Data-driven approaches to biological psychiatry: multimodal data and machine learning in the study of psychiatric disorders

Habets, P.C.

1 General Introduction

The landscape of biological psychiatry encompasses a wide spectrum of diverse data, ranging from gene transcription to neuroimaging, each offering unique, yet interconnected insights into brain functionality (1–6). The study of biomolecular brain processes has been persistently limited by the challenge of obtaining direct in-vivo measurements (7–9). This task has been further complicated over recent years as the emergence of increasingly high-dimensional and multimodal datasets has introduced a new layer of complexity (3). The complexity and vastness of this data poses a significant challenge for classical statistical approaches due to their methodological limitations and the constraints of their underlying assumptions (3). In this thesis, we focus on the intersection of data science and biological psychiatry, employing novel methodologies to dissect and operationalize high-dimensional and multimodal data. Following this introduction, the subsequent chapters of this thesis collectively aim to deepen our understanding of psychiatric disorders related to stress and their symptoms, with a primary focus on depressive and anxiety disorders. This thesis approaches these conditions from multiple perspectives, bridging genetic, proteomic, and psychological aspects of these conditions with an emphasis on the potential for predictive analytics.

Hence, this thesis is primarily focused on two interconnected objectives.
Objective 1: to enhance our understanding of psychiatric disorders of multilevel etiology (i.e. having multiple causes), specifically those related to stress, with a concentrated focus on Major Depressive Disorder (MDD) and anxiety disorders.
Objective 2: to explore data-driven analyses, relying less on specific hypotheses and experiments, and more on the analysis of a comprehensive collection of relevant data. The unifying goal of these objectives is to uncover new and hitherto unexplored patterns that link multiple variables and outcome measurements.

The data-driven method at the heart of this thesis is closely related to predictive modeling, a field that is concerned with finding accurate patterns between multiple

variables and outcomes. Although this might seem similar to traditional statistical analysis, there are key differences. This first chapter will delineate these differences, providing a clear picture of how our method diverges from classical hypothesis-driven epidemiological approaches, and explaining what is meant by 'classical epidemiological' in this context. Next, we will explore how data-driven narratives could potentially offer a solution to a key issue in biological psychiatry: the absence of direct in-vivo brain measurements of high temporal and spatial resolution. Finally, we will introduce the specific topics for every subsequent chapter of this thesis from a data-driven perspective. The initial sections of each following chapter will be devoted to clarifying the unique biological principles and intricate details associated with the particular disorder or biobehavioral phenomenon that is the central focus of the respective chapter. This approach ensures a solid foundation for understanding the subsequent explorations and findings.

1.1 Inference versus prediction

In Leo Breiman's 2001 paper titled "Statistical Modeling: The Two Cultures", he delineates two distinct styles of statistical modeling: one that focuses on inference and another that emphasizes prediction (10). Breiman argues that, traditionally, statistical analysis has favored inference-oriented data models, underpinned by assumptions that often guide the construction of these models. This method, which Breiman refers to as 'data modeling culture,' mirrors the classical approach seen in fields such as epidemiology (referred to as the classical epidemiological approach in the paragraph above). For example, a scientist might presume a probabilistic model for data, such as assuming that in a population sample, individual heights align with a normal distribution given their gender. They would also assume the spread, or variance, of these heights to be consistent across both genders. This framework is suitable for methods like the T-test, which are built upon these probabilistic assumptions and are designed to test specific hypotheses – in this case, whether the average heights between males and females are significantly different.

Inferential statistical models, with their focus on analyzing relationships, patterns, and trends within a dataset, are highly effective for drawing inferences about the broader population from sample data. These models are built on a foundation of specific assumptions, with their principal strength being their capacity for inference, rather than prediction. Unlike predictive models, they do not prioritize prediction accuracy or effect size for prediction, as their primary objective lies in unveiling and understanding the underlying structure of the data and the relationships between variables.

The challenge arises when these models, constructed around particular assumptions, are applied to unseen or future data: they might not yield highly accurate predictions. This discrepancy can occur because the underlying assumptions and relationships, which are valid for the initial dataset, might not be applicable or transferable to new, unseen data. The reason for this misalignment can vary, ranging from differences in data collection conditions and temporal changes to discrepancies in the sampled populations. Consequently, the strength of inferential models is to comprehend the structure of the sampled data and to draw conclusions about the originating population, rather than predicting specific outcomes for new observations.

