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

Habets, P.C.

7 Summary and Discussion

This thesis explores the intersection of data science and biological psychiatry to deepen our understanding of stress-related psychiatric disorders, specifically Major Depressive Disorder (MDD) and anxiety disorders, from a data-driven perspective. The overarching aim is to uncover new insights into the etiology and potential treatment of these complex conditions by leveraging the power of multimodal data analysis and predictive analytics. In the first chapter, the concept of an algorithmic modeling culture is introduced, which prioritizes accurate predictions over inferences based on assumptions. Additionally, data-driven research is contrasted with the traditional hypothesis-driven research approach. The challenge of obtaining direct in-vivo brain measurements with high temporal and spatial resolution is also discussed, emphasizing the importance of multimodal data approaches in neuroscience research. The subsequent chapters exemplify the data-driven, multimodal approach adopted in this thesis.

Chapter 2 investigates the potential of intranasal oxytocin (IN-OXT) to modulate emotional and social processes by altering brain activity patterns. This study, based on neuroimaging findings, takes a data-driven and multimodal approach to understand the role of IN-OXT in these processes. It seeks to clarify the uncertainty surrounding the extent of brain penetration by IN-OXT and its ability to directly bind central oxytocin receptors (OXTRs). The study relies on the analysis of oxytocin pathway gene expression in regions affected by IN-OXT during task-based fMRI, utilizing the publicly available Allen Human Brain Atlas (AHBA) and data from selected IN-OXT fMRI studies. The work suggests that IN-OXT may alter brain functionality by directly activating central OXTRs, based on the finding of higher OXTR expression in affected subcortical regions compared to unaffected ones, regardless of task-type or sex. In terms of methodology, Chapter 2 exemplifies a data-driven approach by integrating task-based fMRI and gene expression data.

Chapter 3 of the thesis presents a data-driven approach to investigate the role of ultradian cortisol pulsatility in emotional processing. The study combines human cell type and transcriptomic atlas data with functional magnetic resonance imaging (fMRI) data to analyze the effects of cortisol rhythmicity on brain regions involved in emotional processing. The research demonstrates that the loss of ultradian cortisol rhythm alters emotional processing in cortical brain areas characterized by GABAergic function. The study identifies specific genes, such as ANXA1 and genes involved in retrograde endocannabinoid signaling, as significantly differentially expressed in brain regions that respond to the loss of cortisol ultradian rhythm. Furthermore, specific cell types, including a specific NPY-expressing GABAergic neuronal cell type, and G protein signaling cascades are implicated in the cerebral effects of the disrupted cortisol rhythm. The findings provide a biological mechanistic understanding of the fMRI results and suggest potential targets for future experimental studies. The chapter employs a range of methods, including a functional MRI study with human subjects, transcriptomic analysis using the Allen Human Brain Atlas, data preprocessing and statistical analysis, probe selection, sample selection, data normalization, testing for differential gene expression, functional and protein-protein interaction enrichment analysis, and cell type enrichment analysis. The data-driven, multimodal approach used in this chapter exemplifies the overall approach of the thesis, aiming to uncover new insights into stress-related psychiatric disorders through the integration of diverse data sources and advanced analytics techniques.

Chapter 4 of the thesis focuses on utilizing a data-driven approach to predict the remission status of Major Depressive Disorder (MDD) at the individual level. The chapter demonstrates the power of multimodal data analysis and predictive analytics in improving our understanding of stress-related psychiatric disorders. The study used a machine learning approach to assess the predictive value of various biological data sets, including whole-blood proteomics, lipid-metabolomics, transcriptomics, and genetics, in combination with clinical baseline variables. Prediction models were trained and cross-validated in a sample of 643 patients with MDD, and subsequently tested in 161 MDD individuals.

The results showed that proteomic data provided the best unimodal predictions for the two-year remission status of MDD (AUROC = 0.68). When combined with clinical data at baseline, the addition of proteomic data significantly improved the prediction performance (AUROC = 0.78) compared to using clinical data alone (AUROC = 0.63). However, the addition of other -omics data did not result in a significant improvement in model performance. Further analysis revealed that proteomic analytes associated with inflammatory response and lipid metabolism, particularly fibrinogen levels, played a crucial role in the predictive models. Machine learning models outperformed the predictions made by clinical psychiatrists in determining the two-year remission status (balanced accuracy = 71% vs. 55%). The study highlights the clinical potential of a multimodal signature consisting of proteomic and clinical data for predicting the disease course of MDD at baseline.

