Machine learning for real-world data from digital mental health

A thesis submitted to the University of Manchester for the degree of

Doctor of Philosophy

in the Faculty of Biology, Medicine and Health

2023

Franziska Günther

School of Health Sciences

Contents

Word count: 36195

List of publications

Günther, F., Wong, D., Elison-Davies, S., & Yau, C. (2023). Identifying factors associated with user retention and outcomes of a digital intervention for substance use disorder: a retrospective analysis of real-world data. *JAMIA Open, 6*(3), Article ooad072. https://doi.org/10.1093/jamiaopen/ooad072

Günther, F., Yau, C., Elison-Davies, S., & Wong, D. (2023). On the difficulty of predicting engagement with digital interventions for substance use disorders. In Hägglund, M., Blusi, M., Bonacina, S., Nilsson, S., Madsen, I. C., Pelayo, S., Moen, A., Benis, A., Lindsköld, L., & Gallos, P. (Eds.), Caring is sharing – exploiting the value in data for health and innovation (pp. 967-971). IOS Press. doi: 10.3233/SHTI230319

Terms and abbreviations

Table 1. Terms and abbreviations.

Term	Abbreviation
substance use disorder	SUD
Breaking Free Online	BFO
randomised controlled trial	RCT
digital intervention	DI
machine learning	ML
ecological momentary assessment	EMA
treatment as usual	TAU
electronic health record	EHR
Recovery Progression Measure (Elison, Davies, and Ward 2016)	RPM
Severity of Dependence Scale (Gossop et al. 1995)	SDS
Patient Health Questionnaire (Kroenke et al. 2009)	PHQ-4
Lifestyle Balance Model (Davies et al. 2015)	LBM
behaviour change technique (Michie et al. 2013)	BCT
cognitive behavioural therapy	CBT

Abstract

Mental healthcare demands are substantial worldwide, overwhelming an understaffed mental healthcare workforce. At the same time, recovery from mental health problems, and its interaction with treatment and person characteristics is still not well enough understood, which has negative consequences for people suffering from mental health problems.

Real-world mental health data is data relating to a person's mental health status, factors which may influence this status, and, potentially, treatment delivery, that is routinely collected, often using digital technology. With the advancing digitisation of daily life, and increasing availability of such data, real-world data have received increasing attention from mental health researchers. Researchers are particularly interested in the ecologically valid, detailed trajectories of mental health and its interaction with treatment for mental illness that these data promise to yield. Due to the volume and dimensionality that these data often have, researchers often call on machine learning methodology to process them.

Such real-world data also accrue from the use of digital mental health interventions, which are researched as alternatives and adjuncts to face-to-face mental healthcare, depending on the severity of the mental illness. Real-world studies using data from such digital interventions are only beginning to be explored as a possibility to generate evidence about digital intervention use and users.

In this thesis, I use real-world data from a digital substance dependence intervention presently available in UK addiction treatment services within three separate studies dealing with different aspects of digital intervention outcomes research. Specifically, I worked with questionnaire, and intervention module completion data from registrants with this intervention.

The first study that I present in this thesis explores initially available data in order to identify associations between digital intervention outcomes, and person and intervention characteristics. In my second study, I evaluate the feasibility of behavioural engagement prediction. My third study comments on the feasibility of explainable outcome prediction modelling with machine learning models using mental health data more generally, which are often characterised by high feature set multicollinearity. Data from the digital substance dependence intervention serve as example data in this study.

Documenting the lessons learned conducting these studies, I hope to contribute to characterising the potential and challenge of using real-world data from digital mental health interventions for research.

Declaration of originality

I hereby confirm that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

i The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made *only* in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see `http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420`), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see `http://www.library.manchester.ac.uk/about/regulations/`) and in The University's policy on Presentation of Theses.

Acknowledgements

I want to thank my fantastic supervisors Professor Christopher Yau, Dr Sarah Elison-Davies, and Dr David Wong. You have been a steady source of support, patience, generosity, and kindness. I will never forget your encouragement and dedication to helping me grow. You are amongst the best teachers I have ever had in my life, and I feel beyond lucky to have been your student.

I want to say a special Thank you to Chris. Your belief in me, a confused Psychology student with negligible experience in data science, changed my life.

I want to thank my phenomenal friends, old and new, who didn‘t leave my side in difficult times, and who made me believe in the longevity and romance of friendship.

I dedicate this thesis first and foremost to my parents Eva-Maria Nillius and Christoph Günther, whose incredibly hard work made my education possible, and whose unwavering love is the most precious gift of all.

I also dedicate this thesis to my grandparents Horst and Hannelore Nillius and Renate Günther who have lovingly accompanied my upbringing and continue to be wonderful role models for me.

Ich möchte meinen fantastischen Betreuern Professor Christopher Yau, Dr Sarah Elison-Davies und Dr David Wong danken, die mir eine ständige Quelle der Unterstützung, Geduld, Großzügigkeit und Freundlichkeit waren. Ich werde es Euch nie vergessen, wie ihr mich ermutigt habt, und wie fest Ihr Euch dem Ziel verschrieben habt, mir dabei zu helfen, mich weiterzuentwickeln. Ihr gehört zu den besten Lehrern die ich je im Leben hatte. Ich schätze mich unglaublich glücklich, eine Eurer Studentinnen gewesen zu sein.

Ganz besonders danken möchte ich Chris. Dein Glaube an mich, eine verwirrte Psychologiestudentin mit geringfügiger Erfahrung im Gebiet data science, war lebensverändernd für mich.

Ich möchte mich bei meinen phänomenalen Freunden bedanken, alten und neuen, die auch in schwierigen Zeiten nicht von meiner Seite gewichen sind, und mich an die Langlebigkeit, und Romantik von Freundschaften haben glauben lassen.

Ich widme diese Doktorarbeit zuallererst meinen Eltern Eva-Maria Nillius und Christoph Günther, deren unglaublich harte Arbeit meine Ausbildung ermöglicht hat, und deren unerschütterliche Liebe das wertvollste Geschenk von allen ist.

Ebenso widme ich diese Arbeit meinen Großeltern Horst und Hannelore Nillius und Renate Günther, die mein Aufwachsen liebevoll begleitet haben und mir wunderbare Vorbilder sind.

Chapter 1

Introduction

Mental health problems affect people in every part of the world. The first World Mental Health Report, issued by the World Health Organization in 2022, estimated that in 2019, one eighth of the global population lived with a mental disorder (World Health Organisation 2022). This number has likely increased in the wake of the global Covid-19 pandemic which has triggered marked increases in the prevalence of anxiety and depressive disorders worldwide (Santomauro et al. 2021). The World Mental Health Report juxtaposed the demand for mental healthcare that is suggested by their prevalence estimate, with the mere 2% of global health budgets that were estimated to be allotted to mental healthcare. Such paucity of budget likely prevents the development of a workforce that could meet the current demand. In a high-income country like the USA, the difference between the currently practicing number of trained mental healthcare professionals and the number needed to meet the demand amounts to 4 million, according to the recently published Behavioural Health Workforce Report from the US Substance Abuse and Mental Health Services Administration (2020). Such gaps between mental healthcare demand and available workforce can be expected to be even larger in middle-, and low-income countries.

Increasing mental health demands worldwide, alongside a stretched mental health workforce, have made the problems of understanding, and effectively treating mental health and illness more pressing. Significant recent technological advances have given rise to several new developments in the field of mental healthcare that may help to address these problems: (1) the use of real-world data, (2) the use of machine learning methodology, and (3) the introduction of digital solutions to mental healthcare. Importantly, these developments do not unfold in isolation; instead, they actively cross-pollinate each other.

In the following, I will describe each of these developments by critically reviewing example studies testifying how these developments are used in concert in order to advance our understanding of mental health and illness, and to improve clinical practice. I begin by describing how real-world data and ma-

chine learning are used together in studies, and further on, make the connection to digital solutions in mental healthcare.

1.1 Real-world data and machine learning in mental healthcare research

Traditionally, mental health researchers have focused on collecting data in the specific setting of controlled clinical trials, and psychological experiments. In these trials and experiments, data are collected in order to answer a particular research question, and researchers arguably have some control over which data are collected, from whom, in which environment, and at what time. Limitations of this approach - one of the most salient of them likely the considerable number of research findings, for example, on the effect of a specific therapeutic, which do not translate readily into practice - are increasingly recognised. Randomised controlled trial (RCT) study schedules and compliance incentives, for example, may obstruct the observation of treatment compliance which can be low for example for certain psychotropic medications in naturalistic settings due to the cardiometabolic alterations these medications bring about (Chapman and Horne 2013). Enforced exclusion criteria may impede the representation of patients in study samples who receive multiple psychotropic medications, or are diagnosed with several psychiatric disorders at any one time (Corrigan-Curay, Sacks, and Woodcock 2018). Similarly, RCT recruiting plans often result in the underrepresentation of individuals living in poverty, individuals of African ancestry, homeless individuals, and undocumented immigrants within study samples (DeAngelis 2021). These limitations can affect research examining a variety of psychotherapeutics and mental health programmes, but mostly those involving substantial levels of patient initiative and self-direction, and patient populations of high clinical complexity.

With the introduction of digital technology to all areas of our lives, for example through the internet, social media, e-health services, and mobile and wearable technology, new so-called real-world data streams, for example, electronic health records, social media posts, search engine trends, insurance claims data, and wearable sensor data, have become available. These can increase in volume fast. Since capacities for storing and processing these data are continually improving, these real-world data could be used fruitfully for mental healthcare research (ibid.). The defining characteristic of these real-world data is that they are, or could be, routinely collected in a particular context or from a particular source, as stipulated also in the definition of the Food and Drug Administration of the USA (US Food and Drug Administration 2023).

One important reason for why these kinds of data spark interest in the mental health research community is that their ecological validity, compared to data collected in RCTs, is deemed to be high. This

means that mental health could be studied under real-world conditions, with - ideally - all the factors present not only in the person and the person's environment that may influence this person's momentary mental health status at the time point of measurement, but also present in datasets (Marsch 2021). Further, real-world data, in comparison to controlled clinical trials, promises to cut costs in data collection, and to make access to an abundance of data immediate (Gehrmann et al. 2023). Necessarily, however, real-world data also bear disadvantages compared to data collected in controlled studies, the most important one likely the difficulty to establish cause-effect relationships, but also the presence of missing data and the difficulty to establish meaningful study endpoints and identify comparator groups (Stern et al. 2022).

Showcases of the informative value of real-world data for the management of a global health crisis have been presented during the course of the Covid-19 pandemic, when researchers have for example used data from smartphones and wearables to differentiate flu and Covid-19 symptoms (Shapiro et al. 2021), and to build real-world social networks with which Covid-19 control strategies could be modelled (Firth et al. 2020). Already prior to the pandemic's prompt of real-world data use, the USA's FDA had recognised the potential of real-world evidence for the approval of a new indication for drugs already approved or to satisfy postapproval study requirements in an evaluation framework published in 2018 (US Food and Drug Administration 2018). This framework supports the use of real-world data for example to generate hypotheses for RCTs, identify biomarkers, assess the impact of inclusion and exclusion criteria, or inform Bayesian prior probability distributions.

Especially in the context of mental health, in which treatment delivery can only be standardised to a certain extent, and a person's mental health status can be directly influenced by a myriad of factors, real-world data may allow researchers to "probe deeper, more nuanced questions about what works for whom and under what context" (DeAngelis 2021).

One example for how real-world data are being collected in mental health research is data recorded through mobile devices. Data collection can be occur passively, recording for example device user screen time, application use, phone call logs, text messages, battery usage, GPS-based location data, and data from device sensors such as accelerometers, or actively, collecting self reports of mental health problems. With regards to active data collection, brief self-report assessments of momentary mental health, called ecological momentary assessments (EMAs), have gained popularity amongst researchers in the past decade, promising high temporal granularity of measurement. EMAs may be delivered at random or fixed (for example in the morning or at bedtime) times during a mobile device user's waking hours, and they may also be initiated by the device user. Whether EMA data can be considered fully real-world data may be a subject of debate.

EMAs are expected to provide naturalistic “snapshots into, for example, [..] context, social interactions, stress, pain, mood, eating, physical activity, mental health symptoms, and substance use” (Marsch, 2021, p. 192) which cannot always be surveyed at this level of granularity in clinical studies, and psychological experiments. Similarly, passively collected real-world data from mobile devices promise to circumvent recall and apprehension biases surfacing in traditional studies, and may additionally alleviate the burden on people to self-report mental health at all.

Analyses of data recorded with mobile devices have been generating interesting findings in the past decade. Ben-Zeev et al. (2017) descriptively analysed such data, collected actively and passively from mobile devices of a small sample of five schizophrenia patients. Their analysis illustrated that relapse into psychosis was preceded by changes in self-reported mental health for some patients. For others, however, only behavioural changes in geospatial activity, such as ceasing to visit a previously frequently visited location, or device use, such as increased night time phone use, preceded the relapse. This may tell us something about the limits of self-report measures of mental health.