In contrast to the data modeling culture, Breiman introduces the 'algorithmic modeling culture.' Here, the primary objective is not to make inferences based on underlying assumptions, but to produce accurate predictions. Algorithmic modeling rapidly developed since the mid 1980s and has been extensively used in the field of speech recognition, image recognition and financial markets (10). In this culture, one might use a machine learning algorithm, such as a decision tree or a neural network, to predict an individual's height given their gender. These models do not necessarily rely on specific probabilistic assumptions but focus on improving predictive accuracy. This is evaluated using metrics like Root Mean Squared Error (for regression tasks) or Area Under the Receiver Operating Characteristic Curve (for classification tasks). Unlike the data modeling culture, which uses statistical tests to assess the probability of observing a particular trend

in the data (or a more extreme trend) given the null hypothesis (i.e., null-hypothesis testing as evaluated by the p-value), the algorithmic modeling culture prioritizes the model's ability to accurately predict individual-level outcomes. This departure from stringent assumptions and null-hypothesis testing allows algorithmic models to more effectively capture the multifaceted nature of observed data. Simultaneously, it warrants the need for robust validation of algorithmic models in practice to ensure that accurate predictions are driven by meaningful signals rather than spurious noise. This entails thorough testing and validation procedures to confirm the reliability and generalizability of the predictive models, safeguarding against overfitting, where the model becomes too specific to the training data and loses its ability to make accurate predictions on new, unseen data (11–14).

While the classical epidemiological approach (the data modeling culture), can yield valuable insights about group trends and associations, it might not always provide reliable information at the individual level. This lack of specificity is where predictive analytics or the algorithmic modeling culture can supplement traditional methods and prove instrumental for enhancing our understanding of disease and refining individualized diagnostic and treatment strategies. By focusing on accurate predictions for individuals rather than inference about groups, this approach can guide personalized decision-making in healthcare, ranging from diagnosing diseases to deciding on the most effective treatment options. Consequently, the balance between these two statistical cultures could significantly impact the future of personalized medicine and individual healthcare outcomes.

Breiman points out that biomedical and psychological research has largely continued to uphold a strong preference for data models (10). In recent years, this tendency has been challenged by the rapid progress in predictive modeling and the subsequent shift towards more data-driven techniques. This shift is driven by an increasing recognition of the potential of algorithmic models to deliver more accurate and informative insights, particularly for large and complex datasets. This confrontation between classic epidemiological data modeling and the burgeoning predictive modeling methods is becoming increasingly difficult to ignore, igniting a

complex dialogue regarding the balance between simplicity and accuracy, interpretability and information validity (15–18). Maintaining the simplicity and interpretability of models does not always yield the most accurate predictions. Generally, greater accuracy requires more complex prediction methods. A shift towards prioritizing predictive accuracy, and subsequently attempting to comprehend the models, could provide a new trajectory for biomedical research. In line with this, Chapter 4 of this thesis employs the principles of the 'algorithmic modeling culture'. Complex machine learning algorithms are applied to maximize the accuracy of individual-level Major Depressive Disorder (MDD) remission status predictions. Subsequently, the attempt to interpret these predictions unveils a multimodal pattern of clinical and biomolecular markers associated with the outcome. This applied methodology in Chapter 4 confirms the potential trajectory for biomedical research, demonstrating the initial prioritization of predictive accuracy, followed by the interpretive analysis of the resulting complex models within a data-driven research framework.

1.2 Hypothesis-driven versus data-driven

In parallel to the distinction between inference and prediction, research often falls on a spectrum between two other distinct paradigms: hypothesis-driven and data-driven research. Hypothesis-driven research, a traditional scientific method, starts with a specific theory - for instance, "Drug A reduces symptom B" - and designs experiments to test this hypothesis. It emphasizes confirming or refuting a previously established theory through structured experimentation. On the other end of the spectrum, data-driven research commences with descriptively analyzing large datasets without predetermined expectations or theories. Researchers employ techniques like machine learning to extract novel insights (19), provide the most accurate depiction of the relations between variables (20), or generate new hypotheses (21). The emergence of this approach has been fueled by the increasingly large and complex datasets available in biomedicine, which require novel techniques to understand their multidimensional nature (3).

