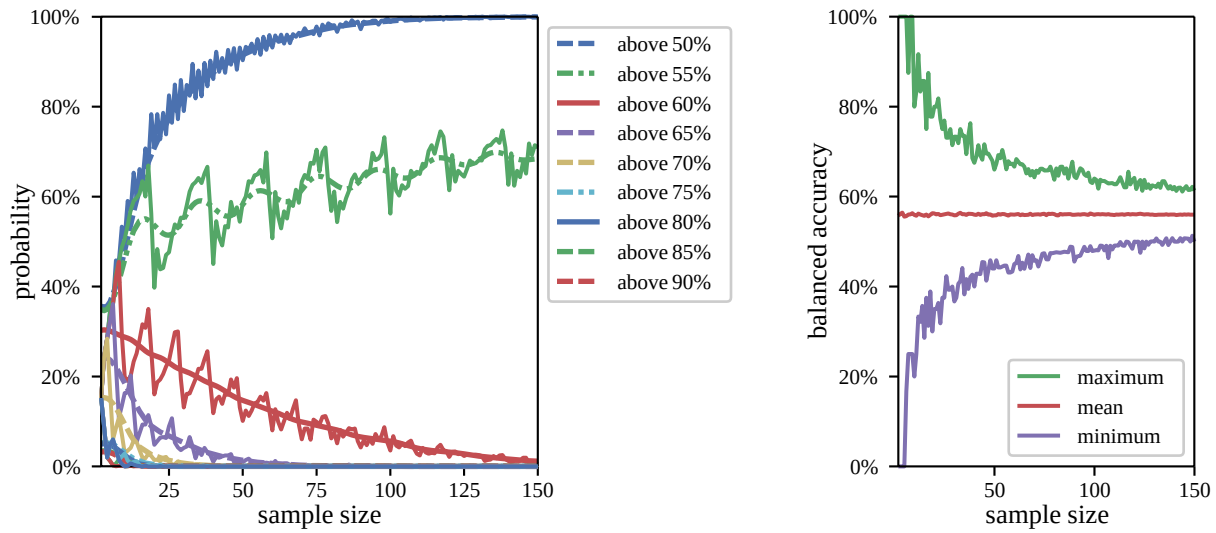
Figure E.42: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)



(a) Probabilities for Random Forest to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the Random Forest as a function of test sample size.

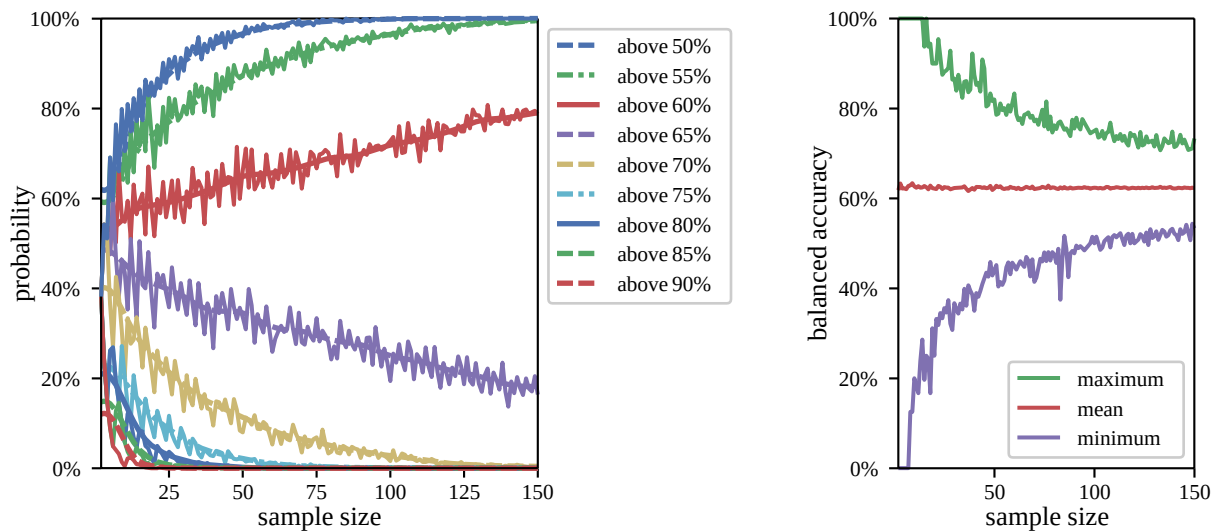
Figure E.43: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 100 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)



(a) Probabilities for SVM with rbf kernel to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the SVM with rbf kernel as a function of test sample size.

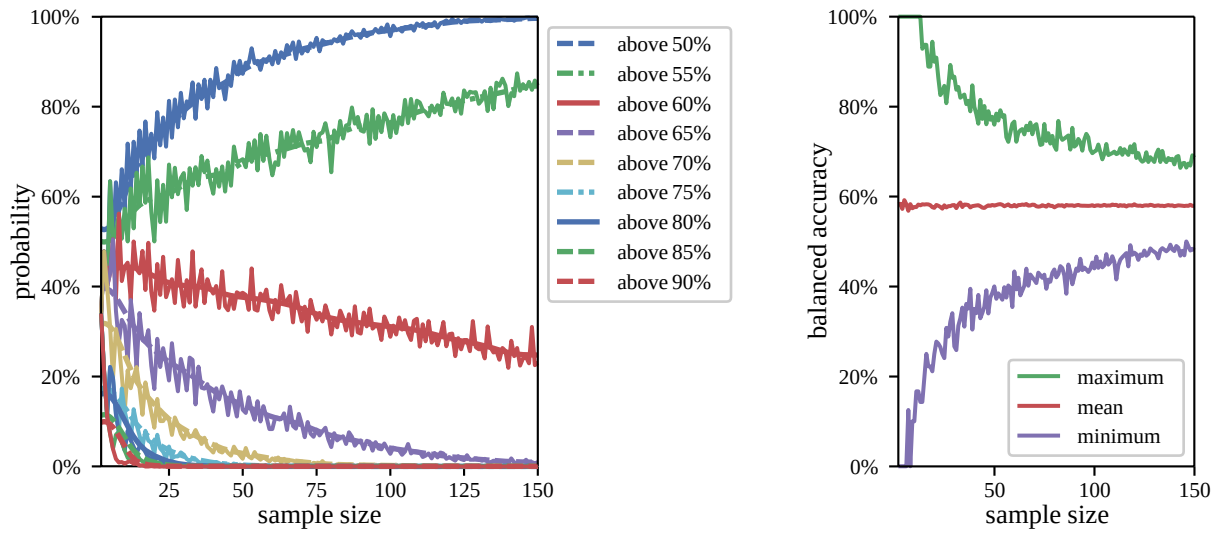
Figure E.44: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. For **feature selection** an ANOVA was computed inside the pipeline and the top 100 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)



(a) Probabilities for linear SVM to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the linear SVM as a function of test sample size.

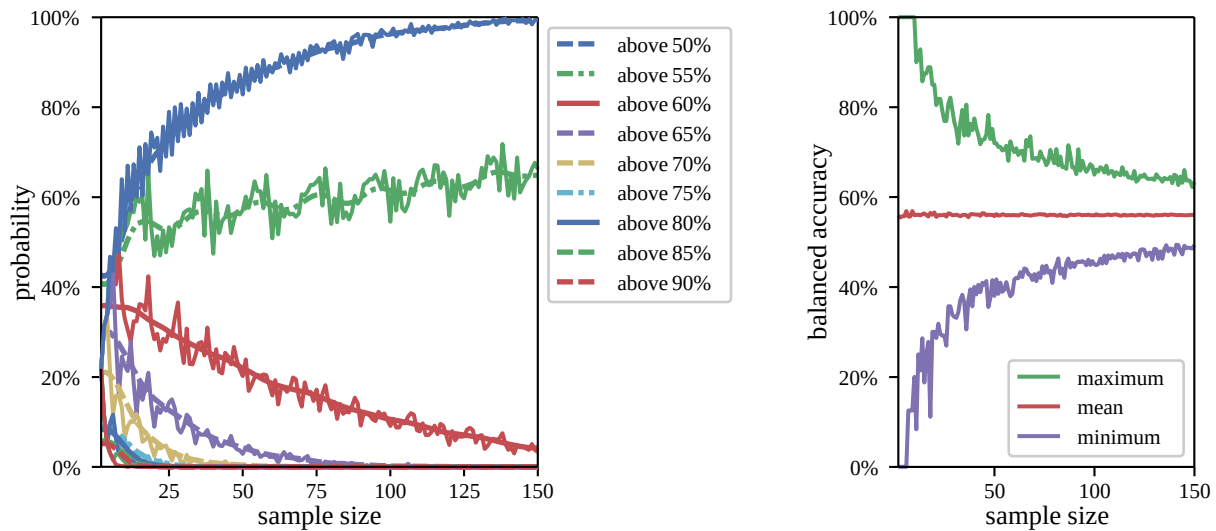
Figure E.45: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. For **feature selection** an ANOVA was computed inside the pipeline and the top 100 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)



(a) Probabilities for Random Forest to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the Random Forest as a function of test sample size.

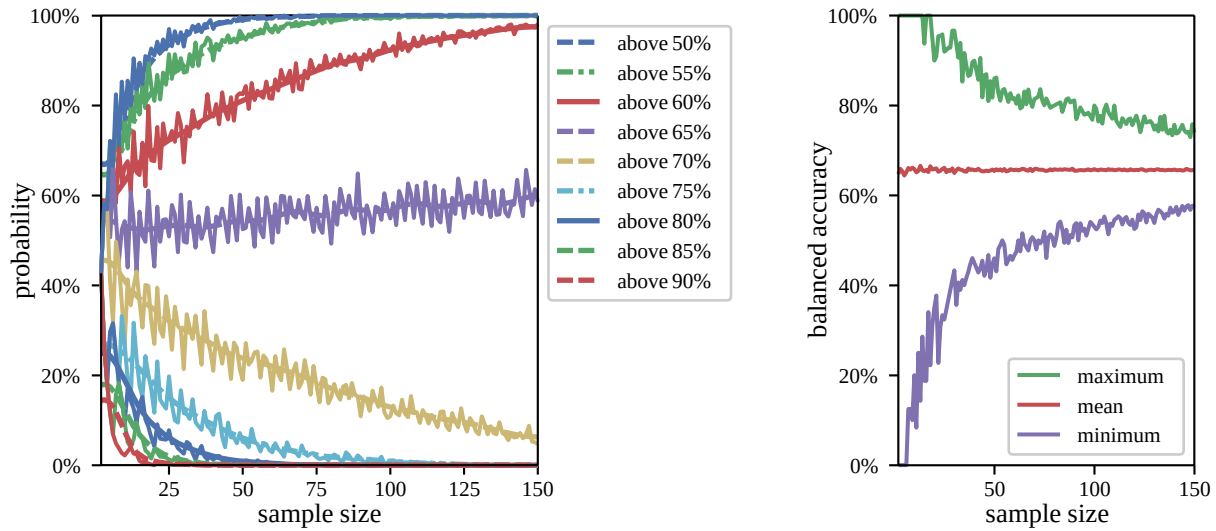
Figure E.46: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 1000 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)



(a) Probabilities for SVM with rbf kernel to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the SVM with rbf kernel as a function of test sample size.

Figure E.47: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 1000 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)



(a) Probabilities for linear SVM to yield an accuracy exceeding a certain threshold as a function of test sample size.

(b) Minimum, maximum and mean results for the linear SVM as a function of test sample size.

Figure E.48: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 1000 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)

E.1.2 Tabular results representation

		$\geq$ accuracy (%)								
test sample size	n	50	55	60	65	70	75	80	85	90
	10	0.79	0.79	0.65	0.54	0.28	0.28	0.08	0.08	0.01
	20	0.92	0.81	0.67	0.47	0.29	0.15	0.07	0.06	0.00
	30	0.96	0.91	0.74	0.58	0.29	0.15	0.03	0.01	0.00
	40	0.98	0.93	0.80	0.50	0.24	0.07	0.01	0.00	0.00
	50	0.99	0.96	0.86	0.59	0.23	0.08	0.00	0.00	0.00
	60	1.00	0.98	0.87	0.55	0.23	0.03	0.00	0.00	0.00
	70	1.00	0.99	0.89	0.61	0.21	0.03	0.00	0.00	0.00
	80	1.00	0.99	0.90	0.54	0.16	0.02	0.00	0.00	0.00
	90	1.00	1.00	0.93	0.61	0.14	0.02	0.00	0.00	0.00
	100	1.00	1.00	0.94	0.59	0.14	0.01	0.00	0.00	0.00
	125	1.00	1.00	0.96	0.64	0.12	0.01	0.00	0.00	0.00
150	1.00	1.00	0.99	0.66	0.07	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.66	0.70	0.20	1.00	0.14
	20	0.67	0.65	0.35	0.95	0.10
	30	0.67	0.67	0.37	0.90	0.08
	40	0.66	0.65	0.42	0.85	0.07
	50	0.66	0.66	0.46	0.84	0.06
	60	0.67	0.67	0.50	0.80	0.05
	70	0.67	0.67	0.51	0.81	0.05
	80	0.66	0.66	0.53	0.81	0.05
	90	0.66	0.67	0.51	0.78	0.04
	100	0.66	0.66	0.55	0.77	0.04
	125	0.66	0.66	0.54	0.76	0.03
	150	0.66	0.66	0.57	0.75	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.2: Results as a function of variable test set sizes with a fixed classifier. A **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)

		≥ accuracy (%)								
test sample size	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.79	0.79	0.65	0.55	0.29	0.29	0.10	0.10	0.01
	20	0.91	0.80	0.65	0.45	0.25	0.12	0.04	0.03	0.00
	30	0.96	0.91	0.71	0.56	0.26	0.14	0.03	0.01	0.00
	40	0.98	0.91	0.79	0.48	0.21	0.04	0.01	0.00	0.00
	50	0.99	0.96	0.84	0.55	0.20	0.07	0.00	0.00	0.00
	60	0.99	0.97	0.83	0.49	0.19	0.04	0.00	0.00	0.00
	70	1.00	0.99	0.86	0.60	0.16	0.03	0.00	0.00	0.00
	80	1.00	0.99	0.91	0.53	0.15	0.02	0.00	0.00	0.00
	90	1.00	0.99	0.92	0.62	0.14	0.02	0.00	0.00	0.00
	100	1.00	1.00	0.95	0.57	0.13	0.00	0.00	0.00	0.00
	125	1.00	1.00	0.96	0.63	0.12	0.01	0.00	0.00	0.00
150	1.00	1.00	0.98	0.64	0.05	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.66	0.70	0.20	1.00	0.15
	20	0.66	0.65	0.30	0.95	0.10
	30	0.66	0.67	0.40	0.90	0.08
	40	0.66	0.65	0.45	0.93	0.07
	50	0.66	0.66	0.46	0.86	0.06
	60	0.66	0.65	0.50	0.82	0.05
	70	0.66	0.66	0.50	0.79	0.05
	80	0.66	0.66	0.53	0.84	0.04
	90	0.66	0.67	0.50	0.80	0.04
	100	0.66	0.66	0.54	0.77	0.04
	125	0.66	0.66	0.56	0.78	0.03
	150	0.66	0.66	0.59	0.75	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.3: Results as a function of variable test set sizes with a fixed classifier. An **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)

		$\geq$ accuracy (%)								
test sample size	n	50	55	60	65	70	75	80	85	90
	10	0.77	0.77	0.63	0.53	0.27	0.27	0.08	0.08	0.01
	20	0.90	0.78	0.65	0.46	0.27	0.14	0.05	0.04	0.00
	30	0.95	0.89	0.69	0.54	0.23	0.12	0.02	0.01	0.00
	40	0.98	0.90	0.78	0.45	0.19	0.06	0.01	0.00	0.00
	50	1.00	0.97	0.83	0.55	0.18	0.06	0.01	0.00	0.00
	60	1.00	0.96	0.82	0.51	0.18	0.02	0.00	0.00	0.00
	70	1.00	0.98	0.84	0.55	0.16	0.03	0.00	0.00	0.00
	80	1.00	0.99	0.89	0.52	0.16	0.01	0.00	0.00	0.00
	90	1.00	1.00	0.89	0.57	0.12	0.01	0.00	0.00	0.00
	100	1.00	1.00	0.93	0.52	0.10	0.00	0.00	0.00	0.00
	125	1.00	1.00	0.96	0.58	0.10	0.00	0.00	0.00	0.00
150	1.00	1.00	0.98	0.61	0.04	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.65	0.70	0.20	1.00	0.15
	20	0.66	0.65	0.30	0.95	0.11
	30	0.66	0.67	0.43	1.00	0.08
	40	0.65	0.65	0.38	0.90	0.07
	50	0.66	0.66	0.46	0.86	0.06
	60	0.66	0.67	0.48	0.82	0.05
	70	0.66	0.66	0.50	0.80	0.05
	80	0.66	0.66	0.50	0.80	0.05
	90	0.66	0.66	0.53	0.78	0.04
	100	0.66	0.66	0.52	0.79	0.04
	125	0.66	0.66	0.54	0.76	0.03
	150	0.66	0.66	0.57	0.75	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.4: Results as a function of variable test set sizes with a fixed classifier. An **linear SVM** was trained with default parameters. ($C = 1.0$)

		≥ accuracy (%)								
test sample size	n	50	55	60	65	70	75	80	85	90
	10	0.67	0.67	0.52	0.41	0.18	0.18	0.06	0.06	0.01
	20	0.81	0.64	0.49	0.28	0.13	0.05	0.01	0.01	0.00
	30	0.87	0.77	0.49	0.34	0.12	0.05	0.01	0.00	0.00
	40	0.93	0.78	0.60	0.25	0.07	0.02	0.00	0.00	0.00
	50	0.95	0.84	0.60	0.28	0.06	0.01	0.00	0.00	0.00
	60	0.97	0.84	0.51	0.20	0.05	0.00	0.00	0.00	0.00
	70	0.98	0.88	0.50	0.23	0.04	0.00	0.00	0.00	0.00
	80	0.99	0.87	0.58	0.17	0.02	0.00	0.00	0.00	0.00
	90	0.99	0.93	0.62	0.21	0.02	0.00	0.00	0.00	0.00
	100	0.99	0.93	0.63	0.15	0.01	0.00	0.00	0.00	0.00
	125	1.00	0.98	0.64	0.14	0.01	0.00	0.00	0.00	0.00
150	1.00	0.98	0.68	0.09	0.00	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.61	0.60	0.10	1.00	0.15
	20	0.61	0.60	0.20	0.90	0.10
	30	0.61	0.60	0.33	0.87	0.09
	40	0.62	0.62	0.40	0.80	0.07
	50	0.61	0.62	0.40	0.84	0.06
	60	0.61	0.62	0.43	0.78	0.06
	70	0.61	0.60	0.43	0.80	0.05
	80	0.61	0.61	0.42	0.77	0.05
	90	0.61	0.61	0.48	0.76	0.04
	100	0.61	0.61	0.49	0.74	0.04
	125	0.61	0.62	0.50	0.72	0.03
	150	0.61	0.61	0.52	0.70	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.5: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)

		≥ accuracy (%)								
test sample size	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.77	0.77	0.64	0.54	0.28	0.28	0.09	0.09	0.01
	20	0.90	0.80	0.66	0.45	0.26	0.12	0.05	0.04	0.00
	30	0.94	0.90	0.67	0.53	0.24	0.13	0.03	0.01	0.00
	40	0.98	0.91	0.75	0.45	0.21	0.05	0.01	0.00	0.00
	50	0.99	0.96	0.85	0.57	0.21	0.08	0.01	0.00	0.00
	60	1.00	0.97	0.82	0.48	0.17	0.03	0.00	0.00	0.00
	70	1.00	0.98	0.83	0.55	0.16	0.03	0.00	0.00	0.00
	80	1.00	0.98	0.89	0.52	0.13	0.01	0.00	0.00	0.00
	90	1.00	0.99	0.90	0.57	0.14	0.01	0.00	0.00	0.00
	100	1.00	1.00	0.94	0.51	0.10	0.01	0.00	0.00	0.00
	125	1.00	1.00	0.95	0.55	0.09	0.00	0.00	0.00	0.00
150	1.00	1.00	0.98	0.59	0.05	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.66	0.70	0.20	1.00	0.15
	20	0.66	0.65	0.30	0.95	0.10
	30	0.66	0.67	0.37	0.90	0.08
	40	0.65	0.65	0.42	0.85	0.07
	50	0.66	0.66	0.48	0.84	0.06
	60	0.66	0.65	0.50	0.82	0.05
	70	0.66	0.66	0.49	0.83	0.05
	80	0.66	0.66	0.53	0.80	0.05
	90	0.66	0.66	0.52	0.78	0.04
	100	0.66	0.66	0.53	0.79	0.04
	125	0.66	0.66	0.56	0.76	0.03
	150	0.66	0.65	0.57	0.73	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.6: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)

		≥ accuracy (%)								
test sample size	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.80	0.80	0.69	0.57	0.30	0.30	0.11	0.11	0.02
	20	0.93	0.84	0.72	0.51	0.32	0.15	0.07	0.05	0.01
	30	0.97	0.94	0.78	0.63	0.32	0.18	0.04	0.01	0.00
	40	0.99	0.94	0.86	0.60	0.30	0.10	0.02	0.00	0.00
	50	1.00	0.99	0.92	0.69	0.30	0.10	0.01	0.00	0.00
	60	1.00	0.99	0.90	0.61	0.27	0.05	0.01	0.00	0.00
	70	1.00	1.00	0.92	0.70	0.28	0.07	0.00	0.00	0.00
	80	1.00	0.99	0.95	0.67	0.23	0.03	0.00	0.00	0.00
	90	1.00	1.00	0.97	0.78	0.23	0.03	0.00	0.00	0.00
	100	1.00	1.00	0.97	0.70	0.20	0.01	0.00	0.00	0.00
	125	1.00	1.00	0.99	0.77	0.23	0.01	0.00	0.00	0.00
150	1.00	1.00	1.00	0.83	0.15	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.67	0.70	0.20	1.00	0.15
	20	0.67	0.70	0.35	1.00	0.10
	30	0.68	0.67	0.40	0.93	0.08
	40	0.68	0.68	0.42	0.88	0.07
	50	0.68	0.68	0.48	0.86	0.06
	60	0.68	0.67	0.52	0.83	0.05
	70	0.68	0.67	0.54	0.83	0.05
	80	0.67	0.68	0.54	0.82	0.05
	90	0.68	0.68	0.56	0.80	0.04
	100	0.67	0.67	0.56	0.80	0.04
	125	0.68	0.67	0.57	0.80	0.03
	150	0.68	0.67	0.60	0.76	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.7: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)

		≥ accuracy (%)								
test sample size	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.73	0.73	0.62	0.49	0.24	0.24	0.09	0.09	0.01
	20	0.86	0.73	0.58	0.37	0.19	0.09	0.03	0.03	0.00
	30	0.92	0.85	0.63	0.47	0.15	0.08	0.01	0.00	0.00
	40	0.96	0.84	0.70	0.37	0.15	0.04	0.01	0.00	0.00
	50	0.98	0.94	0.80	0.47	0.15	0.04	0.00	0.00	0.00
	60	0.99	0.91	0.72	0.38	0.11	0.02	0.00	0.00	0.00
	70	1.00	0.96	0.75	0.43	0.09	0.01	0.00	0.00	0.00
	80	1.00	0.97	0.81	0.39	0.08	0.01	0.00	0.00	0.00
	90	1.00	0.98	0.81	0.43	0.06	0.00	0.00	0.00	0.00
	100	1.00	0.99	0.86	0.36	0.05	0.00	0.00	0.00	0.00
	125	1.00	1.00	0.89	0.38	0.03	0.00	0.00	0.00	0.00
150	1.00	1.00	0.93	0.36	0.01	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.64	0.60	0.20	1.00	0.15
	20	0.64	0.65	0.25	0.95	0.10
	30	0.64	0.63	0.40	0.87	0.08
	40	0.64	0.65	0.40	0.88	0.07
	50	0.65	0.64	0.44	0.82	0.06
	60	0.64	0.63	0.47	0.85	0.06
	70	0.64	0.64	0.49	0.80	0.05
	80	0.64	0.65	0.50	0.81	0.05
	90	0.64	0.64	0.51	0.78	0.04
	100	0.64	0.64	0.52	0.81	0.04
	125	0.64	0.64	0.53	0.75	0.03
	150	0.64	0.64	0.55	0.72	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.8: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 100 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)

		≥ accuracy (%)								
test sample size	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.73	0.73	0.60	0.47	0.20	0.20	0.05	0.05	0.01
	20	0.86	0.71	0.56	0.34	0.19	0.07	0.02	0.01	0.00
	30	0.92	0.85	0.63	0.44	0.17	0.09	0.02	0.01	0.00
	40	0.96	0.85	0.67	0.35	0.15	0.04	0.01	0.00	0.00
	50	0.97	0.91	0.73	0.40	0.12	0.03	0.00	0.00	0.00
	60	0.99	0.92	0.71	0.34	0.09	0.01	0.00	0.00	0.00
	70	1.00	0.96	0.72	0.39	0.09	0.01	0.00	0.00	0.00
	80	1.00	0.95	0.74	0.32	0.06	0.00	0.00	0.00	0.00
	90	1.00	0.99	0.78	0.37	0.06	0.00	0.00	0.00	0.00
	100	1.00	0.99	0.83	0.32	0.04	0.00	0.00	0.00	0.00
	125	1.00	1.00	0.88	0.30	0.03	0.00	0.00	0.00	0.00
150	1.00	1.00	0.90	0.33	0.01	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.63	0.60	0.10	1.00	0.14
	20	0.63	0.65	0.35	0.95	0.10
	30	0.64	0.63	0.37	0.90	0.09
	40	0.64	0.62	0.42	0.88	0.07
	50	0.63	0.64	0.42	0.84	0.06
	60	0.64	0.63	0.48	0.80	0.05
	70	0.64	0.64	0.47	0.79	0.05
	80	0.63	0.64	0.50	0.76	0.05
	90	0.64	0.63	0.51	0.77	0.04
	100	0.64	0.64	0.53	0.75	0.04
	125	0.64	0.64	0.54	0.73	0.03
	150	0.64	0.64	0.57	0.71	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.9: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 100 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)

test sample size	≥ accuracy (%)									
	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.69	0.69	0.56	0.44	0.19	0.19	0.05	0.05	0.00
	20	0.81	0.69	0.54	0.32	0.16	0.06	0.02	0.02	0.00
	30	0.89	0.82	0.53	0.34	0.10	0.04	0.00	0.00	0.00
	40	0.93	0.80	0.59	0.27	0.10	0.02	0.00	0.00	0.00
	50	0.96	0.87	0.61	0.31	0.07	0.01	0.00	0.00	0.00
	60	0.97	0.85	0.54	0.19	0.03	0.00	0.00	0.00	0.00
	70	0.99	0.92	0.58	0.26	0.04	0.00	0.00	0.00	0.00
	80	0.99	0.92	0.67	0.21	0.03	0.00	0.00	0.00	0.00
90	1.00	0.95	0.65	0.22	0.02	0.00	0.00	0.00	0.00	
100	1.00	0.96	0.68	0.18	0.02	0.00	0.00	0.00	0.00	
125	1.00	0.99	0.74	0.17	0.01	0.00	0.00	0.00	0.00	
150	1.00	1.00	0.76	0.12	0.00	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

test sample size	n	mean	median	min	max	std
	10	0.62	0.60	0.10	1.00	0.15
	20	0.62	0.60	0.30	0.95	0.10
	30	0.62	0.63	0.33	0.87	0.08
	40	0.62	0.62	0.38	0.82	0.07
	50	0.62	0.62	0.40	0.84	0.06
	60	0.61	0.62	0.45	0.80	0.05
	70	0.62	0.61	0.46	0.77	0.05
	80	0.62	0.62	0.49	0.77	0.05
	90	0.62	0.62	0.49	0.77	0.04
	100	0.62	0.62	0.47	0.75	0.04
	125	0.62	0.62	0.52	0.72	0.03
	150	0.62	0.62	0.54	0.71	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.10: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 100 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)

		≥ accuracy (%)								
test sample size	n	50	55	60	65	70	75	80	85	90
	10	0.81	0.81	0.71	0.60	0.35	0.35	0.11	0.11	0.02
	20	0.94	0.84	0.72	0.52	0.33	0.16	0.06	0.05	0.01
	30	0.97	0.95	0.78	0.65	0.33	0.20	0.04	0.01	0.00
	40	0.99	0.95	0.86	0.57	0.30	0.09	0.02	0.00	0.00
	50	0.99	0.98	0.89	0.67	0.28	0.10	0.01	0.00	0.00
	60	1.00	0.99	0.88	0.64	0.26	0.05	0.01	0.00	0.00
	70	1.00	0.99	0.92	0.71	0.27	0.06	0.01	0.00	0.00
	80	1.00	0.99	0.95	0.65	0.24	0.03	0.00	0.00	0.00
	90	1.00	1.00	0.97	0.74	0.24	0.04	0.00	0.00	0.00
	100	1.00	1.00	0.98	0.74	0.25	0.03	0.00	0.00	0.00
	125	1.00	1.00	0.99	0.81	0.24	0.01	0.00	0.00	0.00
150	1.00	1.00	1.00	0.82	0.15	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.68	0.70	0.20	1.00	0.15
	20	0.68	0.70	0.35	0.95	0.10
	30	0.68	0.67	0.43	0.93	0.08
	40	0.67	0.68	0.45	0.85	0.07
	50	0.67	0.68	0.48	0.86	0.06
	60	0.68	0.68	0.50	0.85	0.05
	70	0.68	0.67	0.50	0.84	0.05
	80	0.67	0.68	0.50	0.80	0.05
	90	0.68	0.68	0.53	0.81	0.04
	100	0.68	0.68	0.56	0.79	0.04
	125	0.68	0.68	0.58	0.78	0.03
	150	0.68	0.68	0.59	0.76	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.11: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 1000 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)

		≥ accuracy (%)								
test sample size	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.82	0.82	0.69	0.62	0.35	0.35	0.13	0.13	0.03
	20	0.93	0.84	0.72	0.53	0.33	0.17	0.06	0.06	0.00
	30	0.98	0.95	0.78	0.64	0.34	0.20	0.06	0.01	0.00
	40	0.99	0.95	0.87	0.56	0.29	0.12	0.03	0.01	0.00
	50	1.00	0.98	0.92	0.71	0.30	0.10	0.01	0.00	0.00
	60	1.00	0.98	0.91	0.66	0.29	0.06	0.01	0.00	0.00
	70	1.00	1.00	0.94	0.73	0.30	0.08	0.01	0.00	0.00
	80	1.00	1.00	0.97	0.71	0.28	0.04	0.00	0.00	0.00
	90	1.00	1.00	0.98	0.81	0.30	0.04	0.00	0.00	0.00
	100	1.00	1.00	0.97	0.73	0.23	0.03	0.00	0.00	0.00
	125	1.00	1.00	0.99	0.81	0.26	0.02	0.00	0.00	0.00
150	1.00	1.00	1.00	0.88	0.20	0.01	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.69	0.70	0.20	1.00	0.15
	20	0.68	0.70	0.40	0.95	0.10
	30	0.68	0.67	0.33	0.93	0.08
	40	0.68	0.68	0.45	0.93	0.07
	50	0.68	0.68	0.44	0.84	0.06
	60	0.68	0.68	0.52	0.83	0.05
	70	0.68	0.69	0.54	0.84	0.05
	80	0.68	0.68	0.55	0.82	0.04
	90	0.68	0.68	0.54	0.80	0.04
	100	0.68	0.68	0.56	0.79	0.04
	125	0.68	0.68	0.58	0.78	0.03
	150	0.68	0.68	0.60	0.77	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.12: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 1000 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)

		$\geq$ accuracy (%)								
	n	50	55	60	65	70	75	80	85	90
test sample size	10	0.61	0.61	0.44	0.33	0.13	0.13	0.04	0.04	0.00
	20	0.68	0.51	0.35	0.17	0.07	0.03	0.01	0.01	0.00
	30	0.77	0.62	0.32	0.20	0.06	0.02	0.00	0.00	0.00
	40	0.81	0.61	0.40	0.13	0.03	0.00	0.00	0.00	0.00
	50	0.86	0.68	0.41	0.15	0.03	0.00	0.00	0.00	0.00
	60	0.90	0.66	0.33	0.09	0.01	0.00	0.00	0.00	0.00
	70	0.93	0.72	0.32	0.10	0.01	0.00	0.00	0.00	0.00
	80	0.95	0.71	0.33	0.06	0.00	0.00	0.00	0.00	0.00
	90	0.96	0.74	0.30	0.05	0.00	0.00	0.00	0.00	0.00
	100	0.97	0.73	0.31	0.03	0.00	0.00	0.00	0.00	0.00
	125	0.99	0.83	0.27	0.01	0.00	0.00	0.00	0.00	0.00
	150	1.00	0.86	0.23	0.01	0.00	0.00	0.00	0.00	0.00

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.58	0.60	0.10	1.00	0.15
	20	0.57	0.60	0.20	0.90	0.11
	30	0.58	0.57	0.30	0.80	0.08
	40	0.58	0.57	0.38	0.80	0.07
	50	0.58	0.58	0.38	0.78	0.07
	60	0.58	0.58	0.42	0.75	0.06
	70	0.58	0.59	0.43	0.76	0.05
	80	0.58	0.57	0.44	0.71	0.05
	90	0.58	0.58	0.42	0.77	0.04
	100	0.58	0.58	0.46	0.71	0.04
	125	0.58	0.58	0.47	0.68	0.03
	150	0.58	0.58	0.49	0.68	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.13: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 1000 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)

		≥ accuracy (%)								
test sample size	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.84	0.84	0.73	0.61	0.34	0.34	0.13	0.13	0.02
	20	0.94	0.86	0.75	0.56	0.39	0.19	0.08	0.07	0.01
	30	0.98	0.95	0.81	0.69	0.37	0.22	0.05	0.02	0.00
	40	1.00	0.97	0.91	0.65	0.37	0.12	0.03	0.00	0.00
	50	1.00	0.99	0.93	0.74	0.35	0.12	0.01	0.00	0.00
	60	1.00	0.99	0.92	0.72	0.34	0.09	0.01	0.00	0.00
	70	1.00	1.00	0.94	0.76	0.32	0.09	0.01	0.00	0.00
	80	1.00	1.00	0.97	0.77	0.33	0.07	0.00	0.00	0.00
	90	1.00	1.00	0.98	0.81	0.32	0.06	0.00	0.00	0.00
	100	1.00	1.00	0.99	0.80	0.31	0.04	0.00	0.00	0.00
	125	1.00	1.00	1.00	0.87	0.36	0.03	0.00	0.00	0.00
150	1.00	1.00	1.00	0.92	0.27	0.01	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.69	0.70	0.20	1.00	0.15
	20	0.69	0.70	0.30	0.95	0.10
	30	0.69	0.70	0.43	0.93	0.08
	40	0.69	0.68	0.47	0.88	0.07
	50	0.69	0.68	0.50	0.88	0.06
	60	0.69	0.68	0.45	0.85	0.05
	70	0.68	0.69	0.53	0.83	0.05
	80	0.69	0.69	0.53	0.81	0.04
	90	0.69	0.69	0.52	0.81	0.04
	100	0.69	0.69	0.56	0.82	0.04
	125	0.69	0.69	0.60	0.78	0.03
	150	0.69	0.69	0.61	0.77	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.14: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 000 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)

		≥ accuracy (%)								
test sample size	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.82	0.82	0.71	0.61	0.35	0.35	0.13	0.13	0.03
	20	0.94	0.86	0.77	0.58	0.38	0.21	0.10	0.08	0.01
	30	0.98	0.96	0.83	0.71	0.40	0.24	0.05	0.02	0.00
	40	1.00	0.97	0.90	0.67	0.39	0.15	0.04	0.01	0.00
	50	1.00	0.99	0.94	0.77	0.36	0.13	0.01	0.00	0.00
	60	1.00	0.99	0.94	0.76	0.40	0.10	0.01	0.00	0.00
	70	1.00	1.00	0.96	0.79	0.34	0.09	0.01	0.00	0.00
	80	1.00	1.00	0.97	0.79	0.34	0.07	0.00	0.00	0.00
	90	1.00	1.00	0.98	0.86	0.33	0.07	0.00	0.00	0.00
	100	1.00	1.00	0.99	0.82	0.36	0.04	0.00	0.00	0.00
	125	1.00	1.00	1.00	0.89	0.40	0.03	0.00	0.00	0.00
150	1.00	1.00	1.00	0.92	0.30	0.01	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.69	0.70	0.20	1.00	0.15
	20	0.69	0.70	0.35	0.95	0.10
	30	0.69	0.70	0.40	0.90	0.08
	40	0.69	0.70	0.47	0.90	0.07
	50	0.69	0.70	0.48	0.86	0.06
	60	0.69	0.70	0.52	0.85	0.05
	70	0.69	0.69	0.53	0.84	0.05
	80	0.69	0.69	0.54	0.84	0.05
	90	0.69	0.69	0.58	0.81	0.04
	100	0.69	0.69	0.57	0.82	0.04
	125	0.69	0.70	0.58	0.78	0.03
	150	0.69	0.69	0.61	0.79	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.15: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 000 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)

		≥ accuracy (%)								
test sample size	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.77	0.77	0.66	0.55	0.29	0.29	0.09	0.09	0.02
	20	0.89	0.79	0.66	0.43	0.24	0.12	0.04	0.02	0.00
	30	0.95	0.89	0.70	0.54	0.24	0.13	0.02	0.01	0.00
	40	0.98	0.92	0.78	0.45	0.19	0.05	0.01	0.00	0.00
	50	0.99	0.95	0.81	0.54	0.19	0.06	0.00	0.00	0.00
	60	0.99	0.95	0.81	0.48	0.17	0.02	0.00	0.00	0.00
	70	1.00	0.98	0.82	0.51	0.16	0.03	0.00	0.00	0.00
	80	1.00	0.98	0.88	0.47	0.13	0.01	0.00	0.00	0.00
	90	1.00	0.99	0.88	0.56	0.12	0.01	0.00	0.00	0.00
	100	1.00	1.00	0.92	0.49	0.09	0.00	0.00	0.00	0.00
	125	1.00	1.00	0.95	0.53	0.08	0.00	0.00	0.00	0.00
150	1.00	1.00	0.98	0.54	0.04	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.66	0.70	0.20	1.00	0.15
	20	0.66	0.65	0.30	0.95	0.10
	30	0.66	0.67	0.43	0.90	0.08
	40	0.65	0.65	0.42	0.85	0.07
	50	0.66	0.66	0.46	0.82	0.06
	60	0.65	0.65	0.47	0.82	0.05
	70	0.65	0.66	0.47	0.81	0.05
	80	0.65	0.65	0.46	0.81	0.05
	90	0.66	0.66	0.53	0.81	0.04
	100	0.65	0.65	0.53	0.77	0.04
	125	0.65	0.66	0.55	0.75	0.03
	150	0.65	0.65	0.56	0.75	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.16: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 000 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)

		≥ accuracy (%)								
test sample size	n	50	55	60	65	70	75	80	85	90
	10	0.86	0.86	0.75	0.65	0.38	0.38	0.16	0.16	0.03
	20	0.96	0.89	0.80	0.61	0.42	0.24	0.11	0.10	0.01
	30	0.98	0.96	0.83	0.71	0.39	0.26	0.06	0.02	0.00
	40	0.99	0.97	0.91	0.69	0.41	0.17	0.04	0.02	0.00
	50	1.00	0.99	0.95	0.80	0.43	0.20	0.03	0.01	0.00
	60	1.00	0.99	0.95	0.77	0.44	0.14	0.02	0.00	0.00
	70	1.00	1.00	0.96	0.83	0.44	0.14	0.01	0.00	0.00
	80	1.00	1.00	0.99	0.84	0.44	0.07	0.00	0.00	0.00
	90	1.00	1.00	0.99	0.90	0.46	0.11	0.01	0.00	0.00
	100	1.00	1.00	0.99	0.89	0.45	0.08	0.00	0.00	0.00
	125	1.00	1.00	1.00	0.94	0.49	0.06	0.00	0.00	0.00
150	1.00	1.00	1.00	0.97	0.45	0.04	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.70	0.70	0.30	1.00	0.14
	20	0.70	0.70	0.35	0.95	0.10
	30	0.69	0.70	0.43	0.90	0.08
	40	0.70	0.70	0.47	0.93	0.07
	50	0.70	0.70	0.50	0.90	0.06
	60	0.70	0.70	0.48	0.87	0.05
	70	0.70	0.70	0.56	0.86	0.05
	80	0.70	0.70	0.54	0.81	0.04
	90	0.70	0.70	0.58	0.83	0.04
	100	0.70	0.70	0.57	0.82	0.04
	125	0.70	0.70	0.59	0.79	0.03
	150	0.70	0.70	0.62	0.78	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.17: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 50 000 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)

		≥ accuracy (%)								
test sample size	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.77	0.77	0.64	0.51	0.26	0.26	0.09	0.09	0.01
	20	0.89	0.77	0.64	0.40	0.23	0.10	0.03	0.02	0.00
	30	0.96	0.90	0.71	0.54	0.23	0.12	0.02	0.01	0.00
	40	0.98	0.91	0.76	0.47	0.21	0.06	0.01	0.00	0.00
	50	0.99	0.96	0.82	0.50	0.15	0.03	0.00	0.00	0.00
	60	1.00	0.95	0.79	0.48	0.16	0.03	0.00	0.00	0.00
	70	1.00	0.98	0.81	0.52	0.14	0.02	0.00	0.00	0.00
	80	1.00	0.98	0.86	0.48	0.11	0.01	0.00	0.00	0.00
	90	1.00	0.99	0.88	0.52	0.12	0.01	0.00	0.00	0.00
	100	1.00	1.00	0.90	0.48	0.10	0.01	0.00	0.00	0.00
	125	1.00	1.00	0.95	0.52	0.07	0.00	0.00	0.00	0.00
150	1.00	1.00	0.97	0.56	0.04	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.65	0.70	0.20	1.00	0.15
	20	0.65	0.65	0.30	0.90	0.10
	30	0.66	0.67	0.33	0.93	0.08
	40	0.66	0.65	0.42	0.88	0.07
	50	0.65	0.64	0.48	0.82	0.06
	60	0.65	0.65	0.47	0.82	0.06
	70	0.65	0.66	0.41	0.80	0.05
	80	0.65	0.65	0.51	0.80	0.05
	90	0.65	0.66	0.53	0.80	0.04
	100	0.65	0.65	0.54	0.77	0.04
	125	0.65	0.66	0.57	0.75	0.03
	150	0.65	0.65	0.57	0.73	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.18: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 50 000 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)