Other research teams have advanced the explorative case study from Ben-Zeev et al. (ibid.) in clinically meaningful ways: Henson et al. (2021), for example, were able to detect clinician-identified monthly symptom exasperation of schizophrenia patients with a sensitivity of 89% and specificity of 75%, using EMA data and passively collected smartphone data describing anomalous changes in, for example, sleep duration and mobility in the course of the symptom exasperation. This may serve as a more objective description of the outlook of a schizophrenia relapse before a patient sees a clinician than patient self-report.

To note, even though 21% of the patients Henson et al. (ibid.) observed recorded less than two weeks of data with their mobile devices, the ones which passed this two weeks mark contributed as much data as healthy controls, and responded to EMAs at least as often. This may tell us something about the feasibility of monitoring patients using smartphones.

Another example of real-world data used well for mental health research comes from Garriga et al. (2022), who predicted mental health crises from a different data source, namely electronic health records from several UK mental health trusts with an XGBoost model, a performant machine learning technique, with an area under the receiver operating characteristic curve of 0.797. Prediction was better for patients with more service contacts. They noted that clinical teams equipped with the algorithm seemed to often know already about the deterioration of a patient before it was predicted, since these teams reported that the model was clinically valuable in terms of identifying deterioration only in 17% of cases. These 17% of cases may, however, still have made a difference in the teams’ clinical practice.

Real-world data generated by non-patient populations has also been used for mental health research. Kelley and Gillan (2022) found that current depression severity of their sample of active Twitter users - who these findings apply to - is associated with the frequency particular text features, for example, negation and the first person singular, were found in their tweets. They also found that connectivity of networks estimated from these text features (how often text elements were used together in the same tweet) was stronger with increasing depression severity. Significant differences in network connectivity were found between users with regards to their current depression status, and within users with regards to depressed and non-depressed episodes.

Many real-world datasets for mental health research, such as the ones used by Garriga et al. (2022) and Kelley and Gillan (2022), are analysed with machine learning (ML) methodology. ML is an application of artificial intelligence which focuses on teaching a machine through the presentation of data. It entails the machine learning autonomously from this data to perform a circumscribed task, for example, making predictions for new data points, without explicit instruction. Acquired knowledge is represented in the form of an ML model. ML is increasingly used in mental health research to answer a variety of research questions, using both real-world data as well as data collected specifically for the purpose of a research study (Chekroud et al. 2021; Delgadillo, Rubel, and Barkham 2020; Koutsouleris et al. 2016; Paul et al. 2019).

Researchers expect these ML models to account better for the complex patterns possibly present in large datasets. These complex patterns can be expected if these datasets reflect the complexity of mental illness well, which is suggested by heterogeneous phenotypes of people with the same diagnosis, relapse from effective therapies, and the non-response of patients to psychotropic drugs (Stein et al. 2022). The introduction of ML represents a marked shift from traditional data analysis in mental health research which has focused on the analysis of comparably small or, if larger, expensively collected batches of data, testing hypotheses.

More naturalistic data as well as ML methodology are also increasingly being used in the study of substance abuse, for example to understand the interplay and temporal co-occurrence of stressors and problematic consumption patterns, and, ultimately, to predict substance abuse related behaviour. Substance abuse is a behaviour frequently studied in mental health research, and a special focus of this thesis. The fifth edition of the Diagnostic and Statistical Manual of Mental Disorders, the primary diagnostic tool for psychiatric diagnoses at the time of data collection for this thesis, formalises persistent substance use despite negative consequences within the diagnosis of “substance use disorder” (SUD, American Psychiatric Association, 2013). Below, I provide some details on this disorder.

Specific criteria are used by clinicians to diagnose SUD, and can be grouped into criteria revolving around (1) impaired control over substance use, for example great amounts of time spent with substance provision, use, and recovery from use, (2) social impairment, for example the failure to fulfill major role obligations at work or at home, (3) risky substance use, for example recurrent use in situations in which use poses a hazard to health, (4) pharmacological criteria, for example increased tolerance of the substance, and (5) withdrawal symptoms when substance use ceases. Criteria are substance-specific. Each grouping of criteria is related to possible symptoms that may be present for this grouping. Symptoms belonging to (4) and (5) are not necessary for the diagnosis of an SUD. Mild SUD can be diagnosed in a person presenting with two or three symptoms, moderate SUD in a person with four or five symptoms, and severe SUD can be diagnosed in individuals presenting with six symptoms or more.

In the forth edition of the Diagnostic and Statistical Manual of Mental Disorders, the precursor to the currently used fifth version, problematic substance use was formalised within the two separate diagnoses “substance abuse” and “substance dependence” (American Psychiatric Association 2000). Both described problematic patterns of substance use. “Substance dependence” described the more severe of the two, and included withdrawal and tolerance related criteria. As these changes in diagnostic terminology are still reflected in literature that is relevant for this thesis, I occasionally refer to “substance dependence” as the mental disorder described in the forth edition of the manual American Psychiatric Association (ibid.). I will use the term “substance abuse” not in reference to the mental disorder described in the forth edition of the manual American Psychiatric Association (ibid.), but as an umbrella term for problematic substance use that may be categorised as SUD, substance dependence, or less severe patterns of use that are impairing enough for the individual to be relevant for this thesis.

SUD may be compelling to study with real-world data because it has strong behavioural components, and recovery is therefore - to some extent - objectively quantifiable by the amount of substance consumed. Below, I discuss some example studies to illustrate how the new methodological developments in wider mental health research that we have previously described in this Introduction chapter - real world data use and ML - are reflected in studies on substance abuse.

1.1.1 Real-world data and machine learning for the study of substance abuse

Studies on substance abuse working with real-world data have used combinations of passively collected, and self-report EMA data to shed light onto the relationship between stressors internal and external to a person, such as affect or environmental surroundings, substance craving, and substance use relapse (Burgess-Hull and Epstein 2021). Generally, some missing data in response to EMA prompts

is to be expected in people who abuse substances, and incentivisation of EMA responses and general study compliance is common and justifiable even if this may partly undermine the real-world character of the data (Alexander, Sanjuan, and Terplan 2023). Response rates can still be sufficient for analyses, even if no monetary incentives are provided, as suggested in a recent systematic review reporting EMA response rates above 75% for studies conducted with people receiving medication for opioid use disorder who were not incentivised for EMA responses (ibid.).

In an early study, Epstein, Tyburski, Craig, et al. (2014) used randomly timed, geolocated EMAs assessing craving, mood and stress in participants with polydrug use disorder who were admitted to outpatient methadone treatment. These EMAs also prompted participants to rate the disorder of the environment they were currently in. The authors aimed to understand how participant-rated environmental disorder in Baltimore neighbourhoods impacted participants. Their analysis found that neighbourhood physical disorder and drug activity, as found in Baltimore neighbourhoods with the lowest life expectancy in the city, was negatively correlated with stress, negative mood, and cocaine and heroin craving for their participants. With their study, Epstein, Tyburski, Craig, et al. (ibid.) were able to generate quantitative evidence for the qualitative observation that the effect of a particular environmental surrounding on mood and feelings of stress may be dependent on the individual person.

In a related study, Epstein, Tyburski, Kowalczyk, et al. (2020) attempted to predict stress and drug craving from geolocation data, and derived measures of environmental exposure such as property tax level, alone. They used a random forest model, a standard ML model constructing a multitude of "decision trees" which are models whose knowledge can be represented by a tree-like structure, with one branch representing a conjunction of features that leads to a particular decision upon a target variable for a data point. The decisions of all decision trees in a forest are averaged for a specific data point. The random forest's performance indicated high specificity, but low sensitivity resulting from the low prevalence of target events like drug craving. With their study, Epstein, Tyburski, Kowalczyk, et al. highlighted the continued importance of a person's self-reports to detect and predict mental states.

In this vein, some studies have used timed EMA data only, and no passively collected data. Panlilio, Stull, Kowalczyk, et al. (2019) used EMA data, aggregated per week, and were able to associate prior heroin craving, stress, and daily "hassles" of opioid users in buprenorphine or methadone treatment with subsequent treatment dropout. Their result suggests some further work on what these stressors and "hassles" may be, and on how clinicians and social workers could best support individuals at risk of treatment dropout.

In a subsequent study with individuals with opioid use disorder in outpatient treatment, Panlilio, Stull, Bertz, et al. (2021) present another use case of EMAs. The authors included objective measures of

substance use in the form of toxicology screens into their study plan. Negative results from such toxicology screens were incentivised, and also study compliance in terms of completion of prompted EMAs. EMAs were used to self-report substance use and the general context in which use was taking place, as well as stressful events. These EMAs were a combination of self-initiated, randomly prompted, and end-of-day reports on a smartphone application.

Panlilio, Stull, Bertz, et al. (2021) also reported how many of the three randomly triggered EMAs per day were completed, which was the majority (between 2.37 and 2.73, depending on the study arm). They found that of the three opportunities to self-report substance use, self-initiation was used most readily, followed by randomly triggered self-reports. Their study is a good example for research which uses data that lie halfway on the continuum between controlled and real-world collection.

In their primary analysis, Panlilio, Stull, Bertz, et al. performed hierarchical clustering of time-series of toxicology screens, and obtained clusters of patients with generally high, sporadic, or low use of opioids and/or cocaine. They found sporadic users to be most vulnerable to others asking them to use drugs. Further, analysis of EMA data revealed that for high and low users, but not for sporadic users, drug craving levels during the hours leading up to a singular drug use event increased. This may tell us something about internal and external triggers affecting different types of opioid users differently, depending on where they stand on their journey towards recovery.

Other real-world data sources for substance abuse research - beyond actively and passively collected smartphone data - are receiving increased attention in the scientific literature, too, for example in a study by Miller, El-Bahrawy, et al. (2020). Specifically, Miller, El-Bahrawy, et al. modelled monthly trade data for a particular illicit drug on darknet markets - a proxy for the use of this drug - with past drug trade data on the darknet, augmented with monthly Wikipedia page views of illicit drugs. Their augmentation lead to better performance of their cross-validated regression-based prediction model across different drugs as well as countries compared to using past trade data from the darknet alone. With this, they show how different real-world data sources may be combined to predict outcomes relevant for illicit drug regulation.

Another source of real-world data which may be used for mental health related research is data accruing from engagement with digital interventions (DIs) for mental health (Torous, Bucci, et al. 2021). Considered one way of introducing digital solutions into mental healthcare, DIs aim to translate therapeutic elements of face-to-face psychotherapy and counselling into a digital format via web, desktop, or mobile applications and make evidence-based treatment more accessible and scalable. Varying degrees of practitioner involvement and, conversely, automation and user self-direction can be present in these DIs. In this thesis, DIs are referred to as applications which operate with some level of automa-

tion and digitisation present. Therefore, DIs for which digital technology is only used as a medium for practitioners and patients to communicate are mostly excluded from the term as it is used in this thesis.

DIs have shown promise in delivering elements of face-to-face treatment digitally (Torous, Bucci, et al. 2021). They have also been developed to support the recovery from SUD. Below, I outline the maturity of the field of DIs for SUD, and identify a gap in the literature that I give attention to with this thesis.

1.2 Digital interventions to support recovery from substance use disorder

Research on DIs for SUDs is less abundant than research on more prevalent mental disorders like anxiety and depressive disorders (ibid.). Nevertheless, the demand for scalable solutions and workforce relief in the domain of SUD is substantial: In 2019, SUDs were among the 20 leading causes of global disability-adjusted life years, referring to lives lost due to disability, ill health, and death, for adults between 25 and 49 (Vos et al. 2020). For drug use disorders, this burden has been growing since 1990, largely due to the continuing opioid crisis in the USA (ibid.).

In the UK, costs of alcohol and drug use to society have recently been estimated to surpass £66B a year (Black 2020; Burton et al. 2016). Despite this, local authorities in the UK reduced their expenses for drug abuse treatment significantly, with an additional £567.1M needed at year five of a recently proposed plan to counteract this development, i.e. be able to provide treatment to everyone in need (Black 2020, 2021). DIs present themselves as a viable option to continue to deliver evidence-based treatment that can be accessed anywhere, anytime by those who in-person treatment often fails to reach.