While both approaches might appear diametrically opposed, they are not mutually exclusive and instead exist as ends of a spectrum. A researcher's perspective along this spectrum may change depending on the requirements and objectives of the study. The chapters of this thesis are written with an orientation towards the data-driven end of the spectrum. Although data-driven research can revolve around validating hypotheses, data-driven research does not experimentally test a specific hypothesis about a mechanism. Instead, it can facilitate a critical evaluation of hypothesis-driven concepts, answering questions like: does the data support the hypothesis, and to what degree? For example, in chapter 2 of this thesis, a multimodal, data-driven approach is adapted to evaluate a hypothesis-driven concept (i.e. does intranasal oxytocin injection lead to altered brain activity in humans via binding to oxytocin receptors in those brain areas). Another example is chapter 6 of this thesis, that does not directly test a biologically-mechanistic hypothesis. Instead, it utilizes a data-first approach to externally validate a biologically-designed marker that is hypothesized to contribute to stress coping and resilience.

This data-first approach exemplifies a rising trend in biomedical research (22–24). As datasets become increasingly large and complex, data-driven approaches are required because the amount and complexity of data does not permit for simplistic modeling fitted to narrow hypotheses (3). If we consider the premise of systems biology - that biology is a fundamentally complex system where studying its parts in isolation has limited value - then exploring and embracing data-driven approaches might be more fruitful for studying complex, multimodal biological systems. Descriptive analysis of high dimensional datasets complements experimental testing of hypothesis. Or put differently: the use of data-driven methodologies does not invalidate hypothesis-driven research, but rather complements it. When wielded correctly, it could unravel new layers of complexity and provide a more comprehensive understanding of biological phenomena, thereby leading the way to more accurate and personalized treatment options in biological psychiatry.

1.3 The challenge of in-vivo brain measurement

One of the most significant, if not the largest, challenge in neuroscience is the lack of methodologies to directly measure in-vivo brain activity with both high spatial and temporal resolution. Our understanding of the brain is dependent on our ability to examine its processes. Currently available tools often inadequately capture the holistic complexity and speed at which these processes occur. The microscopic world of synapses, neurons, and neurochemical interactions changes on a millisecond scale and within minuscule spatial scales, challenging our capacity to visualize these events in a living, functioning brain. The rise of data-driven methodologies, particularly multimodal approaches (3,25,26), has provided novel avenues to circumnavigate some of these traditional constraints. These approaches offer a way to include multiple proxies of in-vivo brain activity and processes, providing a broader and potentially more nuanced understanding of brain function. The central principle behind this multimodal approach is to leverage the strengths of various measurement techniques while mitigating their individual weaknesses. Each method contributes unique information about brain function and structure, and their combination allows us to extract a more comprehensive picture of in-vivo brain activity.

Chapter 2 and 3 of this thesis are examples of this approach, where the combination of brain functional magnetic resonance imaging (fMRI) data with transcriptomics (data on gene expression in the brain) is used to circumvent the limitations of direct in-vivo measurement of neurochemical brain processes. fMRI offers a measure of brain activity by detecting changes related to blood flow, providing an indirect index of neuronal activity. Despite its widespread use, fMRI has limitations - chiefly, it provides only a coarse approximation of brain activity and is subject to various artifacts. Transcriptomics, on the other hand, profiles the complete set of RNA transcripts produced by the genome, shedding light on the functional elements of the genome and the molecular basis of brain function. The fusion of fMRI and transcriptomics, therefore, can link functional brain activity with underlying biomolecular mechanisms, offering a tentative integrated perspective of the brain at work. Alternatively, in chapter 4 of this thesis, instead of using any

direct brain activity measurement, a range of different data modalities is used in combination to find predictive patterns for duration of depression without direct brain activity measurement.