		≥ accuracy (%)								
test sample size	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.79	0.79	0.64	0.53	0.26	0.26	0.08	0.08	0.01
	20	0.90	0.79	0.66	0.44	0.28	0.14	0.05	0.05	0.00
	30	0.95	0.90	0.70	0.52	0.23	0.12	0.03	0.01	0.00
	40	0.98	0.92	0.79	0.47	0.20	0.06	0.01	0.00	0.00
	50	0.99	0.95	0.84	0.56	0.19	0.06	0.00	0.00	0.00
	60	0.99	0.96	0.81	0.48	0.16	0.03	0.00	0.00	0.00
	70	1.00	0.99	0.83	0.55	0.15	0.02	0.00	0.00	0.00
	80	1.00	0.98	0.89	0.52	0.13	0.01	0.00	0.00	0.00
	90	1.00	1.00	0.90	0.56	0.11	0.01	0.00	0.00	0.00
	100	1.00	0.99	0.92	0.50	0.10	0.00	0.00	0.00	0.00
	125	1.00	1.00	0.95	0.54	0.10	0.00	0.00	0.00	0.00
150	1.00	1.00	0.99	0.61	0.04	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.66	0.70	0.30	1.00	0.14
	20	0.66	0.65	0.30	0.95	0.11
	30	0.65	0.67	0.40	0.90	0.08
	40	0.66	0.65	0.42	0.90	0.07
	50	0.66	0.66	0.48	0.84	0.06
	60	0.65	0.65	0.40	0.83	0.06
	70	0.66	0.66	0.47	0.80	0.05
	80	0.66	0.66	0.51	0.80	0.05
	90	0.66	0.66	0.54	0.77	0.04
	100	0.65	0.65	0.51	0.78	0.04
	125	0.66	0.66	0.55	0.77	0.03
	150	0.66	0.66	0.57	0.77	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.19: Results as a function of variable test set sizes with a fixed classifier. For **feature selection** an ANOVA was computed inside the the pipeline and the top 50 000 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)

		≥ accuracy (%)								
	<i>n</i>	50	55	60	65	70	75	80	85	90
test sample size	10	0.78	0.78	0.64	0.56	0.32	0.32	0.12	0.12	0.02
	20	0.93	0.83	0.70	0.49	0.32	0.16	0.07	0.06	0.00
	30	0.96	0.91	0.75	0.63	0.30	0.16	0.03	0.01	0.00
	40	0.99	0.93	0.83	0.55	0.29	0.09	0.02	0.00	0.00
	50	0.99	0.97	0.87	0.63	0.24	0.09	0.01	0.00	0.00
	60	1.00	0.98	0.87	0.58	0.23	0.05	0.00	0.00	0.00
	70	1.00	0.99	0.90	0.67	0.23	0.05	0.00	0.00	0.00
	80	1.00	1.00	0.94	0.64	0.24	0.02	0.00	0.00	0.00
	90	1.00	0.99	0.93	0.70	0.19	0.03	0.00	0.00	0.00
	100	1.00	1.00	0.96	0.67	0.18	0.01	0.00	0.00	0.00
	125	1.00	1.00	0.98	0.71	0.20	0.01	0.00	0.00	0.00
	150	1.00	1.00	0.99	0.77	0.12	0.00	0.00	0.00	0.00

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.67	0.70	0.10	1.00	0.15
	20	0.67	0.65	0.30	0.95	0.10
	30	0.67	0.67	0.30	0.93	0.08
	40	0.67	0.68	0.45	0.88	0.07
	50	0.67	0.67	0.44	0.84	0.06
	60	0.67	0.67	0.48	0.83	0.05
	70	0.67	0.67	0.49	0.81	0.05
	80	0.67	0.68	0.55	0.80	0.04
	90	0.67	0.67	0.52	0.81	0.04
	100	0.67	0.67	0.56	0.79	0.04
	125	0.67	0.67	0.54	0.76	0.03
	150	0.67	0.67	0.58	0.75	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.20: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 10 **components** were retained. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)

		≥ accuracy (%)								
test sample size	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	20	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	30	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	40	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	50	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	60	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	70	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	80	0.27	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	90	0.28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	125	0.41	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
150	0.51	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.50	0.50	0.50	0.60	0.02
	20	0.50	0.50	0.50	0.55	0.01
	30	0.50	0.50	0.50	0.53	0.01
	40	0.50	0.50	0.50	0.53	0.01
	50	0.50	0.50	0.50	0.52	0.01
	60	0.50	0.50	0.50	0.52	0.01
	70	0.50	0.50	0.50	0.51	0.01
	80	0.50	0.50	0.50	0.51	0.01
	90	0.50	0.50	0.50	0.51	0.01
	100	0.50	0.50	0.50	0.51	0.00
	125	0.50	0.50	0.50	0.51	0.00
	150	0.50	0.51	0.50	0.51	0.00

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.21: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 10 **components** were retained. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)

		≥ accuracy (%)								
	n	50	55	60	65	70	75	80	85	90
test sample size	10	0.58	0.58	0.45	0.32	0.13	0.13	0.04	0.04	0.00
	20	0.70	0.52	0.37	0.17	0.07	0.02	0.00	0.00	0.00
	30	0.74	0.60	0.31	0.18	0.04	0.02	0.00	0.00	0.00
	40	0.81	0.57	0.36	0.11	0.03	0.00	0.00	0.00	0.00
	50	0.88	0.67	0.39	0.12	0.02	0.00	0.00	0.00	0.00
	60	0.88	0.61	0.25	0.07	0.01	0.00	0.00	0.00	0.00
	70	0.92	0.67	0.30	0.07	0.00	0.00	0.00	0.00	0.00
	80	0.92	0.65	0.28	0.03	0.00	0.00	0.00	0.00	0.00
	90	0.92	0.69	0.22	0.04	0.00	0.00	0.00	0.00	0.00
	100	0.95	0.66	0.25	0.02	0.00	0.00	0.00	0.00	0.00
	125	0.99	0.78	0.21	0.01	0.00	0.00	0.00	0.00	0.00
	150	0.99	0.81	0.16	0.00	0.00	0.00	0.00	0.00	0.00

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.58	0.60	0.10	1.00	0.16
	20	0.58	0.60	0.25	0.85	0.10
	30	0.57	0.57	0.30	0.83	0.08
	40	0.57	0.57	0.35	0.78	0.07
	50	0.58	0.58	0.36	0.78	0.06
	60	0.57	0.57	0.42	0.75	0.06
	70	0.58	0.57	0.43	0.73	0.05
	80	0.57	0.57	0.38	0.70	0.05
	90	0.57	0.57	0.41	0.71	0.05
	100	0.57	0.57	0.43	0.70	0.04
	125	0.57	0.58	0.46	0.68	0.03
	150	0.57	0.57	0.49	0.67	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.22: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 10 **components** were retained. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)

		≥ accuracy (%)								
test sample size	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.79	0.79	0.67	0.56	0.31	0.31	0.10	0.10	0.02
	20	0.92	0.84	0.72	0.51	0.30	0.17	0.07	0.06	0.01
	30	0.97	0.92	0.74	0.60	0.30	0.18	0.03	0.01	0.00
	40	0.99	0.94	0.82	0.57	0.28	0.09	0.02	0.00	0.00
	50	0.99	0.98	0.88	0.60	0.25	0.08	0.00	0.00	0.00
	60	0.99	0.98	0.88	0.60	0.25	0.04	0.01	0.00	0.00
	70	1.00	1.00	0.90	0.66	0.23	0.06	0.00	0.00	0.00
	80	1.00	0.99	0.92	0.61	0.21	0.02	0.00	0.00	0.00
	90	1.00	1.00	0.94	0.67	0.20	0.02	0.00	0.00	0.00
	100	1.00	1.00	0.97	0.68	0.18	0.01	0.00	0.00	0.00
	125	1.00	1.00	0.98	0.69	0.17	0.01	0.00	0.00	0.00
150	1.00	1.00	0.99	0.75	0.10	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.67	0.70	0.10	1.00	0.15
	20	0.67	0.70	0.30	0.95	0.11
	30	0.67	0.67	0.37	0.90	0.08
	40	0.67	0.68	0.42	0.90	0.07
	50	0.67	0.66	0.40	0.82	0.06
	60	0.67	0.67	0.47	0.83	0.05
	70	0.67	0.67	0.53	0.81	0.05
	80	0.67	0.68	0.51	0.81	0.05
	90	0.67	0.67	0.54	0.80	0.04
	100	0.67	0.67	0.55	0.78	0.04
	125	0.67	0.67	0.58	0.77	0.03
	150	0.67	0.67	0.59	0.77	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.23: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 50 **components** were retained. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)

		≥ accuracy (%)								
test sample size	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.15	0.15	0.01	0.01	0.00	0.00	0.00	0.00	0.00
	20	0.27	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	30	0.38	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	40	0.42	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	50	0.44	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	60	0.51	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	70	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	80	0.59	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	90	0.61	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.68	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	125	0.79	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
150	0.79	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.51	0.50	0.40	0.70	0.05
	20	0.51	0.50	0.45	0.65	0.03
	30	0.51	0.50	0.43	0.63	0.03
	40	0.51	0.50	0.45	0.60	0.02
	50	0.51	0.50	0.46	0.58	0.02
	60	0.51	0.52	0.47	0.57	0.02
	70	0.51	0.51	0.47	0.56	0.02
	80	0.51	0.51	0.47	0.55	0.01
	90	0.51	0.51	0.48	0.54	0.01
	100	0.51	0.51	0.48	0.55	0.01
	125	0.51	0.51	0.48	0.54	0.01
	150	0.51	0.51	0.49	0.53	0.01

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.24: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 50 **components** were retained. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)

		≥ accuracy (%)								
test sample size	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.67	0.67	0.55	0.43	0.18	0.18	0.06	0.06	0.01
	20	0.85	0.70	0.56	0.34	0.19	0.08	0.02	0.02	0.00
	30	0.90	0.82	0.58	0.41	0.15	0.07	0.01	0.00	0.00
	40	0.96	0.82	0.63	0.31	0.10	0.03	0.00	0.00	0.00
	50	0.95	0.88	0.68	0.35	0.10	0.02	0.00	0.00	0.00
	60	0.99	0.88	0.61	0.27	0.06	0.01	0.00	0.00	0.00
	70	0.98	0.93	0.65	0.32	0.04	0.01	0.00	0.00	0.00
	80	0.99	0.94	0.71	0.27	0.04	0.00	0.00	0.00	0.00
	90	1.00	0.96	0.71	0.28	0.04	0.00	0.00	0.00	0.00
	100	1.00	0.96	0.75	0.24	0.02	0.00	0.00	0.00	0.00
	125	1.00	0.99	0.78	0.22	0.01	0.00	0.00	0.00	0.00
150	1.00	1.00	0.82	0.20	0.00	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.62	0.60	0.20	1.00	0.15
	20	0.63	0.65	0.30	0.90	0.10
	30	0.63	0.63	0.37	0.83	0.09
	40	0.63	0.62	0.40	0.88	0.07
	50	0.63	0.62	0.42	0.86	0.06
	60	0.62	0.62	0.45	0.80	0.05
	70	0.63	0.63	0.44	0.77	0.05
	80	0.63	0.62	0.45	0.75	0.05
	90	0.63	0.62	0.50	0.76	0.04
	100	0.63	0.63	0.47	0.77	0.04
	125	0.63	0.62	0.54	0.74	0.03
	150	0.63	0.63	0.55	0.70	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.25: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 50 **components** were retained. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)

		≥ accuracy (%)								
test sample size	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.72	0.72	0.58	0.45	0.22	0.22	0.08	0.08	0.01
	20	0.88	0.75	0.60	0.38	0.21	0.08	0.02	0.02	0.00
	30	0.94	0.88	0.64	0.46	0.16	0.09	0.01	0.00	0.00
	40	0.97	0.87	0.70	0.37	0.15	0.04	0.01	0.00	0.00
	50	0.98	0.94	0.78	0.47	0.13	0.04	0.00	0.00	0.00
	60	0.99	0.93	0.73	0.39	0.12	0.03	0.00	0.00	0.00
	70	0.99	0.97	0.78	0.47	0.11	0.02	0.00	0.00	0.00
	80	1.00	0.98	0.82	0.39	0.09	0.01	0.00	0.00	0.00
	90	1.00	0.99	0.84	0.43	0.08	0.01	0.00	0.00	0.00
	100	1.00	0.98	0.88	0.39	0.05	0.00	0.00	0.00	0.00
	125	1.00	1.00	0.89	0.37	0.04	0.00	0.00	0.00	0.00
150	1.00	1.00	0.94	0.41	0.02	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.64	0.60	0.20	1.00	0.15
	20	0.64	0.65	0.25	0.95	0.10
	30	0.64	0.63	0.40	0.90	0.08
	40	0.64	0.65	0.40	0.85	0.07
	50	0.65	0.64	0.44	0.84	0.06
	60	0.64	0.65	0.43	0.80	0.06
	70	0.65	0.64	0.47	0.80	0.05
	80	0.64	0.65	0.49	0.78	0.05
	90	0.64	0.64	0.50	0.78	0.04
	100	0.64	0.64	0.52	0.76	0.04
	125	0.64	0.64	0.54	0.76	0.03
	150	0.64	0.65	0.55	0.72	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.26: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 50 **components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)

		$\geq$ accuracy (%)								
test sample size	n	50	55	60	65	70	75	80	85	90
	10	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	20	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	30	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	40	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	50	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	60	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	70	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	80	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	90	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	125	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
150	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.50	0.50	0.40	0.60	0.03
	20	0.50	0.50	0.40	0.55	0.02
	30	0.50	0.50	0.43	0.53	0.02
	40	0.50	0.50	0.45	0.53	0.01
	50	0.50	0.50	0.46	0.52	0.01
	60	0.50	0.50	0.47	0.52	0.01
	70	0.50	0.50	0.47	0.51	0.01
	80	0.50	0.50	0.47	0.51	0.01
	90	0.50	0.50	0.48	0.51	0.01
	100	0.50	0.50	0.48	0.51	0.01
	125	0.50	0.50	0.48	0.51	0.01
	150	0.50	0.50	0.49	0.51	0.01

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.27: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 50 **components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)

		$\geq$ accuracy (%)								
test sample size	n	50	55	60	65	70	75	80	85	90
	10	0.46	0.46	0.34	0.23	0.06	0.06	0.02	0.02	0.00
	20	0.54	0.35	0.23	0.09	0.04	0.01	0.00	0.00	0.00
	30	0.59	0.42	0.19	0.09	0.01	0.01	0.00	0.00	0.00
	40	0.62	0.32	0.17	0.03	0.01	0.00	0.00	0.00	0.00
	50	0.68	0.44	0.16	0.03	0.00	0.00	0.00	0.00	0.00
	60	0.68	0.36	0.11	0.02	0.00	0.00	0.00	0.00	0.00
	70	0.74	0.42	0.09	0.01	0.00	0.00	0.00	0.00	0.00
	80	0.75	0.34	0.09	0.01	0.00	0.00	0.00	0.00	0.00
	90	0.76	0.39	0.06	0.01	0.00	0.00	0.00	0.00	0.00
	100	0.80	0.36	0.07	0.00	0.00	0.00	0.00	0.00	0.00
	125	0.88	0.38	0.04	0.00	0.00	0.00	0.00	0.00	0.00
150	0.87	0.32	0.01	0.00	0.00	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.53	0.50	0.10	1.00	0.16
	20	0.54	0.55	0.15	0.85	0.11
	30	0.54	0.53	0.27	0.80	0.09
	40	0.53	0.53	0.33	0.75	0.07
	50	0.54	0.54	0.34	0.72	0.06
	60	0.54	0.53	0.35	0.75	0.06
	70	0.54	0.54	0.39	0.69	0.05
	80	0.54	0.54	0.38	0.70	0.05
	90	0.54	0.53	0.42	0.69	0.04
	100	0.54	0.54	0.36	0.67	0.04
	125	0.54	0.54	0.43	0.66	0.03
	150	0.54	0.53	0.44	0.65	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.28: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 50 **components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)

		≥ accuracy (%)								
	n	50	55	60	65	70	75	80	85	90
test sample size	10	0.44	0.44	0.19	0.14	0.03	0.03	0.01	0.01	0.00
	20	0.60	0.33	0.14	0.03	0.00	0.00	0.00	0.00	0.00
	30	0.67	0.42	0.09	0.02	0.00	0.00	0.00	0.00	0.00
	40	0.75	0.36	0.09	0.01	0.00	0.00	0.00	0.00	0.00
	50	0.77	0.43	0.08	0.01	0.00	0.00	0.00	0.00	0.00
	60	0.80	0.32	0.04	0.00	0.00	0.00	0.00	0.00	0.00
	70	0.84	0.43	0.05	0.00	0.00	0.00	0.00	0.00	0.00
	80	0.89	0.33	0.04	0.00	0.00	0.00	0.00	0.00	0.00
	90	0.91	0.41	0.02	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.92	0.34	0.02	0.00	0.00	0.00	0.00	0.00	0.00
	125	0.97	0.36	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	150	0.97	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.54	0.50	0.20	0.90	0.10
	20	0.54	0.55	0.30	0.80	0.07
	30	0.54	0.53	0.37	0.73	0.05
	40	0.55	0.55	0.40	0.70	0.05
	50	0.54	0.54	0.40	0.70	0.05
	60	0.54	0.53	0.42	0.68	0.04
	70	0.54	0.54	0.44	0.67	0.04
	80	0.54	0.54	0.45	0.65	0.03
	90	0.54	0.54	0.43	0.64	0.03
	100	0.54	0.54	0.46	0.63	0.03
	125	0.54	0.55	0.47	0.62	0.02
	150	0.54	0.54	0.48	0.60	0.02

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.29: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 500 **components** were retained. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)

		$\geq$ accuracy (%)								
n		50	55	60	65	70	75	80	85	90
test sample size	10	0.12	0.12	0.01	0.01	0.00	0.00	0.00	0.00	0.00
	20	0.21	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	30	0.27	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	40	0.28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	50	0.30	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	60	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	70	0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	80	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	90	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	125	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	150	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.50	0.50	0.30	0.80	0.05
	20	0.50	0.50	0.35	0.65	0.04
	30	0.50	0.50	0.40	0.60	0.03
	40	0.50	0.50	0.42	0.57	0.03
	50	0.50	0.50	0.42	0.56	0.02
	60	0.50	0.50	0.43	0.57	0.02
	70	0.50	0.50	0.44	0.56	0.02
	80	0.50	0.50	0.44	0.55	0.02
	90	0.50	0.50	0.46	0.54	0.02
	100	0.50	0.50	0.45	0.55	0.01
	125	0.50	0.50	0.47	0.54	0.01
	150	0.50	0.50	0.47	0.53	0.01

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.30: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 500 **components** were retained. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)

test sample size	≥ accuracy (%)									
	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.62	0.62	0.49	0.36	0.15	0.15	0.04	0.04	0.01
	20	0.75	0.56	0.41	0.23	0.11	0.05	0.01	0.01	0.00
	30	0.80	0.70	0.43	0.28	0.06	0.03	0.00	0.00	0.00
	40	0.85	0.66	0.43	0.16	0.04	0.01	0.00	0.00	0.00
	50	0.91	0.76	0.52	0.19	0.03	0.01	0.00	0.00	0.00
	60	0.93	0.72	0.40	0.13	0.02	0.00	0.00	0.00	0.00
	70	0.95	0.80	0.39	0.15	0.01	0.00	0.00	0.00	0.00
	80	0.97	0.77	0.41	0.07	0.01	0.00	0.00	0.00	0.00
90	0.98	0.85	0.42	0.09	0.00	0.00	0.00	0.00	0.00	
100	0.98	0.84	0.46	0.07	0.00	0.00	0.00	0.00	0.00	
125	1.00	0.91	0.42	0.05	0.00	0.00	0.00	0.00	0.00	
150	1.00	0.94	0.41	0.02	0.00	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

test sample size	n	mean	median	min	max	std
	10	0.60	0.60	0.10	1.00	0.15
	20	0.59	0.60	0.25	0.95	0.11
	30	0.59	0.60	0.33	0.83	0.09
	40	0.59	0.60	0.38	0.80	0.07
	50	0.60	0.60	0.38	0.78	0.06
	60	0.59	0.58	0.42	0.78	0.06
	70	0.59	0.59	0.41	0.77	0.05
	80	0.59	0.59	0.42	0.74	0.05
	90	0.59	0.60	0.44	0.73	0.04
	100	0.59	0.60	0.47	0.73	0.04
	125	0.59	0.59	0.49	0.70	0.03
	150	0.59	0.59	0.51	0.68	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.31: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 500 **components** were retained. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)

		≥ accuracy (%)								
	n	50	55	60	65	70	75	80	85	90
test sample size	10	0.82	0.82	0.71	0.62	0.35	0.35	0.13	0.13	0.03
	20	0.94	0.85	0.74	0.54	0.34	0.17	0.06	0.05	0.00
	30	0.98	0.95	0.80	0.66	0.34	0.21	0.04	0.01	0.00
	40	0.99	0.95	0.88	0.64	0.34	0.12	0.02	0.01	0.00
	50	1.00	0.98	0.91	0.70	0.31	0.13	0.01	0.00	0.00
	60	1.00	1.00	0.93	0.70	0.31	0.07	0.01	0.00	0.00
	70	1.00	0.99	0.94	0.75	0.33	0.08	0.01	0.00	0.00
	80	1.00	1.00	0.97	0.74	0.31	0.05	0.00	0.00	0.00
	90	1.00	1.00	0.98	0.80	0.31	0.04	0.00	0.00	0.00
	100	1.00	1.00	0.98	0.78	0.30	0.03	0.00	0.00	0.00
	125	1.00	1.00	0.99	0.84	0.31	0.02	0.00	0.00	0.00
	150	1.00	1.00	1.00	0.90	0.23	0.01	0.00	0.00	0.00

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.69	0.70	0.20	1.00	0.15
	20	0.68	0.70	0.35	0.95	0.10
	30	0.68	0.70	0.40	0.90	0.08
	40	0.68	0.68	0.45	0.95	0.07
	50	0.68	0.68	0.50	0.84	0.06
	60	0.68	0.68	0.53	0.83	0.05
	70	0.68	0.69	0.51	0.83	0.05
	80	0.68	0.69	0.55	0.81	0.04
	90	0.68	0.69	0.57	0.83	0.04
	100	0.68	0.68	0.57	0.80	0.04
	125	0.68	0.68	0.56	0.78	0.03
	150	0.68	0.69	0.59	0.76	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.32: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 500 **components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)

		$\geq$ accuracy (%)								
test sample size	n	50	55	60	65	70	75	80	85	90
	10	0.36	0.36	0.11	0.07	0.01	0.01	0.00	0.00	0.00
	20	0.47	0.22	0.07	0.01	0.00	0.00	0.00	0.00	0.00
	30	0.54	0.30	0.03	0.01	0.00	0.00	0.00	0.00	0.00
	40	0.59	0.19	0.03	0.00	0.00	0.00	0.00	0.00	0.00
	50	0.64	0.26	0.02	0.00	0.00	0.00	0.00	0.00	0.00
	60	0.68	0.15	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	70	0.69	0.19	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	80	0.72	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	90	0.75	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.76	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	125	0.86	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00
150	0.86	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.52	0.50	0.20	0.90	0.09
	20	0.52	0.50	0.35	0.75	0.07
	30	0.52	0.53	0.37	0.67	0.05
	40	0.52	0.53	0.40	0.68	0.05
	50	0.52	0.52	0.38	0.64	0.04
	60	0.52	0.52	0.40	0.65	0.04
	70	0.52	0.53	0.41	0.64	0.03
	80	0.52	0.53	0.44	0.61	0.03
	90	0.52	0.52	0.43	0.59	0.03
	100	0.52	0.52	0.45	0.61	0.02
	125	0.52	0.52	0.45	0.58	0.02
	150	0.52	0.52	0.45	0.57	0.02

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.33: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 500 **components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)

		≥ accuracy (%)								
test sample size	n	50	55	60	65	70	75	80	85	90
	10	0.74	0.74	0.61	0.50	0.24	0.24	0.07	0.07	0.01
	20	0.87	0.75	0.59	0.35	0.20	0.09	0.03	0.03	0.00
	30	0.93	0.86	0.63	0.43	0.17	0.09	0.01	0.01	0.00
	40	0.95	0.86	0.70	0.37	0.16	0.03	0.01	0.00	0.00
	50	0.98	0.93	0.76	0.45	0.14	0.04	0.00	0.00	0.00
	60	0.99	0.93	0.72	0.39	0.11	0.01	0.00	0.00	0.00
	70	1.00	0.97	0.74	0.39	0.08	0.01	0.00	0.00	0.00
	80	1.00	0.98	0.81	0.36	0.07	0.00	0.00	0.00	0.00
	90	1.00	0.98	0.81	0.42	0.06	0.01	0.00	0.00	0.00
	100	1.00	0.98	0.84	0.37	0.04	0.00	0.00	0.00	0.00
	125	1.00	1.00	0.88	0.32	0.03	0.00	0.00	0.00	0.00
150	1.00	1.00	0.92	0.35	0.01	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.64	0.65	0.10	1.00	0.15
	20	0.64	0.65	0.30	0.95	0.10
	30	0.64	0.63	0.30	0.93	0.08
	40	0.64	0.65	0.40	0.85	0.07
	50	0.64	0.64	0.48	0.82	0.06
	60	0.64	0.63	0.47	0.82	0.06
	70	0.64	0.64	0.49	0.77	0.05
	80	0.64	0.64	0.51	0.78	0.04
	90	0.64	0.64	0.47	0.77	0.04
	100	0.64	0.64	0.52	0.75	0.04
	125	0.64	0.64	0.54	0.73	0.03
	150	0.64	0.64	0.55	0.73	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.34: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 500 **components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)

		$\geq$ accuracy (%)								
test sample size	n	50	55	60	65	70	75	80	85	90
	10	0.79	0.79	0.66	0.54	0.29	0.29	0.10	0.10	0.02
	20	0.93	0.83	0.73	0.48	0.28	0.12	0.05	0.04	0.00
	30	0.97	0.94	0.76	0.60	0.27	0.16	0.03	0.01	0.00
	40	0.99	0.94	0.83	0.55	0.24	0.06	0.01	0.00	0.00
	50	1.00	0.98	0.88	0.63	0.25	0.09	0.01	0.00	0.00
	60	1.00	0.97	0.88	0.59	0.23	0.04	0.00	0.00	0.00
	70	1.00	0.99	0.87	0.62	0.18	0.03	0.00	0.00	0.00
	80	1.00	0.99	0.91	0.60	0.18	0.02	0.00	0.00	0.00
	90	1.00	1.00	0.95	0.68	0.20	0.02	0.00	0.00	0.00
	100	1.00	1.00	0.97	0.63	0.15	0.01	0.00	0.00	0.00
	125	1.00	1.00	0.99	0.69	0.13	0.00	0.00	0.00	0.00
150	1.00	1.00	1.00	0.75	0.08	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.66	0.70	0.20	1.00	0.15
	20	0.67	0.65	0.35	0.95	0.10
	30	0.67	0.67	0.43	0.93	0.08
	40	0.67	0.68	0.45	0.85	0.06
	50	0.67	0.66	0.48	0.86	0.06
	60	0.67	0.67	0.48	0.82	0.05
	70	0.66	0.67	0.50	0.84	0.05
	80	0.67	0.66	0.51	0.80	0.05
	90	0.67	0.67	0.51	0.81	0.04
	100	0.67	0.67	0.56	0.82	0.04
	125	0.67	0.66	0.57	0.77	0.03
	150	0.67	0.67	0.59	0.74	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.35: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 500 **components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 100 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)

		≥ accuracy (%)								
test sample size	n	50	55	60	65	70	75	80	85	90
	10	0.37	0.37	0.08	0.07	0.01	0.01	0.00	0.00	0.00
	20	0.53	0.19	0.04	0.01	0.00	0.00	0.00	0.00	0.00
	30	0.65	0.35	0.03	0.01	0.00	0.00	0.00	0.00	0.00
	40	0.72	0.19	0.02	0.00	0.00	0.00	0.00	0.00	0.00
	50	0.74	0.27	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	60	0.82	0.18	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	70	0.84	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	80	0.89	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	90	0.93	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.93	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	125	0.98	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00
150	0.99	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.54	0.50	0.30	0.90	0.07
	20	0.53	0.55	0.40	0.70	0.05
	30	0.53	0.53	0.40	0.67	0.04
	40	0.53	0.53	0.45	0.65	0.03
	50	0.53	0.54	0.44	0.66	0.03
	60	0.53	0.53	0.47	0.65	0.03
	70	0.53	0.53	0.46	0.61	0.03
	80	0.53	0.54	0.47	0.61	0.02
	90	0.54	0.53	0.47	0.61	0.02
	100	0.53	0.53	0.47	0.61	0.02
	125	0.53	0.53	0.48	0.60	0.02
	150	0.53	0.53	0.49	0.57	0.01

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.36: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 500 **components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 100 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)

		$\geq$ accuracy (%)								
	n	50	55	60	65	70	75	80	85	90
test sample size	10	0.64	0.64	0.51	0.39	0.15	0.15	0.04	0.04	0.00
	20	0.78	0.64	0.47	0.26	0.12	0.05	0.01	0.01	0.00
	30	0.86	0.75	0.48	0.32	0.09	0.04	0.01	0.00	0.00
	40	0.89	0.72	0.53	0.22	0.06	0.01	0.00	0.00	0.00
	50	0.93	0.81	0.53	0.23	0.05	0.01	0.00	0.00	0.00
	60	0.94	0.77	0.45	0.15	0.03	0.00	0.00	0.00	0.00
	70	0.97	0.85	0.47	0.18	0.02	0.00	0.00	0.00	0.00
	80	0.98	0.83	0.52	0.14	0.01	0.00	0.00	0.00	0.00
	90	0.98	0.88	0.53	0.13	0.01	0.00	0.00	0.00	0.00
	100	0.99	0.88	0.50	0.11	0.01	0.00	0.00	0.00	0.00
	125	1.00	0.95	0.53	0.07	0.00	0.00	0.00	0.00	0.00
	150	1.00	0.97	0.53	0.05	0.00	0.00	0.00	0.00	0.00

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.60	0.60	0.10	1.00	0.15
	20	0.61	0.60	0.30	0.90	0.10
	30	0.61	0.60	0.30	0.90	0.09
	40	0.60	0.60	0.38	0.82	0.07
	50	0.60	0.60	0.40	0.78	0.06
	60	0.60	0.60	0.43	0.77	0.06
	70	0.60	0.60	0.44	0.76	0.05
	80	0.60	0.60	0.45	0.74	0.05
	90	0.60	0.60	0.47	0.74	0.04
	100	0.60	0.60	0.48	0.73	0.04
	125	0.60	0.60	0.50	0.71	0.03
	150	0.60	0.60	0.51	0.69	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.37: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and 500 **components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 100 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)

		≥ accuracy (%)								
test sample size	n	50	55	60	65	70	75	80	85	90
	10	0.67	0.67	0.55	0.42	0.19	0.19	0.04	0.04	0.00
	20	0.83	0.66	0.49	0.27	0.12	0.05	0.01	0.01	0.00
	30	0.88	0.77	0.52	0.35	0.11	0.04	0.01	0.00	0.00
	40	0.92	0.78	0.58	0.24	0.07	0.01	0.00	0.00	0.00
	50	0.95	0.85	0.61	0.28	0.05	0.02	0.00	0.00	0.00
	60	0.96	0.83	0.54	0.21	0.04	0.00	0.00	0.00	0.00
	70	0.97	0.88	0.52	0.22	0.02	0.00	0.00	0.00	0.00
	80	0.98	0.87	0.60	0.17	0.02	0.00	0.00	0.00	0.00
	90	0.99	0.94	0.58	0.17	0.01	0.00	0.00	0.00	0.00
	100	1.00	0.93	0.63	0.16	0.01	0.00	0.00	0.00	0.00
	125	1.00	0.98	0.65	0.13	0.00	0.00	0.00	0.00	0.00
150	1.00	0.99	0.69	0.10	0.00	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.61	0.60	0.20	1.00	0.15
	20	0.61	0.60	0.30	0.90	0.10
	30	0.61	0.60	0.37	0.87	0.08
	40	0.61	0.62	0.35	0.82	0.07
	50	0.61	0.62	0.42	0.86	0.06
	60	0.61	0.62	0.45	0.78	0.06
	70	0.61	0.61	0.43	0.77	0.05
	80	0.61	0.61	0.47	0.75	0.05
	90	0.61	0.61	0.47	0.74	0.04
	100	0.61	0.61	0.50	0.72	0.04
	125	0.61	0.62	0.51	0.73	0.03
	150	0.61	0.61	0.51	0.69	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.38: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)

		≥ accuracy (%)								
test sample size	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.17	0.17	0.03	0.01	0.00	0.00	0.00	0.00	0.00
	20	0.22	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	30	0.19	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	40	0.20	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	50	0.17	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	60	0.15	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	70	0.14	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	80	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	90	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	125	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
150	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.47	0.50	0.20	0.80	0.09
	20	0.47	0.50	0.25	0.65	0.07
	30	0.47	0.47	0.33	0.67	0.05
	40	0.48	0.47	0.35	0.65	0.05
	50	0.48	0.48	0.36	0.60	0.04
	60	0.47	0.47	0.35	0.58	0.03
	70	0.47	0.47	0.39	0.56	0.03
	80	0.47	0.47	0.38	0.56	0.03
	90	0.47	0.48	0.39	0.56	0.03
	100	0.47	0.47	0.40	0.56	0.03
	125	0.47	0.48	0.40	0.56	0.02
	150	0.47	0.47	0.41	0.53	0.02

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.39: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)

test sample size	≥ accuracy (%)									
	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.79	0.79	0.66	0.54	0.27	0.27	0.09	0.09	0.02
	20	0.90	0.79	0.64	0.43	0.24	0.10	0.04	0.03	0.00
	30	0.95	0.91	0.68	0.54	0.23	0.12	0.02	0.01	0.00
	40	0.98	0.91	0.78	0.48	0.21	0.06	0.01	0.00	0.00
	50	0.99	0.95	0.82	0.57	0.19	0.06	0.00	0.00	0.00
	60	0.99	0.97	0.82	0.49	0.17	0.02	0.00	0.00	0.00
	70	0.99	0.98	0.85	0.56	0.16	0.03	0.00	0.00	0.00
	80	1.00	0.99	0.86	0.51	0.14	0.01	0.00	0.00	0.00
	90	1.00	0.99	0.90	0.55	0.11	0.01	0.00	0.00	0.00
	100	1.00	1.00	0.92	0.51	0.11	0.00	0.00	0.00	0.00
125	1.00	1.00	0.95	0.55	0.07	0.00	0.00	0.00	0.00	
150	1.00	1.00	0.98	0.56	0.04	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.66	0.70	0.20	1.00	0.15
	20	0.65	0.65	0.35	0.95	0.10
	30	0.66	0.67	0.40	0.90	0.08
	40	0.66	0.65	0.42	0.85	0.07
	50	0.66	0.66	0.46	0.84	0.06
	60	0.66	0.65	0.48	0.83	0.05
	70	0.66	0.66	0.50	0.80	0.05
	80	0.66	0.66	0.50	0.81	0.05
	90	0.66	0.66	0.52	0.78	0.04
	100	0.66	0.66	0.53	0.76	0.04
	125	0.66	0.66	0.57	0.76	0.03
	150	0.66	0.65	0.56	0.75	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.40: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)

		≥ accuracy (%)								
test sample size	n	50	55	60	65	70	75	80	85	90
	10	0.75	0.75	0.62	0.50	0.24	0.24	0.09	0.09	0.02
	20	0.88	0.76	0.62	0.42	0.24	0.11	0.05	0.04	0.00
	30	0.94	0.87	0.67	0.50	0.20	0.10	0.01	0.00	0.00
	40	0.97	0.89	0.75	0.43	0.20	0.05	0.01	0.00	0.00
	50	0.98	0.94	0.78	0.44	0.13	0.04	0.00	0.00	0.00
	60	0.99	0.94	0.76	0.43	0.14	0.02	0.00	0.00	0.00
	70	1.00	0.97	0.77	0.45	0.11	0.02	0.00	0.00	0.00
	80	1.00	0.97	0.86	0.42	0.09	0.01	0.00	0.00	0.00
	90	1.00	0.98	0.84	0.49	0.09	0.01	0.00	0.00	0.00
	100	1.00	0.99	0.89	0.43	0.07	0.00	0.00	0.00	0.00
	125	1.00	1.00	0.93	0.42	0.05	0.00	0.00	0.00	0.00
150	1.00	1.00	0.95	0.41	0.02	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.65	0.60	0.20	1.00	0.15
	20	0.65	0.65	0.30	0.90	0.11
	30	0.65	0.63	0.40	0.90	0.08
	40	0.65	0.65	0.40	0.90	0.07
	50	0.64	0.64	0.42	0.86	0.06
	60	0.65	0.65	0.43	0.82	0.06
	70	0.64	0.64	0.49	0.80	0.05
	80	0.65	0.65	0.46	0.78	0.05
	90	0.65	0.64	0.51	0.79	0.04
	100	0.65	0.65	0.52	0.76	0.04
	125	0.65	0.65	0.55	0.75	0.03
	150	0.64	0.65	0.55	0.73	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.41: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)

		$\geq$ accuracy (%)								
test sample size	n	50	55	60	65	70	75	80	85	90
	10	0.34	0.34	0.12	0.09	0.01	0.01	0.00	0.00	0.00
	20	0.46	0.17	0.05	0.01	0.00	0.00	0.00	0.00	0.00
	30	0.52	0.26	0.03	0.01	0.00	0.00	0.00	0.00	0.00
	40	0.57	0.17	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	50	0.59	0.20	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	60	0.63	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	70	0.66	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	80	0.69	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	90	0.71	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.70	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	125	0.82	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00
150	0.83	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.52	0.50	0.20	0.80	0.09
	20	0.52	0.50	0.30	0.75	0.06
	30	0.52	0.53	0.37	0.67	0.05
	40	0.52	0.53	0.38	0.68	0.04
	50	0.52	0.52	0.40	0.64	0.04
	60	0.52	0.52	0.42	0.65	0.03
	70	0.52	0.51	0.43	0.63	0.03
	80	0.52	0.53	0.44	0.61	0.03
	90	0.52	0.52	0.44	0.60	0.03
	100	0.52	0.52	0.45	0.59	0.02
	125	0.52	0.52	0.46	0.59	0.02
	150	0.52	0.52	0.47	0.59	0.02

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.42: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)

		$\geq$ accuracy (%)								
	n	50	55	60	65	70	75	80	85	90
test sample size	10	0.11	0.11	0.07	0.03	0.01	0.01	0.00	0.00	0.00
	20	0.06	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	30	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	40	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	60	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	70	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	125	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	150	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.36	0.40	0.00	0.90	0.15
	20	0.36	0.35	0.05	0.70	0.10
	30	0.35	0.37	0.07	0.60	0.08
	40	0.36	0.35	0.15	0.60	0.07
	50	0.35	0.36	0.16	0.58	0.06
	60	0.36	0.35	0.18	0.52	0.06
	70	0.36	0.36	0.23	0.56	0.05
	80	0.36	0.36	0.24	0.51	0.04
	90	0.36	0.36	0.22	0.49	0.04
	100	0.36	0.36	0.23	0.48	0.04
	125	0.36	0.35	0.25	0.45	0.03
	150	0.36	0.36	0.25	0.45	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.43: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 10 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)

		≥ accuracy (%)								
test sample size	n	50	55	60	65	70	75	80	85	90
	10	0.70	0.70	0.58	0.47	0.22	0.22	0.06	0.06	0.01
	20	0.83	0.71	0.57	0.36	0.21	0.10	0.04	0.03	0.00
	30	0.92	0.85	0.59	0.41	0.16	0.08	0.01	0.00	0.00
	40	0.95	0.84	0.67	0.33	0.13	0.03	0.00	0.00	0.00
	50	0.97	0.90	0.68	0.36	0.11	0.03	0.00	0.00	0.00
	60	0.98	0.89	0.65	0.31	0.08	0.01	0.00	0.00	0.00
	70	0.99	0.94	0.66	0.33	0.05	0.00	0.00	0.00	0.00
	80	1.00	0.95	0.73	0.29	0.05	0.00	0.00	0.00	0.00
	90	1.00	0.98	0.72	0.32	0.03	0.00	0.00	0.00	0.00
	100	1.00	0.98	0.79	0.27	0.03	0.00	0.00	0.00	0.00
	125	1.00	0.99	0.81	0.24	0.01	0.00	0.00	0.00	0.00
150	1.00	1.00	0.85	0.22	0.01	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.63	0.60	0.10	1.00	0.15
	20	0.63	0.65	0.30	0.95	0.11
	30	0.63	0.63	0.37	0.87	0.08
	40	0.63	0.62	0.42	0.85	0.07
	50	0.63	0.62	0.44	0.86	0.06
	60	0.63	0.63	0.47	0.82	0.06
	70	0.63	0.63	0.46	0.83	0.05
	80	0.63	0.64	0.49	0.79	0.05
	90	0.63	0.63	0.51	0.74	0.04
	100	0.63	0.63	0.52	0.75	0.04
	125	0.63	0.63	0.54	0.73	0.03
	150	0.63	0.63	0.53	0.73	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.44: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 100 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)

		$\geq$ accuracy (%)								
test sample size	n	50	55	60	65	70	75	80	85	90
	10	0.51	0.51	0.20	0.16	0.03	0.03	0.00	0.00	0.00
	20	0.71	0.40	0.16	0.05	0.01	0.00	0.00	0.00	0.00
	30	0.79	0.55	0.13	0.04	0.00	0.00	0.00	0.00	0.00
	40	0.86	0.45	0.12	0.01	0.00	0.00	0.00	0.00	0.00
	50	0.90	0.59	0.10	0.01	0.00	0.00	0.00	0.00	0.00
	60	0.93	0.51	0.08	0.00	0.00	0.00	0.00	0.00	0.00
	70	0.96	0.63	0.06	0.00	0.00	0.00	0.00	0.00	0.00
	80	0.98	0.57	0.06	0.00	0.00	0.00	0.00	0.00	0.00
	90	0.97	0.64	0.04	0.00	0.00	0.00	0.00	0.00	0.00
	100	0.99	0.56	0.04	0.00	0.00	0.00	0.00	0.00	0.00
	125	1.00	0.64	0.02	0.00	0.00	0.00	0.00	0.00	0.00
150	1.00	0.71	0.01	0.00	0.00	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.56	0.60	0.20	0.80	0.10
	20	0.56	0.55	0.35	0.80	0.07
	30	0.56	0.57	0.37	0.73	0.05
	40	0.56	0.55	0.40	0.68	0.04
	50	0.56	0.56	0.44	0.66	0.04
	60	0.56	0.57	0.45	0.67	0.03
	70	0.56	0.56	0.46	0.64	0.03
	80	0.56	0.56	0.47	0.64	0.03
	90	0.56	0.56	0.46	0.64	0.03
	100	0.56	0.56	0.48	0.64	0.03
	125	0.56	0.56	0.48	0.64	0.02
	150	0.56	0.56	0.50	0.61	0.02

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.45: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 100 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)

		≥ accuracy (%)								
	<i>n</i>	50	55	60	65	70	75	80	85	90
test sample size	10	0.70	0.70	0.55	0.44	0.20	0.20	0.06	0.06	0.01
	20	0.83	0.67	0.52	0.31	0.17	0.07	0.01	0.01	0.00
	30	0.88	0.81	0.55	0.37	0.13	0.07	0.01	0.00	0.00
	40	0.93	0.78	0.61	0.27	0.10	0.03	0.00	0.00	0.00
	50	0.96	0.87	0.66	0.34	0.08	0.02	0.00	0.00	0.00
	60	0.98	0.88	0.62	0.26	0.05	0.01	0.00	0.00	0.00
	70	0.98	0.92	0.62	0.28	0.04	0.00	0.00	0.00	0.00
	80	1.00	0.94	0.68	0.24	0.04	0.00	0.00	0.00	0.00
	90	0.99	0.95	0.69	0.27	0.02	0.00	0.00	0.00	0.00
	100	1.00	0.96	0.73	0.19	0.02	0.00	0.00	0.00	0.00
	125	1.00	0.98	0.75	0.19	0.01	0.00	0.00	0.00	0.00
	150	1.00	0.99	0.79	0.16	0.01	0.00	0.00	0.00	0.00