However, even after more than a decade of research in the area, too little is known about whether DIs are effective at reducing SUD symptoms and substance consumption. Several studies have attempted to synthesise the mixed results obtained in controlled trials (Boumparis et al. 2017; Goldberg, Lam, et al. 2022; Riper et al. 2018; Staiger et al. 2020; Weisel et al. 2019; Whittaker et al. 2019). Due to the fast-moving market of DIs for SUDs, I focus in this introduction on relatively recent studies, summarising the current evidence base. Boumparis et al. (2017) reported in their aggregate data meta-analysis a small effect size (Hedge's g = 0.30) for DIs for people who abuse opioids and other illicit substances, with included studies employing a variety of control groups. Riper et al. (2018) found a mean weekly decrease of 5 standard units of alcohol in their individual patient data meta-analysis of RCTs juxtaposing DIs for adult problem drinking of varying intensity and guided-ness with different control groups. An

aggregate data meta-analysis by Whittaker et al. (2019) found a pooled risk ratio of 1.54 for automated text messaging interventions to effect smoking abstinence compared to minimal smoking cessation support, with a conservative policy of counting dropouts as still smoking. Their finding was supported in a systematic meta-review of meta-analyses by Goldberg, Lam, et al. (2022). Weisel et al. (2019) report a small effect size (Hedge's g = 0.39) of smartphone apps for smoking cessation, and no effect of smartphone apps for alcohol use (both compared to inactive control conditions), however, their aggregate data meta-analysis was based only on a small number of RCTs. A systematic review of Staiger et al. (2020) found less than a third of the controlled trials examining effects of mobile interventions to reduce substance use to obtain significant effects. Taken together, the conducted synthesis studies suggest - at best - small effects of DIs supporting recovery from SUDs.

Statements about the current evidence base on DIs for SUD are complicated by the great diversity of people whose substance use patterns could be considered problematic, DIs developed to target these patterns, and protocols of studies examining these DIs. Specifically, RCT participants differ between studies with regard to their substance abuse severity, and, relatedly, the substance of abuse, ranging from commercially available substances like alcohol and cigarettes to opioids and stimulants such as methamphetamines. They also differ with regards to whether they are in receipt of any medication treatment for their substance abuse at the time point of recruiting.

There is also great variety with regards to the actual DIs evaluated in RCTs. Generally, the ultimate intended use case for a DI evaluated by researchers is rarely specified, and often only implicit in the study design. Use cases I identified within my review of the literature include the increase in care intensity for those currently in receipt of face-to-face care, the (partial) substitution of face-to-face contact with practitioners with the DI, and the support of those individuals currently not in treatment due to limited treatment capacity, or due to an individual's lacking problem awareness. These possible use cases roughly translate into the terms (1) "add-on" (increase in care intensity) and (2) "standalone" (substitution or replacement of in-person treatment) which are actively used in the research community to describe (1) DIs being evaluated as additions to in-person treatment, and (2) DIs which are independent of in-person treatment. Most DIs for SUDs lie in between these two absolutes. A premise I adopt in this thesis is that human facilitation of a DI, while to some extent desirable, and likely connected to greater efficacy, generally limits scalability (Torous, Bucci, et al. 2021).

Related to potential human facilitation are varying degrees of complexity and automation present in DIs, which span multiple dimensions. These include the degree to which users can interact with the DI itself, for example when the DI requests user input, whether a mobile application is involved, whether the DI is conceived as a single-session or multi-session intervention, whether DI components have to be completed in a certain order (termed "linear" DIs) or not ("modular" DIs), and the degree to which

the DI is generally capable of permeating a user's daily life. Examples for such permeation could be push notifications, emails addressed to the user or a mental health professional, integration of support by an individually assigned mental health professional into the DI, or DI integration with external digital technology such as EHRs. To note, some earlier studies summarising the literature on DIs for SUD also review digital technology targeting substance abuse which is only used as a medium for communication between practitioners and patients. These studies are mostly excluded from the concept of a DI of this thesis.

Considerable diversity also exists between design choices of RCTs evaluating the efficacy of DIs for SUD. Control groups can be divided into those intended to be therapeutic (psychosocial and/or pharmacological treatment, usually conceptualised as treatment-as-usual,) and those not intended to be therapeutic (for example waitlist or attentional control) (Goldberg, Sun, et al. 2023). Follow-up assessments to assess longer-term effects of the DI may be conducted or not. Endpoints range from those related to substance use (including hair or urine toxicology screens, self-reports, for example with the Timeline Followback method (Sobell et al. 2001), or complete diagnostic interviews lead by clinicians), endpoints related to general mental health including the presence of comorbid mental disorders, and retention in any kind of treatment. Further, RCTs differ with regards to their treatment of missing assessment data.

Previous research has attempted to address many of these differences. This has been done for example by reviewing only a portion of the literature, for example DIs targeting abuse of a certain substance group (Riper et al. 2018; Whittaker et al. 2019), standalone DIs (Weisel et al. 2019), or mobile DIs (Goldberg, Lam, et al. 2022; Staiger et al. 2020; Whittaker et al. 2019), or through analytical choices, for example individual patient data meta-analyses, quantifying the association of person, DI, and study characteristics with outcomes. Reported effects are often small (Boumparis et al. 2017; Goldberg, Lam, et al. 2022; Staiger et al. 2020), may pertain to a certain substance group or DI type (Weisel et al. 2019), are based on a small number of included studies (ibid.), or could not be established at all (Weisel et al. 2019; Whittaker et al. 2019).

The current evidence base for DIs for SUD is exemplified in a recent meta-review of meta-analyses by Goldberg, Lam, et al. (2022): The authors find small effects for mobile phone based DIs and text-message based DIs for smoking cessation compared to control groups that are not intended to be therapeutic, and very small effects of text-message based DIs for smoking cessation compared to any active controls. They found no effects for DIs on study participants' smoking and drinking when compared to control groups not intended to be therapeutic.

Missing data in RCTs through study dropout are rarely addressed in reviews and meta-analyses, and if they are, conclusions that can be drawn persist to disagree: Riper et al. (2018) conducted a missing-not-at-random meta-analysis of 19 RCTs of DIs targeting alcohol abuse, in which they used statistical methods assuming lower and higher outcome values for study dropouts. Their estimate of 4.8 weekly standard units of alcohol less after DI use was reported robust to this analysis. In contrast, Whittaker et al. (2019) counted dropouts as still consuming the substance in their Cochrane review of RCTs evaluating smartphone based interventions for smoking cessation, and found no evidence of an effect of these interventions on smoking cessation.

RCTs represent the gold standard in efficacy research, however, biases inherent in their design have increasingly come to the attention of researchers. These biases include publication bias and adherence bias, with the latter referring to systematic differences between those who complete a study, and those who do not influencing results. Some meta-analyses, for example, include almost exclusively RCTs which schedule (and partly incentivise) in-person DI use events and assessments with participants (Boumparis et al. 2017; Riper et al. 2018). Individuals who are able to attend in-person DI use and assessment events may be more likely to be less severely impaired. Natural user engagement which may be characterised by ebbs and flows of engagement, and, importantly, dropout, is therefore not accounted for in these syntheses of literature. In fact, dropout from clinical trials of smartphone apps for depressive disorders has been estimated to total 50% when accounting for publication bias (Torous, Lipschitz, et al. 2020), with dropout rates for DIs for SUD likely to be even higher due to the clinical complexity which often characterises SUD (Cross et al. 2022).

In line with the dropout rates already observed in the somewhat controlled setting of clinical trials in which the incentivisation of retention is more feasible, real-world retention in DIs for mental health has also been found to be low: Dropout is estimated to frequently take place within the first 10 days (Baumel and Kane 2018), or after a median of 5.5 days across users (Pratap et al. 2020). The company Pear Therapeutics, which commercialised a comparatively mature DI system for SUD which had uniquely received clearance as a digital therapeutic from the USA Food and Drug Administration filed for bankruptcy in 2023 and sold its assets in an auction (Jennings 2023). It is noteworthy that this DI was almost exclusively supported by RCTs comparing treatment as usual (TAU) to the DI replacing parts or all of TAU and necessitating in-person visits to research centres for DI use and assessment (Campbell et al. 2014; Chaple et al. 2014; Marsch et al. 2014), but by little real-world evidence accounting for user disengagement.

Real-world studies on DIs for SUD, especially severe types, have not been the focus of research in the past. Studies on DIs for other mental health problems, such as Chien et al. (2020), however, bespeak the value that real-world evidence obtained from data naturally accruing from user interactions with

a DI for mental health may have: Chien et al. (2020), investigating real-world data from a DI for moderate anxiety and depression, supported by individually assigned mental health professionals, identified subgroups of users with similar engagement patterns. These subgroups were identified in data describing user initiation of engagement, and the specific therapeutic DI components users were engaging with, for example, components related to psychoeducation, activity scheduling, or goal setting. Subgroups were found to be defined by the quantity of their engagement over time.

While this only applied to users which completed three assessments within the DI, and hence showed considerable engagement in the light of real-world user disengagement rates, Chien et al. (ibid.)'s analysis yielded further interesting insights: Reliable improvement on anxiety symptoms was achieved regardless of engagement intensity of these users, and reliable improvement on depression symptoms were only achieved by a subgroup of users whose engagement intensity was not the highest among all subgroups. Chien et al. (ibid.)'s results may indicate, for their specific DI system, a threshold of engagement intensity above which improvements could be expected, and, once reached, above which the effect of engagement intensity on user benefit levels off.

1.3 Contribution of this thesis

Few real-world studies have been conducted on DIs for SUD, or generally on DIs targeting higher acuity mental health problems (Bell et al. 2020; Ramos et al. 2021). However, this is an area where more research on the interplay of patient clinical complexity, engagement, and improvement - beyond RCTs - is needed to explore the role DIs could play for SUD treatment. Further, the quality of evidence that could be generated with such real-world studies is unknown as yet. This thesis sets out to explore these questions with data from a DI for substance dependence widely available in UK walk-in addiction services, and sustaining on the market for substance dependence DIs since 2012, "Breaking Free Online" (BFO). Uniquely, researchers from TELUS Health, the company which commercialised the BFO programme have published peer-reviewed studies on this intervention; many of them qualitative, but also quantitative, using comparatively small batches of data and standard statistical methodology. Details of the BFO programme and the evidence base behind it are available in Chapter 2. I am interested in what these data can tell us about outcomes, patients and (digital) treatment, and which challenges exist processing them. I am specifically interested in exploring appropriate use of ML methods and non-standard statistical methods for analysis of these data.

This thesis consists of five further chapters: (1) a detailed description of the BFO DI and the evidence base behind it for reference in the following chapters, (2) an exploration of minimally available real-

world data with random forests and multivariable and multivariate regressions, focussing on the association of user heterogeneity with outcomes, (3) a study on ML prediction of BFO engagement from data available before engagement is initiated, and (4) a study on the influence of multicollinearity - a defining characteristic of the datasets which have accumulated from BFO use, and many datasets in mental healthcare - on the reliability of explanations given for ML model predictions in mental healthcare research (5) conclusion and directions for future research.

The main contributions of this thesis are threefold: First, I deliver an analysis of DI data of high ecological validity, collected in real-world clinical settings across the UK. The dataset is, to my knowledge, the largest of its kind. My analysis largely focuses on the prediction of mental health outcomes, and DI adherence. Few studies on DIs for mental health have done this to date. In comparison to other studies, our data comes from a particularly severely impaired, and vulnerable patient group, people with substance dependence, who are often ignored by research.

Secondly, I deliver a careful, detailed analysis of these data: Specifically, I respect the ordinal data type of psychometric scales, the primary measurement instruments in mental health research, and - by including item-level data in my models - use all data available to me instead of aggregating the symptom-level data from questionnaires. Previous research has often aggregated symptom-level data, which may obscure details of a patient's condition and recovery.

Thirdly, I illustrate - against the backdrop of my analyses - the challenges arising when processing such data. This may inform future analyses of real-world data accruing from DIs which may help to address which may be the field's most pertinent problem to date, frequent disengagement from DIs in the real world.

Chapter 2

Breaking Free Online

Breaking Free Online (BFO) is a digital programme based on cognitive behavioural therapy (CBT), a widely used psychotherapeutic approach which has been demonstrated to be effective for a wide range of mental health problems. BFO was developed by clinical psychologists, researchers and clinicians in the field of SUD and behaviour change to support recovery from SUD and concurrent mental health problems. The programme has been available in community- and prison-based UK treatment services since 2012, and as a standard treatment in both Canadian community and US correctional treatment settings since 2019. Developed to target the biopsychosocial and lifestyle factors that underly SUDs more generally, users may address use of a wide number of substances with BFO, including novel psychoactive substances, substitute medications and prescribed medications of abuse.

Researchers at TELUS Health, the company which has commercialised the BFO programme, together with university academics and treatment service clinical staff and managers, have published research on BFO in peer-reviewed journals. In the following description of the BFO programme, I reference this research at appropriate times. A brief summary of available quantitative research specifically on the effectiveness of BFO is provided at the end of this chapter. Before that, I describe BFO's design and intended implementation. Finally, I provide details of my collaboration with researchers from TELUS Health for this thesis.