In conclusion, while multimodal data approaches do not replace direct in-vivo measurements and other techniques, they supplement our understanding of brain function by integrating diverse and complementary sources of information. However, neuroscience research has yet to fully leverage this multimodal approach, potentially due to a tendency to focus on refining specific individual methods. Therefore, a broader utilization of this approach may subtly shift our perspective and contribute to a more comprehensive understanding of the complex mechanisms of the brain.

1.4 Multimodal nature of psychiatry disorders

Psychiatric disorders, particularly stress-related conditions such as Major Depressive Disorder (MDD) and anxiety disorders are intrinsically multimodal in nature, because they manifest through a confluence of various factors—encompassing genetic predispositions, altered brain chemistry and structure, and complex interplays of environmental, cognitive, and behavioral elements (27–32)—all of which interact and influence each other in complex ways. This inherent multimodality reflects the broad array of factors implicated in the genesis and progression of these disorders, including genetic predispositions (30,33,34), environmental stressors (35–37), and neurobiological changes (38,39).

Utilizing multimodal data to study these disorders offers a more comprehensive picture of the underlying systems at play, as it allows for the incorporation of either more relevant data or inclusion of proxies for (potentially unknown) relevant processes. This approach can better account for the multifaceted nature of psychiatric disorders and enhance our capacity to predict and understand their onset, course, and treatment response. As a typical example, in chapter 4 of this thesis, several different types of data are modeled together to find the most accurate information about the relationship between these data types and the

predicted outcome (i.e. 2-year remission in depression). Subsequently, the most predictive pattern is analyzed to explain underlying factors.

1.5 Overview of subsequent chapters from a data-driven viewpoint

Given the availability of large and multifaceted datasets, and considering the multimodal character of psychiatric disorders, the upcoming chapters will underscore a data-driven, multimodal strategy. This approach, which aims to capture the complex reality of psychiatric disorders, will be exemplified in the subsequent chapters. The common thread throughout these chapters is the acknowledgment of the multimodal aspects of psychiatric disorders and the necessity for expansive, data-centric methodologies to apprehend their complexity. Through the consolidation of various data types within these methodologies, and by maintaining a data-centric focus, each chapter aims to achieve a deeper and more comprehensive understanding and interpretation of stress-related psychiatric disorders.

Chapter 2, Oxytocin:

In this chapter, we leverage a multimodal, data-driven approach to reevaluate prior conclusions regarding the biomolecular mechanisms underlying putative behavioral effects of intranasal oxytocin in humans.

Chapter 3, Cortisol:

Here, we employ a multimodal, data-driven strategy that allows for connecting in-vivo effects of abolished cortisol pulsatility on brain activation responses with transcriptomic and cell-type specific brain architectures.

Chapter 4, Multimodal Machine Learning:

We introduce a multimodal predictive modeling approach that initially seeks the most accurate information about the relationship between predictor variables (multimodal data) and predicted outcome (remission status after two years in individuals with depressive disorder), subsequently striving to understand why this information is accurate.

Chapter 5, Resilience:

Here, we adopt a data-driven approach to assess an emerging operationalization of the phenomenon of 'resilience' that conceptualizes resilience as the discrepancy between expected and observed mental health states post-stress exposure in individuals with depression, dysthymia, and/or anxiety disorders. We assess the predictive utility of this novel conceptualization of resilience by focusing on the strength of the association between resilience and expected mental health outcomes.

Chapter 6, Genetic Risk Score Testing:

This chapter uses a data-driven approach to validate the concept of biological vulnerability to stress-related disorders, represented by a novel genetic risk score based on glucocorticoid (GC) effect regulation. Rather than simply testing a hypothesis, the chapter explores the risk score's utility across various phenotypes and biological measures. This allows us to assess the generalizability of its applicability in different contexts of stress-related psychiatric research.

Overall, the multimodal nature of psychiatric disorders necessitates an equally diverse and comprehensive methodological approach. By integrating multiple data modalities and employing a data-driven methodology, this thesis aspires to uncover novel insights into the etiology and potential treatment of these complex conditions.