The chapter highlights the advantage of the data-driven approach over traditional hypothesis-driven research. Rather than relying on preconceived hypotheses, the study explored a wide range of biological and clinical variables to discover the most accurate predictive patterns. This approach allowed for the identification of proteomic analytes involved in inflammatory response and lipid metabolism as crucial predictors of MDD remission status.

Chapter 5 of this thesis exemplifies the data-driven approach employed to investigate resilience in individuals with depression, dysthymia, and/or anxiety disorders. The chapter introduces the residual approach, a novel method in resilience research that quantifies resilience as the discrepancy between expected and observed mental health states following stress exposure. The study utilizes a large multi-center longitudinal dataset (n=1373) and applies linear regression analysis to evaluate the predictive utility of the residual approach for mental health outcomes.

The data-driven methodology employed in this chapter highlights the use of the residual approach as a means to uncover new insights into adaptive stress responses and their impact on mental health. The chapter compares the predictive value of resilience residual scores with symptom severity and an alternate measure of resilience, the symptom-per-stress index, using R-squared values and non-nested likelihood ratio tests. By evaluating the associations between these measures and various mental health outcomes, the study aims to shed light on the effectiveness of the residual approach in predicting key metrics of mental health and well-being.

The results of the study indicate that, although residual resilience scores exhibit a moderate level of stability over time, they are highly correlated with symptom severity at both baseline and the two-year timepoint. This suggests a potential overlap between resilience scores and symptom severity, indicating that residual scores may reflect symptom severity state more than a resilient trait. Additionally, the residual approach demonstrates varying degrees of predictive value, with lower resilience scores associated with an increased burden of anxiety-related conditions, delayed recovery, and, counterintuitively, less severe symptom increases following negative life events and greater relative symptom improvements over a year. However, symptom severity generally outperforms or matches the predictive performance of resilience residual scores, while both measures outperform the symptom-per-stress index in predicting mental health outcomes.

The findings of this chapter emphasize the importance of considering the limitations and potential confounding factors when using residual scores as proxies for resilience. The high correlation between residual scores and symptom severity suggests that caution should be exercised when interpreting and utilizing these scores as indicators of resilience.

In conclusion, the data-driven approach adopted in this chapter allows for a critical appraisal of the intuitive concept of resilience as quantified by the residual

approach. By analyzing the associations between resilience residual scores and mental health outcomes, the study challenges the initial assumptions and reveals the potential limitations of using these scores as true indicators of resilience. This critical evaluation based on empirical data highlights the importance of a rigorous and evidence-based approach to advance our understanding of resilience and its implications for stress-related psychiatric disorders.

Chapter 6 of the thesis focuses on validating a biologically derived risk score as an indicator for the relationship between cumulative stress, adverse life events, and psychiatric disorders such as Major Depressive Disorder (MDD) and anxiety disorders. The chapter adopts a data-driven approach, leveraging a new polygenic risk score (PRS) based on genetic variants associated with glucocorticoid (GC) regulation through regulatory elements (REs). By employing a multimodal data analysis strategy, the study aims to explore the associations between the PRS, clinical and stress-related phenotypes, and biological measures.

The study utilized data from the Dutch depression and anxiety cohort (NESDA) and included individuals with active depressive and/or anxiety diagnoses at baseline, as well as controls without current or lifetime diagnoses. Clinical phenotypes were assessed based on cumulative depressive and anxiety symptom severity, remission rates, chronicity, and lifetime diagnoses. Stress-related phenotypes were evaluated using measures of childhood trauma and resilience, which was calculated as a residual score of symptom severity and cumulative stress exposure. Biological outcome measures included salivary cortisol levels and immune-inflammatory response markers.

The data analysis involved bivariate and multivariate regression models, controlling for covariates such as age, sex, and ancestry principal components. The additive value of the PRS in predicting the outcome variables and gene-environment interactions were also examined. The results revealed a significant association between the PRS and the occurrence of single versus recurrent episodes of MDD and/or anxiety. However, no significant associations were found between the PRS

and clinical or stress-related phenotypes in cases with depression and/or anxiety at baseline. In cases, there were significant associations between the PRS and evening cortisol levels, but these did not survive multiple testing correction. No significant associations were observed in the PRS x cumulative stress interaction analyses.