2.1 Description of the Breaking Free Online programme

An initial working version of BFO was built following a review of the literature around clinically effective approaches for substance abuse (Ward, Davies, et al. 2017) in 2010. The programme was further developed in consultation with treatment service users, treatment service staff, and BFO commissioners. Their involvement is documented in several qualitative studies using semi-structured interviews with

stakeholders which are analysed thematically (Dugdale, Elison, et al. 2016, 2017; Elison, Humphreys, et al. 2014). This thesis explores data from the web-based version of BFO deployed from 2016 - 2019.

BFO is available to users of treatment services commissioning BFO within the bundle of other support options available at these treatment services, which may include face-to-face individual or group therapy. If service users are interested in using BFO, they are provided with an access code which allows them to activate an account on the programme. The BFO website includes an informational video about BFO, accessible without an account, for individuals who have not accessed treatment services yet and hence have not been referred to the programme by service staff.

BFO is made available for self-help, or as an add-on programme to one-to-one or group work sessions facilitated by a practitioner. Researchers at TELUS Health have developed manuals for these practitioner-facilitated sessions.

Human support for BFO is available through treatment service staff, and peer mentors (Ward, Davies, et al. 2017). Treatment service staff are trained by members of TELUS Health to use BFO themselves, support and encourage BFO use in treatment service clients as well as make reference to BFO in their clinical practice. Training takes place within face-to-face, or virtual “train the trainer” sessions which are supposed to give treatment services the capacity to provide “in-house” training. Specific training material as well as a more general e-learning platform are made available to members of staff afterwards. Another source of support and motivation available to users at some treatment services are peer mentors who have experience of using BFO and are trained, equally in “train the trainer” sessions, to support others to engage with the programme.

In the following, the user interface and clinical content of BFO will be described in more detail.

Users can access BFO on personally owned devices, and devices provided at the treatment service. At first contact with the programme, all BFO users must agree to the terms of use of using BFO, and a privacy and cookies policy at the point of account creation. Each person who uses the BFO programme also provides separate consent for their data to be used for (1) storage and processing for the purposes of the functioning of the program, for example to allow personalised feedback to be provided to the participant on the basis of assessments administered within the programme, and (2) for the purposes of research. Before being able to provide their consent for their data to be used for research, participants are first required to read a research participant information screen, which is accessed via a hyperlink. Consent for data to be used for research can be revoked within the BFO user interface without losing access to BFO.

Before accessing the CBT-based content of BFO, individuals are required to complete an assessment of their substance use and dependence, mental health, quality of life, treatment goals, and wider biopsychosocial functioning. The latter aspect, wider biopsychosocial functioning, is assessed with the "Recovery Progression Measure" (RPM, Elison, Davies, and Ward, 2016; Elison, Dugdale, et al., 2017), a 36-item questionnaire measuring functioning in six biopsychosocial domains implicated in substance use and recovery.

The RPM (Elison, Davies, and Ward 2016) has been developed by researchers at TELUS Health with data collected from treatment service users in order to be able to measure changes in psychosocial functioning that may indicate an individual's increasing or decreasing "recovery capital", relating to, for example, stable relationship status, employment, and accommodation (Davies et al. 2015). The contribution of these factors to the development, maintenance, and recovery of SUD is often not captured in commonly used screening instruments for substance abuse.

Baseline assessment data are captured on a backend database for each user. After this initial assessment, answers must be updated at least bi-weekly in order for users to be able to continue accessing the CBT-based content in BFO. A shorter, 6-item version of the RPM, is integrated into these update assessments (Elison, Dugdale, et al. 2017). All updated assessment data is also captured on the backend database. The programme allows users to download any assessment data accumulated through their engagement with the programme, and to launch the process of purging this data if they wish to no longer access BFO.

Once users have completed the baseline assessment, they can access CBT-based content of BFO in the order that they wish – the programme is therefore, modular instead of linear. Their assessment data is used to populate a six-domain model (see **Fig. 2.1**), the "Lifestyle Balance Model" (LBM, Davies et al., 2015). In the BFO programme, the LBM is visualised at a central web page in the programme which users arrive at after their baseline assessment. This visualisation is supposed to help users understand which LBM domains of functioning are implicated in their own substance use, and substance use more generally. It also provides direct access to further CBT-based content of the programme.

On this particular web page, LBM domains are colour-coded based on RPM scores, with green, amber and red indicating, respectively, "little", "moderate" and "significant" impairment. The programme suggests to concentrate on completing clinical content of the programme that will address the domains of a user's functioning in the LBM where greatest levels of impairment are experienced (red domains of the LBM). A one-group pretest-posttest retrospective analysis of BFO users who updated their baseline assessment at least once found that these users follow this advice (Elison, Jones, et al. 2017).

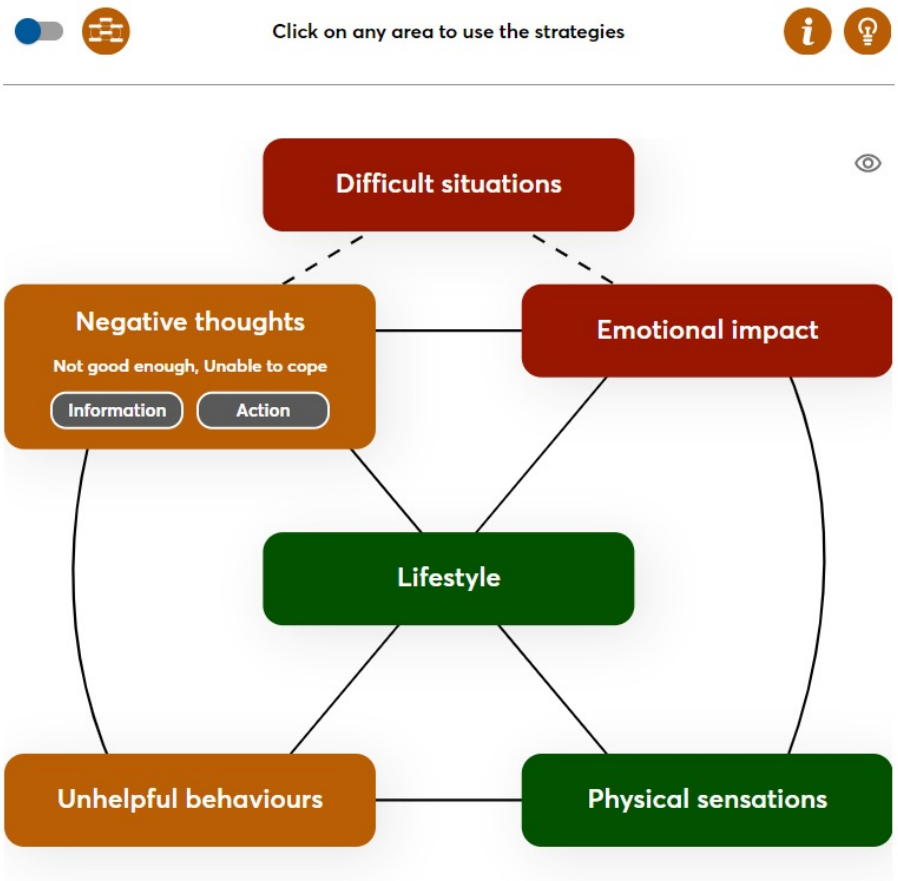

Figure 2.1. The lifestyle balance model, as visualised in the BFO programme.

Individuals can address each of the six domains of functioning included in the LBM by completing a corresponding psychoeducational "information strategy" and interactive skills-building "action strategy". These digital intervention strategies - 12 in total - were mapped in a qualitative study by Dugdale, Ward, et al. (2016) onto so-called "behaviour change techniques" (BCTs, Michie et al., 2013). These BCTs are based on therapeutic approaches such as CBT (Beck et al. 1993), relapse prevention (Marlatt and Donovan 2005), mindfulness (Marlatt, Bowen, et al. 2010), and motivational enhancement (Miller and Rose 2015). Each time a user completes a strategy in the programme, this is captured as a frequency count on the backend database – users can complete each strategy multiple times. **Table 2.1** describes the CBT-based content of BFO.

All BFO web pages are supported by audio or video content. Learnings from psychoeducation and interactive exercises, including user input, can be downloaded. They can also be sent per mail to the user and their BFO recovery supporters which can be nominated in the programme through entering up to three email addresses.

The BFO version used in correctional services is conceived to be delivered as part of a manualised, 8-session structured group or one-to-one intervention that is delivered by trained staff. It is to date the

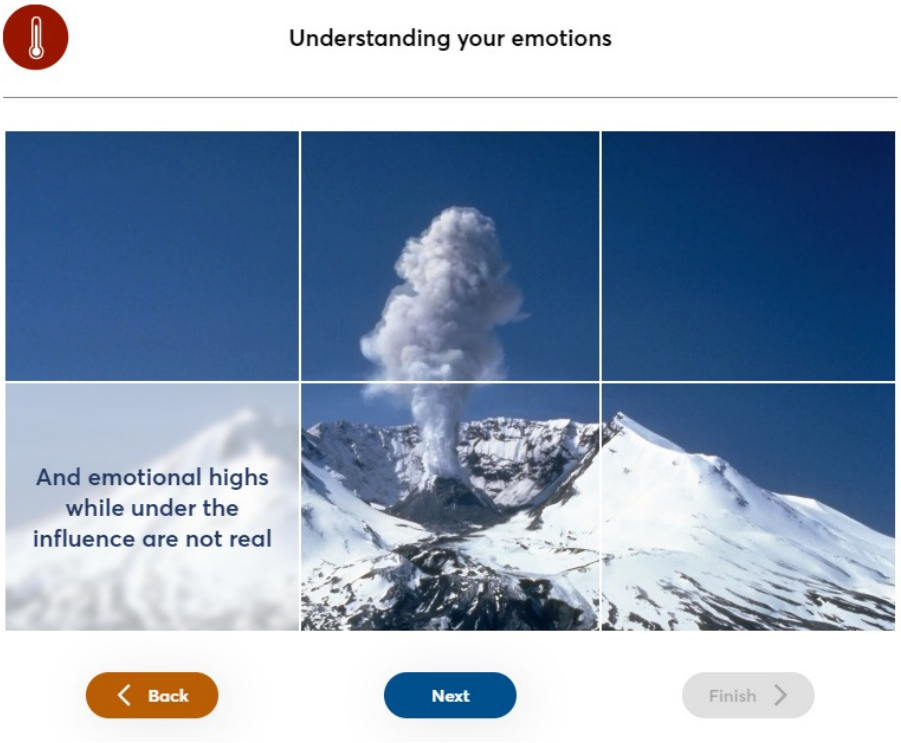

Figure 2.2. Example of an information strategy.

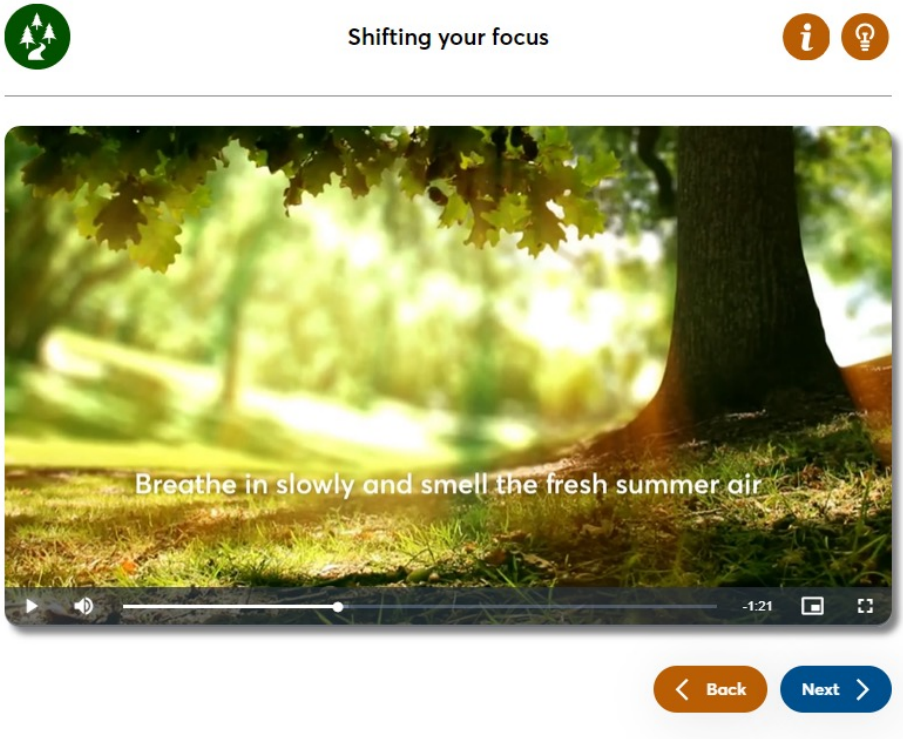

Figure 2.3. Example of a mindfulness video in BFO.