References

1. Glasgow RE, Kwan BM, Matlock DD (2018): Realizing the full potential of precision health: The need to include patient-reported health behavior, mental health, social determinants, and patient preferences data. *J Clin Transl Sci* 2: 183–185.
2. Calhoun VD, Sui J (2016): Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness. *Biol Psychiatry Cogn Neurosci Neuroimaging* 1: 230–244.
3. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ (2022): Multimodal biomedical AI. *Nat Med* 28: 1773–1784.
4. Ardesch DJ, Libedinsky I, Scholtens LH, Wei Y, van den Heuvel MP (2023): Convergence of Brain Transcriptomic and Neuroimaging Patterns in Schizophrenia, Bipolar Disorder, Autism Spectrum Disorder, and Major Depressive Disorder. *Biol Psychiatry Cogn Neurosci Neuroimaging* 8: 630–639.
5. Friedman HS, Kern ML, Hampson SE, Duckworth AL (2014): A new life-span approach to conscientiousness and health: combining the pieces of the causal puzzle. *Dev Psychol* 50: 1377–1389.
6. Ho TC (2022, April): Predicting Depression Risk in Adolescents From Multimodal Data: Current Evidence and Future Directions. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*, vol. 7. pp 346–348.
7. Piatkevich KD, Bensussen S, Tseng H-A, Shroff SN, Lopez-Huerta VG, Park D, *et al.* (2019): Population imaging of neural activity in awake behaving mice. *Nature* 574: 413–417.
8. Ercole A, Magnoni S, Vegliante G, Pastorelli R, Surmacki J, Bohndiek SE, Zanier ER (2017): Current and Emerging Technologies for Probing Molecular Signatures of Traumatic Brain Injury. *Front Neurol* 8: 450.
9. Nguyen Q-T, Schroeder LF, Mank M, Muller A, Taylor P, Griesbeck O, Kleinfeld D (2010): An in vivo biosensor for neurotransmitter release and in situ receptor activity. *Nat Neurosci* 13: 127–132.

10. Breiman L (2001): Statistical Modeling: The Two Cultures. *Stat Sci* 16: 199–215.
11. Poldrack RA, Huckins G, Varoquaux G (2020): Establishment of Best Practices for Evidence for Prediction. *JAMA Psychiatry* 77: 534–540.
12. Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, *et al.* (2012): Reporting and Methods in Clinical Prediction Research: A Systematic Review. *PLoS Med* 9: e1001221.
13. Whelan R, Garavan H (2014): When Optimism Hurts: Inflated Predictions in Psychiatric Neuroimaging. *Biol Psychiatry* 75: 746–748.
14. Yeung AWK, More S, Wu J, Eickhoff SB (2022): Reporting details of neuroimaging studies on individual traits prediction: A literature survey. *Neuroimage* 256: 119275.
15. Rudin C (2019): Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat Mach Intell* 1: 206–215.
16. Shmueli G (2010): To Explain or to Predict? *SSO Schweiz Monatsschr Zahnheilkd* 25: 289–310.
17. Weissler EH, Naumann T, Andersson T, Ranganath R, Elemento O, Luo Y, *et al.* (2021): The role of machine learning in clinical research: transforming the future of evidence generation. *Trials* 22: 537.
18. Cadario R, Longoni C, Morewedge CK (2021): Understanding, explaining, and utilizing medical artificial intelligence. *Nat Hum Behav* 5: 1636–1642.
19. Sayed N, Huang Y, Nguyen K, Krejciova-Rajaniemi Z, Grawe AP, Gao T, *et al.* (2021): An inflammatory aging clock (iAge) based on deep learning tracks multimorbidity, immunosenescence, frailty and cardiovascular aging. *Nat Aging* 1: 598–615.
20. Radhakrishnan A, Friedman SF, Khurshid S, Ng K, Batra P, Lubitz SA, *et al.* (2023): Cross-modal autoencoder framework learns holistic representations of cardiovascular state. *Nat Commun* 14: 2436.
21. Peterson JC, Bourgin DD, Agrawal M, Reichman D, Griffiths TL (2021): Using large-scale experiments and machine learning to discover theories of human decision-making. *Science* 372: 1209–1214.