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	<i>n</i>	mean	median	min	max	std
test sample size	10	0.63	0.60	0.20	1.00	0.14
	20	0.62	0.60	0.25	0.90	0.10
	30	0.62	0.63	0.37	0.87	0.09
	40	0.62	0.62	0.35	0.88	0.07
	50	0.62	0.62	0.44	0.82	0.06
	60	0.62	0.63	0.45	0.83	0.05
	70	0.62	0.63	0.41	0.77	0.05
	80	0.62	0.62	0.47	0.76	0.05
	90	0.62	0.62	0.48	0.76	0.04
	100	0.62	0.62	0.50	0.75	0.04
	125	0.62	0.62	0.50	0.72	0.03
	150	0.62	0.62	0.53	0.73	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.46: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 100 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)

test sample size	≥ accuracy (%)									
	<i>n</i>	50	55	60	65	70	75	80	85	90
	10	0.57	0.57	0.44	0.31	0.12	0.12	0.04	0.04	0.01
	20	0.69	0.50	0.35	0.17	0.09	0.03	0.01	0.01	0.00
	30	0.77	0.65	0.35	0.21	0.06	0.03	0.00	0.00	0.00
	40	0.82	0.61	0.38	0.12	0.03	0.01	0.00	0.00	0.00
	50	0.87	0.68	0.39	0.13	0.02	0.00	0.00	0.00	0.00
	60	0.89	0.65	0.33	0.08	0.02	0.00	0.00	0.00	0.00
	70	0.92	0.72	0.32	0.09	0.01	0.00	0.00	0.00	0.00
	80	0.94	0.65	0.29	0.04	0.00	0.00	0.00	0.00	0.00
90	0.95	0.75	0.29	0.06	0.00	0.00	0.00	0.00	0.00	
100	0.96	0.73	0.29	0.03	0.00	0.00	0.00	0.00	0.00	
125	0.99	0.84	0.28	0.01	0.00	0.00	0.00	0.00	0.00	
150	1.00	0.85	0.23	0.01	0.00	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.58	0.60	0.10	1.00	0.15
	20	0.58	0.57	0.25	0.90	0.11
	30	0.58	0.57	0.27	0.87	0.09
	40	0.58	0.57	0.35	0.80	0.07
	50	0.58	0.58	0.40	0.76	0.06
	60	0.58	0.58	0.43	0.77	0.06
	70	0.58	0.59	0.41	0.76	0.05
	80	0.58	0.57	0.44	0.71	0.05
	90	0.58	0.58	0.42	0.71	0.04
	100	0.58	0.58	0.44	0.70	0.04
	125	0.58	0.58	0.48	0.68	0.03
	150	0.58	0.58	0.48	0.69	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.47: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 1000 **features** were taken based on the ANOVA F-values. Following, a **Random Forest** was trained with default parameters. ($n_{\text{estimators}} = 100$)

		$\geq$ accuracy (%)								
test sample size	n	50	55	60	65	70	75	80	85	90
	10	0.53	0.53	0.33	0.23	0.06	0.06	0.01	0.01	0.00
	20	0.71	0.47	0.28	0.09	0.02	0.01	0.00	0.00	0.00
	30	0.73	0.56	0.23	0.09	0.01	0.00	0.00	0.00	0.00
	40	0.80	0.47	0.22	0.04	0.01	0.00	0.00	0.00	0.00
	50	0.83	0.57	0.20	0.03	0.00	0.00	0.00	0.00	0.00
	60	0.88	0.51	0.16	0.02	0.00	0.00	0.00	0.00	0.00
	70	0.90	0.61	0.13	0.01	0.00	0.00	0.00	0.00	0.00
	80	0.94	0.54	0.13	0.00	0.00	0.00	0.00	0.00	0.00
	90	0.94	0.59	0.09	0.01	0.00	0.00	0.00	0.00	0.00
	100	0.96	0.55	0.09	0.00	0.00	0.00	0.00	0.00	0.00
	125	0.98	0.60	0.07	0.00	0.00	0.00	0.00	0.00	0.00
150	0.99	0.66	0.03	0.00	0.00	0.00	0.00	0.00	0.00	

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.56	0.60	0.20	1.00	0.13
	20	0.57	0.55	0.30	0.85	0.09
	30	0.56	0.57	0.33	0.77	0.07
	40	0.56	0.55	0.40	0.78	0.06
	50	0.56	0.56	0.38	0.72	0.05
	60	0.56	0.57	0.42	0.70	0.05
	70	0.56	0.56	0.43	0.69	0.04
	80	0.56	0.56	0.44	0.68	0.04
	90	0.56	0.56	0.44	0.67	0.03
	100	0.56	0.56	0.45	0.65	0.03
	125	0.56	0.56	0.48	0.65	0.03
	150	0.56	0.56	0.49	0.63	0.02

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.48: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 1000 **features** were taken based on the ANOVA F-values. Following, an **SVM** with a **rbf kernel** was trained with default parameters. ($C = 1.0$; $\gamma = 1/n_{\text{feature}}$)

		≥ accuracy (%)								
	<i>n</i>	50	55	60	65	70	75	80	85	90
test sample size	10	0.77	0.77	0.67	0.54	0.27	0.27	0.10	0.10	0.02
	20	0.91	0.81	0.68	0.44	0.27	0.12	0.04	0.04	0.00
	30	0.93	0.87	0.67	0.50	0.21	0.12	0.03	0.01	0.00
	40	0.97	0.90	0.78	0.47	0.21	0.07	0.01	0.00	0.00
	50	0.99	0.95	0.82	0.53	0.19	0.05	0.00	0.00	0.00
	60	0.99	0.96	0.81	0.49	0.17	0.03	0.00	0.00	0.00
	70	1.00	0.98	0.83	0.56	0.18	0.04	0.00	0.00	0.00
	80	1.00	0.98	0.89	0.49	0.15	0.02	0.00	0.00	0.00
	90	1.00	1.00	0.88	0.56	0.11	0.02	0.00	0.00	0.00
	100	1.00	0.99	0.92	0.55	0.11	0.01	0.00	0.00	0.00
	125	1.00	1.00	0.96	0.55	0.09	0.00	0.00	0.00	0.00
	150	1.00	1.00	0.98	0.59	0.05	0.00	0.00	0.00	0.00

(a) Probabilities to yield an accuracy exceeding a certain threshold related to the test sample size.

	n	mean	median	min	max	std
test sample size	10	0.66	0.70	0.20	1.00	0.15
	20	0.66	0.65	0.35	0.90	0.10
	30	0.65	0.63	0.43	0.97	0.08
	40	0.66	0.65	0.45	0.90	0.07
	50	0.65	0.66	0.46	0.86	0.06
	60	0.66	0.65	0.47	0.82	0.05
	70	0.66	0.66	0.50	0.81	0.05
	80	0.66	0.65	0.49	0.80	0.05
	90	0.65	0.66	0.53	0.81	0.04
	100	0.66	0.66	0.53	0.80	0.04
	125	0.66	0.66	0.56	0.78	0.03
	150	0.66	0.66	0.57	0.74	0.03

(b) Summary of the accuracy value distributions for certain sample sizes.

Table E.49: Results as a function of variable test set sizes with a fixed classifier. To reduce the dimensionality of the feature space a **PCA** was performed and **all components** were retained. For **feature selection** an ANOVA was computed inside the the pipeline and the top 1000 **features** were taken based on the ANOVA F-values. Following, an **linear SVM** was trained with default parameters. ($C = 1.0$)

Statement of Authorship

Title of Paper	From multivariate methods to an AI ecosystem
Publication Status	☐ Published ☐ Accepted for Publication ☒ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Winter, N., Cearn, M., Clark, S., Leenings. R., Dannlowski. U., Baune, B.T., & Hahn. T. (2021). From multivariate methods to an AI ecosystem. <i>Under review.</i>

Principal Author

Name of Principal Author (Candidate)	Micah Cearn		
Contribution to the Paper	Helped formulate the idea and contributed to writing the paper.		
Overall percentage (%)	45%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	15/01/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Nils Winter		
Contribution to the Paper	Helped formulate the idea and contributed to writing the paper. Created figure 1.		
Signature		Date	15/01/2021
Name of Co-Author	Scott Clark		

Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision.		
Signature		Date	15/01/2021
Name of Co-Author	Ramona Leenings		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision.		
Signature		Date	15/01/2021
Name of Co-Author	Udo Dannlowski		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision		
Signature		Date	15/01/2021
Name of Co-Author	Bernhard Baune		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision.		
Signature		Date	15/01/2021
Name of Co-Author	Tim Hahn		
Contribution to the Paper	Idea generation, critical revision of the manuscript, overall guidance and supervision. Communications with journal editors..		
Signature		Date	15/01/2021

Please cut and paste additional co-author panels here as required.

Chapter 5

From Multivariate Methods to an AI Ecosystem

Nils R. Winter^{1*}, Micah Cearns^{2,3*}, Scott R Clark², Ramona Leenings¹, Udo Dannlowski¹,
Bernhard T. Baune^{1,2,3,4}, Tim Hahn¹

¹Department of Psychiatry, University of Muenster, Germany

²Discipline of Psychiatry, School of Medicine, University of Adelaide, Australia

³Department of Psychiatry, Melbourne Medical School, The University of Melbourne, Parkville, Australia

⁴The Florey Institute of Mental Health, The University of Melbourne, Melbourne, VIC Australia

*these authors contributed equally to this work

Molecular Psychiatry (Under review)

A decade ago, at a major international conference, we vividly remember a symposium on psychiatric Artificial Intelligence (AI) drawing a crowd of seven people – including the four speakers. Today one might get the impression that every other funding proposal is required to include at least some degree of AI-based analyses. At a time when 17 of the 29 hot topics listed by the most recent *Gartner Hype Cycle* for emerging technologies¹ – an indicator of perceived innovation – are either AI technologies (e.g. Generative Adversarial Networks) or include AI as a core component (e.g. autonomous driving), there is no shortage of promises. In both psychiatry and medicine in general, expectations to move beyond classical group-level statistical and enter the promising future of personalized medicine are high. Although AI has not yet fully hit mainstream psychiatric research, the availability and advancement of technology and methods have indeed led to a growing adoption of AI methods and agreement on best practice^{2–4}. Despite this progress and some promising first applications (e.g. in suicide prediction⁵), translation to clinical practice has been hampered by a large degree of estimate variability and diagnostic heterogeneity.

In the following, we will argue that the current drawbacks in psychiatry arise not primarily from a lack of methodological advancement and genuine clinical potential for AI in psychiatry, but from fundamental issues pertaining to sample size, model construction, evaluation practice, and the conceptualization of mental disorders. To overcome these challenges, we outline concrete steps towards an AI ecosystem that promote a coordinated collaboration within the field to pave the way for better translation of AI solutions into clinical practice. Built upon four supporting pillars; collection, construction, evaluation and translation, this ecosystem provides a framework for data collection, harmonization and sharing, guidelines for model construction, evaluation and distribution, methods for parsing heterogeneity as well as ethical and transparency standards.

Explaining the Translational Roadblock

In the social and medical sciences, in light of the recent replication crisis, a failure to replicate results has rendered many discoveries from group-based statistical analyses void⁶. While machine learning's goal of optimizing model performance for previously unseen data integrates replication directly into model development, which in theory, should render it more robust to replication failure, analyses drawing on larger samples has revealed substantial overestimation of model performance for many landmark studies. For example, at a recent international machine learning competition, participants sought to classify major depressive disorder (MDD) patients from healthy controls (HC) using structural Magnetic Resonance

Imaging data (N participants = 2,240)⁷. In contrast to numerous smaller studies showing accuracies of more than 80 - 90%, the winning entry ceiled at only 65% accuracy. This effect was demonstrated again on a large structural imaging sample from the UK biobank (N > 3,400), finding accuracies of only 50 – 60% across classification targets⁸.

These findings lead to two important questions. Firstly, where in the analysis pipeline are we most susceptible to overestimating model performance in small sample size studies?

Secondly, are these overestimations and subsequent replication failures distinct to structural imaging studies or should we expect to see this same phenomenon also with other data modalities? A recent study empirically addressed both points⁷. First, the authors demonstrate that predictive structural imaging studies of MDD are prone to systematically overestimating model performance on smaller samples. Second, this effect arises not primarily from small training samples, but is due to too small test sets. Third, this statistical effect is generalizable and contingent on the cross-validation cycle itself rather than the phenotype or data modality used for model training. Thus, the results are likely to generalize across all areas of psychiatric research (from models using imaging, to proteomic, to routine clinical data).

In addition to these statistical properties, the translational abilities of any AI model are dependent upon the clinical utility and construct validity of a diagnostic category, its related illness trajectories, and the degree of heterogeneity contained within each. This latter point partially explains why smaller more homogeneous test samples overestimate the true predictive ability of a model compared to accuracy estimates attained on larger more heterogeneous samples reflective of the true patient population⁹. Therefore, it is likely that the underlying causes of therapeutic response and illness trajectories vary qualitatively and quantitatively between different, more homogeneous sub-samples of these categories. Outside of applied AI, this is already recognized as a major obstacle to translatable clinical research and has sparked many theoretical considerations such as the *Research Domain Criteria*, the *Roadmap for Mental Health Research in Europe*, or the *Hierarchical Taxonomy of Psychopathology*. These problems cannot be eliminated easily but require novel approaches to the classification and stratification of psychiatric disorders, including e.g. normative modelling approaches¹⁰ (see *Construction* below).

Establishing an AI Ecosystem

Collection

While often dismissed as too costly and impractical, feasibility of large-scale data acquisition heavily depends on data modality and existing infrastructure. For example, systematically making routine patient data and electronic health records available will enable dramatic increases in sample size and coverage of potential patient subpopulations, thereby outweighing the often lower quality of this data. In the same vein, building on high-throughput omics platforms or mobile sensing technology (e.g. from smartphones and watches) will allow us to cost-efficiently gather ecologically valid physiological and phenotype data in real-time. We have already seen the great efficacy of data sharing within large consortia in genetics (e.g. PGC) and neuroimaging (e.g. ENIGMA). Similar initiatives and protocols have not yet been established for EHR and eHealth data but will be integral to an AI ecosystem. In cases where data remains limited due to scarcity of the outcome or patient group (e.g. patient subgroups with certain comorbidities), harmonizing data acquisition protocols will be key to training machine learning models based on data from multiple sources – especially with regard to the reduction of epistemic uncertainty. Wherever acquisition protocols differ, we must ensure the development of methods for harmonizing the data itself. However, as data collection is increasingly organized through large initiatives such as the UK Biobank, we must not reduce our efforts in establishing more and secure infrastructure to upload and share data from small studies. This is important since smaller studies often investigate important mental disorders, comorbidities with low incidence, or causal mechanisms that we are less likely to capture in large-scale consortia. Studies by Chekroud¹¹ or Miotto¹² show the great potential large but clinically relevant samples can have but as of now, those are the exception rather than the rule, and similar efforts are needed across all areas of psychiatry.

Construction

In addition to collecting more data, the issue of sample size can also be dealt with at the model construction stage. This is especially relevant when dealing with highly specific questions (e.g. a distinct type of psychotherapeutic intervention) or low-incidence disorders where we cannot collect large amounts of data. Methods for dealing with small training samples have been discussed in previous works⁴. To summarize, one might try to generate synthetic training data through data augmentation or leverage larger training samples from another task through transfer learning. Additionally, formally incorporating (causal) domain knowledge gained from basic research directly (e.g. selecting features that have shown to be relevant for a specific task across studies through meta-analytic approaches¹³) into machine

learning models also holds great promise to dramatically decrease the number of training samples required¹⁴. In addition, federated or collaborative learning allows us to train an algorithm across multiple sites by passing a model trained on one site to the next for further training without the need to exchange sample data, thereby alleviating data protection issues whilst maximizing training sample size. This approach is especially appealing at a time when the importance of protecting sensitive data is increasingly recognized. Easy-to-use platforms like Coinstac [www.coinstac.org] or DataSHIELD [www.datashield.ac.uk] that allow federated learning have started to emerge and will integrated nicely into an AI ecosystem.

Also crucial for an AI ecosystem are ways in which we can parse population heterogeneity. From a methodological perspective, we can differentiate two approaches. First, traditional clustering methods can reveal more homogeneous patient subgroups. Depending on the clustering approach and data used, subjects in these groups might share specific symptoms, a common genotype, or a similar neurobiology that is clinically informative yet evades symptom-based diagnostic categories. At the symptom level, some studies already suggest that symptom clustering might enhance the prediction of antidepressant treatments¹⁵. While easily lending itself to a transdiagnostic view, jointly modelling biological and deep phenotype data with Multi-view methods such as Canonical Correlation Analysis or Partial Least Squares Regression show promise for uncovering transdiagnostic modes of covariation between biological data and behavioural phenotypes. Embedding these low-dimensional representations within supervised prediction frameworks may assist in overcoming population heterogeneity and help decrease our dependency on large training and testing samples.

Second, normative modelling¹⁰ quantifies the degree to which individuals deviate from the expected population distribution, thus, allowing symptoms in individual patients to be recognized as extreme values within a distribution. Unlike clustering approaches, normative modelling does not require a consistent pattern of deviation and can therefore accommodate for the heterogeneity across patients. While normative modelling provides a highly effective and principled approach to the quantification of deviation, a direct translation to clinically useful machine learning models must still be developed.

Finally, clinically useful predictive models need to provide uncertainty estimations for their predictions. We not only need to know how accurate a model is, but how confident it is in its prediction, and whether or not this confidence aligns with the true probability of an event occurring. For example, imagine a model that is highly accurate and deployed in a clinic. It predicts that a high-risk patient will not respond to a first line treatment, but instead, to a first-

generation treatment with a high side-effect profile. In this situation, a clinician needs to make a nuanced decision that weighs up the associated risks against the potential benefits of the treatment. If a model turns out to be only 51% confident in its prediction, even though on average it is 80% accurate, the weighting that a clinician should give to the model should not be high. However, many models currently used in psychiatry (Support vector machines, random forests etc) are poorly calibrated by virtue of their underlying mathematics, meaning that they may be highly confident in their decision when the true probability of an event occurring in the population is actually not much better than chance. This is a seldom discussed issue in computational psychiatry with far reaching implications for clinical decision making¹⁶.

Evaluation

As one of the fundamental problems of overestimating model performance has its roots primarily in the evaluation framework itself, a concise AI ecosystem needs to provide guidelines and an independent infrastructure to adequately evaluate model performance. While the Focus Group on artificial intelligence for health (FG-AI4H) has begun to outline a standardized assessment framework for the evaluation of AI models in partnership with the World Health Organisation (WHO), simple online model-sharing platforms already enable the upload of machine learning models for evaluation by others (e.g. the PHOTON AI model repository; www.photon-ai.com/repo). Extending such initiatives into largely automatized AI assessment infrastructure would allow for highly efficient and transparent model evaluation at scale. To foster transparency, journals and funding agencies ought to ensure that all code is available on request or provided in an online repository upon submission. Finally, the increasing availability of easy-to-use, open-source machine learning frameworks (e.g. the mlr3 R package [mlr3.mlr-org.com], the PHOTON AI framework [www.photon-ai.com] or tidymodels [www.tidymodels.org/]) will continue to simplify robust and replicable machine learning analysis.

In addition to performance-based evaluation (e.g., model accuracy), we also need to assess the clinical utility of a model through practical questions. For example, can we easily use the model without data protection issues? Does the model provide uncertainty estimates and calibrated predictions? Is the model clinically useful, i.e. does it provide us with information we cannot otherwise obtain and/or outperform clinicians in the field? Does the model provide information that effectively supports and augments the clinician? For example, will a costly case/control biomarker model trained on **diagnostic** labels that are already provided by

clinicians actually tell them anything that improves their clinical decision-making process? In situations where biomarkers are of **prognostic/predictive** value, above and beyond that of cheap and easily accessible routine clinical data, is the increase in prognostic certainty justified by the increased health economic costs and invasiveness to obtain these markers at the individual patient level? Overall, questions and measures of clinical utility should be asked and assessed in the same way that we examine a model's predictive performance. Incorporating clinical utility-based evaluation into existing guidelines will be fundamental to progress from proof-of-concept studies to clinical translation.

Translation

Moving predictive models from bench to bedside can be accomplished in a number of ways. Online risk calculators are simple and cost-effective ways of supporting patient care and have already been used for the detection of individuals developing psychosis in secondary mental health care¹⁷. More complex predictive models can be provided through the same infrastructure that is already used for model evaluation. Patients' data could be uploaded to these online models which in turn provide individual predictions. Whenever data privacy is a concern, those models need to also be made available offline. Ultimately, predictive models should be incorporated into AI-based clinical decision support systems (CDSS) for integration into a clinician's workflow. If designed properly, these AI-based CDSS could also facilitate continuous evaluation through feedback loops whilst also improving model performance over time. In addition, providing predictive algorithms with **individual** patient's long-term outcomes and symptom developments would signify a major step towards parsing diagnostic and population heterogeneity. Importantly, as the role of AI in medical care expands, we need to establish ethical standards that balance patient autonomy and foster clinician and patient trust.

Conclusion

Against the background of fundamental issues regarding sample size, model construction, evaluation practice, and the conceptualization of mental disorders, the clinical success of AI in psychiatry will depend on the timely creation of an AI ecosystem that addresses these foundational issues hampering the validity and reliability of the current literature. For the field to emerge from the hype, through the trough of disillusionment and into a plateau of productivity, we propose four foundational pillars to bolster the field. First, building upon existing data collection initiatives and consortia, we need to address the issue of small training

samples by incentivising researchers to harmonize protocols and data acquisition, not only making sure to collect bigger data, but more clinically relevant data for model training. Second, at the model construction phase we need to take advantage of methods such as transfer learning and data augmentation that mitigate some of the detrimental effects of small sample sizes. A simple, safe, and open-sourced federated learning infrastructure would instantly help overcome this problem as it would allow for the training AI models without the need to share data. In addition, transdiagnostic clustering and normative modelling approaches may provide a framework to address the long-standing issue of diagnostic heterogeneity. Further, using algorithms that are well calibrated and thus able to provide estimates of uncertainty are of crucial importance if to ever be deployed clinically. Third, we must ensure reproducibility of machine learning analyses through the use of code sharing and model evaluation infrastructure. Extending existing model repositories would enable large-scale, independent model evaluation and standardized guidelines for rating the clinical utility of AI solutions. This will help ensure their alignment with the needs of patients and clinicians outside of the academic community. Fourth, integrating models into AI-based CDSS will help transform the way we treat and diagnose patients. Ensuring data privacy and patient autonomy at this level is of utmost importance to generate trust towards this new technology. Importantly, in all areas, leading consortia and machine learning groups show that many of the solutions already exist and are actively developed. Building an AI ecosystem based on these advances has grown from a daunting challenge into a unique opportunity for psychiatry in the decade to come.

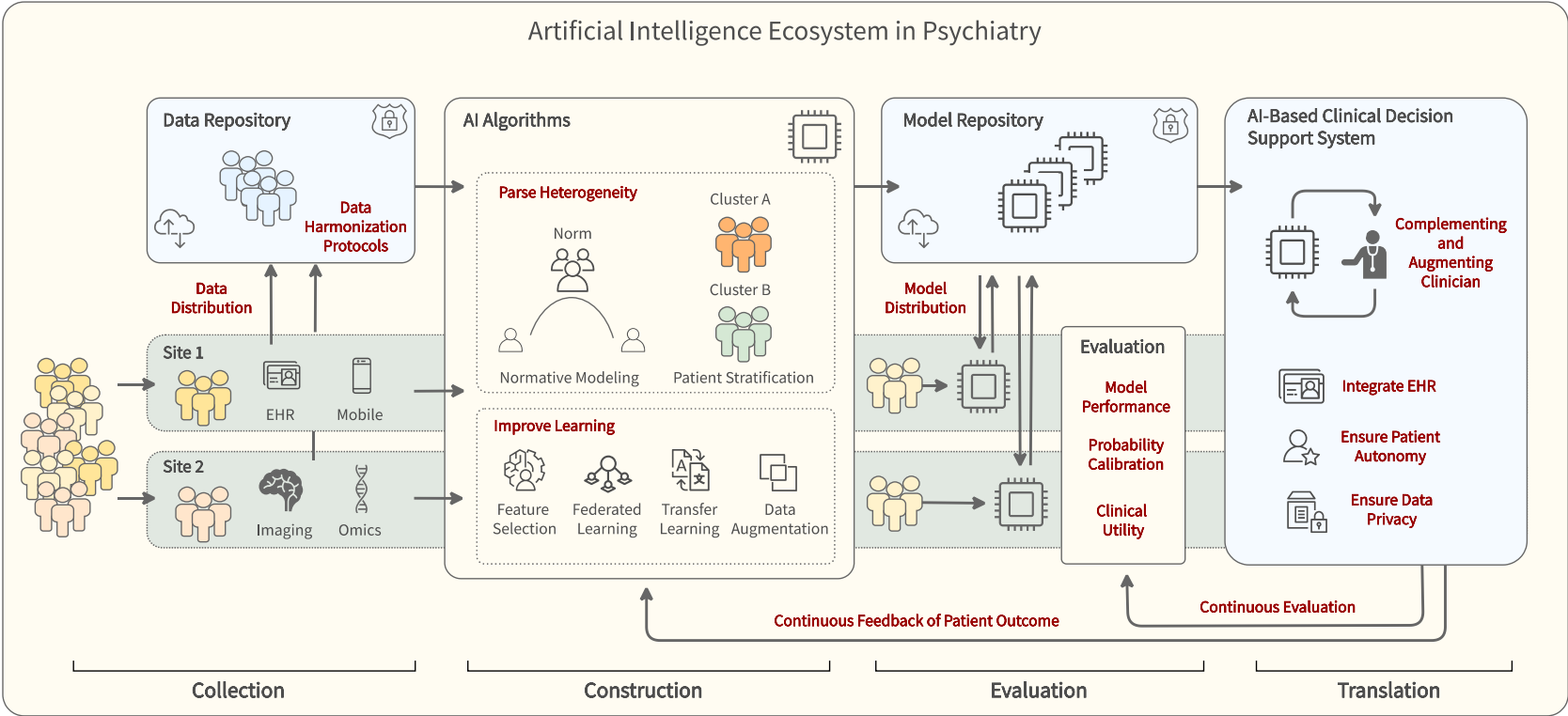
Acknowledgement

This work was funded by the German Research Foundation (DFG grants HA7070/2-2, HA7070/3, HA7070/4 to TH) and the Interdisciplinary Center for Clinical Research (IZKF) of the medical faculty of Münster (grants Dan3/012/17 to UD and MzH 3/020/20 to TH).

Conflicts of Interest

The authors declare no conflicts of interest.

Figure 1. Schematic for an artificial intelligence ecosystem in psychiatry.



References

1. Panetta, K. The Gartner Hype Cycle highlights the 29 emerging technologies CIOs should experiment with over the next year. (2019). Available at: <https://www.gartner.com/smarterwithgartner/5-trends-appear-on-the-gartner-hype-cycle-for-emerging-technologies-2019>.
2. Hahn, T., Nierenberg, A. A. & Whitfield-Gabrieli, S. Predictive analytics in mental health: applications, guidelines, challenges and perspectives. *Mol. Psychiatry* **22**, 37–43 (2017).
3. Poldrack, R. A., Huckins, G. & Varoquaux, G. Establishment of Best Practices for Evidence for Prediction: A Review. *JAMA Psychiatry* 1–7 (2019). doi:10.1001/jamapsychiatry.2019.3671
4. Cearns, M., Hahn, T. & Baune, B. T. Recommendations and future directions for supervised machine learning in psychiatry. *Transl. Psychiatry* **9**, (2019).
5. Kessler, R. C., Bossarte, R. M., Luedtke, A., Zaslavsky, A. M. & Zubizarreta, J. R. Suicide prediction models: a critical review of recent research with recommendations for the way forward. *Mol. Psychiatry* **25**, 168–179 (2020).
6. Stupples, A., Singerman, D. & Celi, L. A. The reproducibility crisis in the age of digital medicine. *npj Digit. Med.* **2**, 2 (2019).
7. Flint, C. *et al.* Systematic Overestimation of Machine Learning Accuracies in Structural Neuroimaging Studies of Depression. *arXiv Prepr. arXiv* (2019).
8. Stolicyn, A. *et al.* Automated classification of depression from structural brain measures across two independent community-based cohorts. *Hum. Brain Mapp.* 1–16 (2020). doi:10.1002/hbm.25095
9. Schnack, H. G. & Kahn, R. S. Detecting neuroimaging biomarkers for psychiatric disorders: Sample size matters. *Front. Psychiatry* **7**, (2016).
10. Marquand, A. F. *et al.* Conceptualizing mental disorders as deviations from normative functioning. *Mol. Psychiatry* **24**, 1415–1424 (2019).
11. Chekroud, A. M. *et al.* Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet Psychiatry* **3**, 243–250 (2016).

12. Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci. Rep.* **6**, 1–10 (2016).
13. Karrer, T. M. *et al.* Brain-based ranking of cognitive domains to predict schizophrenia. *Hum. Brain Mapp.* **40**, 4487–4507 (2019).
14. Schölkopf, B. Causality for Machine Learning. *arXiv* (2019).
15. Chekroud, A. M. *et al.* Reevaluating the efficacy and predictability of antidepressant treatments: A symptom clustering approach. *JAMA Psychiatry* **74**, 370–378 (2017).
16. Cearns, M., Hahn, T., Clark, S. & Baune, B. T. Machine learning probability calibration for high-risk clinical decision-making. *Aust. N. Z. J. Psychiatry* **54**, 123–126 (2020).
17. Fusar-Poli, P. *et al.* Development and validation of a clinically based risk calculator for the transdiagnostic prediction of psychosis. *JAMA Psychiatry* **74**, 493–500 (2017).

Statement of Authorship

Title of Paper	Predicting rehospitalization within 2 years of initial patient admission for a major depressive episode: a multimodal machine learning approach
Publication Status	☒ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Cearns, M., Opel, N., Clark, S., Kaehler, C., Thalamuthu, A., Heindel, W., ... & Baune, B. T. (2019). Predicting rehospitalization within 2 years of initial patient admission for a major depressive episode: a multimodal machine learning approach. <i>Translational psychiatry</i> , 9(1), 1-9.

Principal Author

Name of Principal Author (Candidate)	Micah Cearns
Contribution to the Paper	Formulated the idea for the paper alongside Nils Opel, wrote the manuscript, ran the analysis, and created all figures.
Overall percentage (%)	85%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 15/01/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- the candidate's stated contribution to the publication is accurate (as detailed above);
- permission is granted for the candidate to include the publication in the thesis; and
- the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Nils Opel
Contribution to the Paper	Contributed to study design and revisions. Assisted in the collection and processing of imaging and genetic data.
Signature	Date 15/01/2021
Name of Co-Author	Class Kaehler

Contribution to the Paper	Code reviews.		
Signature		Date	15/01/2021
Name of Co-Author	Scott Clark		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision.		
Signature		Date	15/01/2021
Name of Co-Author	Anbupalam Thalamuthu		
Contribution to the Paper	Contributed to the analysis of genetic data.		
Signature		Date	15/01/2021
Name of Co-Author	Walter Heindel		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision.		
Signature		Date	15/01/2021
Name of Co-Author	Henning Teismann		
Contribution to the Paper	Contributions to BiDirect data collection and processing.		
Signature		Date	15/01/2021
Name of Co-Author	Heike Minnerup		
Contribution to the Paper	Critical revision of the manuscript, contributed to study design.		
Signature		Date	15/01/2021
Name of Co-Author	Udo Dannlowski		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision. Supervision of imaging analyses.		

Signature		Date	15/01/2021
Name of Co-Author	Klaus Berger		
Contribution to the Paper	PI of the BiDirect-Study, supervised data collection and management, critical revision of the manuscript, overall guidance and supervision.		
Signature		Date	15/01/2021
Name of Co-Author	Theresa Winter		
Contribution to the Paper	Coordination of laboratory environment and analysis and preparation of serum biomarkers.		
Signature		Date	15/01/2021
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision. As corresponding author, he made communications with journal editors.		
Signature		Date	15/01/2021

Please cut and paste additional co-author panels here as required.

Chapter 6

Predicting rehospitalization within 2 years of initial patient admission for a major depressive episode: A multimodal machine learning approach

Micah Cearns, BPsy (Hons)^{1 *}, Nils Opel, MD^{2,5 *}, Scott Clark, PhD¹, Claas Kahler, BComSci², Anbupalam Thalamuthu, PhD⁶, Walter Heindel, MD ³, Theresa Winter, PhD^{8,9}, Henning Teismann, PhD ⁴, Heike Minnerup, MD⁴, Udo Dannlowski, PhD², Klaus Berger, MD⁴ and Bernhard T. Baune, PhD^{1,7,10}

¹ Discipline of Psychiatry, School of Medicine, University of Adelaide, Australia

² Department of Psychiatry, University of Münster, Germany

³ Institute of Clinical Radiology, University of Münster, Germany

⁴ Institute of Epidemiology and Social Medicine, University of Münster, Germany

⁵ Interdisciplinary Centre for Clinical Research (IZKF), University of Münster, Germany

⁶ Centre for Healthy Brain Ageing, UNSW, Sydney

⁷ Department of Psychiatry, Melbourne Medical School, The University of Melbourne, Australia

⁸ Institute of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Germany

⁹ Integrated Research Biobank, University Medicine Greifswald, Germany

¹⁰ The Florey Institute of Neuroscience and Mental Health, The University of Melbourne

* Equal contribution: MC and NO contributed equally to the present work

Translational Psychiatry

Abstract

Machine learning methods show promise to translate univariate biomarker findings into clinically useful multivariate decision support systems. At current, works in major depressive disorder have predominantly focused on neuroimaging and clinical predictor modalities, with genetic, blood-biomarker, and cardiovascular modalities lacking. In addition, the prediction of rehospitalization after an initial inpatient major depressive episode is yet to be explored, despite its clinical importance. To address this gap in the literature, we have used baseline clinical, structural imaging, blood-biomarker, genetic (polygenic risk scores), and electrocardiography predictors to predict rehospitalization within 2 years of an initial inpatient episode of major depression. Three hundred and eighty patients from the ongoing 12-year Bidirect study were included in the analysis (Rehospitalized: yes = 102, no = 278). Inclusion criteria was age ≥ 35 and < 66 years, a current or recent hospitalisation for a major depressive episode and complete structural imaging and genetic data. Optimal performance was achieved with a multimodal panel containing structural imaging, blood-biomarker, clinical, medication type and sleep quality predictors, attaining a test AUC of 67.74 ($p = 9.99 \times 10^{-5}$). This multimodal solution outperformed models based on clinical variables alone, combined biomarkers and individual data modality prognostication for rehospitalization prediction. This finding points to the potential of predictive models that combine multimodal clinical and biomarker data in the development of clinical decision support systems.

Introduction

Relapse rates in specialized mental healthcare settings are high in patients with major depressive disorder (MDD) (60% after 5 years, 67% after 10, and 85% after 15)⁽¹⁾. To better predict relapse, previous research has predominantly focused on the presence and magnitude of clinical symptoms, including residual depressive symptoms⁽²⁾, illness severity⁽³⁾, number of prior episodes⁽⁴⁾, age of onset⁽⁴⁾, and comorbid personality disorders⁽⁵⁾. Other studies have explored group level associations between biomarkers and relapse, observing smaller hippocampal volumes⁽⁶⁾, higher levels of post-treatment glucocorticoids⁽⁷⁾, high cortisol response on the combined dexamethasone-CRH test⁽⁸⁾, as well as catecholamine and tryptophan depletion⁽⁹⁾. Such findings have been beneficial in constructing modality specific etiological hypotheses as well as signposts for relapse in clinical practice. However, the elucidation of clinically meaningful predictors of relapse is contingent on the construct validity of the prediction outcome and the size of the analyzed sample. Some of these studies identifying biomarkers as predictors of relapse have used self-reported relapse into a new depressive episode and samples of less than 50 patients^(7, 10). Whether patients have truly relapsed into a new depressive episode or merely have a treatment refractory illness is unknown. These observations may help to explain the lack of empirically validated relapse prediction tools and subsequent intervention strategies for relapse prevention⁽¹⁰⁾.

To overcome this blind spot in patient care, studies of increased size that focus on a well-defined outcome such as rehospitalization are needed to identify clinically robust predictors of this illness trajectory. To facilitate this, a longitudinal, multimodal sample of clinically diagnosed inpatients is required. The BiDirect depression cohort provides such a sample⁽¹¹⁾. In addition, multimodal, multivariate modelling techniques that prognosticate individually, rather than at the group level are needed given the heterogenous nature of MDD and its illness trajectories^(12, 13).

Recent work has shown that machine learning (ML) models are well suited to problems of this nature, demonstrating their efficacy in drug response and functional outcome prediction in MDD^(14, 15), whilst other studies using neuroimaging modalities have shown similar success for ML based MDD diagnostics⁽¹⁶⁾. Interestingly, no studies have applied ML models to relapse prediction, thus, the utility of differing data modalities for relapse prediction remains poorly understood. Regarding illness trajectory modelling, one previous work from Schmaal et al (2015)⁽¹⁷⁾ combined neuroimaging and clinical data to predict MDD remission

trajectories with moderate success (accuracy = 69 – 73%). Further, Koutsouleris et al (2018)⁽¹⁵⁾ used both neuroimaging and clinical data to predict functional outcomes in recent onset MDD (balanced accuracy = 70.3%). However, studies employing a combination of neuroimaging markers with biomarkers from different modalities as well as clinical data are lacking up until now.

Given the apparent gap in multimodal studies of outcome prediction in MDD, we have combined a range of clinical and biomarker predictors that have shown significant associations to MDD in previous works but are yet to be used for ML based rehospitalization trajectory modelling. Modalities used included clinical⁽¹⁸⁾, blood biomarker⁽¹⁹⁻²¹⁾, structural imaging⁽¹⁶⁾, electrocardiography⁽²²⁾, genetic⁽²³⁾, cognitive⁽²⁴⁾, nutritional⁽²⁵⁾, sleep⁽²⁶⁾, and exercise⁽²⁷⁾. Using these modalities, we predicted rehospitalization in a cohort of patients within 2 years of initial hospitalization for an acute episode of MDD.

Materials and Methods

Dataset description

The BiDirect study is an ongoing study of a) patients, hospitalized for an acute episode of major depression at time of recruitment, b) population controls randomly drawn from the register of the city of Münster⁽¹¹⁾ and c) patients 3 months after an acute coronary event or myocardial infarction. Examination of all participants included a computer-assisted face-to-face interview of socio-demographic characteristics and medical history as well as an extensive psychiatric assessment (Supplementary Information 1.1). Only patients in the depression cohort were used in this analysis.

At baseline, 999 MDD patients were recruited and 684 completed their 2-year follow-up and provided their rehospitalization status. Specifically, patients were asked, have you had at least one or more re-admissions to hospital for an acute depressive episode since the initial examination? Patient response was recorded by study assistants. As our aim was to assess both multi and unimodal prediction models of reported rehospitalization, a requirement for inclusion was complete imaging and genetic data. Twenty-nine patients had incomplete genetic data whilst 294 had incomplete imaging data. In addition, 14 participants were excluded due to poor MRI quality, leaving a final sample of 380 participants (Rehospitalized: yes = 102, no = 278). See table 1 for sociodemographic characteristics. In this final sample, 87.9% (334/380) of patients were taking some form of antidepressant medication at their

baseline assessment, 40.3% (153/380) were taking an antipsychotic, whilst 92.6% (352/380) were taking some form of psychotropic medication (see table 2).

Predictor Modalities

Clinical: Detailed information regarding socio-demographic and socio-economic status, lifetime medical diagnoses, current medication use, health care utilization, insurance status, lifestyle and risk behaviour (e.g., diet, physical activity, alcohol consumption, smoking status), and perceived health state was collected via a computer assisted interview. A combination of individual items as well as total scores were included from the Hamilton Depression Rating Scale (HAM-D), the Hamilton Anxiety Rating Scale (HAM-A), the Centre for Epidemiologic Studies Depression Scale (CES-D), the Inventory of Depressive Symptomatology (IDS), the International Physical Activity questionnaire (IPAQ), and the Food Frequency Questionnaire (FFQ) (Supplementary Information 1.1 and 1.2). In addition, we included measures from a cognitive functioning module and several self-report measures (Supplementary Information 1.3). In total, 208 clinical and demographic predictors were included.

Structural imaging: We included imaging data derived from structural magnetic imaging sequences (3D-T1). To reduce the size of the predictor space, we a priori selected 15 regions that have been shown to be significantly associated with MDD in previous ENIGMA meta-analyses^(6, 28). Selected regions included right and left mean hippocampal volume, cortical thickness of the bilateral medial orbitofrontal cortex (OFC), fusiform gyrus, insula, rostral anterior and posterior cingulate cortex and unilaterally in the left middle temporal gyrus, right inferior temporal gyrus and the right caudal anterior cingulate cortex (Supplementary Information 1.4).

Serum and genetic markers: Our serum biomarker panel consisted of 10 measures of high sensitive C-reactive protein⁽¹⁹⁾, free triiodothyronine, thyroxine, thyroid-stimulating hormone⁽²⁹⁻³¹⁾, 17 beta-estradiol^(32, 33), sex hormone-binding globulin, testosterone, the free androgen index⁽³⁴⁻³⁶⁾, total cholesterol⁽³⁷⁾, and high⁽³⁸⁾ density lipoprotein cholesterol. Due to the correlations between genetic variants shared across psychiatric traits and common comorbidities between psychiatric disorders^(39, 40), we also included 7 polygenic risk scores (PGRS) with p-value a threshold of 0.5, for MDD, anxiety, alzheimer's, anorexia, autism spectrum disorder, bipolar, and schizophrenia (Supplementary Information 1.5 - 1.6 and

Supplementary Information 2.1 – 2.6). In addition, we also assessed PGRS with p-value thresholds of 0.05 and 0.01 (Supplementary Information 3.2).

Cardiovascular: We determined cardiovascular and general health status through the assessment of different cardiovascular markers. First, we measured weight (without shoes and heavier clothes), height, and waist circumference. Following, we used bioelectrical impedance measurements (Body Impedance Analyzer BIA 2000-S, Data Input GmbH) including the determination of body fat and water, extracellular mass, body cell mass, and basic metabolic rate to assess general markers of body composition. In addition, a measurement of a standard 3-channel electrocardiogram (ECG) was performed (Supplementary Information 1.7). From these assessments we included the following 7 predictors: heart rate (beats per minute), body mass index (BMI), extracellular mass to body cell mass (ECM/BCM ratio), basic metabolic rate, corrected body fat (kg), total body water (kg), and lean body mass (kg). In total, 247 predictors were included in the analysis (see supplementary table 1 for all predictors).

Machine Learning Pipeline

For our first set of models, we entered all predictor modalities into our pipeline for consideration. To ensure the unbiased approximation of the model's generalisability to new patients, we trained and tested all models using repeated nested cross-validation in a pipeline to prevent information leaking between patients used for training and validation. In the inner cross validation loop, we conducted imputation, standardization, feature selection, hyperparameter optimization, and the fitting of a linear support vector machine (SVM).