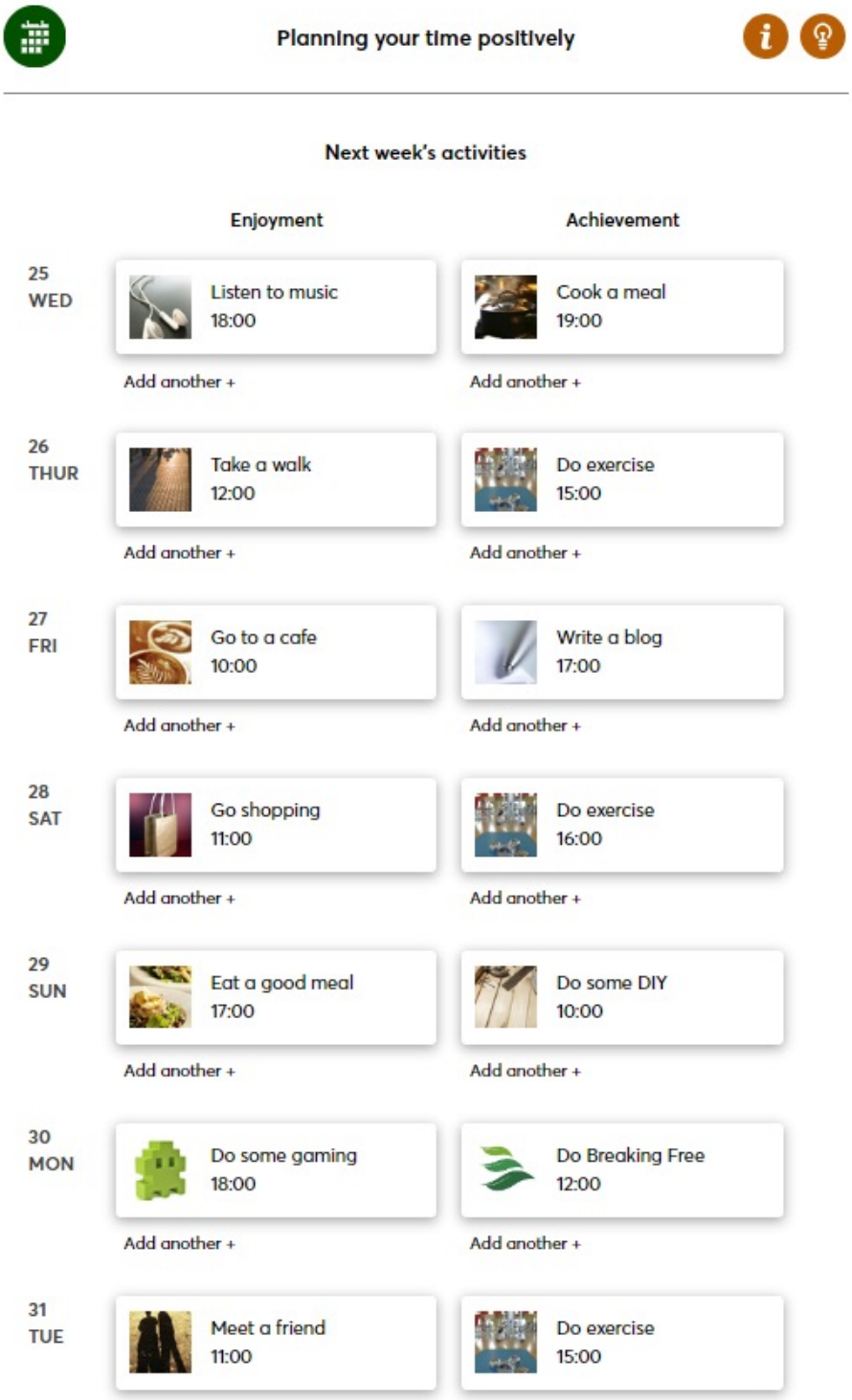

Figure 2.4. Example of a BFO action strategy meant to provide behavioural activation.

Table 2.1. Description of the CBT-based content in Breaking Free.

Content in Breaking Free Online	Clinical Description
Baseline and progress check assessments	Monitor behaviour to provide feedback about progress towards goals; encourage new behaviours via positive feedback
Lifestyle Balance Model	Personalised feedback; case formulation – understand the links between situations, thoughts, emotions, physical sensations, behaviours, and lifestyle
Difficult Situations area of LBM	Psychoeducation on impact of problematic situations; recognise–avoid–cope; relapse prevention for coping with environmental/situational/emotional triggers; creating action plans on how to avoid or cope in high-risk situations; assertiveness and refusal skills training
Negative Thoughts area of LBM	Psychoeducation on impact on negative thoughts; cognitive restructuring; challenge thoughts that may be unhelpful
Emotion Impact area of LBM	Psychoeducation on impact on emotions; attention narrowing; attention switching; emotional regulation; recognise/understand/normalise emotions
Physical Sensations area of LBM	Psychoeducation on impact of physical sensations; understanding the nature of cravings and urges; body scanning; urge surfing; relapse prevention-based techniques
Unhelpful Behaviours area of LBM	Psychoeducation on impact of destructive behaviours; activity scheduling; behavioural activation; encourage new behaviours via planning purposeful activities; increase activity to increase self-esteem and avoid boredom
Lifestyle area of LBM	Psychoeducation on impact of lifestyle; creating SMART systematic goal setting; implementation intentions for goal attainment; increase readiness to change behaviour; positive reinforcement

only DI approved by Her Majesty's Prison and Probation Service as an effective regime intervention, and accredited by the UK Ministry of Justice Correctional Services Advice and Accreditation Panel.

2.2 Evidence base for effectiveness of Breaking Free Online

Several quantitative studies have been conducted on BFO in a retrospective one-group pretest-posttest design (Elison, Davies, and Ward 2015; Elison, Jones, et al. 2017; Elison, Ward, et al. 2017; Elison-Davies, Märtens, et al. 2021; Elison-Davies, Wardell, et al. 2021; Ward, Elison-Davies, et al. 2019). These studies often concentrated on data from community addiction service users who had completed

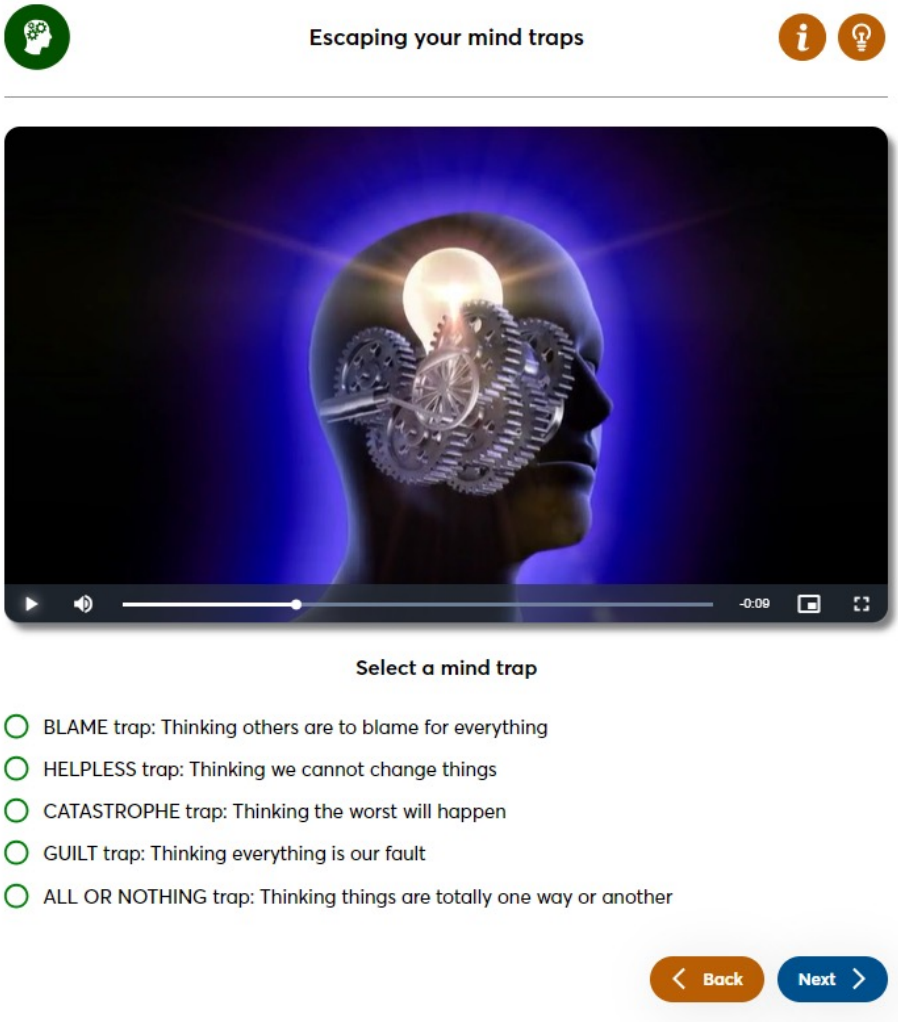

Figure 2.5. Example of a BFO action strategy meant to support cognitive restructuring.

at least one assessment update. Some studies focused on a particular substance using group, for example, alcohol (Ward, Elison-Davies, et al. 2019), cannabis (Elison-Davies, Wardell, et al. 2021), or opioid users (Elison-Davies, Märtens, et al. 2021). Sample sizes ranged from n = 117 (Elison, Ward, et al. 2017) up to n = 2311 (Elison, Jones, et al. 2017). Primary outcomes of these studies were recovery progression (measured through the RPM), quality of life, substance dependence, and anxiety and depression symptoms, assessed through questionnaire sum scores at the last-contact assessment. In individual cases, study outcomes also included questionnaire-threshold-based anxiety and depression (Elison, Ward, et al. 2017), the number of strategy completions (Elison, Jones, et al. 2017; Elison-Davies, Wardell, et al. 2021), treatment goal achievement (Ward, Elison-Davies, et al. 2019), or substance-using days per week (Elison-Davies, Märtens, et al. 2021), all assessed digitally through the BFO programme.

Based on pre-post reductions of questionnaire sum scores and weekly substance using days observed in these studies, the authors conclude that the BFO programme can be effective (Elison, Davies, and

Ward 2015; Elison, Ward, et al. 2017; Elison-Davies, Märtens, et al. 2021; Ward, Davies, et al. 2017). Elison, Ward, et al. (2017). For example, they analyse pre-post symptom data gathered from people triaged into eHealth support with BFO, or with two other, relatively mature DI systems, Sleepio (sleep problems) and Living Life to The Full (anxiety and depression), which are subsequently monitored and followed-up. Authors report that effect sizes for pre-post reductions of sum scores of depression and anxiety questionnaires were in the medium to large range for all three DIs.

Elison, Jones, et al. (2017) and Elison-Davies, Wardell, et al. (2021) further report a "dose response" based on associations of a greater number of module completions with greater changes in psychometric questionnaire sum scores at the time point of the assessment update, and that cognitive restructuring based programme strategies were associated with greatest improvements. On the basis of associations between treatment goals, baseline characteristics, and treatment goal achievement at the last-contact assessment for those who complete an assessment update, Ward, Elison-Davies, et al. (2019) also speculate that specific treatment goals may lead to greater symptom reduction for specific substance user groups. RCTs evaluating BFO with varying degrees of human support against standard treatment are currently underway (Elison-Davies, Davies, et al. 2018; Elison-Davies, Pittard, et al. 2023; Quilty et al. 2022).

Strengths of the evidence base for the effectiveness of BFO include that research on this DI has been published in peer-reviewed journals, which is rarely the case for DIs available for mental health, and that significant symptom reductions were found across several real-world studies and for most outcomes.

Limitations include the pre-post-test design of published studies which does not permit conclusions about BFO efficacy, and which may overstate benefits which may be obtained from use of the programme. Further, analysis concentrates on people who have completed at least one assessment update, and therefore, findings only apply to individuals who have demonstrated some degree of engagement with BFO. Further, symptom reductions are often measured on an aggregate level (questionnaire sum scores), which obscures detail about the specific symptoms reduced, and about whether these reductions translate into clinically significant improvements, especially with regards to substance use.

To clarify how research conducted for Chapter 3 of this thesis is different from the research already conducted on BFO, I list the most pertinent differences in the following. When assessment data were analysed in previous studies on BFO, the focus was on questionnaire sum scores, instead of substance use outcomes, individual symptoms assessed through questionnaire items, or threshold-based outcomes. Conventional statistical methodology used in the field of psychology was used: Hypotheses were tested using general linear model based approaches such as ANOVA and univariate regres-

sions, with the only exception of Elison-Davies, Märtens, et al. (2021) who included a categorical variational autoencoder based consensus score based on baseline assessment data into their analysis. More research on BFO outcomes and engagement is therefore needed to unlock the potential of the programme for improving outcomes for people with substance abuse problems.