22. Yabe T, Rao PSC, Ukkusuri SV, Cutter SL (2022): Toward data-driven, dynamical complex systems approaches to disaster resilience. *Proc Natl Acad Sci U S A* 119. https://doi.org/10.1073/pnas.2111997119

23. Liu H, Chen Q (2023): Computational protein design with data-driven approaches: Recent developments and perspectives. *Wiley Interdiscip Rev Comput Mol Sci* 13. https://doi.org/10.1002/wcms.1646

24. Leonelli S (2012): Introduction: Making sense of data-driven research in the biological and biomedical sciences. *Stud Hist Philos Biol Biomed Sci* 43: 1–3.

25. Stahlschmidt SR, Ulfenborg B, Synnergren J (2021): Multimodal deep learning for biomedical data fusion: a review. *Brief Bioinform* 23: bbab569-.

26. Markello RD, Arnatkeviciute A, Poline J-B, Fulcher BD, Fornito A, Misic B (2021): Standardizing workflows in imaging transcriptomics with the abagen toolbox. *Elife* 10. https://doi.org/10.7554/elife.72129

27. Park S-C, Kim J-M, Jun T-Y, Lee M-S, Kim J-B, Yim H-W, Park YC (2016): How many different symptom combinations fulfil the diagnostic criteria for major depressive disorder? Results from the CRESCEND study. *Nord J Psychiatry* 71: 1–6.

28. Verduijn J, Verhoeven JE, Milaneschi Y, Schoevers RA, van Hemert AM, Beekman ATF, Penninx BWJH (2017): Reconsidering the prognosis of major depressive disorder across diagnostic boundaries: full recovery is the exception rather than the rule. *BMC Med* 15: 215.

29. Jermy BS, Glanville KP, Coleman JRI, Lewis CM, Vassos E (2021): Exploring the genetic heterogeneity in major depression across diagnostic criteria. *Mol Psychiatry* 26: 7337–7345.

30. Schultebraucks K, Choi KW, Galatzer-Levy IR, Bonanno GA (2021): Discriminating Heterogeneous Trajectories of Resilience and Depression After Major Life Stressors Using Polygenic Scores. *JAMA Psychiatry* 78: 744–752.

31. Smith M, Francq B, McConnachie A, Wetherall K, Pelosi A, Morrison J (2020): Clinical judgement, case complexity and symptom scores as predictors of outcome in depression: an exploratory analysis. *BMC Psychiatry* 20: 125.

32. Milaneschi Y, Lamers F, Penninx BWJH (2021): Dissecting Depression Biological and Clinical Heterogeneity—The Importance of Symptom Assessment Resolution. *JAMA Psychiatry* 78: 341–341.

33. Howard DM, Adams MJ, Clarke T-K, Hafferty JD, Gibson J, Shirali M, *et al.* (2019): Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat Neurosci* 22: 343–352.

34. Gandal MJ, Haney JR, Parikshak NN, Leppa V, Ramaswami G, Hartl C, *et al.* (2018): Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* 359: 693–697.

35. Harnett NG, Dumornay NM, Delity M, Sanchez LD, Mohiuddin K, Musey PI, *et al.* (2022): Prior differences in previous trauma exposure primarily drive the observed racial/ethnic differences in posttrauma depression and anxiety following a recent trauma. *Psychol Med* 1–10.

36. Hovens JGFM, Giltay EJ, Wiersma JE, Spinhoven P, Penninx BWJH, Zitman FG (2012): Impact of childhood life events and trauma on the course of depressive and anxiety disorders. *Acta Psychiatr Scand* 126: 198–207.

37. Joormann J, McLean SA, Beaudoin FL, An X, Stevens JS, Zeng D, *et al.* (2022): Socio-demographic and trauma-related predictors of depression within eight weeks of motor vehicle collision in the AURORA study. *Psychol Med* 52: 1934–1947.

38. Pasquini L, Fryer SL, Eisendrath SJ, Segal ZV, Lee AJ, Brown JA, *et al.* (2022): Dysfunctional Cortical Gradient Topography in Treatment-Resistant Major Depressive Disorder. *Biol Psychiatry Cogn Neurosci Neuroimaging*. https://doi.org/10.1016/j.bpsc.2022.10.009

39. Chung A, Jou C, Grau-Perales A, Levy ERJ, Dvorak D, Hussain N, Fenton AA (2021): Cognitive control persistently enhances hippocampal information processing. *Nature* 600: 484–488.