To begin, we imputed predictors using multivariate imputation of chained equations with the 10 nearest predictors used in the imputation process. No predictors entered into the pipeline had more than 20% missing data. Following, all variables were scaled to have a mean of zero and a standard deviation of one. Next, we used the elastic net, a form of penalized logistic regression to select a final subset of variables for prediction. This approach shrinks the coefficients of highly correlated predictors towards each other whilst removing irrelevant predictors from the model^(41, 42). This process was completed simultaneously with an exhaustive grid search to tune model hyperparameters. For the elastic net, the parameters alpha (the amount of penalization) and the l1 ratio (mixing parameter between the l1 and l2 norms) were tuned for predictor selection. For the l1 ratio parameter we searched the values $\lambda \in \{0.1, 0.5, 0.7, 0.9, 0.95, 0.99\}$, with values closer to 1 representing the l1 norm. For

alpha, we searched $a \in \{0.1, 0.2 \dots, 1.0\}$. For the SVM, we tuned four values of the regularization parameter $C \in \{0.001, 0.01, 0.1, 1.0\}$. To accommodate for the class imbalance in the outcome, each C value was weighted by the inverse percentage proportion of each class label (rehospitalized: yes = 102, no = 278). This approach increases the penalty for misclassifying the minority class (rehospitalized) given its relative scarcity. Finally, we used Platt scaling to calibrate the probability estimates for the SVMs binary predictions^(43, 44). The set of predictors and hyperparameters that maximized area under the curve on the receiver operator characteristic (AUC) were selected in the pipeline. See supplementary information 1.8 to 1.8.7 for further details on the pipeline and supplementary information 1.8.5 for the selected hyperparameter values for the elastic net and SVM classifier.

All steps were completed in an inner cross-validation loop with 5 repeats of 10-fold cross validation. This method divides the sample into 10 separate subsets, uses 9 for training, and then makes predictions on the final set. To avoid favourable splits in the data, this process is repeated 5 times, initializing splits uniquely for each repeat. For the testing of our final models, we used 10-fold cross validation in the outer cross validation loop, averaging model performance metrics across test folds. To assess the statistical significance of our final best performing model, we used a permutation test ($m = 10,000$). (see Figure 1).

To assess the predictive capacity of each modality, we trained, clinical, sMRI, cardiovascular, and PGRS only models using the same pipeline as above, as well as a biomarker only model that aggregated together all biomarker modalities. To assess for a significant omnibus effect between models, we used a Kruskal-Wallis H-test. Following, we used Mann-Whitney rank tests to assess for post hoc differences between our multi and unimodal models. All p-values were FDR corrected using the Benjamini and Hochberg method. All ML models were developed using Scikit-learn in Python 3.6.5. All code for analysis is available on request.

Open sourcing of model

To provide transparency, the use of our model to other research groups, and encourage further external validation of our multimodal model, we uploaded our trained model to the Photon AI online model repository (<https://www.photon-ai.com/repo>). This repository allows our multimodal model to be downloaded and tested by other research groups.

Analyses of multimodal predictors

To analyse the direction, magnitude and significance of the models selected predictors after controlling for covariates, we used a non-penalized implementation of logistic regression in Python's statsmodels package. In addition, we included the average SVM weight values from the outer 10-fold cross-validation loop from our ML pipeline (Supplementary Information 3.1 and Supplementary Table 2). Further, we conducted ordinary least squares regression sub-analyses of variables including age, gender, BMI, overall psychotropic medication load, and individual psychotropic medication use on the biomarkers that were selected in the final multimodal model. This was done to elucidate an understanding of variables that may be associated with these biomarkers and help explain their inclusion in the multimodal model. In addition, we conducted an exploratory analysis of patient characteristics for those taking a specific type of medication selected in the multimodal model. This was done to determine: a) whether this medication proxied for the treatment or prevention of a specific comorbid illness, b) was taken as a polytherapeutic treatment strategy for MDD, or c) was acting as a proxy for the presence of unmedicated patients that were relapsing at higher rates than the rest of the sample. For these additional analyses and results, see Supplementary Information 3.3 to 3.5.

Results

Multimodal model

Within two years of hospitalization for an acute episode of MDD 102 patients (26.8%) were rehospitalized for a depressive episode, whilst 278 (73.2%) were not. Our best performing model was our multimodal model (test AUC = 67.74). For our multimodal solution, 10 predictors were selected for optimal predictions by the elastic net. Those with a positive association with rehospitalization included the number of previous inpatient depressive episodes, individual CES-D items 5 (Last week I had trouble concentrating) and 3 (In the last week I could not get rid of my mood, although my friends/family tried to cheer me up), Pittsburgh Sleep Quality Index (PSQI) item 7 (During the past month, how often have you had trouble staying awake while driving, eating meals, or engaging in social activity?), the PSQI sleep quality index (global score), taking diazepines, oxazepines, thiazepines and oxepines, and right hippocampal volume. Predictors with a negative association included cholesterol (mmol/l), taking thyroid medications: yes/no, and how often do you drink alcohol? (At most once a week/more than once a week). See Supplementary Figures 1 – 3.

Our final multimodal model provided an increase in prognostic certainty of 23.21% (prognostic summary index = $(PPV + NPV) - 100$). Further, a patient classified as being rehospitalized by our model was 83% more likely to be rehospitalized than a patient who was not (positive likelihood ratio = $\frac{sensitivity}{1-specificity}$). Finally, our multimodal classifier was statistically significant after permutation testing ($p = 9.99 \times 10^{-5}$) (Figure 1 and Table 3). For individual modality models, see Supplementary Information 3.2.

Discussion

The current study is the first of its kind to investigate the role of multiple predictor modalities for MDD rehospitalization trajectory modelling. Overall, our multimodal model provided a 23.21% increase in prognostic certainty for patient rehospitalization classification whilst providing an isolated subset of multimodal predictors for analysis. Furthermore, our multimodal model was statistically significant after permutation testing.

Of clinical importance is the positive (PPV) and negative predicted values (NPV) of our multimodal model and how they can be used to inform clinical decision making. As PPV was low (PPV = 41.64), using this model to confirm clinician suspicion of rehospitalization is not supported. On the contrary, NPV was modestly high (NPV = 81.57). Suggesting that the model could be used to confirm a clinician's suspicion of low rehospitalization risk, potentially identifying the patient as suitable for less assertive follow up.

As few studies are yet to consider a diverse range of biomarker modalities for rehospitalization trajectory modelling, the discriminative ability of the included modalities, as well as their interaction with clinical markers is of interest. First, our best performing model was multimodal, including a range of clinical, blood-biomarker, and structural imaging predictors. A recent work by Koutsouleris also demonstrated the superiority of multimodal models, finding that a combination of clinical and neuroimaging markers was most predictive of functional outcomes in a cohort of patients at risk for psychosis with recent onset depression⁽⁴⁵⁾. Our findings build on this work as well as provide preliminary evidence for the discriminative ability of blood biomarkers on a well-defined and clinically meaningful dichotomous endpoint.

Given that clinical predictors have consistently demonstrated their discriminative ability in multivariate pattern recognition studies^(12, 14, 46), it is unsurprising that 8 of 10 predictors in our multimodal solution were clinical. On the contrary, PGRS scores were of no use for

rehospitalization classification (test AUC = 50.5) showing the least discriminative ability of all the modelled modalities. Even with lower p-value thresholds ($p = 0.05$ and 0.01), no changes were seen in our multimodal model. In addition, unimodal PGRS discrimination was still low at these thresholds (AUC = 54.83 and 54.17 respectively, Supplementary Information 3.2). PGRS scores have received strong interest in psychiatric research, hoping that the aggregation of multiple single nucleotide polymorphisms may illuminate genetic differences in psychiatric traits as well as parse the heterogeneity of outcomes such as medication response^(47, 48). Significant associations with outcomes of interest have been consistently demonstrated, however, commonly explain less than 1 – 2% of variance, limiting their current clinical use. Given that PGRS scores could not classify rehospitalization any better than chance in the current work, the discovery of more variants as well as non-linear modelling techniques may improve their clinical utility in future works⁽⁴⁹⁾. All other biological modalities performed better, yet still lacked clinically meaningful discriminative ability when modelled without clinical predictors. Given these findings and those of Koutsouleris⁽⁴⁵⁾, it appears that biomarkers may be of prognostic use but likely perform best when modelled with clinical predictors also.

Regarding the selection of diazepines, oxazepines, thiazepines, and oxepines and their positive association with rehospitalization, it is possible that patients on this class of drugs at baseline had a more severe form of illness requiring augmentation with antipsychotic medications, placing them at a greater risk of rehospitalization. In addition, thyroid medication use was negatively associated with rehospitalization. Besides a small handful of patients, those with a past diagnosis of hyper/hypothyroidism who were not currently taking thyroid medications had t3, t4, and TSH levels within healthy reference ranges (Supplementary Information 3.3), suggesting that it was not the presence of unmedicated patients with thyroid disorders relapsing at high rates that were responsible for thyroid medication use and its negative association with rehospitalization. Considering that nearly all patients were taking antidepressant medications at their baseline assessment, it is possible that a synergistic prophylactic relationship between thyroid and antidepressant medication use may exist. Such a polytherapeutic relationship has been robustly demonstrated in the multisite sequenced alternatives to relieve depression trial (STAR*D)⁽²⁹⁾. However, prophylactic effects against rehospitalization should be explored in future works.

Of interest, right hippocampal volume showed a positive association with rehospitalization. This was surprising, given the well-established effect of smaller hippocampal volumes in

MDD patients compared to healthy controls⁽⁶⁾. To better understand this observation, we conducted a range of sub-analysis (Supplementary Information 3.4). In accordance with previous works^(50, 51), we showed that those taking any form of antipsychotic at baseline had larger right hippocampal volumes than those who were not, albeit this effect was not significant after controlling for covariates. In addition, we showed that there was a significant gender/diazepine, oxazepine, thiazepine, and oxepine use interaction effect on right hippocampal volumes, with women currently taking medications from this class having significantly larger right hippocampal volumes than those who were not. Forty four percent of these women ended up being rehospitalized between baseline and their 2 year follow up assessment, compared to only 22% of women who were not. Whilst we had binary usage data for medication, dosage data was not available. Dosage effects were likely prevalent, but not quantifiable in the current work. Overall, it is plausible that changes in right hippocampal volume proxied for a complex gender/medication/dosage specific aetiology not fully represented in our clinical data.

The first limitation of the current work is the models' scope. As our model was trained on a middle-aged European cohort, it is plausible that some of the selected predictors in our pipeline were unique to this demographic. Predictors such as the number of previous inpatient episodes will likely be larger on an older cohort, offering more discriminative ability. Therefore, the clinical utility of the model needs to be considered within this scope. To overcome this limitation, we have provided our trained model online through the Photon AI model repository.

Further limitations include the reporting of nominal significance in our logistic regression model (Supplementary Information 3.1). It is important to consider that the nominally significant effects that were found in our non-penalized logistic regression model did not survive FDR corrections (Supplementary Table 1). However, our primary interest was the emergent multivariate pattern that demonstrated statistically significant class separation between cases and controls at $p = 9.99 \times 10^{-5}$. Given the importance and statistical significance of this aggregated multivariate pattern, we deemed it necessary to conduct an exploratory analysis to illuminate the contribution of each individual predictor after rigorously controlling for known covariates age, gender, smoker status, BMI, and intracranial volume. Given this, we believe the nominal significance of predictors in our logistic regression model to be of use for hypothesis generation in future works.

Regarding clinical use, whilst our multimodal model was statistically significant after permutation testing, its balanced accuracy was relatively low ($BA = 63.05$). Whether this level of performance is sufficient for clinical use is unknown, however, this BA is similar to that attained in a previous work that is now deployed clinically ($BA = 59.60 - 64.6$)⁽¹⁴⁾. Further, we argue that it is not the absolute accuracy of a model that should dictate whether it is of clinical use, but whether or not it outperforms current clinical best practice. For example, as rehospitalization risk is not formally and routinely quantified by clinicians to inform their clinical decision making, we would not expect clinician prognostication of rehospitalization to be any better than chance. Therefore, even a modestly performing model would theoretically confer clinical advantage when deployed at scale. In addition, such a model would be free of well-documented decision-making biases that are known to affect clinician prognostication⁽⁵²⁾ (e.g., anchoring effects), potentially providing further clinical benefit under high workloads commonly seen in inpatient care. Nonetheless, for these questions to be answered, future works that benchmark clinicians' prognostic abilities are needed before such incremental utility can be quantified.

Finally, it is important to note that whilst our multimodal model provided the highest degree of class separation, after testing for post-hoc differences between all models and FDR correcting p-values, our multimodal model did not significantly outperform our model containing only clinical predictors ($p = 0.15$) (supplementary information 3.2). Future works of increased size may be required to elucidate a statistically significant difference given the small predictive discrepancy between the multimodal and clinical model. On the contrary, clinical data of sufficient depth (for example, that captures the previously discussed gender/medication/dosage effects and their association with right hippocampal volumes) may even mitigate the contribution of biomarkers in future works. If so, these models will bestow both ease of use and economic advantage, rendering the inclusion of costly and harder to attain biomarkers ineffectual. To answer these questions a greater dearth of clinical data, discovery samples of increased size, as well as validation samples that are phenotypically and geographically diverse will be required.

In conclusion, the presented findings suggest that the combination of ML techniques with multimodal clinical and biomarker data may lead to an increase in prognostic certainty compared to chance level. Continued research is required, but ML may be of use to derive models for clinicians to make personalised predictions regarding rehospitalization risk as well as better inform prophylactic treatment strategies.

Acknowledgments

The BiDirect Study is funded by the German Federal Ministry of Education and Research (BMBF, (01ER0816 and 01ER1506, and 01ER1205). Additional funding for the analysis was provided by the German Research Foundation (DFG, grant FOR2107 DA1151/5-1 and DA1151/5-2 to UD; SB-TRR58, Projects C09 and Z02), the Interdisciplinary Center for Clinical Research (IZKF) of the medical faculty of Münster (grant Dan3/012/17 to UD, SEED 11/18) and the Deanery of the Medical Faculty of the University of Münster. In addition, this work was supported with supercomputing resources provided by the Phoenix HPC service at the University of Adelaide. The funding sources had no role in study design, collection, analysis, or interpretation of the data, in the writing of the report, or in the decision to submit the article for publication.

Conflicts of interest

None

References

1. Hardeveld F, Spijker J, De Graaf R, Nolen WA, Beekman AT. Prevalence and predictors of recurrence of major depressive disorder in the adult population. *Acta Psychiatr Scand*. 2010 Sep;122(3):184-91. DOI: 10.1111/j.1600-0447.2009.01519.x.
2. Paykel ES, et al. Residual symptoms after partial remission: an important outcome in depression. *Psychol Med*. 1995 Nov;25(6):1171-80.
3. Ramana R, et al. Remission and relapse in major depression: a two-year prospective follow-up study. *Psychol Med*. 1995 Nov;25(6):1161-70.
4. Keller MB, Lavori PW, Lewis CE, Klerman GL. Predictors of relapse in major depressive disorder. *JAMA*. 1983 Dec 23-30;250(24):3299-304.
5. Grilo CM, et al. Personality disorders predict relapse after remission from an episode of major depressive disorder: a 6-year prospective study. *J Clin Psychiatry*. 2010 Dec;71(12):1629-35. DOI: 10.4088/JCP.08m04200gre.
6. Schmaal L, et al. Subcortical brain alterations in major depressive disorder: findings from the ENIGMA Major Depressive Disorder working group. *Mol Psychiatry*. 2016 Jun;21(6):806-12. DOI: 10.1038/mp.2015.69.
7. Appelhof BC, et al. Glucocorticoids and relapse of major depression (dexamethasone/corticotropin-releasing hormone test in relation to relapse of major depression). *Biological psychiatry*. 2006;59(8):696-701.
8. Zobel AW, Yassouridis A, Frieboes RM, Holsboer F. Prediction of medium-term outcome by cortisol response to the combined dexamethasone-CRH test in patients with remitted depression. *Am J Psychiatry*. 1999 Jun;156(6):949-51. DOI: 10.1176/ajp.156.6.949.
9. Berman RM, et al. Transient depressive relapse induced by catecholamine depletion: potential phenotypic vulnerability marker? *Arch Gen Psychiatry*. 1999 May;56(5):395-403.
10. Gili M, Vicens C, Roca M, Andersen P, McMillan D. Interventions for preventing relapse or recurrence of depression in primary health care settings: A systematic review. *Prev Med*. 2015 Jul;76 Suppl:S16-21. DOI: 10.1016/j.ypmed.2014.07.035.

11. Teismann H, et al. Establishing the bidirectional relationship between depression and subclinical arteriosclerosis - rationale, design, and characteristics of the BiDirect Study. *Bmc Psychiatry*. 2014 Jun 13;14. DOI: Artn 174
10.1186/1471-244x-14-174.
12. Chekroud AM, et al. Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. *JAMA psychiatry*. 2017;74(4):370-8.
13. Drysdale AT, et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat Med*. 2017 Jan;23(1):28-38. DOI: 10.1038/nm.4246.
14. Chekroud AM, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry*. 2016 Mar;3(3):243-50. DOI: 10.1016/S2215-0366(15)00471-X.
15. Koutsouleris N, Kambeitz-Ilankovic L, Ruhrmann S, et al. Prediction models of functional outcomes for individuals in the clinical high-risk state for psychosis or with recent-onset depression: A multimodal, multisite machine learning analysis. *JAMA Psychiatry*. 2018. DOI: 10.1001/jamapsychiatry.2018.2165.
16. Kambeitz J, et al. Detecting Neuroimaging Biomarkers for Depression: A Meta-analysis of Multivariate Pattern Recognition Studies. *Biol Psychiatry*. 2017 Sep 1;82(5):330-8. DOI: 10.1016/j.biopsych.2016.10.028.
17. Schmaal L, et al. Predicting the naturalistic course of major depressive disorder using clinical and multimodal neuroimaging information: a multivariate pattern recognition study. *Biological psychiatry*. 2015;78(4):278-86.
18. Moayedoddin B, et al. Prevalence and clinical characteristics of the DSM IV major depression among general internal medicine patients. *Eur J Intern Med*. 2013 Dec;24(8):763-6. DOI: 10.1016/j.ejim.2013.05.016.
19. Lopresti AL, Maker GL, Hood SD, Drummond PD. A review of peripheral biomarkers in major depression: the potential of inflammatory and oxidative stress biomarkers. *Prog Neuropsychopharmacol Biol Psychiatry*. 2014 Jan 3;48:102-11. DOI: 10.1016/j.pnpbp.2013.09.017.

20. Weber B, Lewicka S, Deuschle M, Colla M, Heuser I. Testosterone, androstenedione and dihydrotestosterone concentrations are elevated in female patients with major depression. *Psychoneuroendocrinology*. 2000 Nov;25(8):765-71.
21. Schweiger U, et al. Testosterone, gonadotropin, and cortisol secretion in male patients with major depression. *Psychosom Med*. 1999 May-Jun;61(3):292-6.
22. Musselman DL, Evans DL, Nemeroff CB. The relationship of depression to cardiovascular disease: epidemiology, biology, and treatment. *Archives of general psychiatry*. 1998;55(7):580-92.
23. Wray NR, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet*. 2018 May;50(5):668-81. DOI: 10.1038/s41588-018-0090-3.
24. Knight MJ, Baune BT. Cognitive dysfunction in major depressive disorder. *Current opinion in psychiatry*. 2018;31(1):26-31.
25. Lopresti AL, Hood SD, Drummond PD. A review of lifestyle factors that contribute to important pathways associated with major depression: diet, sleep and exercise. *J Affect Disorders*. 2013;148(1):12-27.
26. Murphy MJ, Peterson MJ. Sleep Disturbances in Depression. *Sleep Med Clin*. 2015 Mar;10(1):17-23. DOI: 10.1016/j.jsmc.2014.11.009.
27. Chekroud SR, Chekroud AM. Efficacy of Resistance Exercise Training With Depressive Symptoms. *JAMA Psychiatry*. 2018 Oct 1;75(10):1091-2. DOI: 10.1001/jamapsychiatry.2018.2084.
28. Schmaal L, et al. Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the ENIGMA Major Depressive Disorder Working Group. *Mol Psychiatry*. 2017 Jun;22(6):900-9. DOI: 10.1038/mp.2016.60.
29. Nierenberg AA, et al. A comparison of lithium and T(3) augmentation following two failed medication treatments for depression: a STAR*D report. *Am J Psychiatry*. 2006 Sep;163(9):1519-30; quiz 665. DOI: 10.1176/ajp.2006.163.9.1519.
30. Joffe RT, Marriott M. Thyroid hormone levels and recurrence of major depression. *Am J Psychiatry*. 2000 Oct;157(10):1689-91. DOI: 10.1176/appi.ajp.157.10.1689.

31. Joffe R, Segal Z, Singer W. Change in thyroid hormone levels following response to cognitive therapy for major depression. *Am J Psychiatry*. 1996 Mar;153(3):411-3. DOI: 10.1176/ajp.153.3.411.
32. Sassarini DJ. Depression in midlife women. *Maturitas*. 2016 Dec;94:149-54. DOI: 10.1016/j.maturitas.2016.09.004.
33. Schiller CE, Meltzer-Brody S, Rubinow DR. The role of reproductive hormones in postpartum depression. *CNS Spectr*. 2015 Feb;20(1):48-59. DOI: 10.1017/S1092852914000480.
34. Rodgers S, et al. Serum testosterone levels and symptom-based depression subtypes in men. *Front Psychiatry*. 2015;6:61. DOI: 10.3389/fpsyt.2015.00061.
35. Almeida OP, Yeap BB, Hankey GJ, Jamrozik K, Flicker L. Low Free Testosterone Concentration as a Potentially Treatable Cause of Depressive Symptoms in Older Men. *Archives of General Psychiatry*. 2008;65(3):283-9. DOI: 10.1001/archgenpsychiatry.2007.33.
36. Kische H, et al. Associations of androgens with depressive symptoms and cognitive status in the general population. *PLoS One*. 2017;12(5):e0177272. DOI: 10.1371/journal.pone.0177272.
37. Parekh A, Smeeth D, Milner Y, Thure S. The Role of Lipid Biomarkers in Major Depression. *Healthcare (Basel)*. 2017 Feb 3;5(1). DOI: 10.3390/healthcare5010005.
38. Maes M, et al. Lower serum high-density lipoprotein cholesterol (HDL-C) in major depression and in depressed men with serious suicidal attempts: relationship with immune-inflammatory markers. *Acta Psychiatr Scand*. 1997 Mar;95(3):212-21.
39. Lee SH, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*. 2013 Sep;45(9):984-+. DOI: 10.1038/ng.2711.
40. Brainstorm C, et al. Analysis of shared heritability in common disorders of the brain. *Science*. 2018 Jun 22;360(6395). DOI: 10.1126/science.aap8757.
41. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2005;67(2):301-20.

42. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw.* 2010;33(1):1.
43. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers.* 1999;10(3):61-74.
44. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: *Proceedings of the 22nd international conference on Machine learning.* 2005: ACM; p. 625-32.
45. Koutsouleris N, et al. Prediction Models of Functional Outcomes for Individuals in the Clinical High-Risk State for Psychosis or With Recent-Onset Depression: A Multimodal, Multisite Machine Learning Analysis. *JAMA Psychiatry.* 2018 Nov 1;75(11):1156-72. DOI: 10.1001/jamapsychiatry.2018.2165.
46. Koutsouleris N, Kahn RS, Chekroud AM. Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach (vol 3, pg 935, 2016). *Lancet Psychiatry.* 2017 Feb;4(2):95-.
47. Amare AT, et al. Association of polygenic score for schizophrenia and HLA antigen and inflammation genes with response to lithium in bipolar affective disorder: a genome-wide association study. *JAMA psychiatry.* 2018;75(1):65-74.
48. Amare AT, et al. Association of the polygenic scores for personality traits and response to selective serotonin reuptake inhibitors in patients with major depressive disorder. *Frontiers in psychiatry.* 2018;9:65.
49. Zheutlin AB, et al. Multivariate Pattern Analysis of Genotype–Phenotype Relationships in Schizophrenia. *Schizophrenia Bull.* 2018;44(5):1045-52.
50. Bodnar M, et al. The effect of second-generation antipsychotics on hippocampal volume in first episode of psychosis: longitudinal study. *BJPsych Open.* 2016 Mar;2(2):139-46. DOI: 10.1192/bjpo.bp.115.002444.
51. Chakos MH, et al. Duration of illness and treatment effects on hippocampal volume in male patients with schizophrenia. *Br J Psychiatry.* 2005 Jan;186:26-31. DOI: 10.1192/bjp.186.1.26.

52. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. *BMC Med Inform Decis Mak.* 2016 Nov 3;16(1):138. DOI: 10.1186/s12911-016-0377-1.

Figure 1. Left: Area under the receiver operator characteristic for all classifiers. Right: Null distribution (blue gaussian distribution) and classifier performance (green dashed line) for our multimodal model after permutation testing ($m = 10,000$).

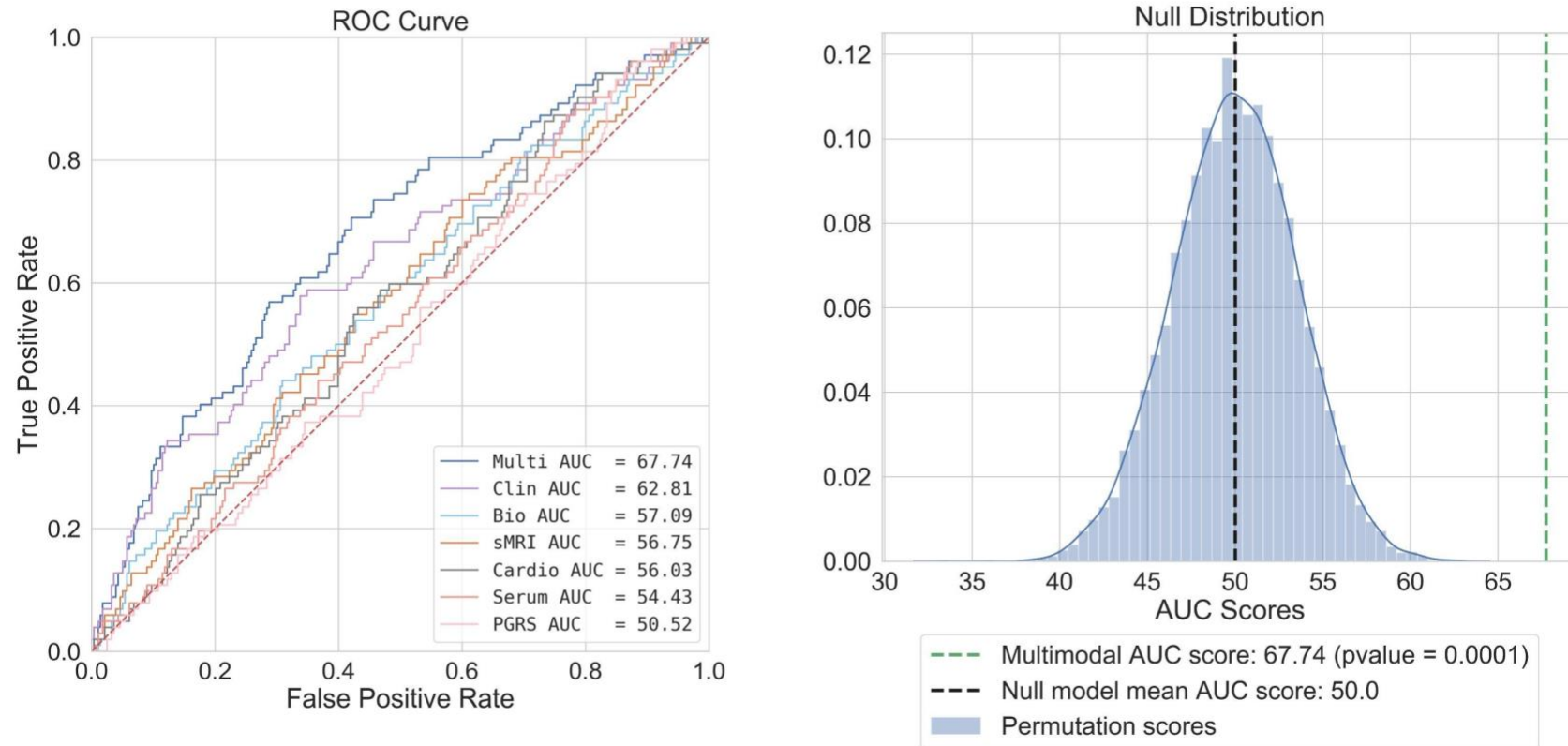


Table 1. Summary statistics for the final study sample. Means, standard deviations (SD), minimal (Min) and maximal (Max) values are presented. Significance testing between groups was conducted with independent samples t-tests. “Total inpatient episodes” includes the baseline assessment inpatient episode as well as all previous inpatient episodes. Abbreviations: HAM-D total – total score for the first 17 items of the HAM scale. CES-D total – total score with the inversion of positive items 4, 8, 12, and 16 taken into account.

	Mean	SD	Min	Max		Mean	SD	Min	Max	P
Rehospitalized?					Rehospitalized?					
Sex (m/f) (n = 43/59)					Sex (m/f) (n = 108/170)					
Age (n = 102)	49.03	7.32	34.96	63.96	Age (n = 278)	49.91	7.38	35.15	65.37	0.3
HAM-D total (n = 101)	15.33	6.59	0.00	27.00	HAM-D total (n = 278)	12.71	6.33	0.00	33.00	< 0.01
CES-D total (n = 102)	31.30	12.97	1.00	56.00	CES-D total (n = 276)	25.40	11.50	0.00	48.00	< 0.01
Total inpatient (n = 101)	2.06	2.00	0.00	10.00	Total inpatient (n = 274)	1.42	0.90	0.00	6.00	< 0.01

Table 2. Percentage proportions and total counts for psychotropic medication use in each rehospitalization outcome group. Significant differences between groups were assessed using chi-square tests.

Medication	Rehospitalized? No (N = 278)	Rehospitalized? Yes (N = 102)	P
Selective serotonin reuptake inhibitors	29.50% (N = 82)	22.55% (N = 23)	0.18
Beta blocking agents	18.71% (N = 52)	10.78% (N = 11)	0.07
Non-selective monoamine reuptake inhibitors	10.43% (N = 29)	13.73% (N = 14)	0.37
Other antidepressants	57.91% (N = 161)	71.57% (N = 73)	0.06
Benzodiazepines	30.58% (N = 85)	53.92% (N = 55)	< 0.01
Butyrophenone derivatives	3.96% (N = 11)	8.82% (N = 9)	0.06
Diazepines, oxazepines, thiazepines, oxepines	24.10% (N = 67)	41.18% (N = 42)	< 0.01
Lithium	2.52% (N = 7)	3.92% (N = 4)	0.47
Other antipsychotics	6.47% (N = 18)	17.65% (N = 18)	0.01

Table 3. Performance metrics for all classifiers. All classifiers used a Linear Support Vector Machine with Platt scaling, only predictor modalities varied across models. Mean (SD) scores from the outer 10-fold cross-validation loops are presented. Model abbreviations: Multi – our multimodal model (all biomarker modalities, clinical, and demographic variables), Clinical – clinical and demographic predictors only, Bio – model with all biomarker modalities (no clinical or demographic data), sMRI: structural imaging predictor model only, Cardio: electrocardiography and bioelectrical impedance analysis predictor model only, Serum – blood biomarkers only, PGRS: PGRS model only. Metric abbreviations: AUC – Area-under-the-Curve, F1 – Harmonic mean of Sens + Spec, BAC – Balanced accuracy, Acc – Accuracy, Sens – Sensitivity, Spec – Specificity, PPV – Positive predicted value, NPV – Negative predicted value, PSI – Prognostic summary index, PLR – Positive likelihood ratio, NLR – Negative likelihood ratio, DOR – Diagnostic odds ratio.

	Train	Test	SVM Results										
	AUC	AUC	F1	BAC	Acc	Sens	Spec	PPV	NPV	PSI	PLR	NLR	DOR
Multi	78.86 (2.81)	67.74 (13.86)	67.15	63.05	65.72	57.45	68.65	41.64	81.57	23.21	1.83	0.62	2.96
Clinical	73.59 (1.64)	62.81 (11.14)	64.30	60.10	62.62	54.73	65.44	37.04	79.79	16.83	1.58	0.69	2.29
Bio	63.12 (1.04)	57.09 (11.47)	55.80	51.47	53.64	46.90	56.03	27.44	74.78	2.22	1.07	0.95	1.13
sMRI	64.53 (1.74)	56.75 (10.46)	56.94	52.03	55.31	45.00	59.06	29.16	74.96	4.12	1.10	0.93	1.18
Cardio	61.44 (1.48)	56.03 (13.12)	56.15	54.22	53.65	55.55	52.90	29.62	77.18	6.80	1.18	0.84	1.40
Serum	60.70 (0.75)	54.43 (8.31)	51.81	50.00	63.97	20.00	80.00	5.41	58.57	-36.02	1.02	0.98	1.04
PGRS	59.72 (1.26)	50.52 (13.62)	53.32	50.52	50.85	49.91	51.18	27.88	73.30	1.18	1.83	0.62	2.96

Supplementary Information

Supplementary Information SI1:

1.1 Recruitment of participants and ethics statement

Cohort 1 consisted of 1016 patients, who suffered from an episode of depression at the time of recruitment. Recruitment took place at the Institute of Epidemiology and Social Medicine at the University of Münster, Germany. The recruitment of outpatients was limited to those who had been hospitalized due to depression at least once during the 12 months period prior to inclusion into the study. Inclusion criteria were age (≥ 35 and < 66 years) and current in- or outpatient treatment due to acute depression. Exclusion criteria were compulsory admission, comorbid dementia, and comorbid drug abuse (including alcohol). Potential participants were ascertained by trained and certified study psychologists, and eligible patients were invited to participate in the BiDirect-Baseline assesment¹. Appointments were scheduled via telephone or email. All MDD patients included in the present study suffered from an episode of major depression at the time of recruitment and were either currently hospitalized or had been hospitalized due to depression at least once during the 12-month period prior to inclusion into the study according to hospital service records. Further information on hospitalization was obtained via the total number of inpatient depressive episodes. MDD diagnosis was confirmed in all MDD patients with a structured diagnostic interview administered by trained and certified study psychologists. To assess baseline levels of depressive symptoms the Hamilton Rating Scale for Depression (HAM-D) and the Centre for Epidemiologic Studies Depression Scale (CES-D) were administered (appendix methods 1 and 2). The study was approved by the ethics committee of the University of Münster and the Westphalian Chamber of Physicians. All participants provided written informed consent. Further details on rationale, design and recruitment procedures of the BiDirect study have been extensively described in previous publications¹¹.

1.2 Diagnosis of depression

For diagnostic purposes, all participants received selected modules (i.e. modules A, A', B, D, and O) of the M.I.N.I. International Neuropsychiatric Interview (German version 5.0.0)², which assessed whether a participant exhibited acute (first or recurrent) major depression with or without melancholic features, acute dysthymia, acute/ former manic/hypomanic episodes, or acute generalized anxiety disorder. Further, for all patients from cohort 1 and for those participants from cohorts 2 and 3 who had received a M.I.N.I. diagnosis of acute major depression, it was clarified whether atypical depression features were currently present by means of six selected items of the Inventory of Depressive Symptomatology (IDS)³ (items 8, 12, 14, 27, 29, and 30). In addition, all patients with depression were assessed using the 17 items version of the Hamilton Depression Rating Scale (HAM-D-17)⁴ and the 14 items version of the Hamilton Anxiety Rating Scale (HAM-A-14)⁵.

1.3 Cognitive function module and self-report predictors

To measure cognitive function, we included time scores from the trail making test (parts A and B), immediate recall of emotional words from a 12-word passage of positive, neutral, and negative terms, verbal production of different animal names in a one-minute interval, the Purdue Pegboard test, and an 11-point patient self-evaluated attention score. In addition, several self-report questionnaires were available including the Pittsburgh

Sleep Quality Index (PSQI), the Pain Sensitivity Questionnaire (PSQ), the Childhood Trauma Screener (CTS) and the EQ-5D, which assesses perceived health-related quality of life. See supplementary table 1 below for all included predictors.

1.4 Structural MRI procedure

All structural T1 images were automatically preprocessed using FreeSurfer (Version 5.3) with default parameters. Cortical thickness measures were based on the Desikan–Killiany atlas. Segmentation quality was assessed by visual inspection and statistically evaluated for outliers following a standardized protocol provided by the ENIGMA consortium (<http://enigma.ini.usc.edu/protocols/imaging-protocols>). T1 structural images were acquired at a single 3.0 T MRI scanner (Intera with Achieva update; Philips Medical Systems, Best, The Netherlands) using a three-dimensional fast gradient echo sequence with a repetition time of 7.26 milliseconds, echo time=3.56 milliseconds, flip angle=9°, two signal averages, inversion prepulse every 404 milliseconds, acquired over a field of view of 256 x 256 mm, phase encoding in AP and RL direction, reconstructed to voxels of 1 mm x 1 mm x 1 mm.

1.5 Serum biomarker analysis

Blood was collected from all patients during examination and serum samples were prepared within two hours after collection and subsequently stored at -80°C in the local biobank until analysis. All laboratory analyses were performed at the Institute of Clinical Chemistry and Laboratory Medicine at the University Medicine Greifswald according to the instructions of the manufacturer using the Dimension Vista® 1500 System (Siemens Healthineers, Eschborn, Germany). Serum samples were stored at -80°C in the Integrated Research Biobank Greifswald after the shipment and until the analysis.

1.6 Genetic Markers

The genotyping for the BiDirect cohort was done for 2129 samples using the Infinium PsychArray-24 Kit from Illumina (<https://www.illumina.com/products/by-type/microarray-kits/infinium-psycharray.html>). The genotype data that passed the above QC criteria (n=2115) were imputed in the Michigan server⁶ <https://imputationserver.sph.umich.edu> using the 1000 Genome Phase 3 version 5, following the ENIGMA protocol (ENIGMA 1KGP_p3v5 Cookbook_20170713.pdf). The genotype data for PGRS scores used were extracted from the imputed dosage using PLINK⁷. See PGRS methods below for more information.

1.7 Electrocardiogram procedure

We non-invasively determined vascular status through the measurement of a standard 3-channel electrocardiogram (ECG) using extremity leads, brachial systolic and diastolic blood pressure by oscillometry, the ankle brachial index by photoplethysmography, pulse wave velocity, the augmentation index (Vascular Explorer, enverdis GmbH), and the intima-media-thickness of the far wall of the carotid arteries by ultrasound (Acuson X300, Siemens).

1.8 Machine learning pipeline architecture

A Scikit learn pipeline acts as a composite estimator allowing for the chaining of multiple estimators into one. Pipeline architecture affords several advantages. Firstly, parameters across multiple estimators can be tuned in

unison, increasing the efficiency of the hyperparameter optimization process. Second, and most importantly, pipelines help avoid leakage from test data into a trained model in the cross-validation process. This is achieved by ensuring that the same samples are used to train the transformers and predictors^{8,9}. Within our pipeline we conducted imputation, standardization, feature selection, and hyperparameter optimization.

1.8.1 Imputation

Imputation was completed using multivariate imputation by chained equations (MICE). Many multiple imputation techniques assume a large joint model for all variables such as a joint normal distribution. In datasets with n variables of different types (continuous, categorical etc), this assumption is not appropriate. MICE is an alternative approach, whereby regression models are run where each variable with missing data is modelled conditional upon the other variables in the data. This approach affords the ability to model each variable according to its own distribution, for example, binary variables are modeled with logistic regression whilst continuous variables are modelled with linear regressions¹⁰.

1.8.2 Standardization

Standardization, also known as zero mean, unit variance ($x' = \frac{x - \bar{x}}{\sigma}$) is a common requirement for many machine learning models. Certain models such as radial basis function support vector machines and the l1 and l2 regularizes of linear models (as used in our elastic net feature selection) may perform poorly if the predictors are not centered around zero with variance in the same order. If one predictor has variance that is disproportionately larger than others, it may exert undue influence over the objective function and make the model unable to learn from other predictors as expected¹¹. For this reason, all predictors were standardized within our pipeline.

1.8.3 Predictor selection

For predictor selection, we used the elastic net, a form of penalized regression that combines the l1 and l2 regularizes from lasso and ridge regression. This combination allows the learning of a sparse model where some highly correlated coefficients are shrunk to zero as is done in lasso, whilst lesser correlated coefficients are shrunk but non-zeroed¹². This approach is ideal when a large degree of multicollinearity exists between coefficients, as is common in psychiatric datasets.

1.8.4 Estimators

When dealing with small to moderate samples, low complexity (high bias) models tend to provide better fits to the training data. When sample size is constrained, only a small number of hypotheses that map our x matrix to our y vector can be learnt¹³⁻¹⁵. For this reason, we focused on a low complexity linear model, a linear support vector machine. For information on model formulation, see the SVC classification section of the Scikit learn documentation.

1.8.5 Parameter tuning

As the overall hyperparameter space was small, we conducted an exhaustive grid search to tune the SVM C and l1 ratio hyperparameters. A grid search allows us to search all specified hyperparameters (parameters not directly learnt within a model) that maximize our cross-validation score according to a specified criterion. In our

case, our criterion was the maximization of area under the curve on the receiver operator characteristic. See the [Scikit](#) learn grid search documentation for more information. For the elastic net, the final hyperparameter values were $\alpha = 0.3$, $\text{l1\_ratio} = 0.1$, $\text{eps} = 0.001$, $\text{max\_iter} = -1$, $\text{tol} = 0.0001$). For the SVM, the final hyperparameters were $C = 0.001$, $\text{kernel} = \text{linear}$, $\text{class\_weight} = \text{balanced}$, $\text{probability} = \text{True}$).

1.8.6 Cross Validation

Cross-validation was conducted using 5 repeats of 10-fold cross-validation. Research suggests that 10-fold cross-validation may better balance the bias-variance trade-off be compared to the more expensive leave-one-out cross-validation. In addition, repeated runs are suggested to avoid favorable splits in the data that may lead to overly optimistic accuracy estimates. Alternatively, bootstrap methods tend to have low variance but extremely high bias on certain datasets. Thus, were avoided in the current analysis. For detailed information, see the work of [Kohavi](#)¹⁶.

1.8.7 Permutation Testing

To test whether a model's classification score is statistically significant, a technique called permutation testing is used. Permutation tests randomly shuffle the class labels and then attempt to make a prediction with the given predictor space. This process is repeated n times until a null distribution is attained. Model performance is then compared to that of the null distribution.