2.3 Collaboration with TELUS Health for this thesis

The PhD project resulting in this thesis was set up as a collaboration between the University of Manchester and TELUS Health. It meant to address the lack of research on real-world, post-deployment use and outcomes of mental health DIs with machine learning and non-standard statistical methodology, and on interventions for less prevalent but equally debilitating mental illnesses such as substance dependence. Dr Sarah Elison-Davies, research director at TELUS Health, co-supervised my PhD and provided me with the real-world datasets analysed in Chapter 3, Chapter 4 and Chapter 5. Supervisory meetings took place on a needs basis. They initially involved answering questions on the functioning of the BFO programme, introduction of literature on substance dependence, and, one time, the suggestion of an analysis target, the implication of gender in answering patterns before and after use of the BFO programme. A few months into my PhD, supervisory meetings primarily involved advice on academic writing, and progress reports from my side. Additionally, my attendance of the conference of the European Federation for Medical Informatics 2023 was partially funded by TELUS Health.

Chapter 3

Identifying factors associated with user retention and outcomes of a digital intervention for substance use disorder: a retrospective analysis of real-world data

Due to the relative novelty of digital interventions (DIs) in mental healthcare, and the fact that research resources in this area tend to get distributed towards the more prevalent anxiety and depressive disorders instead of substance use disorder (SUD), factors implicated in engagement and outcomes of DIs for SUD are still insufficiently understood. Real-world data may be especially fit to identify these factors, having recently been proposed to bear potential to ”identify patient subgroups who may respond differently to treatment” (Kympouropoulos, 2023, p. 2).

Factors implicated in engagement and outcomes of DIs for SUD may differ from those found to impact outcomes of face-to-face treatment. Factors impacting outcomes of face-to-face treatment include patient-related factors such as socioeconomic status, the primary substance abused, the presence of poly-drug use, and individual adaptability to slips and relapses, as well as treatment-related factors such as treatment provider proactivity and focus on long-term recovery (Argyriou et al. 2023; Lappan, Brown, and Hendricks 2020; Svendsen et al. 2021). Factors implicated in real-world outcomes from

DIs for SUD are less well known, and understanding them better may lead to better DI interface design and general DI implementation, and eventually greater patient benefit.

The following chapter describes the exploration of real-world data from registrants with “Breaking Free Online” (BFO, see Chapter 2) in terms of descriptive and inferential statistics as well as predictive modelling, and information about digital mental health use(rs) that can be obtained from it. I take a detailed look at the association of individual and treatment service related factors with outcomes on an item and aggregate level.

In this study, I have chosen to pay special attention to individual gender as such a factor associated with outcomes. Prevalence of alcohol and drug use, as well as representation in face-to-face substance abuse treatment, is higher for men than women in most countries, including the UK (Bankiewicz and Robinson 2020; European Monitoring Centre for Drugs and Drug Addiction 2020; National Drug Treatment Monitoring System, Office for Health Improvement & Disparities 2023). Further, initiation of substance use, for example alcohol use, starts earlier for men (Diehl et al. 2007; Keyes et al. 2010; Lewis, Hoffman, and Nixon 2014). However, research suggests that women abusing substances face unique clinical challenges (Khan, Okuda, et al. 2013; Khan, Secades-Villa, et al. 2013; Lewis and Nixon 2014; National Drug Treatment Monitoring System, Office for Health Improvement & Disparities 2023).

This chapter was published as an article in the JAMIA Open journal as “Günther, F., Wong, D., Elison-Davies, S., & Yau, C. (2023). Identifying factors associated with user retention and outcomes of a digital intervention for substance use disorder: a retrospective analysis of real-world data. *JAMIA Open, 6*(3), Article ooad072. https://doi.org/10.1093/jamiaopen/ooad072”. For this thesis, small changes were made to the original text at relevant points, including the insertion of the supplementary materials of the published article into the chapter.

3.1 Abstract

3.1.1 Objective

Successful delivery of digital health interventions is affected by multiple real-world factors. These factors may be identified in routinely collected, ecologically valid data from these interventions. I propose ideas for exploring these data, focussing on interventions targeting complex, comorbid conditions.

3.1.2 Materials and Methods

This study retrospectively explores pre-post data collected between 2016 and 2019 from users of digital cognitive behavioural therapy (CBT) - containing psychoeducation and practical exercises - for substance use disorder (SUD) at UK addiction services. To identify factors associated with heterogenous user responses to the technology, I employed multivariable and multivariate regressions and random forest models of user reported questionnaire data.

3.1.3 Results

The dataset contained information from 14,078 individuals of which 12,529 reported complete data at baseline and 2,925 did so again after engagement with the CBT. 93% screened positive for dependence on one of 43 substances at baseline, and 73% screened positive for anxiety or depression. Despite pre-post improvements independent of user sociodemographics, women reported more frequent and persistent symptoms of SUD, anxiety and depression. Retention - minimum two use events recorded - was associated more with deployment environment than user characteristics. Prediction accuracy of post-engagement outcomes was acceptable (Area Under Curve, AUC: 0.74-0.79), depending non-trivially on user characteristics.

3.1.4 Discussion

Traditionally, performance of digital health interventions is determined in controlled trials. My analysis showcases multivariate models with which real-world data from these interventions can be explored and sources of user heterogeneity in retention and symptom reduction uncovered.

3.1.5 Conclusion

Real-world data from digital health interventions contain information on natural user-technology interactions which could enrich results from controlled trials.

3.2 Background and Significance

Substance use disorders (SUDs) are widely recognised as a major contributor to global disease burden (Vos et al. 2020). The impact of SUDs is intensified by consistently low SUD treatment rates (Degenhardt et al. 2017), which result from a variety of factors. Among these are variable access to evidence-based treatment due to financial barriers (Ali, Teich, and Mutter 2017), place of residence (Browne et al. 2016), treatment service workforce capacity (Black 2020), and recently, the circumstances of a global pandemic (Russell et al. 2021). Other factors are linked to the unavailability of treatment services at times of increased risk of substance use or relapse, for example out of treatment service operating hours (Phillips, Epstein, and Preston 2013) or during periods of transition between treatment settings, for example when re-entering the community after a custodial sentence (Pelissier, Jones, and Cadigan 2007). Further, fear of stigmatisation (Hammarlund et al. 2018) and an unstructured lifestyle centering around substance provision and consumption may prevent attendance of treatment appointments.

In contrast to traditional face-to-face delivery, digital delivery promises to widen access to SUD treatment by being anonymous, scalable, and accessible anywhere and anytime. So-called digital interventions (DIs) make therapeutic content available on the internet, via telephone or video chat, or within a smartphone app. DIs for mental and physical health may accumulate large quantities of individual-level usage data, often in order to personalise content, but these data can also provide information on real-time user experience and recovery progression (Marsch 2021).

Further, the ecological validity of these data contrasts with that collected in highly controlled studies, such as randomised controlled trials (RCTs), in which the natural interaction of a user with a DI may be perturbed by study participant selection, monitoring and incentives (Baumel, Edan, and Kane 2019; Fleming et al. 2018). As sustainable uptake of many digital health interventions remains low despite rigorous study of their efficacy (Torous, Nicholas, et al. 2018), post-deployment exploration of ecologically valid data may produce novel insights into real-world effectiveness of the DI.

Little research in this area exists to date. Typical research goals are the identification of determinants of DI success, including sociodemographic user characteristics, baseline case severity or engagement frequency. Success has been defined in terms of diagnostic outcomes, programme completion or retention. Chien et al. (2020) used a probabilistic latent variable model to identify user subgroups within longitudinal engagement data from a DI for anxiety and depression. Bell et al. (2020) employed k-modes clustering with a similar agenda and describe the role of notifications for daily engagement patterns of users of a DI for the reduction of harmful and hazardous alcohol use in the general popula-

tion. Titov et al. (2020) reported yearly trends in user characteristics and outcomes for a national DI for anxiety and depression. Ramos et al. (2021) predicted goal attainment and completion of a linear DI programme for alcohol, marijuana and tobacco use with a random forest, using engagement data from the first three days of use.

More research into the impact of user heterogeneity on intervention outcomes, however, is warranted, especially analyses of data from interventions targeting individuals with complex and transdiagnostic symptom profiles common in mental health contexts.

3.3 Objectives

In this study, I conduct a retrospective analysis of real world data from a DI to understand the factors associated with its deployed performance. I use the example of BFO, a CBT-based DI widely established in SUD treatment services in the United Kingdom. My analyses focus on static user characteristics as a source of heterogenous responses to the DI, and the prediction of post-engagement symptomatology.

3.4 Materials and Methods

This study is a retrospective, exploratory, one-group pretest-posttest study of individuals accessing 477 SUD treatment services in England, Scotland and Wales between January 4, 2016 and December 6, 2019, with each agreeing to create a BFO account.

BFO is a tailorable, CBT-based DI programme offered by commissioning SUD treatment services in the community or in correctional environments (Elison, Ward, et al. 2017; Elison-Davies, Märtens, et al. 2021; Elison-Davies, Wardell, et al. 2021; Ward, Elison-Davies, et al. 2019). Unlimited access to clinical content on personally or service owned devices is provided free of charge. Users are required to complete an integrated digital questionnaire-based assessment of their recovery progression and sociodemographic background before first access and at least fortnightly after that. Care was taken to ensure assistance with the programme is available to users at participating services, and to closely integrate the programme with face-to-face clinical practice (Ward, Davies, et al. 2017).

3.4.1 Clinical model of BFO

BFO content was developed by researchers and clinicians in the field of SUD and behaviour change, in consultation with multiple stakeholders, including people accessing SUD treatment and people in long-term recovery from SUD, practitioners, service managers and commissioners (Dugdale, Elison, et al. 2016, 2017; Elison, Humphreys, et al. 2014). BFO is endorsed by the UK National Institute for Health and Care Excellence (NICE) and the content of the programme complies with NICE guidance around treatment for SUD.

Access to BFO is provided after account creation and completion of an assessment battery of the user's demographic characteristics, substance use, substance dependence, mental health, quality of life, biopsychosocial functioning and treatment goals. Account creation requires an access code from an SUD treatment service commissioning BFO, and an email address. Service staff are trained to assist during the entire first contact with BFO and beyond. Additionally, the BFO website features an informational video about BFO, accessible without an account, for individuals who have not accessed treatment services yet and hence have not been referred to the programme by service staff. Users can access BFO on personally owned devices and devices provided at the service.

BFO can be integrated with face-to-face interventions at commissioning SUD treatment services. Practitioners and clinicians are trained via an e-learning platform to interlock clinical practice with BFO by referring to and reflecting on learnings, and offering support. Another source of support and motivation available to users at some treatment services are peer mentors who have experience of using BFO and are trained to support others to engage with the programme. The BFO version used in correctional services is delivered as part of a manualised, 8-session structured group or one-to-one intervention that is delivered by trained staff. It is to date the only DI approved by Her Majesty's Prison and Probation Service as an Effective Regime Intervention, and accredited by the UK Ministry of Justice Correctional Services Advice and Accreditation Panel.

BFO builds on the six-domain Lifestyle Balance Model (Davies et al. 2015) which details domains of functioning implicated in substance use disorder: "impact of emotions", "unhelpful behaviours", "physical sensations", "difficult situations", "negative thoughts", and "lifestyle". BFO visualises a user's degree of functioning in each of these domains based on their answers from the assessment battery included in the programme. After the initial pre-engagement assessment, these answers must be updated at least bi-weekly in order for users to be able to continue accessing the clinical content in BFO. In this study, users who have updated their assessment at least once are considered retained in treatment, and otherwise, dropouts.

Each domain is associated with slide series based psychoeducation on the impact of this domain on functioning ("information strategy"), and an interactive, skills building exercise ("action strategy"). Domain modules can be accessed in any desired order and pace. The interactive exercises make use of a range of evidence-based behavioural change techniques, including refusal and assertiveness skills, emotional regulation, coping strategy enhancement, mindfulness-based cognitive therapy, motivational enhancement, cognitive restructuring, reward and reinforcement, harm reduction and crisis management (Dugdale, Ward, et al. 2016). All BFO pages are supported by audio or video content. Learnings from psychoeducation and interactive exercises, including user input, can be downloaded. They can also be sent per mail to the user and their BFO recovery supporters which can be nominated in the programme through entering up to three email addresses.

To tailor the LBM visualisation in the BFO programme to an individual user's recovery progression and challenges, users are required to provide consent for the collection, storage, and use of their non-identifiable data. Specific consent to use of data for research is not required for access to BFO. The programme allows users to download any assessment data accumulated through their engagement with the programme, and to launch the process of purging their data if they wish to no longer access BFO.