The findings suggest that individuals with recurrent episodes of MDD and/or anxiety may have a stronger genetic vulnerability related to GC-dependent genetic variants compared to those experiencing a single episode. The study highlights the potential importance of GC-dependent regulatory elements in understanding vulnerability to depression, anxiety, and stress-related outcomes. However, further research is needed to elucidate the precise molecular mechanisms and confirm the role of the PRS in these conditions.

Overall, Chapter 6 exemplifies the data-driven approach employed throughout the thesis by integrating diverse data sources, including genetic, clinical, stress-related, and biological measures. By using advanced statistical techniques, such as regression modeling and gene-environment interaction analyses, the chapter provides new insights into the genetic factors influencing the risk and course of MDD and anxiety disorders. The data-driven approach enables a comprehensive exploration of complex relationships and contributes to our understanding of the underlying mechanisms, ultimately guiding future research and potential interventions in stress-related psychiatric disorders.

The common thread throughout these chapters is the recognition of the multimodal nature of psychiatric disorders and the need for comprehensive, data-driven methodologies to capture their complexity. By integrating multiple data modalities, these approaches provide a more comprehensive understanding of the underlying systems and potential treatment options.

7.1 Conclusion

In conclusion, this thesis contributes to the growing recognition of the potential of data-driven approaches in biological psychiatry. It emphasizes the importance of striking a balance between simplicity and accuracy, interpretability and information validity. By embracing data-driven methodologies and multimodal data analysis, researchers can unravel new layers of complexity and enhance our understanding of biological phenomena, paving the way for more accurate and personalized treatment options in psychiatry. The findings presented in this thesis highlight the promise of data-driven approaches in advancing our understanding of stress-related psychiatric disorders and provide a foundation for future investigations into the complex interplay of factors contributing to these disorders.

7.2 Future Prospects

Building on the methodologies applied in Chapters 2 and 3, an apparent trajectory for future investigation could be extending the principles of multimodal data integration towards predictive modeling (1). This endeavor would signify a further expansion of the multimodal approach illustrated in Chapter 4, by not only integrating psychological, biological, and clinical data but also incorporating brain activity data, such as those derived from fMRI and EEG techniques. Harnessing this complex data interplay may further augment both the predictive accuracy and explanatory capacity of machine learning models, thereby improving their relevance in clinical settings (1–6). Still, the realization of such advancements will necessitate larger sample sizes in imaging studies to ensure the robustness and broad applicability of the findings (7).

This thesis emphasizes the crucial role of contemporary data science and machine learning methods in psychiatric research. A key avenue for future studies involves conducting external validation to ascertain the generalizability of the results. As indicated in Chapter 4 (Supplementary Figure 2), selecting simpler, more heavily regularized models to minimize overfitting may inadvertently lead to a significant decrease in performance. Conversely, rigorous cross-validation and external testing are vital to avoid overestimation of outcomes. A delicate equilibrium is

necessary to prevent both overfitting and underfitting of models, with the definitive test residing in external validation, which ascertains the wider applicability of an algorithm to unseen, unrelated data.

In the progression towards clinical application and personalized medicine, extensive external validation becomes increasingly paramount. To certify high clinical value of algorithms, their applicability to large-scale, external datasets must be validated. Further validation of their effectiveness as clinical decision-making tools in real-world settings through randomized controlled trials (RCTs) is not merely an augmentation of their clinical significance, but a prerequisite if we are genuinely intent on harnessing the full potential of these algorithms.

The recommendations for external validation and generalizability testing, alongside the exploration of predictive utility and external validity in Chapters 5 and 6, underscore the necessity of a robust and versatile data science approach in biomedical research.

References

1. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ (2022): Multimodal biomedical AI. *Nat Med* 28: 1773–1784.
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. Boehm KM, Khosravi P, Vanguri R, Gao J, Shah SP (2022): Harnessing multimodal data integration to advance precision oncology. *Nat Rev Cancer* 22: 114–126.
4. Schmaal L, Marquand AF, Rhebergen D, van Tol M-J, Ruhé HG, Wee NJA van der, *et al.* (2015): Predicting the Naturalistic Course of Major Depressive Disorder Using Clinical and Multimodal Neuroimaging Information: A Multivariate Pattern Recognition Study. *Biol Psychiatry* 78: 278–286.
5. Koutsouleris N, Kambeitz-Ilankovic L, Ruhrmann S, Rosen M, Ruef A, Dwyer DB, *et al.* (2018): 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* 75: 1156.
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. Whelan R, Garavan H (2014): When Optimism Hurts: Inflated Predictions in Psychiatric Neuroimaging. *Biol Psychiatry* 75: 746–748.