References

1. Teismann H, Wersching H, Nagel M, et al. Establishing the bidirectional relationship between depression and subclinical arteriosclerosis - rationale, design, and characteristics of the BiDirect Study. *Bmc Psychiatry*. 2014;14.
2. Ackenheil M, Stotz-Ingenlath G, Dietz-Bauer R, Vossen A. MINI mini international neuropsychiatric interview, German version 5.0. 0 DSM IV. *Psychiatrische Universitätsklinik München, Germany*. 1999.
3. Rush AJ, Giles DE, Schlesser MA, Fulton CL, Weissenburger J, Burns C. The inventory for depressive symptomatology (IDS): preliminary findings. *Psychiat Res*. 1986;18(1):65-87.
4. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 1960;23:56-62.
5. Hamilton M. The assessment of anxiety states by rating. *Br J Med Psychol*. 1959;32(1):50-55.
6. Das S, Forer L, Schonherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48(10):1284-1287.
7. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575.
8. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in Data Mining: Formulation, Detection, and Avoidance. *Acm T Knowl Discov D*. 2012;6(4).
9. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
10. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Meth Psych Res*. 2011;20(1):40-49.
11. Géron A. *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. "O'Reilly Media, Inc."; 2017.
12. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2005;67(2):301-320.
13. Beleites C, Neugebauer U, Bocklitz T, Krafft C, Popp J. Sample size planning for classification models. *Anal Chim Acta*. 2013;760:25-33.
14. Guo Y, Graber A, McBurney RN, Balasubramanian R. Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms. *Bmc Bioinformatics*. 2010;11.
15. Raudys SJ, Jain AK. Small Sample-Size Effects in Statistical Pattern-Recognition - Recommendations for Practitioners. *Ieee T Pattern Anal*. 1991;13(3):252-264.

16. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Paper presented at: Ijcai1995.
17. Ojala M, Garriga GC. Permutation Tests for Studying Classifier Performance. *J Mach Learn Res.* 2010;11:1833-1863.

Supplementary Information SI2:

The genetic profile scores (GPSs) were computed for 33 traits including the major cardiometabolic and psychiatric disorders, metabolic syndrome, the big five personality traits, educational attainment, smoking behavior and childhood behavioural problems.

2.1 Genotyping and quality control

For each sample, we implemented quality control (QC) procedures for the genotype data using PLINK¹ and samples with low genotype rates <95%, sex inconsistencies (X-chromosome heterozygosity), and genetically related individuals were excluded. We also excluded SNPs that had poor genotyping rate <95%, an ambiguity (A/T and C/G SNPs), a minor allele frequency (MAF) < 0.01 or that showed deviation from Hardy-Weinberg Equilibrium ($p < 10^{-6}$).

2.2 Imputations

The genotype data that passed the above QC criteria were imputed in the Michigan server² <https://imputationserver.sph.umich.edu> separately for each study samples using the Haplotype Reference Consortium (HRC) reference panel³.

After excluding the low-frequency SNPs (MAF<0.1); low-quality variants (imputation INFO < 0.9; and indels, the imputed dosages were converted to best guess genotypes with PLINK¹. The subsequent genetic profile score analyses were performed using the best guess genotypes.

2.3 Discovery GWAS data

The genetic profile scores were calculated using the approach previously described by the International Schizophrenia Consortium⁴. This approach requires a discovery and target datasets. The discovery data which refers to the summary data, i.e., effect sizes (beta, a log of odds ratio or log of hazard ratio) in our study were obtained from previous GWASs (see below). The different disorders and phenotypes included were MDD⁵, bipolar disorder⁶, schizophrenia⁷, autism spectrum disorders, anxiety disorders⁸, anorexia nervosa⁹, Alzheimer's disease¹⁰, type 2 diabetes¹¹, coronary artery disease¹², myocardial infarction¹², lipids level¹³, body mass index¹⁴, waist to hip ratio¹⁵, glycemic phenotypes (fasting glucose¹⁶ and fasting insulin¹⁶, indices of beta-cell function-HOMA-B and insulin resistance-HOMA-IR¹⁷, fasting proinsulin¹⁸, and hemoglobin A_{1c}(C)¹⁹, big Five personality traits (extraversion, openness to experience, agreeableness, conscientiousness and neuroticism)^{20,21}, subjective wellbeing²¹, depressive symptoms²¹, loneliness²², educational attainment²³, smoking behavior²⁴, and Childhood behavioural problems²⁵.

2.4 Target sample: Study Samples

Genetic data from three German cohorts were used to compute the genetic profile scores: FOR_2107 (n=922), MS_Imaging_MDD_BD_controls_joint_sample (n=1084) and Marburg_PsychChip (n=374).

2.5 Genetic profile scoring

We computed GPS for each of the 33 target traits using the HRC imputed data and GWAS summary statistics of the respective GWASs, separately for the three samples. The GPSs were calculated within a P-value threshold

(P_T) as the sum of reference SNP alleles multiplied by the effect size (β - coefficient or log (OR) or log (HR)) weighted by the sum of effect sizes derived from the GWAS summary statistics. The major histocompatibility complex region was excluded from the GPS calculation because of its complex linkage disequilibrium structure. Quality-controlled SNPs were clumped for linkage disequilibrium based on P-value informed clumping using $r^2 = 0.1$ within a 250-kb window to create a SNP-set in linkage equilibrium using PLINK software run on Linux; *plink-clump-p1 1 -clump-p2 1 -clump-r2 0.1 -clump-kb 250*). The genetic profile scores were computed at different P-value thresholds in the three samples ($<5 \times 10^{-8}$, $<1 \times 10^{-6}$, $<1 \times 10^{-5}$, $<1 \times 10^{-4}$, $<1 \times 10^{-3}$, <0.01 , <0.05 , <0.1 , <0.2 , <0.3 , <0.4 , <0.5 , <0.6 , <0.7 , <0.8 , <0.9 , <1)

2.6 List of traits for which genetic profile score was calculated and references for the summary data

2.6.1 Cardiometabolic diseases

- Coronary artery disease-cad_additive_model_2015¹²
- Type 2 diabetes-diagram_mega_type2DM_meta¹¹
- Myocardial infarction-mi_additive_model_2015¹²

2.6.2 Measures of Obesity

- Body mass index-Obesity_all¹⁴,
- Waist to hip ratio-GIANT_2015_WHR_COMBINED_AllAncestries¹⁵

2.6.3 Lipids level¹³

- jointGwasMc_TG
- jointGwasMc_TC
- jointGwasMc_LDL
- jointGwasMc_HDL

2.6.4 Glucose and insulin traits

- Fasting proinsulin: MAGIC_proinsulin¹⁸
- MAGIC_Manning_et_al_InFastingInsulin_MainEffect_Oct112012¹⁶
- MAGIC_Manning_et_al_FastingGlucose_MainEffect_Oct112012¹⁶
- MAGIC_In_HOMA_IR¹⁷
- MAGIC_In_HOMA_B¹⁷
- Hemoglobin A_{1c}(C)- MAGIC_HbA1C¹⁹

2.6.5 Psychiatric disorders

- Anxiety disorders- Anxiety_case_control_2016⁸
- Bipolar disorder-pgc_bip_full_2012⁶
- MDD- pgc_mdd_full_2012⁵ and MDD23andme²⁶
- Schizophrenia-PGC_SCZ2_2014⁷
- Alzheimer's disease- IGAP_Alzheimer¹⁰

- Anorexia nervosa: PGC_anorexia_snp_all_13May2016⁹
- PGC_ASD_AUD_5Mar2015: *Autism Spectrum Disorder Working Group of the Psychiatry Genomics Consortium. Dataset: PGC-ASD summary statistics from a meta-analysis of 5,305 ASD-diagnosed cases and 5,305 pseudocontrols of European descent (March 2015). (available at: <http://www.med.unc.edu/pgc/results-anddownloads>)*

2.6.6 Psychological traits

- Subjective wellbeing-SWB_Full ²¹
- Depressive symptoms-DS_Full ²¹
- Loneliness-in continuous and dichotomous scales ²²

2.6.7 Big Five personality traits

- GPC_1_NEO_AGREEABLENESS²⁰
- GPC_1_NEO_CONSCIENTIOUSNESS²⁰
- GPC_1_NEO_EXTRAVERSION²⁰
- GPC_1_NEO_OPENNESS²⁰ and
- NEUROTICISM_Full²¹

2.6.8 Smoking behavior²⁴

- Cigarettes_per_day_continuous
- Cigarettes_Ever_never_Categorical
- Cigarettes_Current_former_Categorical
- Cigarettes_Ageofonset_continuous

2.6.9 Educational attainment and Childhood behavioral problems

- EduYears_2016²³
- Preschool_Internalizing_Problem²⁵

PGRS references

- 1 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575, doi:10.1086/519795 (2007).
- 2 Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284-1287, doi:10.1038/ng.3656
<http://www.nature.com/ng/journal/v48/n10/abs/ng.3656.html#supplementary-information> (2016).
- 3 McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-1283, doi:10.1038/ng.3643
<http://www.nature.com/ng/journal/v48/n10/abs/ng.3643.html#supplementary-information> (2016).
- 4 Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752, doi:10.1038/nature08185 (2009).
- 5 Ripke, S. *et al.* A mega-analysis of genome-wide association studies for major depressive disorder. *Mol Psychiatry* **18**, 497-511, doi:10.1038/mp.2012.21 (2013).
- 6 Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat Genet* **43**, 977-983, doi:10.1038/ng.943 (2011).
- 7 Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427, doi:10.1038/nature13595 (2014).
- 8 Otowa, T. *et al.* Meta-analysis of genome-wide association studies of anxiety disorders. *Mol Psychiatry*, doi:10.1038/mp.2015.197 (2016).
- 9 Boraska, V. *et al.* A genome-wide association study of anorexia nervosa. *Mol Psychiatry* **19**, 1085-1094, doi:10.1038/mp.2013.187 (2014).
- 10 Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* **45**, 1452-1458, doi:10.1038/ng.2802 (2013).
- 11 Mahajan, A. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* **46**, 234-244, doi:10.1038/ng.2897 (2014).
- 12 Nikpay, M. *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* **47**, 1121-1130, doi:10.1038/ng.3396 (2015).
- 13 Global Lipids Genetics, C. Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274-1283, doi:10.1038/ng.2797
<http://www.nature.com/ng/journal/v45/n11/abs/ng.2797.html#supplementary-information> (2013).
- 14 Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206, doi:10.1038/nature14177 (2015).

- 15 Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187-196, doi:10.1038/nature14132 (2015).
- 16 Manning, A. K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet* **44**, 659-669, doi:<http://www.nature.com/ng/journal/v44/n6/abs/ng.2274.html#supplementary-information> (2012).
- 17 Dupuis, J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* **42**, 105-116, doi:10.1038/ng.520 (2010).
- 18 Strawbridge, R. J. *et al.* Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes* **60**, 2624-2634, doi:10.2337/db11-0415 (2011).
- 19 Soranzo, N. *et al.* Common variants at 10 genomic loci influence hemoglobin A(1)(C) levels via glycemic and nonglycemic pathways. *Diabetes* **59**, 3229-3239, doi:10.2337/db10-0502 (2010).
- 20 de Moor, M. H. *et al.* Meta-analysis of genome-wide association studies for personality. *Mol Psychiatry* **17**, 337-349, doi:10.1038/mp.2010.128 (2012).
- 21 Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat Genet* **48**, 624-633, doi:10.1038/ng.3552
<http://www.nature.com/ng/journal/v48/n6/abs/ng.3552.html#supplementary-information> (2016).
- 22 Gao, J. *et al.* Genome-Wide Association Study of Loneliness Demonstrates a Role for Common Variation. *Neuropsychopharmacology*, doi:10.1038/npp.2016.197 (2016).
- 23 Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539-542, doi:10.1038/nature17671
<http://www.nature.com/nature/journal/v533/n7604/abs/nature17671.html#supplementary-information> (2016).
- 24 Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* **42**, 441-447, doi:http://www.nature.com/ng/journal/v42/n5/supinfo/ng.571_S1.html (2010).
- 25 Benke, K. S. *et al.* A genome-wide association meta-analysis of preschool internalizing problems. *Journal of the American Academy of Child and Adolescent Psychiatry* **53**, 667-676.e667, doi:10.1016/j.jaac.2013.12.028 (2014).
- 26 Hyde, C. L. *et al.* Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat Genet* **48**, 1031-1036, doi:10.1038/ng.3623
<http://www.nature.com/ng/journal/v48/n9/abs/ng.3623.html#supplementary-information> (2016).

Supplementary Information SI3:

3.1 Logistic regression and SVM weights

To understand the contributions of the 10 selected predictors in our multimodal model, we used logistic regression and controlled for demographic covariates age and gender and general health covariates, smoker status and BMI (of which may have confounding effects on cholesterol and right hippocampal volumes). In addition, we controlled for intracranial volume to adjust for the natural variability of brain size in this analysis. Due to the multicollinearity between variables, we standardised all variables to have a mean of zero and a standard deviation of one. As this analysis is exploratory/hypothesis generating and secondary to the emergent multivariate pattern that was statistically significant in our multimodal model, we only report on variables that attained or were approaching nominal significance in this section (with nominal p-values reported). FDR corrected p-values and all other variables from our multivariate pattern can be found in table 2.

For every 1 standard deviation increase in the number of previous hospital admissions recorded at baseline, patients became 50% more likely to be rehospitalized within the following two-years post assessment [OR: 1.50, 95% CI: 1.11 – 2.03, $p = 0.009$]. In addition, for every 1 standard deviation increase in right hippocampal volume, patients were 72% more likely to be rehospitalized in the following 2 years [OR: 1.72, 95% CI: 1.13 – 2.62, $p = 0.01$]. Of note, the following predictors approached nominal significance and may be worthy of investigation in future works. Those taking thyroid medications at baseline were 59% less likely to be rehospitalized in the following 2 years [OR: 0.41, 95% CI: 0.15 – 1.10, $p = 0.08$], whilst those who had trouble concentrating ‘mostly or all of the time’ on CES-D item 5, were more than 3 times as likely to be rehospitalized in the following two years [OR: 3.12, 95% CI: 0.83 – 11.77, $p = 0.09$]. For all other results, see Table 2.

Finally, whilst the odds ratios and associated p-values from our logistic regression model allow us to interpret the univariate relationships between each predictor and rehospitalization after controlling for clinically relevant covariates, it is not analogous to the SVM weight values for each predictor that were used to construct the hyperplane in our final multimodal classification model. Therefore, we have also taken the final average weight values for each selected predictor across the outer 10-fold cross-validation loop from our multimodal ML pipeline. In descending order, SVM weights were: CES-D 5 (Last week I had trouble concentrating) = 0.1015, PSQI Sum: PSQI sleep quality index (global score) = 0.0948, Number of inpatient depressive episodes = 0.0925, PSQI 7 (Difficulties staying awake) = 0.0871, CES-D 3 (Could not get rid of my troubling mood) = 0.0847, Take Diazepines, oxazepines, thiazepines and oxepines? = 0.0717, Right hippocampal volume = 0.0560, Total cholesterol = -0.0450, Thyroid therapy? = -0.0511, How often do you drink an alcoholic beverage? = -0.0650.

3.2 Individual modality models

We re-trained our classifier with individual data modalities only, including clinical, sMRI, serum markers, cardiovascular markers, PGRS, as well as a combination of all biomarker modalities. The pipeline used was the same as that in our multimodal analysis. However, as the size of the predictor space was small for our individual biological modalities, we used all a-priori selected predictors with no elastic net predictor selection for the sMRI, serum, cardiovascular and PGRS models. Due to the higher number of predictors in our clinical modality and combined biomarker model, a smaller subset was selected with the elastic net as per the multimodal model

pipeline. Results for the following models were: clinical modality model (train AUC = 73.59, test AUC = 62.81) our combined biomarker model (train AUC = 63.12, test AUC = 57.09), sMRI model (train AUC = 64.53, test AUC = 56.75), cardiovascular model (train AUC = 61.44, test AUC = 56.03), serum model (train AUC = 60.70, test AUC = 54.43), and PGRS model (train AUC = 59.72, test AUC = 50.52). Significant differences were found between models after correcting for ties ($H = 11.42$, $p = 0.04$). Following FDR corrections, our multimodal model performed significantly better than our combined biomarker model ($p = 0.04$), serum model ($p = 0.03$), sMRI model ($p = 0.05$), cardio model ($p = 0.04$), and our PGRS model ($p = 0.03$), but not our clinical model ($p = 0.15$). For the unimodal PGRS models with lowered p-value thresholds (0.05 and 0.01), we attained test AUCs of 54.83 and 54.17.

3.3 Were thyroid medications protective against rehospitalization?

To investigate the negative association between thyroid medication and rehospitalization, we looked at the proportions of those with and without a past thyroid disorder diagnosis relative to those taking thyroid medications at their baseline assessment. Fifty-eight patients had a past thyroid diagnosis at baseline and were taking thyroid medications at their assessment. In addition, 29 patients had a past diagnosis of a thyroid disorder and were not taking thyroid medications at their baseline assessment. Of these patients, 25 had t3 levels greater than 3.5 mmol/l, whilst four had levels that went down to 3.08 mmol/l. No patients exceeded 5.24 mmol/l. For t4, all 29 patients were in the healthy reference range (9 mmol/l to 25 mmol/l). For TSH, 26 had levels in the healthy range (0.04 mU/L to 4.0 mU/L) whilst 3 dropped as low as 0.02 mU/L. Finally, 5 patients were taking thyroid medications with no diagnosis of a thyroid condition, presumably as a polytherapeutic treatment strategy. Besides 1 patient with a TSH of 0.01 mU/L, all thyroid markers for all 5 patients were in their healthy reference ranges.

3.4 Were right hippocampal volumes effected by medication use?

Previous studies have consistently demonstrated smaller hippocampal volumes in MDD patients compared to healthy controls⁶. Surprisingly, larger right hippocampal volumes were predictive of rehospitalization in our logistic regression model (Rehospitalized? **Yes**: $M = 4141.54$, $SD = 430.16$, **No**: $M = 4002.74$, $SD = 484.74$). To elucidate an explanation for this effect, we conducted a range of sub-analysis, controlling for age, gender, and intracranial volume in all analyses.

As nearly all patients were on an antidepressant medication at their baseline inpatient assessment, teasing out neurogenic effects of antidepressants on hippocampal volume was not possible. However, previous studies have shown that the use of antipsychotics may also affect hippocampal volumes⁴⁶⁻⁴⁸. Of interest, 40.3% (153/380) patients were taking an antipsychotic medication as augmentation to normal antidepressant therapy. To assess the effect of these medications on right hippocampal volume we used ordinary least squares regression.

According to the minimization of the Akaike information criterion⁴⁹, we found a significant gender/medication use interaction for the antipsychotic class diazepines, oxazepines, thiazepines, and oxepines. Medications that fall into this class include Olanzapine, Clozapine, Loxapine, Quetiapine, and Asenapine. Specifically, we found that women on this class of medication had significantly larger right hippocampal volumes ($M = 4017.9$, $SD = 397.03$) than women who were not ($M = 3846.58$, $SD = 336.51$) ($b = 270.09$, $CI = 63.89 - 476.3$, $p = 0.01$).

Further, 44.7% of women who were on this class of medication at baseline were re-hospitalized between

baseline and their 2 year follow up assessment, whilst only 22% of women who were not were rehospitalized within this period. However, the main effect for diazepam, oxazepam, thiazepam, and oxepin use at baseline was negative and non-significant ($b = -95.39$, CI: $-248.25 - 57.46$, $p = 0.22$). In aggregate, those taking any form of other antipsychotic medication (classes include phenothiazines with piperidine structure, butyrophenone derivatives, thioxanthene derivatives, benzamides, and “other antipsychotics”) had larger hippocampal volumes ($M = 4139.13$, $SD = 475.47$) compared to those who were not ($M = 4028.48$, $SD = 473.28$), however, this difference was not significant ($b = 85.58$, CI: $-55 - 226.18$, $p = 0.23$). Finally, we found smaller right hippocampal volumes in women ($M = 3875.26$, $SD = 352.3$) compared to men ($M = 4288.14$, $SD = 524.3$), however, these differences were not significant after controlling for intracranial volume ($b = -46.63$, CI: $-141.35 - 48.1$, $p = 0.33$). In addition, we found a significant relationship between age and right hippocampal volume. As age increased, right hippocampal volumes significantly decreased ($b = -8.5$, CI: $-13.5 - -3.5$, $p = 0.001$).

3.5 Were cholesterol levels affected by medication use?

Twenty-eight patients were taking lipid modifying agents at their baseline assessment. To assess the effects of these medications on cholesterol levels we controlled for covariates age, gender, and BMI and ran an ordinary least squares regression onto cholesterol levels (mmol/l). For patients taking lipid modifying agents, we found a statistically significant decrease of 0.80 mmol/l in cholesterol levels compared to those who were not taking these medications ($b = 0.80$, CI: $-1.19 - -0.41$, $p = < 0.0001$).

Supplementary Table 1. Variable table for all considered predictors entered into the multimodal pipeline for selection by the Elastic Net and subsequent SVM training.

Structural imaging	Freesurfer - volume and thickness
Lhippo	Left hippocampal volume
Rhippo	Right hippocampal volume
L_medialorbitofrontal_thickavg	Left medial orbitofrontal thickness average
R_medialorbitofrontal_thickavg	Right medial orbitofrontal thickness average
L_fusiform_thickavg	Left fusiform thickness average
R_fusiform_thickavg	Right fusiform thickness average
L_insula_thickavg	Left insula thickness average
R_insula_thickavg	Right insula thickness average
L_rostralanteriorcingulate_thickavg	Left rostral anterior cingulate thickness average
R_rostralanteriorcingulate_thickavg	Right rostral anterior cingulate thickness average
L_posteriorcingulate_thickavg	Left posterior cingulate thickness average
R_posteriorcingulate_thickavg	Right posterior cingulate thickness average
L_middletemporal_thickavg	Left middle temporal thickness average
R_inferiortemporal_thickavg	Right inferior temporal thickness average
R_caudalanteriorcingulate_thickavg	Right caudal anterior cingulate thickness average
Blood biomarkers	
andrix	Free androgen index
b17o	17 beta estradiol
shbg	Sex hormone binding hormone
testosteron	Testosterone
cholesterol	Total cholesterol
ft3	Free triiodothyronine (pmol)
hdl	High density lipoprotein cholesterol (mmol/l)
ft4	Free thyroxine (pmol)
tsh	Thyroid stimulating hormone (mU)
hscrp	CRP highly sensitive (mg/l)
Polygenic risk scores	
anxiety_pgrs_score	p-value threshold p = 0.5
alzheimer_pgrs	p-value threshold p = 0.5
anorexia_pgrs	p-value threshold p = 0.5
asd_pgrs	p-value threshold p = 0.5

pgc_bip_pgrs	p-value threshold p = 0.5
pgc_mdd_pgrs	p-value threshold p = 0.5
pgc_scz_pgrs	p-value threshold p = 0.5
Cardiovascular health	Electrocardiogram and bioelectrical impedance analysis
ekg_hr	Heart rate beats per minute
bia_bmi	Calculated body mass index kg / m 2
bia_ecm_bcm_index	ECM BCM Ratio
bia_grundumsatz	Basic metabolic rate kcal
bia_kfett_k_kg	Corrected body fat in kg
bia_koerperwasser	Total body water kg
bia_magermasse	Lean body mass kg
Habit change	
s_lsstair	Changed habits? Use the stairs more often
s_lswalk	Changed habits? Do frequent shopping on foot
s_lsweight	Changed habits? decrease weight loss
s_lsalc	Changed habits? Drink less alcohol
s_lssmoke	Changed habits? give up smoking
Demographic	
sex	Gender
sf1	How would you describe your state of health at the moment?
job_f1	Are you currently employed?
szlage	How would you assess your own economic situation at the moment?
szeink	What is the total monthly net income of your household?
rfdiat	Keep a specific diet or diet. If yes, which ones?
sm_f1	Currently smoke cigarettes pipes or cigars?
sz_mde1	If you have a disability card, please state the severity of your disability in
minia1	A1 In the last 2 weeks, you felt depressed or depressed almost every day,
minia2	A2 In the last 2 weeks have you almost always had the feeling that you
alter1	Age at first contact?
weight	Measured weight kg
edu_f1	Which is your highest school or university degree?
edu_f2	Which is your highest vocational degree?
Clinical symptoms	HAM-D, HAM-A, CES-D, IDS, EQ-5D
ph_hamd1	Depressed mood
ph_ids8	Mood Reaction

ph_hamd2	Work and other activities
ph_ids4	Hypersomnia
ph_hamd3	Sleep Disorders / Difficulties to fall asleep
ph_hamd4	Sleeping disturbances
ph_hamd5	Early awakening
ph_hama6	Sleep disorders
ph_hama7	Depressed mood
ph_hama8	Intellectual skills
ph_hamd9	Genital tract related symptoms
ph_hama10	Urogenital symptoms
ph_hamd11	Somatic symptoms Gastrointestinal
ph_ids12	Appetite Increases
ph_ids14	Weight gain within the last 14 days
ph_hamd12	Weight loss assessment according to the medical history
ph_hamd13	General somatic symptoms
ph_ids30	Lead Lore Physical Energy
ph_hamd14	Guilty
ph_hamd15	Suicidal thoughts
ph_hama16	Anxious mood
ph_hamd17	Anxiety Mental symptoms
ph_hama18	Inner Tension
ph_hama19	Somatic muscle discomfort
ph_hama20	Somatic Sensory Discomfort
ph_hama21	Fear
ph_hama22	Gastrointestinal symptoms
ph_hama23	Vegetative symptoms
ph_hama24	Cardiovascular symptoms
ph_hama25	Respiratory symptoms
ph_hamd26	Anxiety somatic symptoms
ph_ids29	Interpersonal Sensitivity
ph_hamd27	Hypochondria
ph_hamd28	Disease insight
ph_hamd29	Agitation
ph_hama30	Behavior during the interview
ph_hamd31	Slowdown

pi_n_epi_inpatient	Number of inpatient episodes?
pi_n_episodes	Total number of depressive episodes?
pi_n_inpatient	Total number of hospitalizations?
pi_age_1episode	Age at first depressive episode
pi_dat_prevepi	Month and Year of onset of last episode?
pi_dat_1inpatient	Month and year of first hospitalization?
pi_curepi_year	Year of onset of current depressive episode
hamd_17total	HAM D total score 17 items
hamd_17n	HAM D answers to the 17 items
hama_14total	HAM A total score 14 items
hama_14n	HAM A answers to the 14 items
pm_a2	In the last 2 weeks you have almost always had the feeling that you are
pm_lifa7_a	Have you experienced the feeling of depression differently than the
pm_lifa7_b	Did you feel worse in the morning compared to evening?
pm_lifa7_c	Wake up almost twice an hour or earlier than usual and then have trouble
pm_lifa7_d	Have you talked slower than usual almost every day, or were you restless
pm_lifa7_e	Have lost or had increases in your appetite almost every day? Gain or lose
pm_lifa7_f	Feelings of excess guilt?
pm_lifepis_n	Melancholic symptoms occurred in episode?
pm_lif_age	Age of this melancholic episode?
pm_d1_a	Have you ever had a time when you felt so overflowing and energetic that
pm_d2_a	Have you ever had a time in which you constantly felt so irritable for
pm_o1_a	In the past 6 months, have you been overly anxious and worried about
pm_a4	Episode of major depression current ???
cesd1	In the last week things have worried me that usually do not worry me
cesd2	I hardly had any appetite during the last week
cesd3	In the last week I could not get rid of my troublesome mood although my
cesd4	Last week, I felt as good as others
cesd5	In the last week I had trouble to concentrate
cesd6	In the last week I was depressed
cesd7	In the last week everything was exhausting for me
cesd8	Last week, I thought of the future with hope
cesd9	Last week, I thought my life was a single failure
cesd10	I was often scared in the last week
cesd11	I slept badly last week

cesd12	Last week I was happy
cesd13	In the last week, I've talked less than usual
cesd14	In the last week I felt lonely
cesd15	In the last week, the people were unkind to me
cesd16	In the last week, I enjoyed life
cesd17	I had to cry sometimes in the last week
cesd18	I was sad the last week
cesd19	In the last week I had the feeling that people do not like me
cesd20	In the last week, I could not get going
cesdsum	CES D Sum
cesd2c	CES D 0 16 points 1 16 points
cesd3c	CES D 0 15 points 1 15 and 21 points 2 21 points
cesd_dp	CES D depression items 3 6 9 10 14 17 18 Subscale
cesd_wb	CES D well-being items 4 8 12 16 REVERSED Subscale
cesd_so	CES D somatic items 1 2 5 7 11 13 20 Subscale
cesd_ip	CES D interpersonal items 15 19 Subscale
eq5d_2	EQ 5D 3L Item 2 take care of yourself
eq5d_3	EQ 5D 3L Item 3 general activities
eq5d_4	EQ 5D 3L Item 4 pain physical discomfort
eq5d_5	EQ 5D 3L Item 5 Fear depression
eq5d_vas	EQ 5D 3L thermometer today's state of health
eq5dscore	EQ 5D Score European VAS value set
Childhood trauma	Childhood trauma screener
cts_1	CTS Childhood Item 1 was loved
cts_2	CTS Childhood Item 2 physical violence in the family
cts_3	CTS Childhood Items 3 Hatred in the family
cts_4	CTS Childhood Item 4 sexually debased
cts_5	CTS Childhood Item 5 taken to doctor if needed
ctssum	CTS Sumscore about Items 1 to 5 Inversion of Items 1 and 5 takes into
Sleep	Pittsburg sleep quality index
psqi_1	When have you usually gone to bed?
psqi_2_	How long (in minutes) has it taken you to fall asleep each night?
psqi_3	When have you usually gotten up in the morning?
psqi_4	How many hours of actual sleep do you get at night? (This may be
psqi_5_sum	During the past month, how often have you had trouble sleeping

psqi_6	During the past month, how often have you taken medicine (prescribed or
psqi_7	During the past month, how often have you had trouble staying awake
psqi_8	During the past month, how much of a problem has it been for you to
psqi_9	During the past month, how would you rate your sleep quality overall?
psqilaten	PSQI subcomponent Sleep latency
psqidurat	PSQI subcomponent sleep duration
psqihse	PSQI subcomponent habitual sleep efficiency
psqidistb	PSQI subcomponent sleep disturbances
psqidaydys	PSQI subcomponent Daytime dysfunction
psqi_sum	PSQI sleep quality index global score
Cognition	
d_ewo_d1n	Memory (~VLMT) for emotional words: 1st pass, number correct
d_ewo_d1w01	Memory (~VLMT) for emotional words: 2nd pass, number correct
d_ewo_d1w02	Memory (~VLMT) for emotional words: 3rd pass, number correct
d_flue_r	Word fluidity – Animal words stated in 1 min interval
d_peg_li	Pegboard test, left hand, number of correctly set pins
d_peg_re	Pegboard test, right hand, number of correctly set pins
d_tmt_a	Trail making test, part A, duration
d_tmt_b	Trail making test, part B, duration
d_self	Attention self-evaluation: 0 - 10
Diet and alcohol consumption	
diet_score	Diet Quality Score (range 0 to 30)
al_f1	How often do you drink an alcoholic beverage of beer, spirits, liquor
al_f3_b	How much beer, wine or liquor did you drink yesterday in liters?
al_f4_b	How much beer and wine did you drink on the last weekend in liters?
al_f5	How often does it happen that you drink 6 or more alcoholic drinks on
al_f8	Have your spouse or other close relatives or friends ever worried or
basicsum	BASIC alcohol FB sum score points
alkohol	average alcohol consumption last week grams
Physical pain	Pain sensitivity questionnaire
ks_allks	Headache General Headache
ks_mig	Headache Migraine
psq_mean	Mean Pain Sensitivity Questionnaire includes Items PSQ1 to PSQ7
dx_chschmerz	Chronic Pain Diagnosis?
Medication	

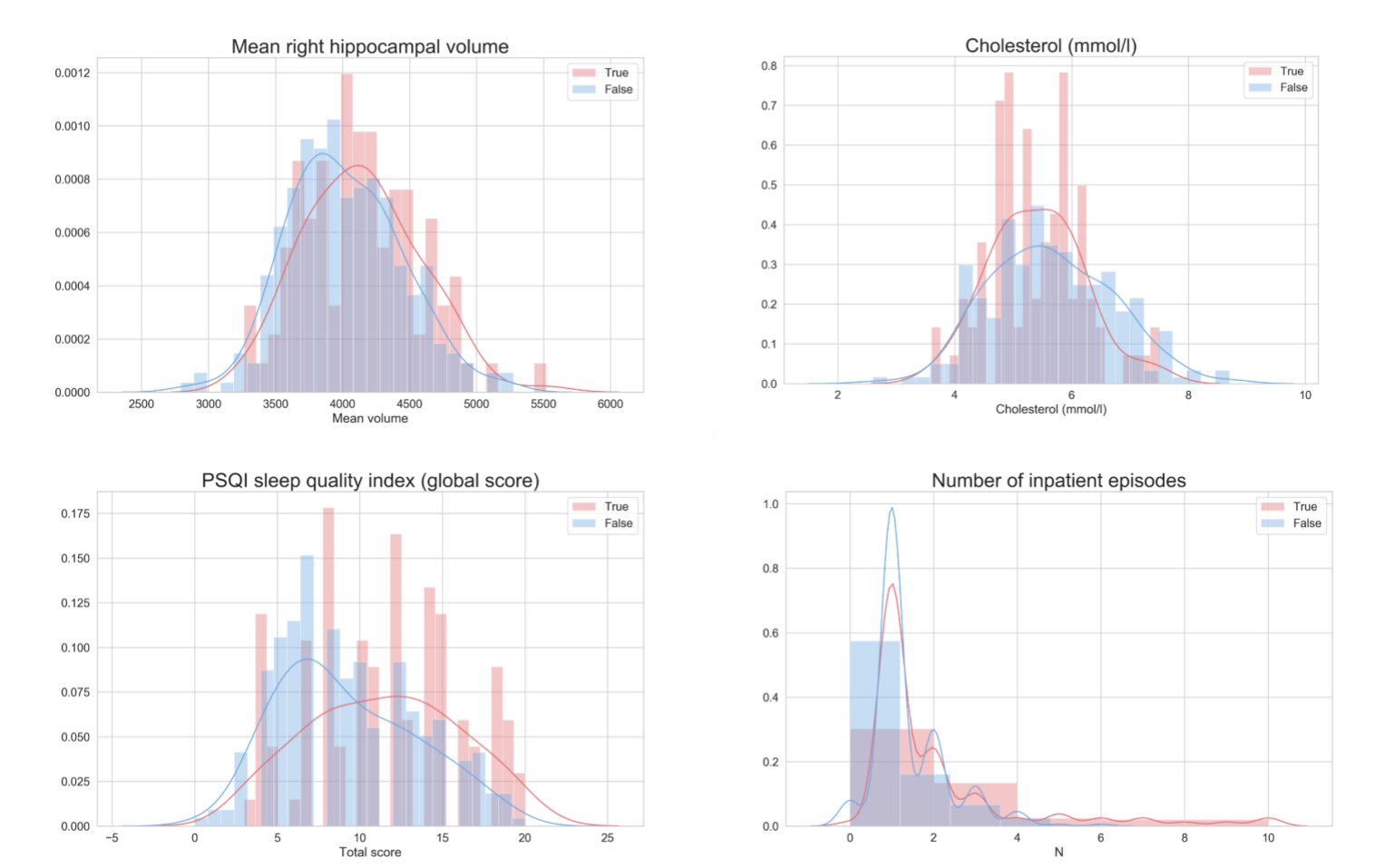
med_a03	Drugs for functional gastrointestinal disorder
med_a10	Drugs used in diabetes
med_a11	vitamins
med_a12	Mineral supplements
med_c07	Beta blocking agents
med_c08	Calcium channel blockers
med_c09	Agents acting on the renin angiotensin system
med_c10	Lipid modifying agents
med_g03	Sex hormones and modulators of the genital system
med_g04	Urologicals
med_h02	Corticosteroids for systemic use
med_h03	Thyroid therapy
med_l04	immunosuppressants
med_m03	Muscle relaxants
med_n02aa	Natural opium alkaloids
med_n02ab	Phenylpiperidines derivatives
med_n02ae	Oripavine derivatives
med_n02ax	Other opioids
med_n02ba	Salicylic acid and derivatives
med_n02bb	Pyrazolones
med_n02be	Anilides
med_n02bg	Other analgetics and antipyretics
med_n02cc	Selective serotonin agonists
med_n03ae	Benzodiazepine derivatives
med_n03af	Carboxamide derivatives
med_n03ag	Fatty acid derivatives
med_n03ax	Other antiepileptics
med_n05ab	Phenothiazines with piperazine structure
med_n05ad	Butyrophenone derivatives
med_n05af	Thioxanthene derivatives
med_n05ah	Diazepines oxazepines thiazepines and oxepines
med_n05al	Benzamide
med_n05an	Lithium
med_n05ax	Other antipsychotics
med_n05ba	Benzodiazepine derivatives

med_n05cd	Benzodiazepine derivatives
med_n05cf	Benzodiazepines related drugs
med_n05ch	Melatonin receptor agonists
med_n05cm	Other hypnotics and sedatives
med_n06aa	Non-selective monoamine reuptake inhibitors
med_n06ab	Selective serotonin reuptake inhibitors
med_n06af	Monoamine oxidase inhibitors non selective
med_n06ax	Other antidepressants
med_v06	General nutrients
Exercise	
ipaq_f1d	On how many of the past 7 days have you been doing strenuous physical
ipaq_f2h	How much time did you usually spend on one of these days of strenuous
ipaq_f3d	On how many of the past 7 days do you have moderate physical
ipaq_f4h	How much time did you usually spend on one of these days with
ipaq_f5d	On how many of the past 7 days have you walked at least 10 minutes
ipaq_f6h	How much time did you usually spend walking on one of these days?
ipaq_f7h	How much time have you spent sitting over the last 7 days? hours
ipaq_total_met	IPAQ total score
ipaq_category	IPAQ categorical score 3 levels of physical activity

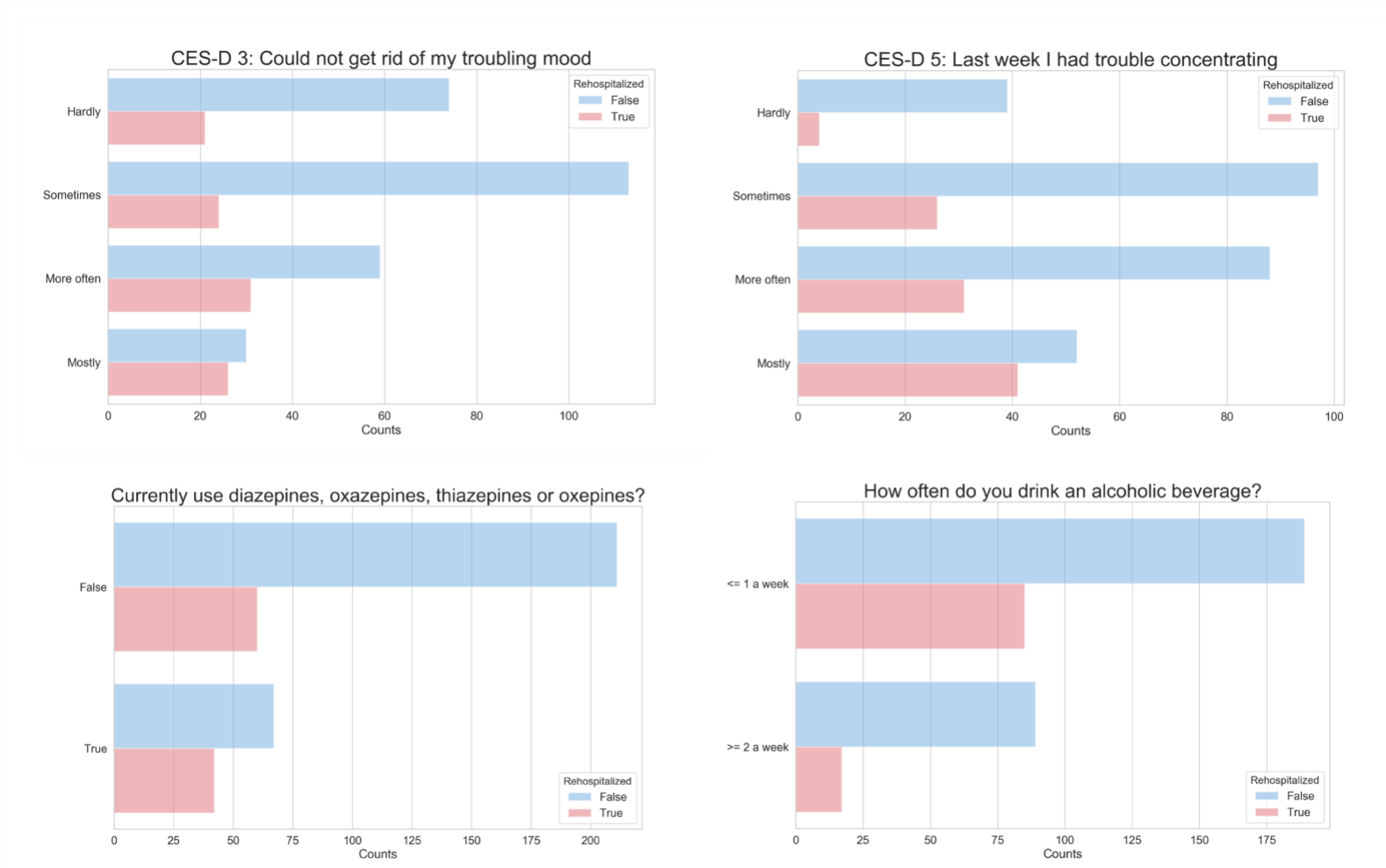
Supplementary Table 2. Summary of our logistic regression model for the final 10 predictors of rehospitalization after controlling for age, gender, BMI, smoker status, and intracranial volume.