This study analyses user self-reported data from the first- and, if available, last-contact assessment. In the following, assessment times will be referred to as "pre-engagement" and "post-engagement". This relates to my decision to consider those dropouts who have not completed the assessment at least twice and hence did not engage with the programme in the recommended way. Those who did not dropout are subsequently referred to as "retained". Note that this study was conducted in the first year of my PhD when first- and last-contact assessment data, but no strategy completion data were available to me.

Ethical approval for collection, storage and use of data accumulating from routine use of BFO by clients in participating treatment services, was obtained from an NHS Research Ethics Committee (London - South East, 16 May 2012 and 22 May 2017, references 12/LO/0076 and 12/LO/0287).

3.4.2 Data description

The BFO assessment combines a variety of items associated with validated and standardised psychometric scales, the Severity of Dependence Scale (SDS), the Patient Health Questionnaire 4 (PHQ-4), the World Health Organization Quality of Life measure (WHOQOL-BREF) and the Recovery Progression Measure (RPM), and additionally collects data on the substance-using days in the preced-

ing week. Over the course of this chapter, I will refer to “recovery progression” as the overall construct measured within the BFO assessment rather than the RPM.

The SDS (Gossop et al. 1995) is a five-item scale from 0 to 3 measuring the degree of dependence on a substance. If the total of the answers on these items ≥ 3, an individual is considered dependent. Users addressing both alcohol and drug use with BFO are presented with the SDS twice to assess their dependence on each substance. Pre-engagement, users are presented with an additional item on the impact their target substances have on them.

The PHQ-4 (Kroenke et al. 2009) is a four-item screener for depression and anxiety, with items scored on a Likert scale from 0 to 3. The total of the answers on the first, or last two items, meeting a cutoff of 3 is considered an indicator of anxiety, or depression, respectively. The PHQ-4 is presented together with an item from the RPM on the impact of emotions. Pre-engagement, an additional item on negative self-perception is presented to users alongside the PHQ-4. Additional items are presented to align with the RPM subscale format, i.e. five items on a specific aspect of functioning plus one item on this aspect’s impact.

Five items (number 1, 2, 17, 18, 20) from the WHOQOL-BREF (Skevington, Lotfy, and O’Connell 2004) are presented alongside two items on overall satisfaction with life and ability to cope with life’s difficulties.

From the RPM (Elison, Davies, and Ward 2016), a 36-item scale measuring biopsychosocial functioning during SUD recovery, subscales, “unhelpful behaviours”, “lifestyle”, “physical sensations”, “difficult situations” and “negative thoughts” are used. Post-engagement, a six-item screening version of the RPM (Elison, Dugdale, et al. 2017) is used.

3.4.3 Statistical analysis

I included individuals who completed the pre-engagement assessment, defined as answering at least one item associated with a scale, and sub-scale in the case of the RPM.

Sociodemographic characteristics of the population for which BFO was an attractive option when accessing an SUD treatment service, and their time of engagement with BFO in weeks - if a post-engagement assessment was completed - were analysed descriptively.

Several regression models were used to understand the association of participant characteristics with retention and post-engagement outcomes. Among these was a multivariate, multivariable ordinal logistic regression model inspired by Hirk, Hornik, and Vana (2020) which is faithful to the ordinal

Table 3.1. Models used to understand associations of participant characteristics with retention and post-engagement outcomes. §To ensure identifiability of the models and their parameters, I followed recommendations in Hirk, Hornik, and Vana (2020) to fix intercept parameters to zero and error variance to unity for all models. I also assumed a general correlation structure of correlated outcomes, proportional odds, and a multivariate logit link function. †Included in model of pre-engagement assessment. ‡Included in model of pre- and post-engagement assessment.

Regression model	Outcome	Inputs	Example
univariate multivariable binary logistic	meeting of cutoff for anxiety, depression or substance dependence	age, gender, ethnicity, service, target substance, retention†, assessment time‡, gender * assessment time‡	"pre-engagement meeting of cutoff for anxiety ∼ age + gender + ethnicity + service + target substance + retention"
univariate multivariable ordinal logistic	number of substance-using days per week		"number of alcohol-using days per week ∼ age + gender + ethnicity + service + target substance + assessment time + gender * assessment time"
multivariate multivariable ordinal logistic§	items originating from or presented alongside with SDS, PHQ-4, WHOQOL-BREF, RPM subscale†, Rapid RPM‡		"SDS ∼ age + gender + ethnicity + service + target substance + assessment time + gender * assessment time"
univariate multivariable binary logistic	retention		"retention ∼ age + gender + ethnicity + service + target substance"

measurement level of data self-reported with questionnaires (Anvari 2023; Liddell and Kruschke 2018), and the high correlations which often exist between questionnaire items (see **Fig. 3.1**).

I apply this methodological approach to describing the association of participant characteristics and post-engagement outcomes through the example characteristic of gender. Gender may serve as a suitable example characteristic for this study as its role in SUD treatment continues to be specified in the scientific literature (McHugh et al. 2018). However, the approach can be adapted to explore the interplay of other participant characteristics and DI retention and outcomes. To avoid bias of estimates of the effect of gender, as well as of statements about transgender individuals with SUD, I did not include transgender individuals in my analysis, pertaining to their small proportion of the sample (0.4%). Details of statistical models can be found in **Table 3.1**. An example for one of my multivariate, multivariable models would take age, gender, ethnicity, service, target substance, assessment time, and the interaction between gender and assessment time as inputs to model answering patterns on all SDS items pre- and post-engagement for retained participants.

Following practices by Titov et al. (2020), reliable recovery was defined as the proportion of participants meeting the cutoff for substance dependence, depression or anxiety on the PHQ-4 and SDS screening instruments pre-engagement, but not post-engagement; or abstinence in terms of zero substance-using days in the preceding week post-engagement, but not pre-engagement. Reliable deterioration was defined as the opposite.

I predicted participant retention and post-engagement outcomes with random forests, using the complete questionnaire and sociodemographic data available from the pre-engagement assessment where possible (see **Table 3.2**). For binary outcomes, the R package “randomForest” was used, for ordinal outcomes the R package “ordinalForest”. Predictions were made in ten repetitions of 10-fold cross validation. Predictor importance was assessed by computing the mean decrease in accuracy at random permutations of a predictor in each fold for binary outcomes. For ordinal outcomes, predictor importance uses the ranked probability score as an error measure, accounting for the ordinal measurement scale of the outcome (Epstein 1969). Resulting ranks were aggregated for each predictor across folds and outcomes predicted by the same set of predictors.

As this visualisation method of the effects of predictor values on the outcome in ML models is reported to account for correlations between predictors, accumulated local effects (Apley and Zhu 2020) were used to visualise average effects of pre-engagement answers on the 10 most important drivers of prediction, on post-engagement anxiety as an example outcome. In principle, accumulated local effects visualisations illustrate how the prediction of an ML model would change if only the predictor value of interest were changed within a quantile-defined window. While accurate prediction would be of clinical

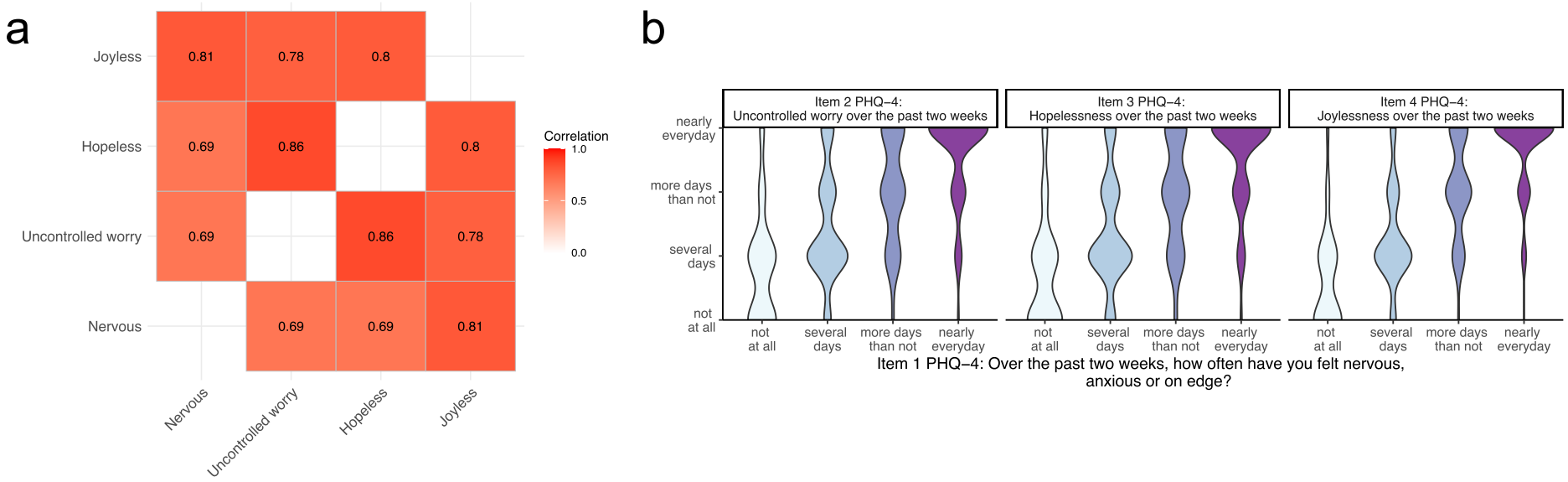

Figure 3.1. Correlations between PHQ-4 items. Correlation matrices as in (a) are estimated alongside other parameters of my regression model inspired by Hirk, Hornik, and Vana (2020). This allows accounting for the high correlation between items of a scale estimated (range ρ for PHQ-4 items: 0.69 - 0.86) when modelling single items. High correlation is also suggested by a high rate of empirical co-occurrence of the same answer category on two items of the same scale in the data, as shown for the PHQ-4 in (b).

utility to allocate targeted support at individuals with negative prognoses, examining the importance of predictors also allows us to gain a better understanding of the relative importance of factors implicated in DI outcomes at the level of the individual user.

Details of random forest prediction

Pre-engagement answers on items associated with the SDS are used as predictors only if post-engagement answers on these items are predicted. The rationale behind this was that the SDS was administered twice for participants using BFO to target alcohol and drug use and their predictor set would be larger or primacy of one substance would have to be determined which may conceal dependence on both. 500 trees were grown in an individual random forest. The number of predictors randomly sampled at each node split to increase tree diversity and, ultimately, reduce overfitting, corresponded to the square root of the number of predictors included to predict an outcome. Random forest performance was assessed by calculating ROC curves and areas under them which were averaged across folds. Accumulated local effects, used to illustrate the effects of specific values on prediction-driving predictors on post-engagement anxiety, describe the averaged main effect of an answer on one of these drivers on the predicted probability of anxiety.

To better understand the characteristics of users for which my prediction model fails for post-engagement anxiety, depression, drug and alcohol dependence, I employed binary logistic regression models to identify associations of repeated participant misclassification (model output) with participant sociodemographic characteristics and their meeting of cutoffs for substance dependence, depression or anxiety screeners (model input). I defined repeated misclassification as incorrect prediction in 100% of cases across cross-validation folds.

I encounter missing data on some variables which I planned to integrate into my data models. Missingness in my case is a result of technical failure to record participant-input answers or to remind participants of missing input. The strategies for dealing with missing data are illustrated in **Table 3.2**. As my complete case analyses never result in the exclusion of more than 5% of cases, I deem my decision for them appropriate (multivariate imputation will likely not influence results). When I exclude the weekly number of substance using days from predictor sets or decide not to use the variable as an outcome in the model, I do this knowing that meeting the cutoff for clinical SUD gives us a reasonably good proxy for an individual's substance dependence.

Table 3.2. Missing values on variables and strategies of dealing with them. *Outcome in ordinal regression model. †Predictor in random forest model. ‡Outcome in random forest model. Model tolerance of substantial amounts of missing outcome data can be assumed for the multivariate, multivariable ordinal regression models by Hirk, Hornik, and Vana (2020). Two numbers of missing data for a variable indicate different numbers of missing data for the datasets with and without dropouts. Complete variables may be excluded from models because their equivalents for another target substance are excluded (see Drug using days).