	Estimate	Std Error	Statistic	OR	2.5% CI	97.5% CI	P Val	FDR P Val
Intercept	-1.64	0.72	-2.27	0.19	0.05	0.80	0.02	0.18
CES-D 5 (trouble concentrating?) - Sometimes	0.67	0.60	1.11	1.96	0.60	6.40	0.27	0.60
CES-D 5 (trouble concentrating?) - More often	0.29	0.64	0.46	1.34	0.38	4.68	0.64	0.82
CES-D 5 (trouble concentrating?) - Mostly	1.14	0.68	1.68	3.12	0.83	11.77	0.09	0.37
Diazepines oxazepines thiazepines oxepines - Yes	0.36	0.33	1.09	1.43	0.75	2.71	0.28	0.60
PSQI 7 (Trouble staying awake?) - Less than once a week	0.18	0.58	0.32	1.20	0.39	3.71	0.75	0.83
PSQI 7 (Trouble staying awake?) - One or two times a week	0.19	0.62	0.31	1.21	0.36	4.05	0.76	0.83
PSQI 7 (Trouble staying awake?) - Three or more times a week	0.44	0.39	1.12	1.56	0.72	3.37	0.26	0.60
CES-D 3 (could not get rid of mood) - Sometimes (1-2 days)	-0.73	0.43	-1.71	0.48	0.21	1.11	0.09	0.37
CES-D 3 (could not get rid of mood) - More often (3-4 days)	0.14	0.45	0.30	1.15	0.47	2.79	0.76	0.83
CES-D 3 (could not get rid of mood) - Mostly (5-7 days)	0.24	0.52	0.46	1.27	0.46	3.55	0.65	0.82
Thyroid medication at baseline? Yes	-0.89	0.50	-1.78	0.41	0.15	1.10	0.08	0.37
Drink alcohol more than once a week? Yes	-0.33	0.36	-0.91	0.72	0.35	1.46	0.36	0.67
Sex - Female	0.08	0.42	0.20	1.09	0.47	2.49	0.85	0.85
Smoke? No, but I used to	-0.22	0.37	-0.61	0.80	0.39	1.65	0.54	0.82
Smoke? Yes	-0.19	0.36	-0.52	0.83	0.41	1.67	0.60	0.82
Number of previous hospital admissions?	0.41	0.16	2.61	1.50	1.11	2.03	0.009*	0.13
PSQI (sleep quality index global score)	0.28	0.19	1.48	1.32	0.91	1.92	0.14	0.42
Mean right hippocampal volume	0.54	0.21	2.54	1.72	1.13	2.62	0.011*	0.13
Intracranial volume	-0.18	0.23	-0.79	0.84	0.54	1.30	0.43	0.74
Total cholesterol	-0.23	0.15	-1.53	0.80	0.60	1.07	0.13	0.42
Age	0.03	0.16	0.18	1.03	0.76	1.39	0.85	0.85
BMI	-0.08	0.17	-0.46	0.93	0.67	1.29	0.65	0.82
Medication index	0.15	0.17	0.93	1.17	0.84	1.62	0.35	0.67

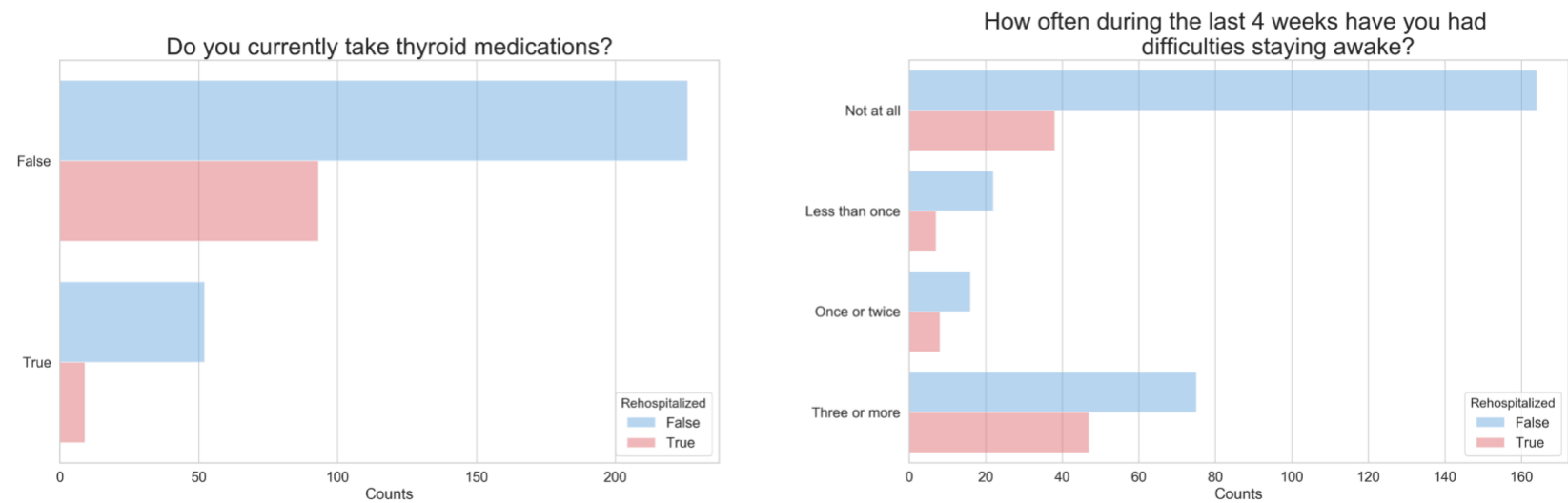
Supplementary Figure 1. Mean right hippocampal volume, cholesterol, PSQI sleep quality global score, and previous number of inpatient episodes stratified by rehospitalization status. Notes: Red: Rehospitalized = True, Blue: Rehospitalized = False.



Supplementary Figure 2. CES-D 3: Could not get rid of my troubling mood, CES-D 5: Last week I had trouble concentrating, currently use diazepines, oxazepines, thiazepines, oxepines, and alcohol consumption stratified by rehospitalization status. Notes: Red: Rehospitalized = True, Blue: Rehospitalized = False.



Supplementary Figure 3. Thyroid medication and PSQI item 7 stratified by rehospitalization status. Notes: Red: Rehospitalized = True, Blue: Rehospitalized = False.



Statement of Authorship

Title of Paper	Stratification of bipolar patients using polygenic scores increases the accuracy of lithium response prediction: A machine learning approach
Publication Status	☐ Published ☐ Accepted for Publication ☒ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Cearns, M., Amare, A., Schubert, O., Clark, S., & Baune, B.T. (2021). Stratification of bipolar patients using polygenic scores increases the accuracy of lithium response prediction: A machine learning approach. <i>Submitted.</i>

Principal Author

Name of Principal Author (Candidate)	Micah Cearns		
Contribution to the Paper	Formulated the idea for the paper, conducted all machine learning analysis, wrote the manuscript.		
Overall percentage (%)	85%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	15/01/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Azmeraw Amare		
Contribution to the Paper	Calculated the schizophrenia and major depression PRS. Provided feedback on the manuscript.		
Signature		Date	15/01/2021

Name of Co-Author	Scott Clark		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision.		
Signature		Date	15/01/2021
Name of Co-Author	Oliver Schubert		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision.		
Signature		Date	15/01/2021
Name of Co-Author	International Consortium on Lithium Genetics (ConLi*gen) co-authors		
Contribution to the Paper	Provided clinical and genetic data, reviewed, commented, and approved the manuscript.		
Signature		Date	15/01/2021
Name of Co-Author	Bernhard Baune		
Contribution to the Paper	Critical revision of the manuscript, overall guidance and supervision. As corresponding author, he made communications with journal editors.		
Signature		Date	15/01/2021

Please cut and paste additional co-author panels here as required.

Chapter 7

Stratification of bipolar patients using polygenic scores increases the accuracy of lithium response prediction: A machine learning approach

Micah Cearns BSc (Hons)¹, Azmeraw T. Amare, MPH, MSc, PhD^{11,2}, Klaus Oliver Schubert, MD, PhD^{1,3}, Anbupalam Thalamuthu, PhD⁷⁸; Joseph Frank, PhD⁷¹; Fabian Streit, PhD⁷¹; Mazda Adli, MD¹³; Nirmala Akula, PhD⁴; Kazufumi Akiyama, MD¹⁴; Raffaella Arda, MD¹⁵; Bárbara Arias, PhD¹⁶; Jean-Michel Aubry, MD¹⁷; Lena Backlund, MD, PhD¹⁸; Abesh Kumar Bhattacharjee, MD¹²; Frank Bellivier, MD, PhD¹⁹; Antonio Benabarre, MD, PhD²⁰; Susanne Bengesser, MD²¹; Joanna M. Biernacka, PhD^{22,23}; Armin Birner, MD²¹; Clara Brichant-Petitjean, MD¹⁹; Pablo Cervantes, MD²⁴; Hsi-Chung Chen, MD, PhD²⁵; Caterina Chillotti, MD¹⁵; Sven Cichon, PhD^{8,26}; Cristiana Cruceanu, PhD²⁷; Piotr M. Czerski, PhD²⁸; Nina Dalkner, MSc²¹; Alexandre Dayer, MD¹⁷; Franziska Degenhardt, MD⁸; Maria Del Zompo, MD²⁹; J. Raymond DePaulo, MD³⁰; Bruno Étain, MD, PhD¹⁹; Peter Falkai, MD³²; Andreas J. Forstner, MD^{8,26,33}; Louise Frisen, MD¹⁸; Mark A. Frye, MD²³; Janice M. Fullerton, PhD^{34, 35}; Sébastien Gard, MD³⁶; Julie S. Garnham, BN³⁷; Fernando S. Goes, MD³⁰; Maria Grigoriu-Serbanescu, PhD³⁸; Paul Grof, MD, PhD³⁹; Ryota Hashimoto, MD, PhD^{40,41}; Joanna Hauser, MD²⁸; Urs Heilbronner, PhD^{5,7}; Stefan Herms, Dipl.Biol.^{8,26}; Per Hoffmann, PhD^{8,26}; Andrea Hofmann, PhD⁸; Liping Hou, PhD⁴; Yi-Hsiang Hsu, PhD^{10,11}; Stephane Jamain, PhD³¹; Esther Jiménez, PhD²⁰; Jean-Pierre Kahn, MD, PhD⁴²; Layla Kassem, PhD⁴; Po-Hsiu Kuo, PhD⁴³; Tadafumi Kato, MD, PhD⁴⁴; John Kelsoe, MD¹²; Sarah Kittel-Schneider, MD⁴⁵; Sebastian Kliwicki, MD⁴⁶; Barbara König, MSc⁴⁷; Ichiro Kusumi, MD⁴⁸; Gonzalo Laje, MD⁴; Mikael Landén, MD^{49,50}; Catharina Lavebratt, PhD¹⁸; Marion Leboyer, MD, PhD⁵¹; Susan G. Leckband, BSc⁵²; Mario Maj, MD⁶³; Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium⁷⁴; Mirko Manchia, MD, PhD^{54,55}; Lina Martinsson, MD⁵⁶; Michael J. McCarthy, MD, PhD^{12,57}; Susan McElroy, MD⁵⁸; Francesc Colom, PhD^{59,60}; Marina Mitjans, PhD^{60,61}; Francis M. Mondimore, MD³⁰; Palmiero Monteleone, MD⁶²; Caroline M. Nievergelt, PhD¹²; Markus M. Nöthen, MD⁸;

Tomas Novák, MD⁶⁴; Claire O'Donovan, MB³⁷; Norio Ozaki, MD⁶⁵; Urban Ösby, MD, PhD⁶⁶; Sergi Papiol, PhD^{5,6}; Andrea Pfennig, MD⁶⁷; Claudia Pisanu, MD²⁹; James B. Potash, MD, MPH³⁰; Andreas Reif, MD⁴⁵; Eva Reininghaus, MD²¹; Guy A. Rouleau, MD⁶⁸; Janusz K. Rybakowski, MD⁴⁶; Martin Schalling, MD¹⁸; Peter R. Schofield, PhD, DSc^{34,35}; Barbara W. Schweizer, RN³⁰; Giovanni Severino, MD²⁹; Tatyana Shekhtman, MSc¹²; Paul D. Shilling, PhD¹²; Katsumasa Shimoda, MD⁶⁹; Christian Simhandl, MD⁷⁰; Claire M. Slaney, RN³⁷; Alessio Squassina, PhD²⁹; Thomas Stamm, MD¹³; Pavla Stopkova, MD⁶⁴; Fasil Tekola-Ayele, PhD⁹; Alfonso Tortorella, MD⁵³; Gustavo Turecki, MD²⁷; Julia Veeh, PhD⁴⁵; Eduard Vieta, MD, PhD²⁰; Stephanie H. Witt, PhD⁷¹; Gloria Roberts, PhD⁷²; Peter P. Zandi, PhD⁷³; Martin Alda, MD³⁷; Michael Bauer, MD, PhD⁶⁷; Francis J. McMahon, MD⁴; Philip B. Mitchell, MD⁷²; Thomas G. Schulze, MD^{4,5,7,30,71}; Marcella Rietschel, MD⁷¹; Scott Clark¹; Bernhard T. Baune, MD, PhD⁷⁵⁻⁷⁷

¹Discipline of Psychiatry, School of Medicine, University of Adelaide, Adelaide, SA, Australia

²South Australian Academic Health Science and Translation Centre, South Australian Health and Medical Research Institute (SAHMRI), Adelaide, South Australia, Australia

³Northern Adelaide Local Health Network, Mental Health Services, Adelaide, SA, Australia

⁴Intramural Research Program, National Institute of Mental Health, National Institutes of Health, US Department of Health & Human Services, Bethesda, MD, USA

⁵Institute of Psychiatric Phenomics and Genomics (IPPG), University Hospital, LMU Munich, Munich, Germany

⁶Department of Psychiatry and Psychotherapy, Ludwig-Maximilian-University Munich, Munich, Germany

⁷Department of Psychiatry and Psychotherapy, University Medical Center (UMG), Georg-August University Göttingen, Göttingen, Germany

⁸Institute of Human Genetics, University of Bonn and Department of Genomics, Life & Brain Center, Bonn, Germany

⁹Epidemiology Branch, Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD, USA

¹⁰HSL Institute for Aging Research, Harvard Medical School, Boston, MA, United States

¹¹Program for Quantitative Genomics, Harvard School of Public Health, Boston, MA, United States

¹²Department of Psychiatry, University of California San Diego, San Diego, CA, United States

- ¹³Department of Psychiatry and Psychotherapy, Charité - Universitätsmedizin Berlin, Campus Charité Mitte, Berlin, Germany
- ¹⁴Department of Biological Psychiatry and Neuroscience, Dokkyo Medical University School of Medicine, Mibu, Tochigi, Japan
- ¹⁵Unit of Clinical Pharmacology, Hospital University Agency of Cagliari, Cagliari, Italy
- ¹⁶Unitat de Zoologia i Antropologia Biològica (Dpt. Biologia Evolutiva, Ecologia i Ciències Ambientals), Facultat de Biologia and Institut de Biomedicina (IBUB), University of Barcelona, CIBERSAM, Barcelona, Spain
- ¹⁷Department of Psychiatry, Mood Disorders Unit, HUG - Geneva University Hospitals, Geneva, Switzerland
- ¹⁸Department of Molecular Medicine and Surgery, Karolinska Institute, Stockholm, Sweden, and Center for Molecular Medicine, Karolinska University Hospital, Stockholm, Sweden
- ¹⁹INSERM UMR-S 1144, Université Paris Diderot, Département de Psychiatrie et de Médecine Addictologique, AP-HP, Groupe Hospitalier Saint-Louis-Lariboisière-F.Widal, Paris, France
- ²⁰Bipolar Disorder Program, Institute of Neuroscience, Hospital Clinic, University of Barcelona, IDIBAPS, CIBERSAM, Barcelona, Catalonia, Spain
- ²¹Department of Psychiatry and Psychotherapeutic Medicine, Research Unit for bipolar affective disorder, Medical University of Graz, Graz, Austria
- ²²Department of Health Sciences Research, Mayo Clinic, Rochester, MN, United States
- ²³Department of Psychiatry and Psychology, Mayo Clinic, Rochester, MN, United States
- ²⁴The Neuromodulation Unit, McGill University Health Centre, Montreal, Canada
- ²⁵Department of Psychiatry & Center of Sleep Disorders, National Taiwan University Hospital, Taipei, Taiwan
- ²⁶Human Genomics Research Group, Department of Biomedicine, University Hospital Basel, Basel, Switzerland
- ²⁷Douglas Mental Health University Institute, McGill University, Montreal, Canada
- ²⁸Psychiatric Genetic Unit, Poznan University of Medical Sciences, Poznan, Poland
- ²⁹Department of Biomedical Sciences, University of Cagliari, Cagliari, Italy
- ³⁰Department of Psychiatry and Behavioral Sciences, Johns Hopkins University, Baltimore, MD, United States
- ³¹Univ Paris Est Créteil, INSERM, IMRB, Translational Neuropsychiatry, Fondation FondaMental, Créteil, France
- ³²Department of Psychiatry and Psychotherapy, Ludwig-Maximilian-University Munich, Munich, Germany
- ³³Department of Psychiatry (UPK), University of Basel, Basel, Switzerland
- ³⁴Neuroscience Research Australia, Sydney, NSW, Australia
- ³⁵School of Medical Sciences, University of New South Wales, Sydney, NSW, Australia

- ³⁶Service de psychiatrie, Hôpital Charles Perrens, Bordeaux, France
- ³⁷Department of Psychiatry, Dalhousie University, Halifax, Nova Scotia, Canada
- ³⁸Biometric Psychiatric Genetics Research Unit, Alexandru Obregia Clinical Psychiatric Hospital, Bucharest, Romania
- ³⁹Mood Disorders Center of Ottawa, Ontario, Canada
- ⁴⁰Molecular Research Center for Children's Mental Development, United Graduate School of Child Development, Osaka University, Osaka, Japan
- ⁴¹Department of Psychiatry, Osaka University Graduate School of Medicine, Osaka, Japan
- ⁴²Service de Psychiatrie et Psychologie Clinique, Centre Psychothérapique de Nancy - Université de Lorraine, Nancy, France
- ⁴³Department of Public Health & Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan
- ⁴⁴Department of Psychiatry & Behavioral Science, Juntendo University Graduate School of Medicine, Tokyo, Japan
- ⁴⁵Department of Psychiatry, Psychosomatic Medicine and Psychotherapy, University Hospital Frankfurt, Frankfurt, Germany
- ⁴⁶Department of Adult Psychiatry, Poznan University of Medical Sciences, Poznan, Poland
- ⁴⁷Department of Psychiatry and Psychotherapeutic Medicine, Landesklinikum Neunkirchen, Neunkirchen, Austria
- ⁴⁸Department of Psychiatry, Hokkaido University Graduate School of Medicine, Sapporo, Japan
- ⁴⁹Institute of Neuroscience and Physiology, the Sahlgrenska Academy at the Gothenburg University, Gothenburg, Sweden
- ⁵⁰Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
- ⁵¹Univ Paris Est Créteil, INSERM, IMRB, Translational Neuropsychiatry, AP-HP, Mondor University Hospital, DMU Impact, Fondation FondaMental, Créteil, France
- ⁵²Office of Mental Health, VA San Diego Healthcare System, San Diego, CA, United States
- ⁵³Department of Psychiatry, University of Perugia, Italy
- ⁵⁴Section of Psychiatry, Department of Medical Sciences and Public Health, University of Cagliari, Cagliari, Italy
- ⁵⁵Department of Pharmacology, Dalhousie University, Halifax, NS, Canada
- ⁵⁶Department of Clinical Neurosciences, Karolinska Institutet, Stockholm, Sweden
- ⁵⁷Department of Psychiatry, VA San Diego Healthcare System, San Diego, CA, United States
- ⁵⁸Department of Psychiatry, Lindner Center of Hope / University of Cincinnati, Mason, OH, United States

- ⁵⁹Mental Health Research Group, IMIM-Hospital del Mar, Barcelona, Catalonia, Spain
- ⁶⁰Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Instituto de Salud Carlos III, Madrid, Spain
- ⁶¹Clinical Neuroscience, Max Planck Institute of Experimental Medicine, Göttingen, Germany
- ⁶²Department of Medicine, Surgery and Dentistry “Scuola Medica Salernitana”, University of Salerno, Salerno, Italy
- ⁶³Department of Psychiatry, University of Campania “Luigi Vanvitelli”, Naples, Italy
- ⁶⁴National Institute of Mental Health, Klecany, Czech Republic
- ⁶⁵Department of Psychiatry & Department of Child and Adolescent Psychiatry, Nagoya University Graduate School of Medicine, Nagoya, Japan
- ⁶⁶Department of Neurobiology, Care Sciences, and Society, Karolinska Institutet and Center for Molecular Medicine, Karolinska University Hospital, Stockholm, Sweden
- ⁶⁷Department of Psychiatry and Psychotherapy, University Hospital Carl Gustav Carus, Medical Faculty, Technische Universität Dresden, Germany
- ⁶⁸Montreal Neurological Institute and Hospital, McGill University, Montreal, Canada
- ⁶⁹Department of Psychiatry, Dokkyo Medical University School of Medicine, Mibu, Tochigi, Japan
- ⁷⁰Bipolar Center Wiener Neustadt, Sigmund Freud University, Medical Faculty, Vienna, Austria
- ⁷¹Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany
- ⁷²School of Psychiatry, University of New South Wales, Sydney, Australia
- ⁷³Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States
- ⁷⁴For a full list of major depressive disorder working group of the PGC investigators, see the Supplementary Material
- ⁷⁵Department of Psychiatry and Psychotherapy, University of Münster, Münster, Germany
- ⁷⁶Department of Psychiatry, Melbourne Medical School, University of Melbourne, Parkville, Victoria, Australia
- ⁷⁷The Florey Institute of Neuroscience and Mental Health, The University of Melbourne Parkville, VIC, Australia
- ⁷⁸Centre for Healthy Brain Ageing (CHeBA), School of Psychiatry, University of New South Wales, Sydney, Australia

Corresponding author

Prof. Bernhard Baune, PhD, MD, MPH, MBA, FRANZCP
Head, Department of Mental Health,

Director, Lab division of Molecular Neurobiology of Mental Health
University of Münster, Albert-Schweitzer Campus 1, Building A 9, 48149 Münster, Germany
T +49251-83-56664, Email: Bernhard.Baune@ukmuenster.de

Word count: 4,090

Abstract

Background. Response to lithium in patients with bipolar disorder is associated with clinical and genetic factors. However, the predictive combination of these variables is lacking. This study aimed to assess whether a combined analysis of genetic and clinical factors improves lithium response prediction when compared to genetic or clinical factors alone. In addition, we assessed whether the *a-priori* genetic stratification of patients improved the predictive ability of machine learning models trained with only clinical data.

Methods. This study utilised genetic and clinical data (N=1,034) collected as part of the International Consortium on Lithium Genetics (ConLi+Gen) project. Polygenic scores (PRS) were computed for schizophrenia (SCZ) and major depressive disorder (MDD) and then combined with clinical variables using a cross-validated machine learning (ML) regression approach. Unimodal, multimodal, and genetically stratified models were trained and validated using ridge, elastic net, and random forest regression on 692 bipolar patients from 10 study sites. All models were then tested using leave site out cross-validation on an independent test set of 342 patients. The best performing models were then also tested in a classification framework.

Results. The best performing linear model explained 5.1% ($p = 0.0001$) of variance in lithium response and was composed of clinical variables, PRS variables, and interaction terms between them. The best performing non-linear model used only clinical variables and explained 8.1% ($p = 0.0001$) of variance in lithium response, while the best performing model overall explained 13.7% ($p = 0.0001$) of variance in lithium response. This model *a-priori* stratified patients based on their meta-polygenic loadings for major depression and schizophrenia and was trained using clinical data. This model also improved the binary classification of lithium response.

Conclusions. Using PRS to first stratify patients genetically and then train ML models with clinical predictors led to large improvements in lithium response prediction.

Introduction

Lithium remains a commonly used first line treatment for bipolar disorder (BD)¹⁻³, highly effective for both acute manic episodes^{4,5} and maintenance treatment³. However, in around 65% of acute manic presentations, response is incomplete, whilst 35% of patients do not respond to treatment at all^{6,7}. In maintenance treatment, approximately 30% of patients report an excellent long-term response, around 30% report an intermediate response, and 30% respond poorly⁸. In addition, lithium has a range of serious acute and chronic side effects, including increased risk of renal failure and suppression of thyroid and parathyroid function². Moreover, lithium can be toxic at high doses so plasma levels often need to be monitored⁹. These varying response rates and side-effect profiles suggest the need to better tailor lithium treatment for individual patients, ensuring timely prescription of the right drug for the right patient at the right time. Better understanding the link between genetic and clinical factors may assist in achieving this personalised approach.

Regarding associated clinical factors, variables reflecting an episodic pattern of mania-depression intervals, a later age of onset and fewer hospitalizations preceding treatment have shown significant associations with lithium response^{10,11}. Using a large range of clinical factors, Nunes et al¹² demonstrated the ability of machine learning (ML) models to classify lithium responders from non-responders. Beyond these clinical factors, a genetic basis to lithium response has also been found. In genome wide association studies (GWAS), multiple genetic variants have been associated with lithium response. However, the effects of these variants have been too small to facilitate lithium response prediction⁸. Combining these variants into polygenic risk scores (PRS) has improved their performance, however, these scores still only explain ~ 1% of variance in lithium response¹³.

Given this biopsychosocial basis to lithium response, one solution may be to combine clinical factors with PRS to improve lithium response prediction. This genotype-phenotype approach was recently used by Antonucci et al¹⁴ to classify schizophrenia patients from healthy controls. Using environmental and genetic data, they *a-priori* stratified patients into tertiles (thirds) based on decision scores from two support vector machines (SVM). Making predictions with patients in the lower and upper tertiles led to an increase in balanced accuracy from 77.7% to 89.4%.

In the context of lithium response prediction, Amare et al found evidence for a non-linear stratified relationship between schizophrenia (SCZ) PRS and lithium response¹³. In particular,

patients in the lowest decile of the schizophrenia (SCZ) PRS distribution were 3.46 times more likely to be lithium responders when compared to patients in the 10th decile. In addition, a link between polygenic risk for major depressive disorder (MDD) and lithium response has also recently been discovered¹⁵. This transdiagnostic and polygenic basis to lithium response is not without precedent. For example, a multitude of studies have shown significant genetic overlap and shared biological pathways between SCZ, MDD, and BD¹⁶⁻²². Exploiting this genetic overlap, Maier et al²¹ used multi-trait models to utilise correlations between, as well as a subjects individual risk for SCZ, MDD, and BD. This multivariate approach led to an equivalent increase in sample size of 34% for SCZ, 68% for BD, and 76% for MDD when compared to single trait models.

Given these findings, we have conducted a range of analyses to test the predictive ability of combined genetic and clinical data for lithium response prediction. First, to measure the predictive contribution of PRS alongside clinical variables, we trained both uni- and multimodal prediction models of lithium response in BD patients. Second, we trained uni- and multimodal models containing interaction terms within and across variables from each data modality to measure non-linear and biopsychosocial effects between each modality. Fourth, to measure the effects of a patient's PRS loadings on clinical model accuracy¹⁴, we used MDD and SCZ PRS, as well as a combined MDD and SCZ meta-PRS to *a-priori* stratify patients according to their polygenic loadings prior to the supervised prediction of lithium response with clinical data. This approach was then directly compared to the traditional method of including PRS and clinical data modalities together in a single predictive model²³. Finally, to test the effects of model linearity on prediction, we compared the use of linear and non-linear ML models for regression analyses and validated all findings on a geographically stratified test set. We then assessed the best performing models in a classification framework.

Methods

The International Consortium on Lithium Genetics (ConLi⁺Gen)

The ConLi⁺Gen Consortium (www.ConLiGen.org) is an initiative by the National Institute of Mental Health (NIMH) and the International Group for the Study of Lithium-Treated Patients (IGSLI) (www.IGSLI.org) and was established with the aim of studying the genetic basis of lithium treatment response in patients with BD²⁴. The ConLi⁺Gen study involved patients with BD from Europe, South America, USA, Asia, and Australia²⁵ who have been treated

with lithium. A series of quality control procedures were implemented on the genotype data before and after imputation as described below. Sample characteristics have been published in previous works^{8,13}.

Computing polygenetic scores

PRS were computed for schizophrenia (SCZ) and major depressive disorder (MDD), two traits previously associated with lithium response^{13,15}. Each PRS was calculated using individual genetic data from the International Consortium on Lithium Genetics (ConLi+Gen; www.ConLiGen.org)⁸ and summary statistics from the previous largest GWASs available for MDD²⁶, and SCZ¹⁹. PRS were calculated at different GWAS p-value thresholds, however, the best predictive score was selected for each trait based on previous analysis. More details on PRS calculation, genotyping, imputation and quality control steps can be found elsewhere^{13,15,27}.

Study sample

As our aim was to assess both uni- and multimodal regression models of lithium response, a requirement for inclusion was complete PRS data and no more than 20% missingness on clinical predictors. Any clinical predictors above this threshold were removed. Only ConLi+Gen GWAS 1 was used for analysis (N = 1,163) as it contained both clinical and genetic data, whereas GWAS 2 only contained genetic data. Fifteen demographic, clinical, substance use, and comorbid psychiatric illness predictors were used in analyses. To ensure geographic homogeneity, only samples of European descent were used (Halifax N = 240, University California San Diego N = 216, Cagliari N = 196, Poznan N = 97, Wuerzburg N = 91, Geneva N = 46, Prague N = 45, Dresden N = 43, National Institute for Mental Health = 36, John Hopkins University = 24), leaving a final sample of N = 1,034 patients. To ensure the unbiased approximation of the model's generalisability to new patients, we partitioned 33% (N = 342) of our sample into a holdout set for model testing, leaving 692 observations for training and validation. This partitioning was stratified by data collection site to ensure the same distribution of sites across partitions.

Regression and classification target

Lithium treatment response was assessed using the validated "Retrospective Criteria of Long-Term Treatment Response in Research Subjects with Bipolar Disorder" scale, also known as the *Alda* scale²⁸⁻³⁰. To arrive at a total Alda score, this scale measures symptom improvement

over the course of treatment (A score, range 0–10), which is then weighted against five criteria (B score) that assesses the quality of evidence for the response score⁸. For the predictive regression analyses, the total Alda score was used³¹. For a subset of models assessed in a classification framework, patients with a score of greater than or equal to seven were coded as responders, while patients scoring less than seven were coded as non-responders³¹.

Unimodal and multimodal machine learning pipelines

On our training sample of 692 patients, we fit pipelines that conducted imputation, polynomial feature engineering (interaction terms only), standardization, feature selection, hyperparameter optimization, and the fitting of linear regression (LR) (regularized with ridge and the elastic net) and random forest (RF) models. For the unimodal LR models, we imputed predictors using multivariate imputation by chained equations (MICE) with the 10 nearest predictors used in the imputation process³². As regularized LR models may perform poorly if variables are on different scales, we standardized all predictors to have a mean of zero and a standard deviation of one. As the number of predictors were low in the unimodal and multimodal PRS and clinical models (PRS = 9, clinical = 15), all predictors were included in analyses. Finally, we fit a LR model with ridge regularization to the training sample. For the interaction term models, interaction terms between all predictors were engineered in the pipeline prior to feature selection and hyperparameter tuning. As the number of predictor variables grew exponentially in these analyses, we now conducted feature selection with the elastic net, a form of penalized regression that removes highly correlated predictors from the model whilst retaining the most predictive subset for model fitting³³. This same procedure was then repeated using an RF model, however, as RF models are scale invariant, we did not scale the predictors prior to model training³⁴. Instead of regularized linear predictor selection, we conducted non-linear predictor selection according to the mean decrease in variance provided by each predictor in the RF model³⁴. For the subset of classification models, equivalent classifiers replaced the regression models in each pipeline.

For all pipelines, we used a random search of 60 iterations to tune model hyperparameters. When less than 60 hyperparameter combinations were present, we used an exhaustive grid search. See supplementary Table 1 for the tuned hyperparameter values. This process and all steps inside the pipelines were completed using leave site-out cross-validation. This method trains on all data collection sites minus one. The excluded site is then used to assess the

selected features and hyperparameters, with the combination that minimises the root mean squared error on the held-out site selected³⁵. As there were 10 collection sites, this equates to 10-fold cross-validation for model selection. This site-based stratification protects against the optimization of hyperparameters and selection of features that may proxy for disparities in feature and outcome distributions across sites and result in ungeneralizable estimates³⁶. All tuned and selected models from training and validation were then further tested in the *a-priori* held-out set of 342 patients.

Stratified meta-PRS analyses

Stratification was conducted using the MDD, SCZ, and a combined MDD and SCZ meta-PRS. For this combined meta-PRS, the MDD and SCZ PRS were first standardised to have a mean of zero and a standard deviation of one and then summed together into a single composite PRS score. These scores were then re-standardised resulting in a meta-PRS representing polygenic loading for MDD and SCZ. Patients with polygenic loadings in the lower and upper quartiles of each PRS distribution were then selected (25th percentile N = 259, 75th percentile N = 259). As per the previous analyses, 33% (N = 171) of patients were partitioned for testing, leaving 347 patients for training and validation. Following this genetic stratification, we re-ran regression analyses using only clinical predictors and then tested the statistical significance of each model using permutation tests ($m = 10,000$). All p-values were FDR corrected using the Benjamini and Hochberg method³⁷.

Supplementary analyses

As previous works have shown that smaller samples have a higher risk of misestimating model performance^{38,39}, we used a Monte-Carlo procedure to control for the decreases in sample size in the stratified analyses (Supplementary methods 1). In addition, we used Pearson's r to measure whether changes in the number of predictor variables were influencing model R² scores⁴⁰ (Supplementary methods 2).

Results

Sample characteristics

The final analysis cohort contained 1,034 patients with an average age (SD) of 47.7 (14) years and an average bipolar disorder onset age of 24.9 (11) years. Of these patients, 627 (60.6%) were male and 801 (77.7%) were classified as having bipolar disorder 1. The average ALDA score for Lithium response was 4.3 (3.3) out of 10. See Figure 2 (bottom left)

for the full distribution of ALDA scores. See Table 1 for more information on sample characteristics.

Unimodal and multimodal models

According to the coefficient of determination (R^2), the unimodal LR PRS and clinical models explained 1.2% and 1.8% of variance in lithium response respectively, whilst the combined multimodal model explained 4.7% of variance in lithium response. Re-running the 3 models including interaction terms between all variables resulted in 1.4%, 4.5%, and 5.1% explained variance. For the non-linear RF models, the unimodal PRS and clinical models explained 2% and 8.1% of variance in lithium response, whilst the combined multimodal model explained 7.4% of variance in lithium response. Re-running the 3 models and including interaction terms between all variables resulted in -0.9%, 6.7%, and 5.2% explained variance.

Stratified PRS analyses

For the stratified analysis using patients in the upper and lower quartiles of the MDD PRS distribution, the clinical linear and clinical linear interaction models explained -2.8% and 2.7% of variance in Lithium response, whilst the non-linear RF and RF interaction models explained 3.5% and 1.8% of variance. For the stratified SCZ PRS analyses, the clinical linear and clinical linear interaction models explained 7.1% and 9% of variance in Lithium response, whilst the non-linear RF and RF interaction models explained 7.2% and 9.3% of variance. Finally, for the stratified meta-PRS analyses, the clinical linear and clinical linear interaction models explained 12.1% and 9.2% of variance in lithium response, whilst the non-linear RF and RF interaction models explained 13.7% and 4.5% of variance. All models were statistically significant after FDR corrections (Table 2).

Completing 1000 runs of the Monte-Carlo sampling procedure to control for decreases in sample size on the stratified meta-PRS model, we attained an average $R^2 = 2.7\%$ ($sd = 5$, $p = 0.002$). Therefore, the superior performance of our meta-PRS stratified model was not explainable by increased performance variability resulting from decreased sample size³⁸. In addition, increases in R^2 were not explained by changes in the number of predictor variables across models (Pearson's $r = 0.17$, $p = 0.44$) (Figure 2, top right). See table 2 for results and supplementary tables 2 – 5 for all model metrics. After controlling for these confounds, the best performing meta-PRS stratified model explained 69% more variance ($R^2 = 13.7\%$, $p = 0.0001$) than the equivalent model containing no *a-priori* meta-PRS stratification ($R^2 = 8.1\%$, $p = 0.0001$). In this model, all clinical variables were retained in model selection (See Table 4

for the variables and their Gini importance scores). See Figure 2 for all a visualisation of model R2 scores. Re-running these two best performing models in a classification framework led to balanced accuracies of 58.95% and 63.65% respectively. See supplementary table 6 for all classification metrics.

Patient characteristics in the genetically stratified cohort

After FDR corrections, significant differences in clinical characteristics were found between those in quartiles 1 (low meta-PRS load) and 4 (high meta-PRS load) of the combined meta-PRS distribution for binary lithium response (ALDA ≥ 7) ($X^2 = 12.214$, $p = 0.005$), bipolar 1 vs rest (bipolar 2 and schizo-affective) ($X^2 = 12.755$, $p = 0.005$), and DSM diagnosis (bipolar 1, bipolar 2, and schizo-affective) ($X^2 = 13.33$, $p = 0.027$). In quartile 1 of the meta-PRS distribution, 70% had bipolar 1, 26% had bipolar 2, and 4% had schizo-affective bipolar. In total, 39% of these patients were lithium responders. In quartile 4, 86% had bipolar 1, 12% had bipolar 2, and 3% had schizo-affective bipolar. In total, 22% of these patients were lithium responders. Overall, those in quartile 1 were 67.7% more likely to be lithium responders compared to those in quartile 4 (OR = 1.677, CI = 1.14 – 2.47, $p = 0.009$). For all other clinical characteristics, including variables that attained nominal significance, see Table 3.

Discussion

This is the first study to provide evidence for the combined predictive ability of routine clinical data and PRS for lithium response. Specifically, we show that first using PRS to stratify patients according to their polygenic loadings, followed by training with clinical data explains more variance in lithium response and improves model accuracy in a classification setting^{13,14}. Interestingly, the combination of PRS with clinical data performed best in the linear models, but not in the non-linear models. Outside of the best performing stratified meta-PRS model, neither of these multimodal models performed best overall. Moreover, unimodal clinical models outperformed their PRS equivalents.

This observation of clinical variables outperforming their biological counterparts has been repeatedly demonstrated across a range of multimodal machine learning studies^{23,41}. The most intuitive explanation is that the small effect sizes yielded by biological variables, when compared to clinical variables, leads to overfitting, and/or their lack of selection in cross-validation, resulting in underperformance for biomarker models when tested out of sample. The next consideration is why effects are smaller for biological variables? To answer this, we

need to consider how psychiatric traits are constructed and the implications this has for studies attempting to elucidate a biological basis for psychiatric phenomenon. In comparison to other disorders, psychiatric phenotypes are defined by deviations from normative behaviours and emotional-cognitive experiences, rather than from well-defined physiological processes³⁶. Therefore, it is plausible that they bias toward larger effect sizes for clinical variables that correlate with clinical data already used in the construction of phenotypes and illness trajectories. Consequently, this tautology in the formation of diagnostic and prognostic constructs may limit the predictive contribution of biological data. In theory, this problem could be circumvented by first parsing patient heterogeneity at the biological level, and then using clinical variables in secondary analyses.

This rationale informed our stratified analyses where we *a-priori* partitioned patients based on their polygenic loadings for major depression, schizophrenia, and their combination in the form of a standardised meta-PRS. Interestingly, this method was most predictive of lithium response overall, explaining 69% more variance than the equivalent model with no *a-priori* meta-PRS stratification. These results support the view that first parsing biological heterogeneity, and specifically in our case, underlying genetic heterogeneity, may improve the prediction of BD lithium response with clinical data.

When assessing the best performing non-stratified model - the clinical random forest model, and the best performing overall model - the meta-PRS stratified model, we also observed an increase in model performance in a classification framework, albeit a smaller percentage change. This observation warranted an inspection and interpretation of the lithium response distributions between patients in the low and high meta-PRS quartiles (Figure 2, bottom left and right). Here, we observe disproportionate densities of very low (0) and moderate response (5 – 7) scores for patients with high meta-PRS loadings. Conversely, for patients with low meta-PRS loadings, we observe disproportionate densities of very low (0) and very high response (7 – 10) scores. Both quartiles of meta-PRS loadings demonstrated high densities of very low response, yet differences between moderate and high response scores were evident. More specifically, patients with low meta-PRS loadings belonged to a continuous bimodal response distribution, whereas those with high meta-PRS loadings appeared to be mixed across the distribution and skewed towards lower response. When dichotomising lithium response, this nuanced understanding between a patient's genetic loadings and lithium response was lost.

This observation is interesting in light of recent work that quantified the asymmetrical reliability of the Alda scale, finding higher inter-rater reliability in the upper tail of the response distribution⁴². Therefore, a dichotomous representation of lithium response was generally argued for, even after considering the resultant loss in statistical power. However, rather than deciding *a-priori* to discretise this distribution, an alternative approach would be to tune and select models in a leave-site out cross-validation framework, as was done in the current work. This is because we would expect to see the highest amount of inter-rater disagreement between data collection sites, as purported by Nunes et al⁴², and if it was high enough to warrant *a-priori* discretisation, these across site models would not generalise due to their disagreement in lithium response. However, in the current work nearly all models generalised across sites to the out of sample test sets that were excluded from model construction, demonstrating that the use of leave-site out cross-validation ensured that each model was tuned to learn parameters and relationships that generalised regardless of any disagreement between raters across sites. In addition, this established that there was enough agreement between raters to learn meaningful, informative, and generalisable patterns in the continuous lithium response distribution.

In future works, an alternative to dichotomising the Alda scale would be to use the full scale and run analyses using spline regression⁴³. With this technique, we would not build one model for the entire dataset, but instead, divide the dataset into multiple bins and fit each bin with its own model. Some of these models may be linear, whilst others may be polynomial. This approach would allow us to fit PRS to the lithium response distribution and account for the linear and non-linear relationships between different strata of the PRS and lithium response distributions⁴³.

Regarding the clinical characteristics of patients in each meta-PRS quartile, we observed significant differences between the types of psychiatric diagnosis. Quartile 1 (low meta-PRS load) had lower proportions of BD 1 diagnoses and higher proportions of BD 2 and SAD, whereas the opposite was true for those in quartile 4 (high meta-PRS load). Given that higher meta-PRS loadings are associated with poorer lithium response, and that people with ‘purer’ forms of BD1 are considered better lithium responders^{2,44,45}, this is an unexpected finding: one might have hypothesized that there would be higher proportion of BD1 in the low, and therefore less ‘contaminated’, meta-PRS group. Our finding suggests that the relationship between meta-polygenic disposition for SCZ and MDD and actual phenotypical expression of bipolar spectrum disorders is more complex⁴⁶, and that people with seemingly unfavourable

genetic constellations may still benefit from lithium once other clinical and environmental parameters come into play. Similarly, patients with seemingly less favourable diagnoses for lithium response (i.e., BD2 and SAD) may still benefit if their polygenic disposition points towards better responsiveness.

A number of limitations exist in the current work. Firstly, there was a limited amount of clinical data available for analysis in this cohort. Future studies should aim to collect a larger amount of clinical data (e.g., symptom scales) to elucidate the relationship between PRS and clinical characteristics, as well as their combined predictive ability. Ideally, prospective studies of lithium response will be required in the future to quantify the predictive ability of machine learning models in an environment that is analogous to clinical practice. Second, correctly operationalising BP1, BP2, and SABP DSM phenotypes is difficult in real world practice. Relying on patient's retrospective reporting of symptoms and past episodes to form these diagnoses, as was done in the current study, can lead to misdiagnosis of BP subtypes^{2,3,47}. A more informative approach might be to assess symptom dimensions (e.g., the presence of psychosis, mania, anxiety etc), however, such deep phenotyping is yet to be implemented in this cohort. In addition, the strength of these findings is limited by the validity of the Alda scale, which has demonstrated vulnerability to measurement error in the composite B score used to derive the final total score³¹.

Third, it has been shown that heterogeneity between sites presents as both a strength and a limitation³⁹. For example, training models to learn generalisable patterns across sites results in a model with high external validity and generalisability, but less predictive power.

Alternatively, models trained on individual sites will likely be more homogenous in their clinical presentation and have a greater consistency in variables that moderate response (e.g., demographic, geographic, and socio-economic effects, as well as selection effects for certain patients, i.e., some clinics/sites may specialise in only treating certain BP subtypes). Models trained on these unique profiles will likely contain less noise and be more predictive, however, will only generalise to patients that meet the required clinical parameters. With enough data and a larger range of clinical predictors, multiple predictive models could be trained for different patient subsets, helping to overcome this problem^{36,48}.

Another consideration is the selection of quantile based PRS stratification over tertile¹² and decile¹³ stratification used in previous works. Choosing the number of PRS strata involves considering the trade-offs between a higher number of bins (i.e., decile stratification) that

would likely contain larger differences in clinical characteristics, lithium response, and polygenic risk, but result in a smaller sample (only 20% of the original sample would be retained when taking the extreme deciles). Alternatively, a lower number of bins (i.e., tertiles) would likely result in the opposite being true. To balance this trade-off, we chose quartile-based stratification. When taking the two extreme quartiles, we retained 50% of the original sample, while removing the middle of the PRS distributions that shows the smallest genetic differences in lithium response^{13,15}. However, future studies could attempt to find the optimal number of strata through the use of cross-validation with spline regression, where the optimal number of strata could be tuned and selected according to the minimisation of a loss function. If completed in a leave-site out framework, between site rater disagreement would be controlled for and the full lithium response distribution could be modelled.

The next limitation pertains to the patients that do not fall in the tails of the PRS distributions and who would therefore be excluded from prediction with this model. However, such a stratified model that confers superior predictive ability could first be used for patients that fall within these strata, and for patients that do not, models without stratification could be used¹², or other stratifying biomarkers could be incorporated⁴⁹. Through this lens, we envision a stepwise process in clinical deployment where the choice of model itself would be tailored to individual patients depending on their unique clinical and biological characteristics (Figure 4).

An alternative approach to parse biological heterogeneity would be to use unsupervised ML models⁵⁰⁻⁵². However, the disproportionately small effect sizes afforded by PRS^{53,54}, the large risk of overfitting on unlabelled data⁵⁵, the high level of polygenic collinearity across psychiatric traits^{16,56}, and the resultant demands these considerations impose on statistical power⁵⁷, led us to take a simpler approach informed by previous findings^{13,14}.

In conclusion, using PRS to stratify patients genetically and then train ML models with clinical predictors led to large improvements in lithium response prediction over other forms of unimodal and multimodal modelling. Clinical data explained the most variance and both clinical and PRS data showed non-linear relationships with lithium response. To adequately model the linear and non-linear relationships between these PRS and lithium response across different genetic strata, future works should consider modelling these relationships using spline regression. Moreover, engineering a direct lithium response PRS and using this to parse heterogeneity may further improve model performance. In addition, parsing

heterogeneity with biomarkers from neuroimaging and omics domains should also be considered. Finally, datasets with a larger range of clinical variables will likely improve prediction following genetic stratification.

Funding: Dr Amare has received 2020-2022 NARSAD Young Investigator Grant from the Brain & Behaviour Research Foundation.

Table 1. Characteristics of bipolar disorder patients included in analyses stratified by their study site. Statistics were calculated using Fischer exact tests for categorical variables and one-way ANOVA for continuous variables. All p-values were FDR corrected.

	Total	Cagliari /Sardinian	Dresden	Geneva	Halifax	JHU	NIMH	Poznan	Prague /Czech	San Diego /UCSD	Wuerzburg	Statistic	P
N	1,034	196	43	46	240	24	36	97	45	216	91	-	-
Age at interview, Mean (sd)	47.7 (14.0)	46.0 (14.7)	43.5 (14.4)	46.9 (9.2)	48.5 (13.4)	44.0 (10.0)	45.6 (12.8)	61.9 (10.4)	44.8 (14.6)	45.0 (12.9)	-	15.4	4.5x10 ⁻²¹
Age of onset BD, Mean (sd)	24.9 (11.0)	27.1 (11.0)	25.4 (9.1)	21.4 (9.0)	26.1 (9.8)	21.7 (9.6)	19.9 (6.8)	31.5 (10.2)	28.8 (10.7)	17.6 (9.7)	28.8 (11.0)	21.9	1.3x10 ⁻³³
Age of onset MDD, Mean (sd)	26.2 (12.2)	27.1 (14.3)	26.9 (11.5)	26.6 (10.1)	28.6 (11.0)	25.6 (12.0)	22.2 (10.7)	32.4 (10.2)	29.4 (10.9)	19.4 (10.1)	30.0 (10.4)	13.5	3.7x10 ⁻²⁰
Alda scale total score, Mean (sd)	4.3 (3.3)	4.1 (3.1)	3.7 (3.2)	3.5 (3.0)	5.9 (3.6)	4.9 (3.1)	4.2 (2.7)	6.2 (2.8)	4.2 (3.0)	2.9 (2.6)	2.8 (3.0)	19.4	9.1x10 ⁻³⁰
Lithium responder (Yes)	31.9	28.1	25.6	21.7	57.5	33.3	27.8	48.5	28.9	10.6	16.5	144.3	3.1x10 ⁻²⁶
Bipolar 1 Vs Rest,	77.7	75.0	72.1	78.3	72.1	100.0	91.7	82.5	86.7	86.6	58.2	49.9	1.4x10 ⁻⁰⁷
Bipolar 1 SABD Vs rest,	79.4	75.0	72.1	78.3	73.8	100.0	97.2	82.5	86.7	92.1	58.2	70.0	2.3x10 ⁻¹¹
Alcohol dep (Yes),	17.5	9.7	7.3	30.4	2.9	100.0	16.7	12.5	20.0	37.0	0.0	170.1	3.7x10 ⁻³²
Substance dep (Yes),	16.8	8.7	2.4	28.3	0.4	100.0	5.6	52.5	2.2	30.5	0.0	231.4	1.0x10 ⁻⁴⁴
OCD (Yes),	5.1	0.0	7.3	8.7	2.5	100.0	13.9	7.5	0.0	9.4	0.0	69.8	8.6x10 ⁻¹²
PTSD (Yes),	7.2	0.0	0.0	4.4	0.0	0.0	0.0	20.0	2.2	20.3	0.0	112.9	1.1x10 ⁻²¹
Panic disorder (Yes),	13.8	0.0	12.2	13.0	6.3	100.0	25.0	53.8	8.9	17.5	0.0	189.1	4.9x10 ⁻³⁶
Sex (Male),	60.6	67.9	53.5	54.3	57.9	70.8	77.8	61.9	66.7	50.5	69.2	25.1	2.8x10 ⁻⁰³
Polarity 1st episode (depression onset),	63.3	66.3	58.1	47.8	62.5	54.2	63.9	52.6	84.4	61.6	76.9	28.0	1.0x10 ⁻⁰³
Polarity 1st episode (mania onset),	17.9	31.1	16.3	6.5	13.8	45.8	30.6	12.4	8.9	15.3	11.0	55.4	1.4x10 ⁻⁰⁸
Polarity 1st episode (hypomania onset),	6.5	2.6	7.0	4.3	3.3	0.0	0.0	1.0	4.4	21.3	0.0	103.0	6.4x10 ⁻¹⁸
Any suicidal features (No),	41.9	63.8	0.0	19.6	43.3	62.5	58.3	62.9	31.1	19.0	47.3	154.7	2.2x10 ⁻²⁸
Any suicidal features (Yes),	43.4	35.7	41.9	73.9	17.1	37.5	41.7	19.6	68.9	75.9	52.7	220.8	7.8x10 ⁻⁴²
Any suicidal features (Unknown),	14.7	0.5	58.1	6.5	39.6	0.0	0.0	17.5	0.0	5.1	0.0	267.4	1.9x10 ⁻⁵¹
DSM Dx (BD1),	77.7	75.0	72.1	78.3	72.1	100.0	91.7	82.5	86.7	86.6	58.2	49.9	1.4x10 ⁻⁰⁷
DSM Dx (BD2),	19.6	25.0	27.9	21.7	26.3	0.0	2.8	17.5	13.3	6.9	33.0	58.3	4.1x10 ⁻⁰⁹
DSM Dx (SABD),	1.7	0.0	0.0	0.0	1.7	0.0	5.6	0.0	0.0	5.6	0.0	31.0	3.1x10 ⁻⁰⁴
DSM Dx (BD3),	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.3	31.2	3.1x10 ⁻⁰⁴
DSM Dx (BDNOS),	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	5.5	36.6	3.7x10 ⁻⁰⁵
BD family history (Y),	33.9	42.3	72.1	0.0	18.3	0.0	33.3	43.3	22.2	59.7	0.0	213.5	2.23x10 ⁻⁴⁰
BD family history (N),	37.3	57.7	20.9	0.0	64.2	0.0	0.0	39.2	77.8	17.1	0.0	300.1	3.42x10 ⁻⁵⁸
BD family history (Unknown),	28.7	0.0	7.0	100.0	17.5	100.0	66.7	17.5	0.0	23.1	100.0	555.9	1.65x10 ⁻¹¹²

Table 2. Train/validation (N = 692), and test (N = 342) results across all models. Unimodal, multimodal, and interaction term predictors spaces were measured using both linear regression (ridge and the elastic net) and random forest regression models. In addition, PRS stratified models composed of MDD PRS, SCZ PRS, and their standardised combinations in the form of a meta-PRS were assessed across model types and feature interaction combinations. All p-values were FDR corrected with the Benjamini and Hochberg method.

Linear (Ridge and Elastic Net Regression)				
	Train Mean (SD)	Validate (SD)	Test	P-Value
PRS	0.01 (0.0)	0.20 (0.18)	0.012	0.0131
Interaction PRS	0.01 (0.0)	0.20 (0.18)	0.014	0.01
Clinical	0.12 (0.02)	0.32 (0.17)	0.018	0.0004
Interaction Clinical	0.12 (0.02)	0.26 (0.19)	0.045	0.0001
Clinical and PRS	0.09 (0.02)	0.20 (0.17)	0.047	0.0001
Interaction Clinical and PRS	0.13 (0.02)	0.27 (0.19)	0.051	0.0001
Random Forest Regression				
	Train Mean (SD)	Validate (SD)	Test	P-Value
PRS	0.09 (0.01)	0.22 (0.18)	0.02	0.0001
Interaction PRS	0.09 (0.01)	0.22 (0.19)	0.009	0.0001
Clinical	0.21 (0.02)	0.21 (0.11)	0.081	0.0001
Interaction Clinical	0.24 (0.02)	0.23 (0.14)	0.067	0.0001
Clin and PRS	0.22 (0.02)	0.21 (0.12)	0.074	0.0001
Interaction Clinical and PRS	0.25 (0.02)	0.23 (0.16)	0.052	0.0001
A-Priori PRS Stratified Regression (Trained with clinical predictors)				
	Train Mean (SD)	Validate (SD)	Test	P-Value
Linear MDD stratified	0.12 (0.01)	0.12 (0.31)	-0.028	0.0171
Linear interaction MDD-PRS stratified	0.13 (0.01)	0.21 (0.36)	0.027	0.0013
Linear SCZ-PRS stratified	0.12 (0.01)	0.41 (0.42)	0.071	0.0004
Linear interaction SCZ-PRS stratified	0.12 (0.01)	0.31 (0.21)	0.09	0.0001
Linear Meta-PRS stratified	0.09 (0.01)	0.67 (0.39)	0.121	0.0001
Linear interaction Meta-PRS stratified	0.10 (0.01)	0.51 (0.34)	0.092	0.0001
Random Forest MDD-PRS stratified	0.18 (0.01)	0.16 (0.22)	0.035	0.0001
Random Forest interaction MDD-PRS stratified	0.23 (0.01)	0.20 (0.29)	0.018	0.0001
Random Forest SCZ-PRS stratified	0.24 (0.01)	0.24 (0.14)	0.072	0.0001
Random Forest interaction SCZ-PRS stratified	0.27 (0.02)	0.26 (0.14)	0.093	0.0001
Random Forest Meta-PRS stratified	0.23 (0.02)	0.40 (0.29)	0.137	0.0001
Random Forest interaction Meta-PRS stratified	0.70 (0.01)	0.39 (0.35)	0.045	0.0001

Table 3. Descriptive statistics for the clinical profiles of bipolar patients in the lower and upper quartiles of the meta-PRS distribution. Nominal and FDR corrected p-values in bold.

	Quartile 1 (Low Met PRS)	Quartile 4 (High Met PRS)	Statistic	P Value	FDR P Value
ALDA scale total score	4.852 (3.412)	4.078 (3.171)	2.412	0.016	0.065
Lithium response (ALDA ≥ 7) %	61/39	78/22	12.214	0.001	0.005
Age of onset bipolar, mean (sd)	25.723 (11.606)	23.548 (10.784)	1.993	0.047	0.104
Age of onset depression, mean (sd)	27.520 (12.587)	25.790 (12.010)	1.427	0.154	0.237
Age at interview, mean (sd)	48.277 (14.052)	48.434 (14.010)	-0.114	0.909	0.909
Bipolar 1 vs rest (no/yes) %	30/70	14/86	12.755	0.000	0.005
Polarity first episode (depression/mania/hypomania onset)	74/17/10	67/21/12	1.961	0.375	0.491
Alcohol dependence (no/yes) %	87/13	78/22	4.567	0.033	0.093
Substance dependence (no/yes) %	87/13	82/18	1.916	0.166	0.238
OCD (no/yes) %	95/5	94/06	0.270	0.603	0.670
PTSD (no/yes) %	96/4	91/09	3.726	0.054	0.107
Panic disorder (no/yes) %	92/8	86/14	2.980	0.084	0.153
Any suicidal features (no/yes/unknown) %	46/39/15	39/52/09	6.986	0.030	0.093
Sex (female/male) %	42/58	46/54	0.552	0.458	0.539
DSM diagnosis (BD1/BD2/SADP) %	70/26/4	86/12/03	13.330	0.004	0.027
Bipolar family history (no/yes/unknown) %	42/46/12	42/41/16	1.869	0.393	0.491

*Calculated using independent samples t-tests for continuous variables and Chi-square tests for categorical variables. All p-values were FDR corrected using the Benjamini and Hochberg method.

Table 4. Clinical predictor variables from the best performing meta-PRS stratified model and the Gini feature importance coefficients for each predictor.

Predictor variable	Gini-importance
Age at interview	0.2238
Age of onset bipolar	0.2030
Any suicidal features	0.1486
Age of onset depression	0.1013
OCD	0.0481
Polarity 1 st episode	0.0401
Alcohol dependence	0.0368
Panic disorder	0.0335
Substance dependence	0.0331
Bipolar family history	0.0320
PTSD	0.0316
Bipolar 1 and schizo-affective bipolar vs rest	0.0242
Sex	0.0226
Bipolar 1 vs rest	0.0128
DSM diagnosis	0.0084

Figure 1. Analysis pipeline for the stratified analysis of Lithium response prediction.

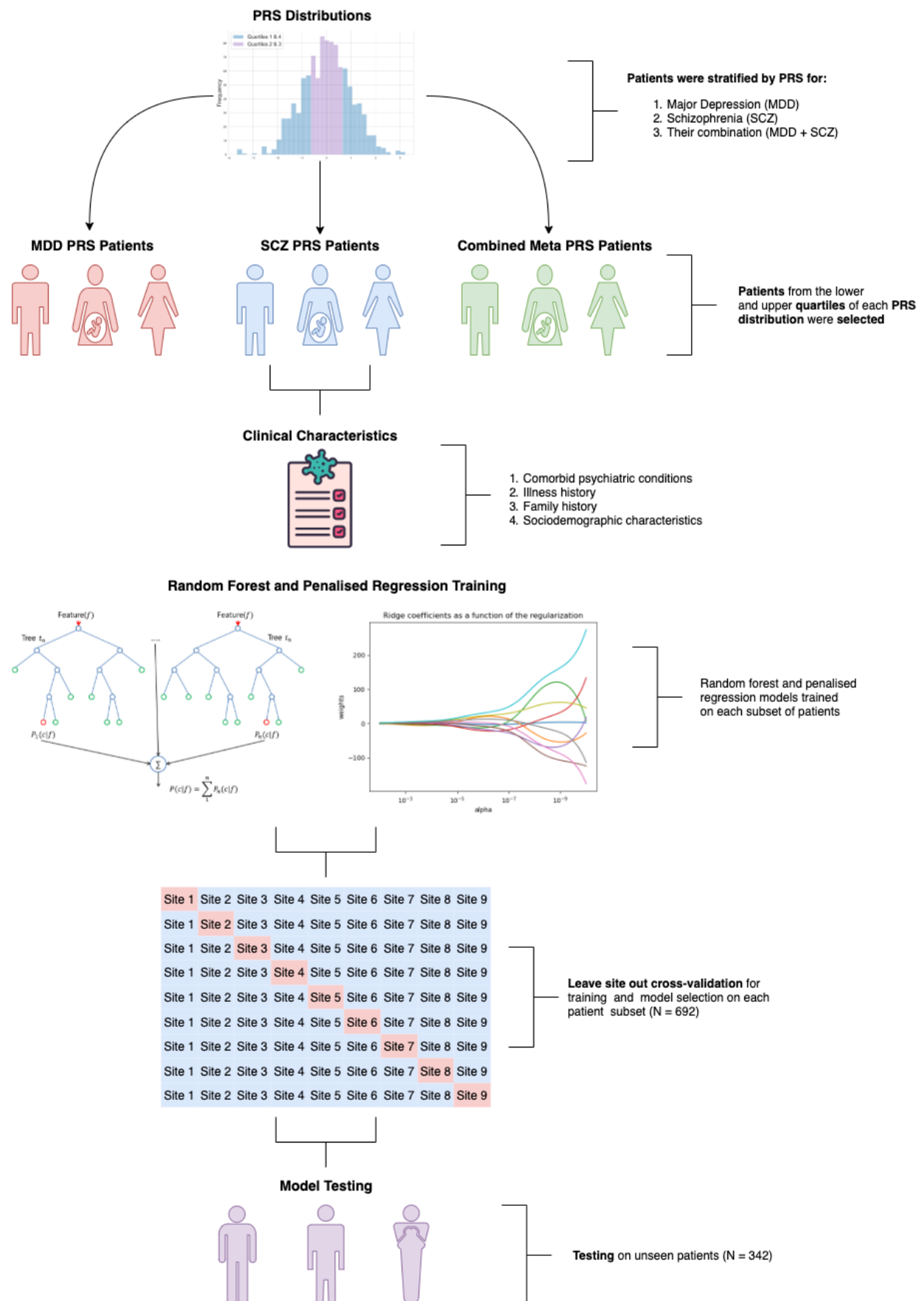


Figure 2. Top left: Meta-PRS loading distribution stratified into quartiles. Top right: Scatterplot and Pearson's r correlation between the number of predictor variables (x-axis) and test set R2 scores (y-axis) across all analyses. Bottom left: ALDA total score distribution used as the regression target in analyses. Bottom right: ALDA total score distribution stratified by combined meta-PRS load for patients in quartiles 1 and 4 of the distribution.

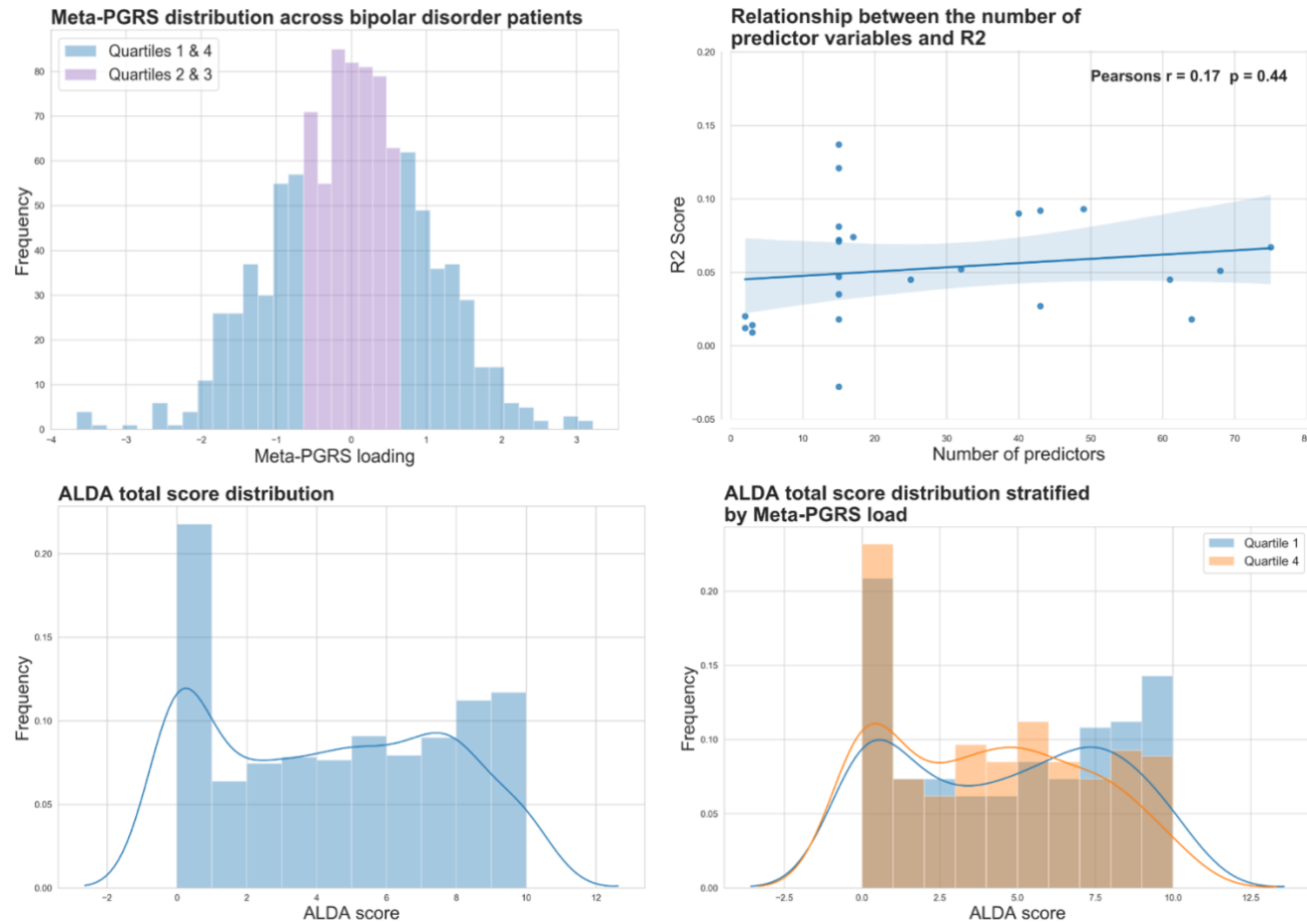


Figure 3. Line plot for all model R2 scores. Red dashed line: R2 benchmarks for Schizophrenia and Major depressive disorder PRS and their relationship with lithium response in previous works from Amare et al^{13,16}. Blue line: R2 scores for the non-PRS stratified models. Orange line: R2 scores for the PRS stratified models. Abbreviations: Inter = Interactions terms, Strat = Stratified, LR = Linear regression, RF = Random forest.

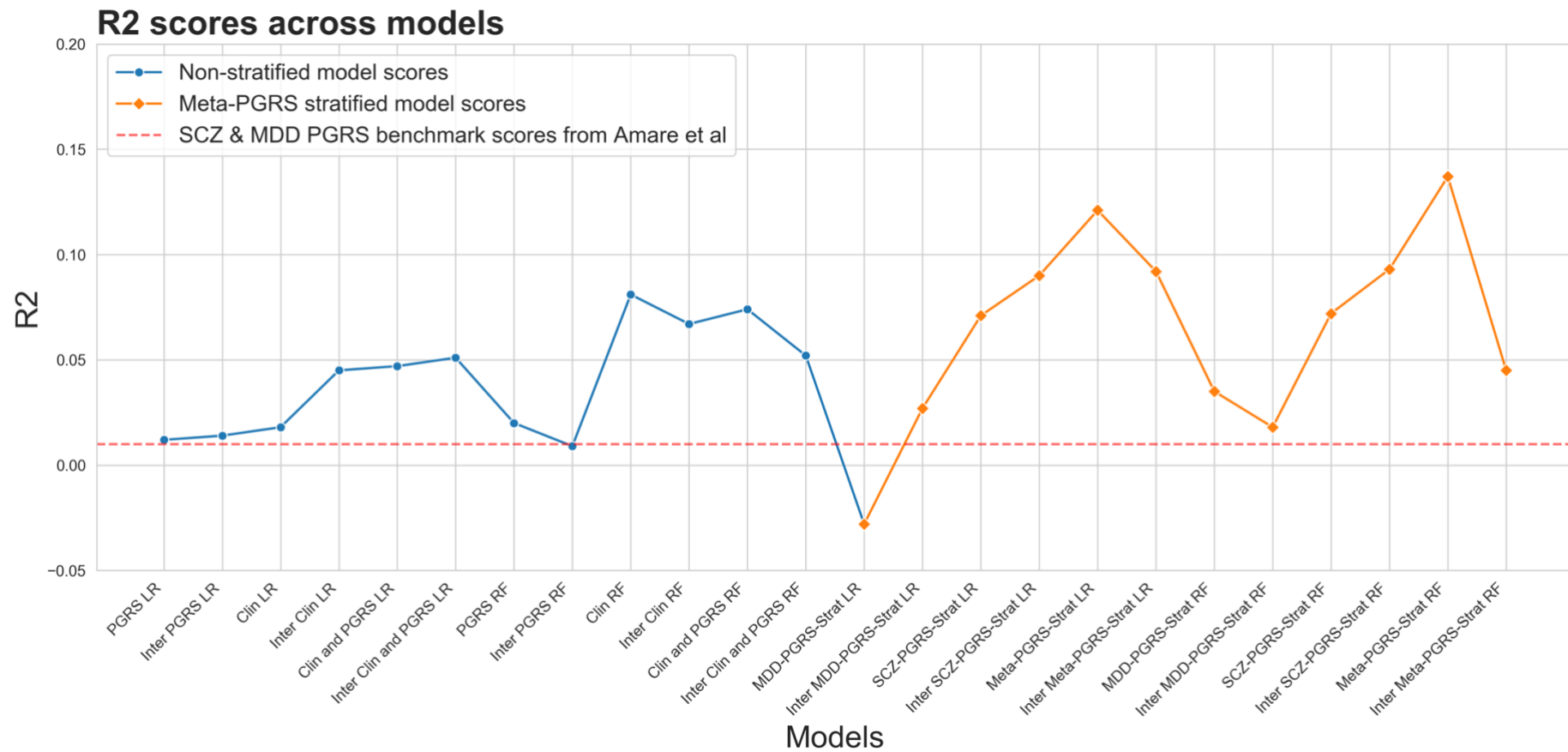
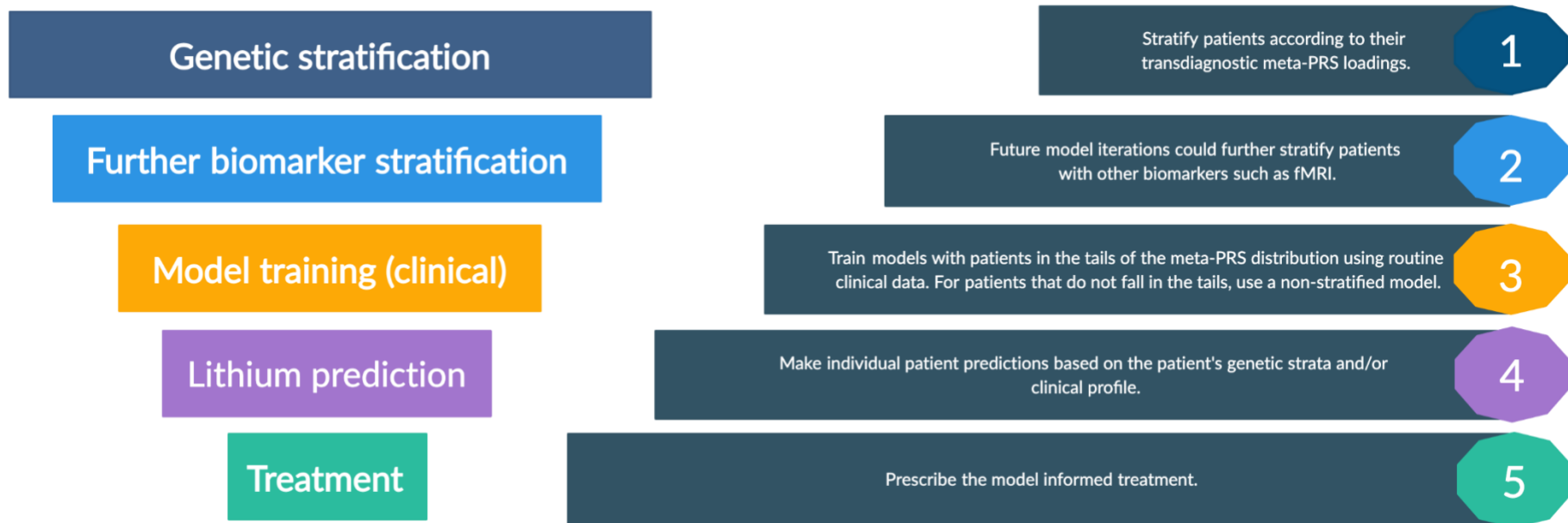


Figure 4. Schematic representation of the stepwise procedure to parse the transdiagnostic genetic heterogeneity of bipolar disorder using meta-PRS followed by the prediction of lithium response with clinical data. Level 2 of the figure (Further biomarker stratification) introduces the idea of using additional biomarkers (e.g., sMRI and fMRI) to further parse patient heterogeneity. In theory, this step would be secondary to genetic stratification as it would most closely follow the biological order of the system (i.e., genetic polymorphisms x environment > central nervous system abnormalities > emergent clinical symptoms / deviations in behavioural norms).



References

1. Miura T, Noma H, Furukawa TA, et al. Comparative efficacy and tolerability of pharmacological treatments in the maintenance treatment of bipolar disorder: a systematic review and network meta-analysis. *Lancet Psychiatry*. 2014;1(5):351-359.
2. Malhi GS, Tanious M, Das P, Berk M. The science and practice of lithium therapy. *Aust N Z J Psychiatry*. 2012;46(3):192-211.
3. Malhi GS, Adams D, Berk M. Is lithium in a class of its own? A brief profile of its clinical use. *Australian & New Zealand Journal of Psychiatry*. 2009;43(12):1096-1104.
4. Cipriani A, Barbui C, Salanti G, et al. Comparative efficacy and acceptability of antimanic drugs in acute mania: a multiple-treatments meta-analysis. *Lancet*. 2011;378(9799):1306-1315.
5. Yildiz A, Vieta E, Leucht S, Baldessarini RJ. Efficacy of antimanic treatments: meta-analysis of randomized, controlled trials. *Neuropsychopharmacology*. 2011;36(2):375-389.
6. Miller F, Tanenbaum JH, Griffin A, Ritvo E. Prediction of treatment response in bipolar, manic disorder. *J Affect Disord*. 1991;21(2):75-77.
7. Machado-Vieira R, Luckenbaugh DA, Soeiro-de-Souza MG, et al. Early improvement with lithium in classic mania and its association with later response. *J Affect Disord*. 2013;144(1-2):160-164.
8. Hou L, Heilbronner U, Degenhardt F, et al. Genetic variants associated with response to lithium treatment in bipolar disorder: a genome-wide association study. *Lancet*. 2016;387(10023):1085-1093.
9. Schubert KO, Wisdom A. Should the Australian Therapeutic Goods Administration recommend rapid dosing of lithium carbonate in acute mania? *Australian & New Zealand Journal of Psychiatry*. 2018;52(4):387-387.
10. Kleindienst N, Engel R, Greil W. Which clinical factors predict response to prophylactic lithium? A systematic review for bipolar disorders. *Bipolar disorders*. 2005;7(5):404-417.
11. Tighe SK, Mahon PB, Potash JB. Predictors of lithium response in bipolar disorder. *Ther Adv Chronic Dis*. 2011;2(3):209-226.
12. Nunes A, Ardaub R, Berghöfer A, et al. Prediction of lithium response using clinical data. *Acta Psychiatrica Scandinavica*. 2020;141(2):131-141.

13. Amare AT, Schubert KO, Hou L, et al. Association of polygenic score for schizophrenia and HLA antigen and inflammation genes with response to lithium in bipolar affective disorder: a genome-wide association study. *JAMA psychiatry*. 2018;75(1):65-74.
14. Antonucci LA, Pergola G, Pigoni A, et al. A Pattern of Cognitive Deficits Stratified for Genetic and Environmental Risk Reliably Classifies Patients With Schizophrenia From Healthy Control Subjects. *Biol Psychiatry*. 2020;87(8):697-707.
15. Amare AT, Schubert KO, Hou L, et al. Association of polygenic score for major depression with response to lithium in patients with bipolar disorder. *Mol Psychiatry*. 2020.
16. Lee SH, Ripke S, Neale BM, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature genetics*. 2013;45(9):984.
17. Consortium C-DGotPG. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet*. 2013;381(9875):1371-1379.
18. Consortium B. Analysis of shared heritability in common disorders of the brain. *Science (New York, NY)*. 2018;360(6395).
19. Consortium SWGotPG. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. 2014;511(7510):421-427.
20. Gandal MJ, Haney JR, Parikshak NN, et al. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science*. 2018;359(6376):693-697.
21. Maier R, Moser G, Chen G-B, et al. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *The American Journal of Human Genetics*. 2015;96(2):283-294.
22. Van Rheenen W, Peyrot WJ, Schork AJ, Lee SH, Wray NR. Genetic correlations of polygenic disease traits: from theory to practice. *Nature Reviews Genetics*. 2019;20(10):567-581.
23. Cearns M, Opel N, Clark S, et al. Predicting rehospitalization within 2 years of initial patient admission for a major depressive episode: a multimodal machine learning approach. *Transl Psychiatry*. 2019;9(1):285.
24. Schulze TG, Alda M, Adli M, et al. The International Consortium on Lithium Genetics (ConLiGen): an initiative by the NIMH and IGSLI to study the genetic basis of response to lithium treatment. *Neuropsychobiology*. 2010;62(1):72-78.

25. Malhi GS, Bassett D, Boyce P, et al. Royal Australian and New Zealand College of Psychiatrists clinical practice guidelines for mood disorders. *Aust N Z J Psychiatry*. 2015;49(12):1087-1206.
26. Wray NR, Ripke S, Mattheisen M, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet*. 2018;50(5):668-681.
27. Amare AT, Schubert KO, Klingler-Hoffmann M, Cohen-Woods S, Baune BT. The genetic overlap between mood disorders and cardiometabolic diseases: a systematic review of genome wide and candidate gene studies. *Transl Psychiatry*. 2017;7(1):e1007.
28. Garnham J, Munro A, Slaney C, et al. Prophylactic treatment response in bipolar disorder: results of a naturalistic observation study. *J Affect Disord*. 2007;104(1-3):185-190.
29. Duffy A, Alda M, Milin R, Grof P. A consecutive series of treated affected offspring of parents with bipolar disorder: is response associated with the clinical profile? *Can J Psychiatry*. 2007;52(6):369-376.
30. 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(6):e65636.
31. Scott J, Etain B, Manchia M, et al. An examination of the quality and performance of the Alda scale for classifying lithium response phenotypes. *Bipolar Disord*. 2020;22(3):255-265.
32. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Meth Psych Res*. 2011;20(1):40-49.
33. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2005;67(2):301-320.
34. Biau G, Devroye L, Lugosi G. Consistency of random forests and other averaging classifiers. *J Mach Learn Res*. 2008;9(Sep):2015-2033.
35. Cearns M, Hahn T, Clark S, Baune BT. Machine learning probability calibration for high-risk clinical decision-making. *Aust N Z J Psychiatry*. 2020;54(2):123-126.
36. Cearns M, Hahn T, Baune BT. Recommendations and future directions for supervised machine learning in psychiatry. *Translational psychiatry*. 2019;9(1):1-12.

37. Green GH, Diggle PJ. On the operational characteristics of the Benjamini and Hochberg False Discovery Rate procedure. *Statistical applications in genetics and molecular biology*. 2007;6(1).
38. Flint C, Cearns M, Opel N, et al. Systematic Overestimation of Machine Learning Performance in Neuroimaging Studies of Depression. *arXiv preprint arXiv:191206686*. 2019.
39. Schnack HG, Kahn RS. Detecting Neuroimaging Biomarkers for Psychiatric Disorders: Sample Size Matters. *Front Psychiatry*. 2016;7:50.
40. Helland IS. On the interpretation and use of R² in regression analysis. *Biometrics*. 1987;61-69.
41. Dinga R, Marquand AF, Veltman DJ, et al. Predicting the naturalistic course of depression from a wide range of clinical, psychological, and biological data: a machine learning approach. *Transl Psychiatry*. 2018;8(1):241.
42. Nunes A, Trappenberg T, Alda M, Genetics iCoL. Asymmetrical reliability of the Alda Score favours a dichotomous representation of lithium responsiveness. *PloS one*. 2020;15(1):e0225353.
43. Durrleman S, Simon R. Flexible regression models with cubic splines. *Statistics in medicine*. 1989;8(5):551-561.
44. Sportiche S, Geoffroy PA, Brichant-Petitjean C, et al. Clinical factors associated with lithium response in bipolar disorders. *Aust N Z J Psychiatry*. 2017;51(5):524-530.
45. Nunes A, Stone W, Ardaur R, et al. Exemplar scoring identifies genetically separable phenotypes of lithium responsive bipolar disorder. *Transl Psychiatry*. 2021;11(1):36.
46. Bipolar D, Schizophrenia Working Group of the Psychiatric Genomics Consortium. Electronic address drve, Bipolar D, Schizophrenia Working Group of the Psychiatric Genomics C. Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes. *Cell*. 2018;173(7):1705-1715 e1716.
47. Parker G, Tavella G, Macqueen G, et al. Revising Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, criteria for the bipolar disorders: Phase I of the AREDOC project. *Aust N Z J Psychiatry*. 2018:4867418808382.
48. Huang SH, LePendur P, Iyer SV, Tai-Seale M, Carrell D, Shah NH. Toward personalizing treatment for depression: predicting diagnosis and severity. *J Am Med Inform Assoc*. 2014;21(6):1069-1075.

49. Clark S, Baune B, Schubert K, et al. Prediction of transition from ultra-high risk to first-episode psychosis using a probabilistic model combining history, clinical assessment and fatty-acid biomarkers. *Translational psychiatry*. 2016;6(9):e897-e897.
50. Toenders YJ, Schmaal L, Harrison BJ, Dinga R, Berk M, Davey CG. Neurovegetative symptom subtypes in young people with major depressive disorder and their structural brain correlates. *Translational Psychiatry*. 2020;10(1):1-11.
51. 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(4):370-378.
52. Dwyer DB, Kalman JL, Budde M, et al. An investigation of psychosis subgroups with prognostic validation and exploration of genetic underpinnings: the PsyCourse Study. *JAMA psychiatry*. 2020;77(5):523-533.
53. Clementz BA, Sweeney JA, Hamm JP, et al. Identification of Distinct Psychosis Biotypes Using Brain-Based Biomarkers. *Focus (Am Psychiatr Publ)*. 2018;16(2):225-236.
54. Ivleva EI, Clementz BA, Dutcher AM, et al. Brain Structure Biomarkers in the Psychosis Biotypes: Findings From the Bipolar-Schizophrenia Network for Intermediate Phenotypes. *Biol Psychiatry*. 2017;82(1):26-39.
55. Buhmann JM, Held M. Unsupervised learning without overfitting: Empirical risk approximation as an induction principle for reliable clustering. Paper presented at: International Conference on Advances in Pattern Recognition1999.
56. Martin J, Taylor MJ, Lichtenstein P. Assessing the evidence for shared genetic risks across psychiatric disorders and traits. *Psychol Med*. 2018;48(11):1759-1774.
57. Kasiulevičius V, Šapoka V, Filipavičiūtė R. Sample size calculation in epidemiological studies. *Gerontologija*. 2006;7(4):225-231.

Supplementary Appendix

Stratification of bipolar patients using polygenic scores increases the accuracy of lithium response prediction: A machine learning approach

Micah Cearns¹, Azmeraw T. Amare^{1,2}, K. Oliver Schubert^{1,2}, Scott Clark¹, International Consortium on Lithium Genetics (ConLi+Gen), Bernhard T. Baune³

1. The discipline of Psychiatry, School of Medicine, University of Adelaide, Adelaide, Australia
2. South Australian Health and Medical Research Institute (SAHMRI), Adelaide, SA, Australia
3. Northern Adelaide Local Health Network, Mental Health Services, Adelaide, Australia
4. Department of Mental Health, University Hospital of Münster, Münster, Germany

Supplementary methods.

Table 1. Missing data table

Table 2. Searched hyperparameters for all models

Table 3. Extended ridge and elastic net regression results

Table 4. Extended random forest regression results

Table 5. Extended ridge and elastic net PRS stratified model results

Table 6. Extended random forest PRS stratified model results

Table 7. Extended classification metrics for the clinical random forest and meta-PRS random forest models

Table 8. Missing data table by study site

Supplementary methods

Assessing sample size effects

As previous works have shown that smaller samples have a higher risk of systematically misestimating model performance^{1,2}, we devised a Monte-Carlo based statistical test to measure whether the observed changes in performance were attributable to the polygenic stratifications or the subsequent decreases in sample size that resulted from only using patients in the lower and upper quartiles of the PRS distributions. In this procedure, we took 1000 random sub-samples from our full sample ($N = 1,034$) equal to the N attained in our stratified meta-PRS analysis samples ($N = 518$) and re-ran the full pipeline on each sub-sample. Following, we used each R^2 score to derive a null distribution of scores attainable for each train/validate ($N = 347$) and test ($N = 171$) sample combination. This process allowed us to derive a null distribution of scores that were only attributable to decreases in sample size without any PRS stratification. We then took each test score from this distribution that was equal to or greater than the score from our stratified meta-PRS models. We divided the number of occurrences that were greater than or equal to the meta-PRS score by the total number of sub-samples ($m = 1000$) plus 1 (for the meta-PRS model scores), allowing us to derive p-values for the stratified meta-PRS clinical models. Significant p-values represented a difference in scores that were not attributable to changes in sample size.

Measuring the relationship between the number of predictor variables and R^2

In traditional statistical analysis, every time a predictor variable is added to a model, the model's R^2 increases, therefore, the model R^2 needs to be adjusted³. A better approach is to estimate R^2 on a held-out test set as was done in the current analyses. However, to further rule out artificial R^2 inflation resulting from the increasing number of predictor variables across analyses (as the size of the feature space increases across uni, multi, and interaction term models), we used Pearson's r to measure the relationship between the number of variables in each model and its subsequent R^2 score.

Supplementary references

1. Flint C, Cearns M, Opel N, et al. Systematic Overestimation of Machine Learning Performance in Neuroimaging Studies of Depression. *arXiv preprint arXiv:191206686*. 2019.
2. Schnack HG, Kahn RS. Detecting Neuroimaging Biomarkers for Psychiatric Disorders: Sample Size Matters. *Front Psychiatry*. 2016;7:50.
3. Helland IS. On the interpretation and use of R^2 in regression analysis. *Biometrics*. 1987:61-69.

Table 1. Percentage proportions of missing data stratified by data collection site.

[illegible]

DSM Dx (BP2)	0	0	0	0	0	0	0	0	0	0
DSM Dx (SABP)	0	0	0	0	0	0	0	0	0	0
DSM Dx (BP3)	0	0	0	0	0	0	0	0	0	0
DSM Dx (BPNOS)	0	0	0	0	0	0	0	0	0	0
BPD family history (Y)	0	0	0	0	0	0	0	0	0	0
BPD family history (N)	0	0	0	0	0	0	0	0	0	0
BPD family history (Unknown)	0	0	0	0	0	0	0	0	0	0

Table 2. Searched hyperparameter combinations from each model used in analyses.

Ridge												
model__alpha	0.001	0.01	0.1	1	2	3	4	5	6	7	8	9
Elastic net												
model__alpha	0.001	0.01	0.1	0.3	0.5	0.7	1	-	-	-	-	-
model__l1_ratio	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1	-	-
Unimodal random forest												
model__n_estimators	25	50	75	100	-	-	-	-	-	-	-	-
model__max_features	sqrt	log2	-	-	-	-	-	-	-	-	-	-
model__max_depth	4	6	8	10	-	-	-	-	-	-	-	-
model__min_samples_leaf	2	5	10	-	-	-	-	-	-	-	-	-
Clinical interaction random forest												
model__n_estimators	25	50	75	100	-	-	-	-	-	-	-	-
model__max_features	sqrt	log2	-	-	-	-	-	-	-	-	-	-
model__max_depth	4	6	8	10	-	-	-	-	-	-	-	-
model__min_samples_leaf	2	5	10	-	-	-	-	-	-	-	-	-
selector__estimator__n_estimators	25	50	75	100	-	-	-	-	-	-	-	-
selector__estimator__max_features	sqrt	log2	-	-	-	-	-	-	-	-	-	-
selector__estimator__max_depth	4	6	8	10	-	-	-	-	-	-	-	-
selector__estimator__min_samples_leaf	2	5	10	-	-	-	-	-	-	-	-	-
selector__max_features	25	50	75	100	125	150	175	200	225	250	-	-
PRS interaction random forest												
model__n_estimators	25	50	75	100	-	-	-	-	-	-	-	-
model__max_features	sqrt	log2	-	-	-	-	-	-	-	-	-	-
model__max_depth	4	6	8	10	-	-	-	-	-	-	-	-
model__min_samples_leaf	2	5	10	-	-	-	-	-	-	-	-	-
selector__estimator__n_estimators	25	50	75	100	-	-	-	-	-	-	-	-
selector__estimator__max_features	sqrt	log2	-	-	-	-	-	-	-	-	-	-
selector__estimator__max_depth	4	6	8	10	-	-	-	-	-	-	-	-
selector__estimator__min_samples_leaf	2	5	10	-	-	-	-	-	-	-	-	-

selector__max_features	10	20	30	40	50	60	70	80	90	-	-	-
Multimodal interaction random forest												
model__n_estimators	25	50	75	100	-	-	-	-	-	-	-	-
model__max_features	sqrt	log2	-	-	-	-	-	-	-	-	-	-
model__max_depth	4	6	8	10	-	-	-	-	-	-	-	-
model__min_samples_leaf	2	5	10	-	-	-	-	-	-	-	-	-
selector__estimator__n_estimators	25	50	75	100	-	-	-	-	-	-	-	-
selector__estimator__max_features	sqrt	log2	-	-	-	-	-	-	-	-	-	-
selector__estimator__max_depth	4	6	8	10	-	-	-	-	-	-	-	-
selector__estimator__min_samples_leaf	2	5	10	-	-	-	-	-	-	-	-	-
selector__max_features	62	124	186	248	310	372	434	496	558	620	-	-
Multimodal random forest												
model__n_estimators	25	50	75	100	-	-	-	-	-	-	-	-
model__max_features	sqrt	log2	-	-	-	-	-	-	-	-	-	-
model__max_depth	4	6	8	10	-	-	-	-	-	-	-	-
model__min_samples_leaf	2	5	10	-	-	-	-	-	-	-	-	-
selector__estimator__n_estimators	25	50	75	100	-	-	-	-	-	-	-	-
selector__estimator__max_features	sqrt	log2	-	-	-	-	-	-	-	-	-	-
selector__estimator__max_depth	4	6	8	10	-	-	-	-	-	-	-	-
selector__estimator__min_samples_leaf	2	5	10	-	-	-	-	-	-	-	-	-
selector__max_features	2	4	6	8	10	12	14	16	18	20	22	24

Table 3. Train/validation (N = 692), and test (N = 342) results using linear regression with Ridge regularization. Unimodal, multimodal, and interaction effect models were considered. All p-values were FDR corrected with the Benjamini and Hochberg method.

	Train Mean	Validation	Test	P-Value
PRS R2	0.01 (0.0)	0.2 (0.18)	0.012	0.0131
PRS RMSE	3.27 (0.09)	3.19 (0.45)	3.365	-
PRS MSE	10.67 (0.58)	10.37 (3.04)	11.324	-
PRS MAE	2.86 (0.09)	2.79 (0.44)	2.969	-
Interaction PRS R2	0.01 (0.0)	0.2 (0.18)	0.014	0.01
Interaction PRS RMSE	3.26 (0.09)	3.2 (0.46)	3.3624	-
Interaction PRS MSE	10.66 (0.58)	10.43 (3.08)	11.306	-
Interaction PRS MAE	2.86 (0.08)	2.8 (0.44)	2.9623	-
Clin R2	0.12 (0.02)	0.32 (0.17)	0.018	0.0004
Clin RMSE	3.06 (0.07)	3.33 (0.33)	3.356	-
Clin MSE	9.38 (0.4)	11.22 (2.25)	11.260	-
Clin MAE	2.64 (0.06)	2.85 (0.34)	2.875	-
Interaction Clin R2	0.12 (0.02)	0.26 (0.19)	0.045	0.0001
Interaction Clin RMSE	3.07 (0.07)	3.26 (0.41)	3.309	-
Interaction Clin MSE	9.41 (0.4)	10.83 (2.77)	10.951	-
Interaction Clin MAE	2.66 (0.06)	2.82 (0.41)	2.867	-
Clin and PRS R2	0.09 (0.02)	0.2 (0.17)	0.047	0.0001
Clin and PRS RMSE	3.12 (0.07)	3.2 (0.42)	3.306	-
Clin and PRS MSE	9.72 (0.44)	10.39 (2.78)	10.927	-
Clin and PRS MAE	2.72 (0.07)	2.77 (0.43)	2.875	-
Interaction Clin and PRS	0.13 (0.02)	0.27 (0.19)	0.051	0.0001
Interaction Clin and PRS	3.06 (0.07)	3.27 (0.43)	3.299	-
Interaction Clin and PRS	9.34 (0.41)	10.9 (2.86)	10.883	-
Interaction Clin and PRS	2.66 (0.06)	2.82 (0.43)	2.856	-

Table 4. Train/validate (N = 692), and test (N = 342) results using non-linear random forest regression. Unimodal, multimodal, and interaction effect models were considered. All p-values were FDR corrected with the Benjamini and Hochberg method.

	Train Mean	Validation	Test	P-Value
PRS R2	0.09 (0.01)	0.22 (0.18)	0.020	0.0001
PRS RMSE	3.13 (0.1)	3.21 (0.45)	3.352	-
PRS MSE	9.81 (0.6)	10.54 (3.03)	11.24	-
PRS MAE	2.74 (0.09)	2.81 (0.45)	2.953	-
Interaction PRS R2	0.09 (0.01)	0.22 (0.19)	0.009	0.0001
Interaction PRS RMSE	3.12 (0.1)	3.22 (0.46)	3.371	-
Interaction PRS MSE	9.75 (0.61)	10.59 (3.09)	11.360	-
Interaction PRS MAE	2.72 (0.09)	2.81 (0.44)	2.973	-
Clin R2	0.21 (0.02)	0.21 (0.11)	0.0812	0.0001
Clin RMSE	2.91 (0.07)	3.2 (0.35)	3.245	-
Clin MSE	8.45 (0.38)	10.38 (2.33)	10.531	-
Clin MAE	2.51 (0.06)	2.77 (0.38)	2.801	-
Interaction Clin R2	0.24 (0.02)	0.23 (0.14)	0.0667	0.001
Interaction Clin RMSE	2.86 (0.06)	3.23 (0.4)	3.270	-
Interaction Clin MSE	8.19 (0.36)	10.59 (2.63)	10.695	-
Interaction Clin MAE	2.47 (0.06)	2.8 (0.42)	2.809	-
Clin and PRS R2	0.22 (0.02)	0.21 (0.12)	0.074	0.0001
Clin and PRS RMSE	2.9 (0.07)	3.21 (0.37)	3.258	-
Clin and PRS MSE	8.41 (0.37)	10.44 (2.43)	10.615	-
Clin and PRS MAE	2.51 (0.06)	2.77 (0.39)	2.798	-
Interaction Clin and PRS	0.25 (0.02)	0.23 (0.16)	0.052	0.0001
Interaction Clin and PRS	2.84 (0.06)	3.23 (0.42)	3.296	-
Interaction Clin and PRS	8.04 (0.33)	10.61 (2.8)	10.862	-
Interaction Clin and PRS	2.45 (0.05)	2.8 (0.43)	2.828	-

Table 5. Train/validate (N = 347), and test (N = 171) results running ML regression analyses on only patients in the lower and upper quartiles of the meta-PRS distribution. Linear models composed of clinical predictors and interaction terms between them were considered. All p-values were FDR corrected with the Benjamini and Hochberg method.

	Train Mean (SD)	Validation Mean (SD)	Test	P-Value
Linear MDD stratified R2	0.12 (0.01)	0.23 (0.31)	-0.028	0.0171
Linear MDD stratified RMSE	3.11 (0.07)	3.37 (0.34)	3.309	
Linear MDD stratified MSE	9.69 (0.41)	11.48 (2.27)	10.947	
Linear MDD stratified MAE	2.71 (0.07)	2.93 (0.37)	2.903	
Linear interaction MDD stratified R2	0.13 (0.01)	0.21 (0.36)	0.027	0.0011
Linear interaction MDD stratified RMSE	3.09 (0.07)	3.35 (0.45)	3.22	
Linear interaction MDD stratified MSE	9.52 (0.43)	11.43 (2.87)	10.368	
Linear interaction MDD stratified MAE	2.71 (0.07)	2.96 (0.43)	2.81	
Linear SCZ stratified R2	0.12 (0.01)	0.41 (0.42)	0.071	0.0003
Linear SCZ stratified RMSE	3.07 (0.07)	3.42 (0.31)	3.326	
Linear SCZ stratified MSE	9.44 (0.41)	11.79 (2.1)	11.059	
Linear SCZ stratified MAE	2.63 (0.07)	2.89 (0.36)	2.936	
Linear interaction SCZ stratified R2	0.12 (0.01)	0.31 (0.21)	0.087	0.0001
Linear interaction SCZ stratified RMSE	3.07 (0.07)	3.36 (0.47)	3.293	
Linear interaction SCZ stratified MSE	9.43 (0.4)	11.52 (3.27)	10.843	
Linear interaction SCZ stratified MAE	2.65 (0.07)	2.87 (0.48)	2.937	
Linear (combined Meta-PRS stratified) R2	0.09 (0.01)	0.67 (0.39)	0.121	0.0001
Linear (combined Meta-PRS stratified) RMSE	3.15 (0.08)	3.62 (0.38)	3.075	
Linear (combined Meta-PRS stratified) MSE	9.94 (0.5)	13.25 (2.75)	9.456	
Linear (combined Meta-PRS stratified) MAE	2.73 (0.08)	3.14 (0.38)	2.681	
Linear interaction (combined Meta-PRS stratified) R2	0.1 (0.01)	0.51 (0.34)	0.092	0.0001
Linear interaction (combined Meta-PRS stratified) RMSE	3.13 (0.08)	3.46 (0.43)	3.124	
Linear interaction (combined Meta-PRS stratified) MSE	9.79 (0.48)	12.17 (2.96)	9.758	
Linear interaction (combined Meta-PRS stratified) MAE	2.73 (0.07)	3.03 (0.41)	2.724	

Table 6. Train/validate (N = 347), and test (N = 171) results running ML regression analyses on only patients in the lower and upper quartiles of the meta-PRS distribution. Non-linear models composed of clinical predictors and interaction terms between them were considered. All p-values were FDR corrected with the Benjamini and Hochberg method.

	Train Mean (SD)	Validation Mean (SD)	Test	P-Value
Random Forest MDD stratified R2	0.18 (0.01)	0.16 (0.22)	0.035	0.0001
Random Forest MDD stratified RMSE	3.01 (0.07)	3.29 (0.29)	3.206	
Random Forest MDD stratified MSE	9.05 (0.39)	10.9 (1.99)	10.281	
Random Forest MDD stratified MAE	2.64 (0.07)	2.9 (0.33)	2.826	
Random Forest interaction MDD stratified R2	0.23 (0.01)	0.2 (0.29)	0.018	0.0001
Random Forest interaction MDD stratified RMSE	2.9 (0.06)	3.33 (0.31)	3.234	
Random Forest interaction MDD stratified MSE	8.43 (0.36)	11.18 (2.07)	10.458	
Random Forest interaction MDD stratified MAE	2.54 (0.07)	2.93 (0.33)	2.859	
Random Forest SCZ stratified R2	0.24 (0.01)	0.24 (0.14)	0.072	0.0001
Random Forest SCZ stratified RMSE	2.86 (0.07)	3.28 (0.43)	3.324	
Random Forest SCZ stratified MSE	8.16 (0.4)	10.96 (2.77)	11.048	
Random Forest SCZ stratified MAE	2.46 (0.07)	2.82 (0.47)	2.971	
Random Forest interaction SCZ stratified R2	0.27 (0.02)	0.26 (0.14)	0.093	0.0001
Random Forest interaction SCZ stratified RMSE	2.8 (0.07)	3.3 (0.46)	3.286	
Random Forest interaction SCZ stratified MSE	7.84 (0.39)	11.13 (2.95)	10.798	
Random Forest interaction SCZ stratified MAE	2.4 (0.07)	2.81 (0.51)	2.935	
Random Forest (combined Meta-PRS stratified) R2	0.23 (0.02)	0.4 (0.29)	0.137	0.0001
Random Forest (combined Meta-PRS stratified) RMSE	2.9 (0.07)	3.35 (0.44)	3.047	
Random Forest (combined Meta-PRS stratified) MSE	8.42 (0.4)	11.39 (3.01)	9.282	
Random Forest (combined Meta-PRS stratified) MAE	2.52 (0.07)	2.92 (0.44)	2.650	
Random Forest interaction (combined Meta-PRS stratified) R2	0.7 (0.01)	0.39 (0.35)	0.045	0.0001
Random Forest interaction (combined Meta-PRS stratified) RMSE	1.81 (0.07)	3.33 (0.5)	3.204	
Random Forest interaction (combined Meta-PRS stratified) MSE	3.28 (0.24)	11.31 (3.42)	10.262	
Random Forest interaction (combined Meta-PRS stratified) MAE	1.51 (0.06)	2.86 (0.46)	2.760	

Table 7. Extended classification metrics for the clinical random forest and meta-PRS random forest models. AUC = Area under the receiver operator characteristic, ACC = Accuracy, BAC = Balanced accuracy, Sens = Sensitivity, Spec = Specificity, PPV = Positive predictive value, NPV = Negative predictive value.

	AUC	ACC	BAC	F1	Sens	Spec	PPV	NPV
Clinical RF model	65.63	63.16	58.95	63.24	46.09	71.81	45.30	72.44
Meta-PRS clinical RF model	69.88	69.59	63.65	68.93	46.43	80.87	54.17	75.61

Chapter 8

General discussion and future perspectives

General discussion

The premise of this thesis was simple. First, if we treat patients as individuals, not averages, we may be able to better cater to their clinical needs. Second, if we incorporate biomarkers from diverse biological modalities alongside clinical data, we may be able to improve predictions by capturing more of the biopsychosocial structure underpinning their illness. To test these premises, we needed a modelling framework that could make predictions for individuals, rather than tell us something about the ‘average’ patient. Such a framework should be able to deal with multidimensional data without the shortcomings of classical modelling techniques. For example, it must be able to deal with multicollinearity across predictor variables and handle datasets containing more variables than observations. Indeed, machine learning methods are well suited to these problems and have risen in both popularity and accessibility due to advances in computing power and open-source programming languages. Therefore, I have focussed on using supervised machine learning models as the main modelling tool in this thesis.

However, before proceeding to use these models, I understood that the results and subsequent conclusions from this work were only as strong as the methodology itself and its idiosyncrasies when applied to psychiatric research. Therefore, I first sought to establish best practices in the field and address issues of sample size and model performance variability. Moving from the code to the clinic, I also deemed it important to build a case for considering the clinical decision-making process and how doctors may use predictive machine learning models in the future. These considerations have implications for how we construct models now and is a topic worth further disseminating. Therefore, this thesis expanded, and became composed of two larger themes: the advancement of the methodological application of machine learning methods to psychiatric traits, and the individual prediction and prognosis of patient outcomes using multimodal biological and clinical data. In the following sections, I will discuss the findings of each paper and their contribution to the field.

In **Chapter 1**, I built the case for this thesis, and then progressed to address our observations through peer-reviewed studies in **Chapters 2-7**. In **Chapter 2**, the applied methodological foundation for the rest of this thesis is laid out. In particular, I review literature highlighting the effects of sample size on model performance, and the counterintuitive inverse relationship between decreasing sample size and increasing model performance. While beyond the scope of this piece to further quantify this effect, I explored the related issue of cross-validation

scheme selection and its influence on the variability and bias of model performance estimates. In particular, I discuss commonly used schemes and their specific use cases in psychiatric studies, dependent on factors such as sample size, the presence of multiple data collection sites, and model selection and hyperparameter tuning. More specifically, I considered how CV scheme choice and structure, alongside the correct specification of pipeline architectures within these schemes has implications for data leakage and overfitting, two widespread problems in the field¹⁻⁵.

In the second half of this piece, I establish a framework to measure the maturity of applied machine learning publications, composed of three mutually exclusive categories: generalisation, model scope, and incremental utility. For each category, I devised a scoring system to assess the maturity of applied machine learning publications on a grading scale, representing studies along a spectrum from ‘proof of concept’ to ‘ready for clinical application’. This development was motivated by observations that many editors and funding bodies lacked a measured way to interpret the maturity and impact of predictive psychiatric studies. Finally, this section considered the topic of explainability in machine learning models, the effects of human bias (e.g., gender and race), and a combination of solutions and trade-offs to balance these considerations.

After establishing these foundations, I stopped and considered how machine learning applications may be used by clinicians in **Chapter 3**. Given the dynamic and interpersonal nature of psychiatric care, I deemed it unlikely that current machine learning systems would replace clinicians in the near future. Moreover, given the regulated and litigatory nature of clinical medicine and the incremental utility conferred by current models, I deemed it unlikely that a clinician would solely trust a model to make decisions for them, even if given the choice⁶⁻⁹. Therefore, I considered it likely that a clinician would use a machine learning system to complement their own decision-making process. In this context, without knowing how confident a model is in its decision, the categorical decisions of a model are not enough to inform a clinician about a model’s decision. To accurately measure model confidence, it must be well-calibrated. Specifically, its probability estimate of an event (patient responding to a drug) needs to be well aligned to the true probability of the event actually happening (patient actually responding to a drug)¹⁰. However, I observed that most authors of published models did not assess calibration, and the clinical implications of this were seldom considered, even when models were proposed for use in clinical decision support systems. As a result, we analysed the calibration properties of commonly used models, and discussed the

implications for model confidence and clinician use. In addition, we discussed easy to use post-processing methods to overcome poor calibration when non-linear models performed better than their simpler well-calibrated linear counterparts. Given the need to disseminate these ideas to a clinical readership, I chose to communicate these ideas within an applied clinical context. However, for readers interested in the mathematical underpinnings of different model loss functions and scoring rules and their effects on model calibration, see the work of Gneiting, Balabdaoui, and Raftery¹¹. While real time prospective studies are yet to assess this specific collaborative clinician/model workflow in psychiatry, recent work by Koutsouleris et al¹² provided evidence for improved prediction when taking independent clinician and model estimates and later combining them to derive meta-estimates. Retrospectively they combined clinical and biological data with clinicians estimates of transition to psychosis in high-risk patients. They found that clinicians high false-negative rate of 38.5% could be decreased to 15.4% by combining their own predictions with that of a multimodal machine learning model. Future studies should prospectively assess this workflow across different psychiatric phenotypes in everyday clinical settings, as well as assess the performance of a dynamic real time collaboration between clinicians and well calibrated models.

In **Chapter 4**, I proceeded to address the elephant in the room. Specifically, the inverse relationship between sample size and model accuracy in psychiatric machine learning studies^{4,13}. This effect is well-established; thus, I did not seek to acquire further evidence for its existence, but to explain where in the cross-validation cycle this effect was arising. Interestingly, I found that the risk of misestimating model performance predominantly arose from small test set sizes. Moreover, we demonstrated through the use of dummy classifiers that this effect was not specific to any data modality or psychiatric phenotype but was a generalisable statistical effect applicable to all machine learning applications. Moreover, by measuring train set and test set effects across 1000 samples for each value of N, I was able to tease apart the effects of how different sample sizes effect different train and test set pairings, resulting in the emulation of every possible K-Fold scheme constitution, further demonstrating the robustness of this effect. To overcome this misestimation effect, I suggest that future studies require larger testing samples, or conduct further external testing across diverse data collection sites. This can be facilitated through consortiums or online model repositories where trained models can be downloaded and further tested by different research groups. In situations where there is a finite amount of data on a novel problem,

train/validate/test partitions can be partitioned in a way that uses less data for training and validation, and more data for final model testing. While a simple solution to a major problem, this conclusion is in contrast to traditional 70/30 and 80/20 (e.g., use 80% of data for training and validation, and 20% for testing) data partition recommendations.

Building on these findings, my work was recently corroborated and extended by research at Google¹⁴. Here, 40 machine learning engineers found that models trained, validated, and tested across computer vision, medical imaging, natural language processing, medical genomics, and clinical risk prediction based on electronic medical records, often failed to perform as expected in real world deployment. The authors concluded that this phenomenon was due to a phenomenon known as ‘Underspecification’, a statistical problem whereby observed effects can have many possible causes. Specifically, the training and model selection process can produce many different models that all appear to perform equally well, however, will differ in small and arbitrary ways, such as differences in the random values used to initialise the nodes of a neural network, the way the training data is selected, parsed, and represented, and the number of training runs, or iterations used to tune a model. These small and often random differences are ignored if they do not affect how a model performs in training and validation. Nonetheless, it turns out that they can lead to large variation in real-world performance.

For example, the authors trained 50 versions of an image recognition model on ImageNet¹⁵, a dataset comprised of everyday objects. The only differences between training runs were the random values used to initialise the neural network. Despite all the models performing similarly, their performance varied extensively when tested on ObjectNet¹⁵, a dataset comprised of images of everyday objects in unusual poses (chairs on their backs, upside-down teapots, t-shirts hanging from hooks etc). Some models did well with these poses, others did not, and some performed much better than the average model. However, as far as the traditional train/validate/test framework was concerned, all models were the same. While, I have advocated for further rigour through forward testing in the previous chapters, train/validate/test schemes often constitute the totality of model production in many domains. In accordance with our findings and recommendations in **Chapters 2 and 4**, Google now also suggests that these routine schemes are not enough to establish the real-world utility of a model, and instead, more consistent post-deployment monitoring is required (akin to a forward test in a psychiatric machine learning study or post marketing surveillance of drug treatments). In addition, they suggest that multiple models should be selected and forward

tested, rather than the single best model, as models that appear close to equal in training and validation will demonstrate different characteristics in the real world. For example, models trained and validated on ImageNet performed equally well, yet differed in their performance in real-world testing on ObjectNet. This was not only in their overall ability, but on specific performance characteristics. For example, some models were good at recognising pixelated images and bad at recognising images with high contrast, other models demonstrated the opposite characteristics.

This lack of translation is particularly concerning in psychiatry, where the clinical utility of a model could seemingly change if the positive and negative predicted values were to invert in real world performance. Moreover, given their heterogeneity, this underspecification effect could plausibly be stronger in psychiatric applications compared with other fields. Therefore, in **Chapter 5**, we proceeded to lay the foundation for an ecosystem where these problems could be addressed. This ecosystem was built on four pillars: collection, construction, evaluation and translation, and provided a foundation for data collection, harmonization and sharing, guidelines for model construction, evaluation and distribution, methods for parsing heterogeneity as well as ethical and transparency standards. For more general works on the handling and processing of Big Data, see¹⁶⁻¹⁸. While this framework gets to the heart of each issue that contributes to systematic overestimation and underspecification and attempts to provide a standardised framework for the field, significant investment and further research will be required to facilitate its establishment. For example, each component of the ecosystem introduces new considerations, some of which risk introducing new biases into the modelling process. For instance, while data augmentation strategies help increase sample size in neuroimaging studies, it is also possible for a researcher to flip the data in ways that do not respect the natural asymmetry of the brain, introducing artificial noise. In addition, this ecosystem proposes that federated learning will help facilitate model training and testing at scale, whilst overcoming issues pertaining to privacy and data protection. However, the more recent adoption of federated learning techniques has led to a new set of problems, particularly in relation to covariate shift. For example, differences in class frequencies and patterns of comorbidity will arise across different datasets, and methods to deal with these issues are required. Some methodologies already exist, such as class label resampling procedures that both up sample and down sample a datasets class label distribution^{19,20}, and methods such as Bayes Adjusted Positive and Negative predicted values afford the re-calculation of Positive and Negative predicted values across any prevalence rate of an outcome label²¹. Moreover,

methodological advancements in the field of “robust” federated learning appear promising and will help to directly tackle these issues at scale²², although, further work is still needed. In the meanwhile, online open-source applications are being built that will help facilitate the implementation of early iterations of this ecosystem. For example, Photon AI²³, a Python API built by a collaborating research group, provides an easy-to-use interface that allows researchers to build machine learning pipelines and conduct analyses in a no-code environment. In addition, their online model repository allows a research group to upload a trained model which can then be downloaded and tested by others, without the need for data sharing. Given the importance of further model testing, this feature is particularly useful in the face of data sharing constraints imposed by the Health Insurance Portability and Accountability Act (HIPAA)²⁴ and the General Data Protection Regulations (GDPR)²⁵.

With this methodological foundation laid, we progressed onto the application of machine learning methods in **Chapters 6** and **7**. Here, we addressed the contribution of biomarkers, and in particular, their contribution when used alongside clinical data. First, we predicted rehospitalisation within 2 years of initial patient admission for a major depressive episode (MDD) using a combination of clinical, structural imaging, blood-biomarker, genetic (polygenic risk scores), bioelectrical impedance, and electrocardiography predictors. Optimal performance was achieved with a multimodal biomarker panel containing structural imaging, blood-biomarker, clinical, medication type, and sleep quality predictors, attaining a test AUC of 67.74 ($p = 0.0001$), followed by a model trained with only clinical predictors, attaining a test AUC = 62.81 ($p = 0.0001$). While the multimodal model performed best overall, it only led to a small increase in predictive performance. When considering biomarker acquisition costs, processing times, and the invasiveness to a cohort of patients already experiencing distress, this increase in predictive performance is small, and arguably, not worth the trade-off^{26,27}. However, progress in the field is likely to improve the contribution of biomarkers in the future. First, acquisition times and processing times will likely decrease, and in addition, less invasive forms of biomarker acquisition may develop²⁶. Furthermore, while this work considered a larger range of biomarkers for this particular problem than previous studies, more advanced and precise markers have since been collected on larger datasets, demonstrating superior predictive performance than those considered in this work. For example, in the context of transition to psychosis in a cohort of adolescents in a clinical high-risk state, Mongan et al²⁸ provided preliminary evidence that a multivariate pattern of clinical and proteomic markers could predict transition status (test AUC = 95%, PPV = 75%, NPV =

98.6%). Moreover, Antonucci et al²⁹, showed that using genetic data to stratify schizophrenia patients according to their genetic loadings for schizophrenia, in combination with environmental factors associated with schizophrenia onset, improved the performance of their case/control classifier later trained with only cognitive predictors (balanced accuracy of 77.7% to 89.4%). In comparison, their model only trained with genetic predictors attained a balanced accuracy of 55.5%.

This approach of first using biomarkers to parse patient heterogeneity may help circumvent some of the circularity in the construction of psychiatric traits, helping to explain its efficacy³⁰. For example, as psychiatric phenotypes are defined by deviations from behavioural norms, rather than well-defined biological bases, it is plausible that they may bias toward larger effect sizes for clinical variables that correlate with clinical characteristics already used in the construction of phenotypes and illness trajectories. Therefore, this tautology in the formation of diagnostic and prognostic constructs may limit the predictive contribution of biological modalities. Thus, first parsing patients based on their underlying biology may help sidestep this problem, making the best use of biological markers to increase the predictive performance of models later trained with clinical predictors. This may be particularly important when considering response to biological treatments.

Motivated by this rationale and these findings, I set out in **Chapter 7** to test this approach with the goal of predicting lithium response in a cohort of patients with bipolar disorder. Here, I looked at the predictive contribution of basic clinical data (demographic, clinical, substance use, and comorbid psychiatric traits)³¹, as well as the predictive contribution of polygenic risk scores (PRS) for schizophrenia and MDD, traits that have both shown associations with lithium response in previous works^{32,33}. We assessed uni (only PRS and only clinical predictor models) and multi modal models (PRS plus clinical predictor models), as well as models that genetically stratified patients based on their genetic loadings. In accordance with the findings of Antonucci et al²⁹ in their schizophrenia case/control classification, this strategy worked best. Given that no predictive improvements were seen in our uni and multimodal analyses using PRS, and the limited predictive ability from psychiatric PRS more generally³⁴, this is a compelling finding. Moreover, this dovetails to a broader consideration. Specifically, given these improvements in supervised model performance, does more work first need to be done on parsing the biological and symptomatic heterogeneity of psychiatric traits before proceeding to supervised analyses? Given our findings and those of Antonucci et al²⁹, this may be a useful way forward.

Likewise, a handful of emerging studies support this rationale using both clinical and biological variables.

Using unsupervised machine learning methods to parse clinical heterogeneity at the level of self-reported mood, anxiety, and stress symptomology, Grisanzio et al³⁵ found evidence for 6 transdiagnostic symptom subtypes across subjects diagnosed with major depressive disorder, panic disorder, generalized anxiety disorder, obsessive compulsive disorder, posttraumatic stress disorder, and bipolar disorder type 2. These subtypes mapped onto differences in neurocognitive behaviour, neurophysiological brain activation, and daily function, and were replicated in an independent sample. Parsing symptom level heterogeneity within major depressive disorder, Chekroud et al³⁶ found evidence for 3 distinct patient subtypes (core-emotional, atypical, and sleep) across two multisite samples of 7,221 patients. Across these subtypes, they found that antidepressants worked best for treating core emotional and sleep subtypes, and within these subtypes, different treatments worked better. For example, combined escitalopram and bupropion treatment was significantly more effective in treating core emotional symptoms than citalopram, venlafaxine with mirtazapine outperformed citalopram for sleep/insomnia symptoms, and high-dose duloxetine outperformed escitalopram for core emotional symptoms. In addition, the predictive performance of supervised machine learning models improved after parsing subjects into their respective computational subtypes, providing further support for this rationale. Starting at the level of biology, Drysdale et al³⁷ provided evidence for four distinct neurophysiological subtypes of depression defined by resting-state connectivity biomarkers. These biotypes could not be differentiated solely on the basis of clinical features but were associated with differing symptom profiles. In addition, they predicted different levels of responsiveness to transcranial magnetic stimulation therapy on a subset of patients.

Overall, between our findings within this thesis, and those above that have arisen while this thesis has been formulated, there has been an arising body of evidence supporting the parsing of symptom level and biological heterogeneity before proceeding to predictive supervised analyses. However, as was the observation at the start of this thesis, there appears to be a range of methodological inconsistencies and a lack of established best practices that raise questions around the validity and reliability of these findings. For example, Dinga et al³⁸ set out to replicate the Drysdale et al³⁷ findings in an independent cohort of 187 patients. Moreover, they attempted to extend the original findings by testing whether the subjects were sampled from a single continuous distribution with no underlying clusters. Specifically, were

the cluster indices significantly higher than what would be expected under the null hypothesis of data with no underlying clusters? In addition, they tested the stability of the clusters by making small perturbations to the data. This procedure assessed whether it was likely to identify similar clusters, and whether the same subjects would be assigned to the same clusters if the analysis was repeated using similar datasets. This is an important and different question than whether the clusters were statistically significant, as cluster assignment could be stable even if there were no real clusters in the data. Interestingly, they found that none of the original clusters identified in the original Drysdale analysis were statistically significantly, and in addition, all of them were unstable³⁸. In review of the previously discussed studies, it appears that none of them tested the statistical significance of their derived clusters, and only Grisanzio et al³⁵ tested their stability. This does not render these findings null, however, it does dampen their validity and reliability. Future works should aim to address these issues and replicate these findings.

Looking through a broader lens, even if these supervised and unsupervised approaches prove effective, particularly in combination, their efficacy is still constrained by the validity and fidelity of the biomarkers used in their construction. One consideration, if we are to move beyond incremental progress and into a framework that facilitates a paradigm shift in the way that we understand and treat psychiatric pathology, is to invest more money into the technology that underlies the science that we do, rather than only funding studies asking new questions using the same technology. This approach aims to lift the ceiling of what is possible through technological breakthroughs, rather than just asking new questions and testing new hypotheses that edge us closer to the ceiling dictated by our current technologies (e.g., present MRI and genetic hardware). This viewpoint is championed by Mark Zuckerberg and Priscilla Chan, who through their Chan-Zuckerberg foundation plan on investing 3 billion US dollars into advancing medical technologies that facilitate large scale breakthroughs³⁹. For example, one of the projects funded through this initiative aims to develop tools that can directly measure the chemistry of the brain and then use the output of these tools to test the effectiveness of drugs prescribed for mental health conditions⁴⁰⁻⁴². Such paradigm shifts in the technology that creates the input variables for machine learning models would potentially mitigate the need for, or greatly improve the performance of models in the future.

Future directions

It is early days for personalised psychiatry. Over the last few years, the evidence base for machine learning applications and prognostic biomarkers has risen, yet as is common, we are now left with more unanswered questions than when we began. This thesis has provided solutions to some of these problems and developed a set of new questions for the field.

Firstly, this work has laid the foundation for applied machine learning works in psychiatry through establishing best practices, model maturity frameworks, and considering clinical context and decision making in the model construction process. Moreover, it has addressed a major problem in the field, the systematic overestimation of model performance on small samples. Specifically, we delineated where in the cross-validation cycle we are most likely to observe overestimation of model performance, providing solutions to this problem through larger test sets, alternative data partitioning strategies, consortia, and online model repositories. In addition to these solutions, future research may consider the possibility of developing a false discovery rate correction for model performance estimates derived on small test sets. This would provide another solution to the problem of small data sets in psychiatric biomarker research.

To integrate these findings into a structured framework and assist in overcoming these problems, a hypothetical AI ecosystem was introduced. Currently, our collaboration partners are developing components of this ecosystem through automated machine learning pipelines and model repositories²³. However, further work is needed to develop practical no code infrastructure for all elements of this ecosystem (collection, construction, evaluation, and translation). Moving forward, researchers and developers may use our blueprint as a guide to develop further components of this ecosystem.

Furthermore, we provided evidence for the potential of multimodal machine learning models to make individual and personalised patient predictions, and the possibility of improving this further through the *a-priori* genetic stratification of patients prior to model training with clinical data. Likewise, we considered the *a-priori* parsing of patient heterogeneity with symptom level and biological data, particularly through the use of unsupervised machine learning methods. Future research should focus its efforts on further parsing patient heterogeneity prior to the use of supervised machine learning methods as well as consider a broader range of biomarkers and models. This approach may assist in better personalising psychiatric treatment in the face of heterogeneity, helping to circumvent many of the highlighted problems in this thesis.

While this thesis predominantly focused on simpler models, this was done on the basis of model fit. For example, whilst out of sample performance was significantly better than chance on the models I applied, overfitting was still present, even after the use of regularisation. Future works on larger samples might consider increasing model complexity, however, this decision should be informed by the broader clinical context and the presence of underfitting. In addition, whilst I have taken a strong position on the utility of disclosing the mathematical underpinnings of machine learning models in this thesis, recent work has begun to demonstrate the benefit of techniques such as layer-wise relevance propagation to visualise the decision-making process of neural networks⁴³. Future works should explore these avenues further.

Finally, we concluded this thesis by considering the case for investing more in the technology that underpins our science, and how this may progress us from a framework of incremental progress to a true paradigm shift in the understanding of psychiatric pathology. Looking forward, this view should not only inform individual researchers, but both funding bodies and investors.

References

1. Kambeitz J, Cabral C, Sacchet MD, et al. Detecting Neuroimaging Biomarkers for Depression: A Meta-analysis of Multivariate Pattern Recognition Studies. *Biol Psychiatry*. 2017;82(5):330-338.
2. Kambeitz J, Cabral C, Sacchet MD, et al. Reply to: Sample Size, Model Robustness, and Classification Accuracy in Diagnostic Multivariate Neuroimaging Analyses. *Biol Psychiatry*. 2018.
3. Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage*. 2017;145:166-179.
4. Neuhaus AH, Popescu FC. Sample Size, Model Robustness, and Classification Accuracy in Diagnostic Multivariate Neuroimaging Analyses. *Biol Psychiatry*. 2018;84(11):e81-e82.
5. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in Data Mining: Formulation, Detection, and Avoidance. *Acm T Knowl Discov D*. 2012;6(4).
6. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*. 2020;3(1):1-10.
7. Kaplan B. Evaluating informatics applications—clinical decision support systems literature review. *International journal of medical informatics*. 2001;64(1):15-37.
8. Kesselheim AS, Cresswell K, Phansalkar S, Bates DW, Sheikh A. Clinical decision support systems could be modified to reduce ‘alert fatigue’ while still minimizing the risk of litigation. *Health affairs*. 2011;30(12):2310-2317.
9. Sittig DF, Wright A, Osheroff JA, et al. Grand challenges in clinical decision support. *Journal of biomedical informatics*. 2008;41(2):387-392.

10. Nixon J, Dusenberry M, Zhang L, Jerfel G, Tran D. Measuring Calibration in Deep Learning. *arXiv preprint arXiv:190401685*. 2019.
11. Gneiting T, Balabdaoui F, Raftery AE. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2007;69(2):243-268.
12. Koutsouleris N, Dwyer DB, Degenhardt F, et al. Multimodal Machine Learning Workflows for Prediction of Psychosis in Patients With Clinical High-Risk Syndromes and Recent-Onset Depression. *JAMA psychiatry*. 2020.
13. Schnack HG, Kahn RS. Detecting Neuroimaging Biomarkers for Psychiatric Disorders: Sample Size Matters. *Front Psychiatry*. 2016;7:50.
14. D'Amour A, Heller K, Moldovan D, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:201103395*. 2020.
15. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. Paper presented at: 2009 IEEE conference on computer vision and pattern recognition2009.
16. Ji C, Li Y, Qiu W, Awada U, Li K. Big data processing in cloud computing environments. Paper presented at: 2012 12th international symposium on pervasive systems, algorithms and networks2012.
17. Zaharia M, Xin RS, Wendell P, et al. Apache spark: a unified engine for big data processing. *Communications of the ACM*. 2016;59(11):56-65.
18. Dittrich J, Quiané-Ruiz J-A. Efficient big data processing in Hadoop MapReduce. *Proceedings of the VLDB Endowment*. 2012;5(12):2014-2015.
19. He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. Paper presented at: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)2008.

20. He H, Garcia EA. Learning from imbalanced data. *Ieee T Knowl Data En.* 2009;21(9):1263-1284.
21. Abu-Akel A, Bousman C, Skafidas E, Pantelis C. Mind the prevalence rate: overestimating the clinical utility of psychiatric diagnostic classifiers. *Psychological medicine.* 2018;48(8):1225-1227.
22. Reisizadeh A, Farnia F, Pedarsani R, Jadbabaie A. Robust federated learning: The case of affine distribution shifts. *arXiv preprint arXiv:200608907.* 2020.
23. Leenings R, Winter NR, Sarink K, et al. The PHOTON Wizard--Towards Educational Machine Learning Code Generators. *arXiv preprint arXiv:200205432.* 2020.
24. Act A. Health insurance portability and accountability act of 1996. *Public law.* 1996;104:191.
25. Tikkinen-Piri C, Rohunen A, Markkula J. EU General Data Protection Regulation: Changes and implications for personal data collecting companies. *Computer Law & Security Review.* 2018;34(1):134-153.
26. Lakhan SE, Vieira K, Hamlat E. Biomarkers in psychiatry: drawbacks and potential for misuse. *International Archives of Medicine.* 2010;3(1):1.
27. Oosterhoff M, van der Maas ME, Steuten LM. A systematic review of health economic evaluations of diagnostic biomarkers. *Applied health economics and health policy.* 2016;14(1):51-65.
28. Mongan D, Focking M, Healy C, et al. Development of Proteomic Prediction Models for Transition to Psychotic Disorder in the Clinical High-Risk State and Psychotic Experiences in Adolescence. *JAMA Psychiatry.* 2020.
29. Antonucci LA, Pergola G, Pigoni A, et al. A Pattern of Cognitive Deficits Stratified for Genetic and Environmental Risk Reliably Classifies Patients With Schizophrenia From Healthy Control Subjects. *Biol Psychiatry.* 2020;87(8):697-707.

30. Williams LM. Defining biotypes for depression and anxiety based on large-scale circuit dysfunction: a theoretical review of the evidence and future directions for clinical translation. *Depression and anxiety*. 2017;34(1):9-24.
31. Nunes A, Ardaou R, Berghöfer A, et al. Prediction of lithium response using clinical data. *Acta Psychiatrica Scandinavica*. 2020;141(2):131-141.
32. International Consortium on Lithium G, Amare AT, Schubert KO, et al. Association of Polygenic Score for Schizophrenia and HLA Antigen and Inflammation Genes With Response to Lithium in Bipolar Affective Disorder: A Genome-Wide Association Study. *JAMA Psychiatry*. 2018;75(1):65-74.
33. Amare AT, Schubert KO, Hou L, et al. Association of polygenic score for major depression with response to lithium in patients with bipolar disorder. *Mol Psychiatry*. 2020.
34. Murray GK, Lin T, Austin J, McGrath JJ, Hickie IB, Wray NR. Could Polygenic Risk Scores Be Useful in Psychiatry?: A Review. *JAMA psychiatry*. 2020.
35. Grisanzio KA, Goldstein-Piekarski AN, Wang MY, Rashed Ahmed AP, Samara Z, Williams LM. Transdiagnostic Symptom Clusters and Associations With Brain, Behavior, and Daily Function in Mood, Anxiety, and Trauma Disorders. *JAMA Psychiatry*. 2018;75(2):201-209.
36. 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(4):370-378.
37. Drysdale AT, Grosenick L, Downar J, et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat Med*. 2017;23(1):28-38.

38. Dinga R, Schmaal L, Penninx B, et al. Evaluating the evidence for biotypes of depression: Methodological replication and extension of. *Neuroimage Clin.* 2019;22:101796.
39. Crow D. Silicon Valley Meets Biomedical Research in the Chan Zuckerberg Initiative. *Cell.* 2017;169(5):767-769.
40. Beyene AG, McFarlane IR, Pinals RL, Landry MP. Stochastic simulation of dopamine neuromodulation for implementation of fluorescent neurochemical probes in the striatal extracellular space. *ACS chemical neuroscience.* 2017;8(10):2275-2289.
41. Beyene AG, Delevich K, Del Bonis-O'Donnell JT, et al. Imaging striatal dopamine release using a nongenetically encoded near infrared fluorescent catecholamine nanosensor. *Science advances.* 2019;5(7):eaaw3108.
42. Jeong S, Yang D, Beyene AG, et al. High-throughput evolution of near-infrared serotonin nanosensors. *Science advances.* 2019;5(12):eaay3771.
43. Böhle M, Eitel F, Weygandt M, Ritter K. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Frontiers in aging neuroscience.* 2019;11:194.

Acknowledgements

I would like to thank Bernhard Baune and Scott Clark for all the opportunities they have provided me over the years. From starting out as a research assistant in Bernhard Baune's lab in undergraduate Psychology, to progressing to a PhD student under their supervision, Bernhard and Scott have both provided and entrusted me with the opportunity and freedom to pursue my academic interests whilst guiding me in my development. I would also like to thank all the researchers involved in Bidirect and ConLiGen, without their efforts none of this work would have been possible. Finally, I would like to thank those at Muenster who I have had the pleasure of collaborating with. All our conversations and work has significantly contributed to my thinking and development.