Variable name	Assessment time	Number of missing data	Strategy
Impact of alcohol	pre-engagement	1*	tolerance*
Alcohol using days	pre-engagement	5*, 1*†	complete case analysis*, exclusion†
Alcohol using days	post-engagement	3*‡	complete case analysis*, exclusion‡
Drug using days	pre-engagement	10*, 0†	complete case analysis*, exclusion†
Drug using days	post-engagement	0‡	exclusion‡
Coping with life's difficulties	post-engagement	495*‡	tolerance*, exclusion‡

3.5 Results

3.5.1 Participant description

Across the study period, 14,078 users of addiction services in England, Scotland and Wales created a BFO account. From these registrations, 12,529 (90%) initially completed BFO's assessment battery which is required to access clinical content, and were eligible for statistical analysis (see **Fig. 3.2**). This cohort was characterised by a mean age of 40.28 years (SD: 11.67, range: 18 - 99), and a proportion of 43% (5,346) identifying as women. The proportion of participants identifying as white was 93% (11,703), 2% (243) identified as Asian/Asian British, 2% (214) identified as Black/Black British, 2% (299) as having a mixed ethnic background, and 1% (70) as having another ethnicity than the ones specified. A total of 93% (11,606) of participants accessed BFO through community services while 7% (923) obtained access through correctional addiction services. Users sought help for their use of alcohol (57%, 7,090), drugs (22%, 2,708) and for the use of more than one substance (22%, 2,731). Within the cohort, 66% (8,235) identified alcohol as their most problematic substance, 9% (1,132) named cocaine, 8% (960) heroin, 6% (805) marijuana and 5% (567) crack.

The sociodemographic characteristics of and substances used by the cohort reflect those of the population seeking help for their substance use in England and Wales, which 99% of BFO users in this study participate from, except that BFO attracts more women and more individuals seeking help for alcohol use (National Drug Treatment Monitoring System, Office for Health Improvement & Disparities 2023; NHS Wales Informatics Service 2019). Despite that, sociodemographic characteristics differ from those of the general population in the UK, especially with regards to a greater proportion of male

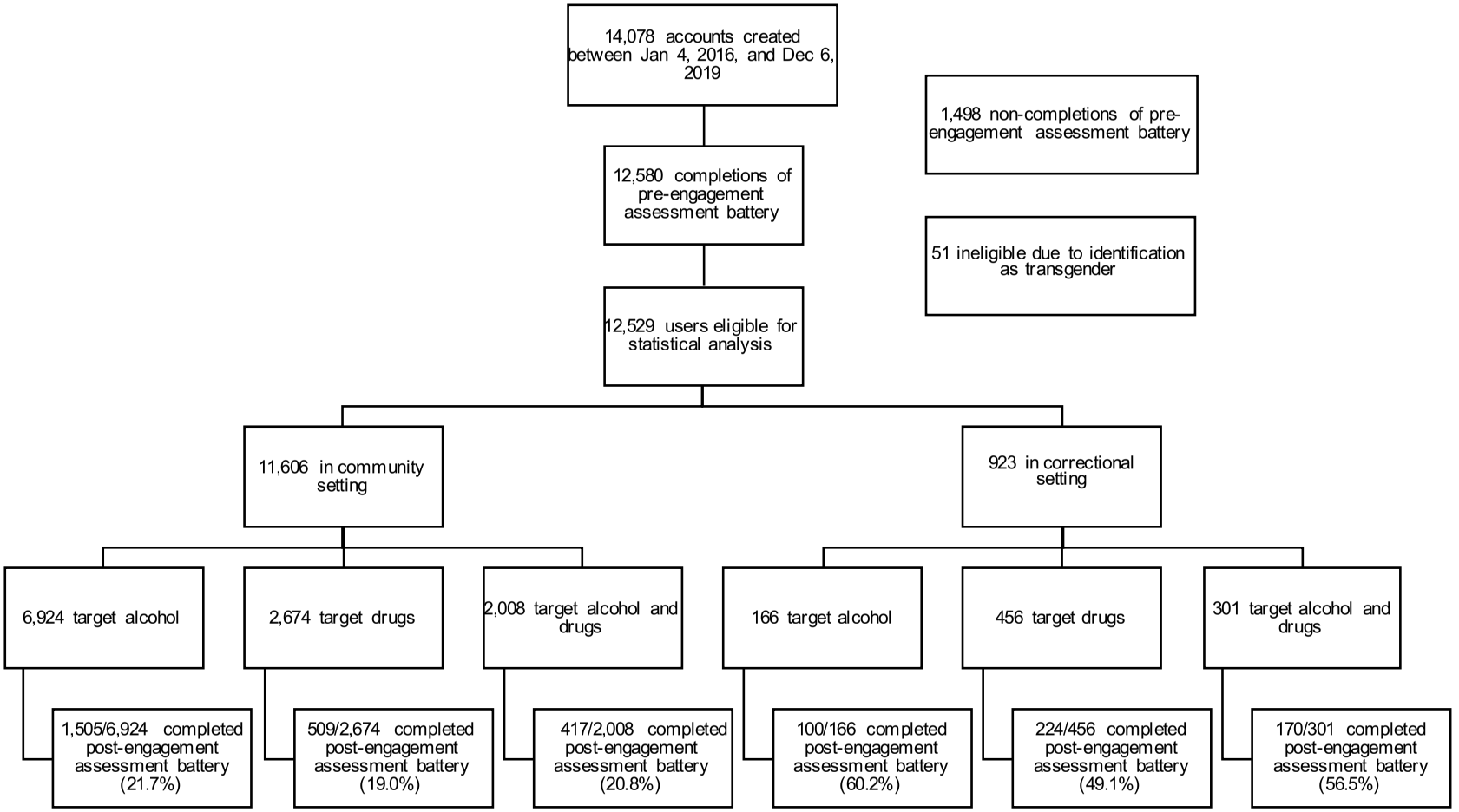

Figure 3.2. Participant flow through BFO and eligibility. Completion of the pre-engagement assessment battery is defined as providing an answer for at least one item per psychometric (sub-) scale. Information about missing data in the dataset used for statistical analysis is provided in Table 2.

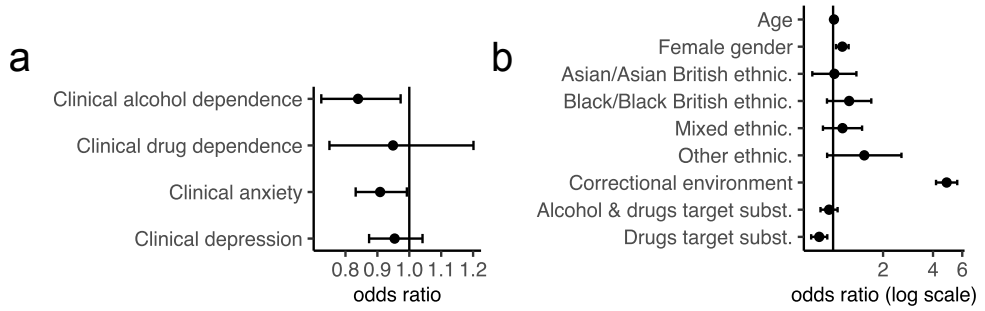

Figure 3.3. Association of participant retention with pre-engagement recovery progression and sociodemographic characteristics. Association analysis showing (a) odds ratios of participant retention against four clinical outcome measures; (b) odds ratios of association with sociodemographic membership. Reference groups are male gender, white ethnicity, community environment, and alcohol as a target substance. ORs are shown with their 95% confidence intervals.

(Office for Health Improvement and Disparities 2021) and white (Office for National Statistics 2021) users.

3.5.2 User characteristics associated with participant retention

Outcomes of DIs have to be evaluated in the light of how many users these interventions were able to retain. Retention can be low, especially when these DIs cater to individuals whose condition involves strong cognitive pre-occupations and high risk of relapse like SUD, which can interfere with regular engagement. I found BFO retained 23% (2,925) of participants for an average of 9 weeks (SD: 17 weeks, range: 0 days - 142 weeks).

I suspected participant dropout to be non-random and partly driven by specific, quantifiable differences between participants at baseline and found statistically significant associations between dropout and answering patterns on 13 out of a total of 57 items measuring participants' recovery progression at baseline (see **Fig. 3.5**). **Fig. 3.3** a shows that meeting the cutoff for clinical anxiety or alcohol dependence ($OR_{anxiety}$ = 0.91, p = 0.033, $OR_{alcohol\ dependence}$ = 0.84, p = 0.020) was associated with dropout. While male gender (OR = 1.14, p = 0.004) and intending to address drug use with BFO (OR = 0.82, p = 7.29 $\times 10^{-4}$) were also associated with participant dropout, accessing BFO from a community SUD service (OR = 4.83, p = 9.22 $\times 10^{-98}$) showed the strongest association (**Fig. 3.3**b).

Retrospective analysis of BFO data has therefore identified that participant dropout is highly linked with certain deployment environments, with certain user characteristics and baseline symptoms playing a smaller or no statistically significant role.

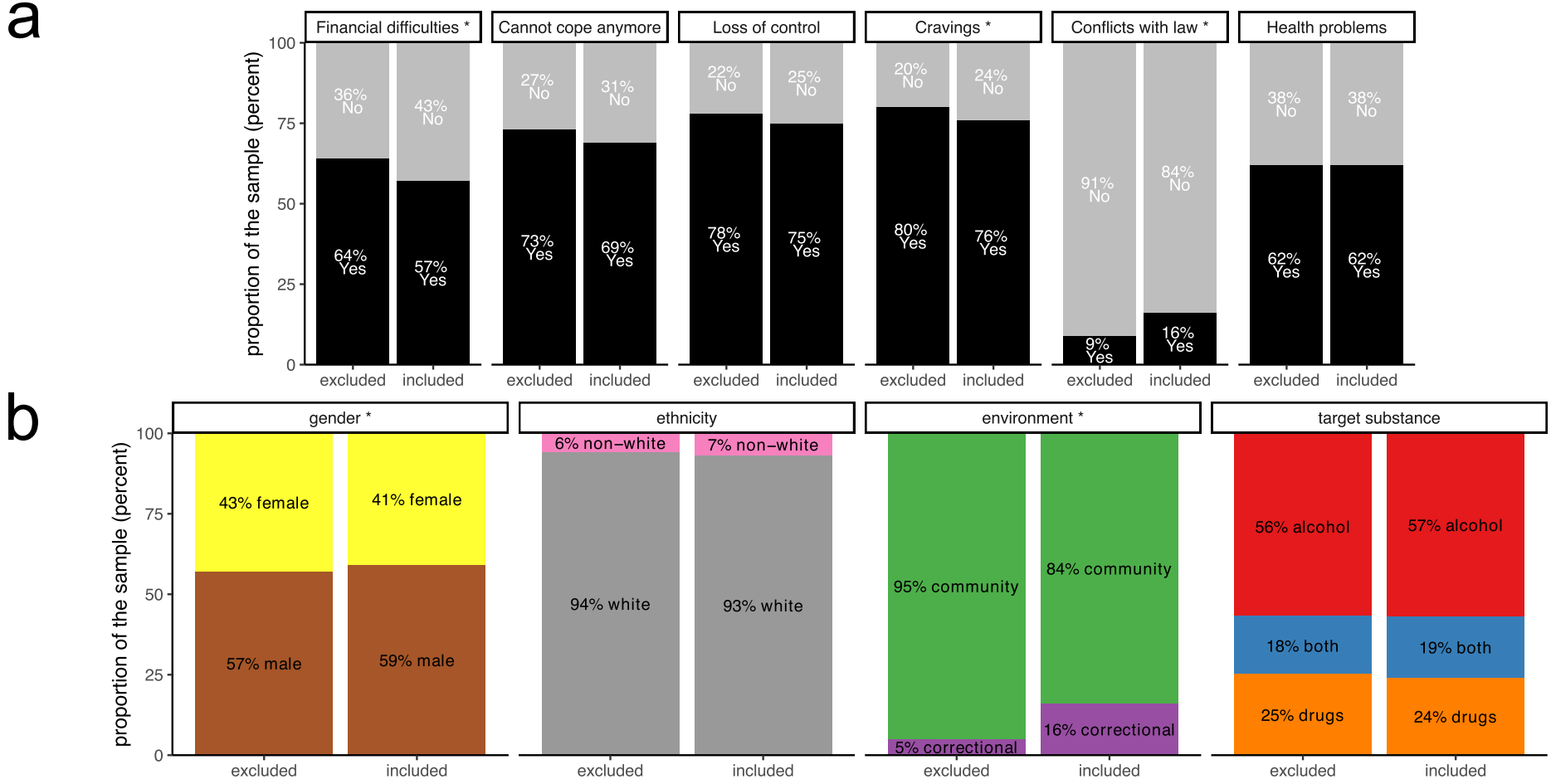

Figure 3.4. Differences between participants in- and excluded of analyses of post-engagement outcomes on basis of availability of a post-engagement assessment. Stacked bar plots showing differences in responses between participants with and without post-engagement assessment data for (a) a selection of questionnaire items and (b) sociodemographic characteristics. Each color block reflects a different categorical level for the questionnaire item displayed. An asterisk indicates statistical significance of the difference between participants with and without post-engagement assessment data, suggested by regression models.

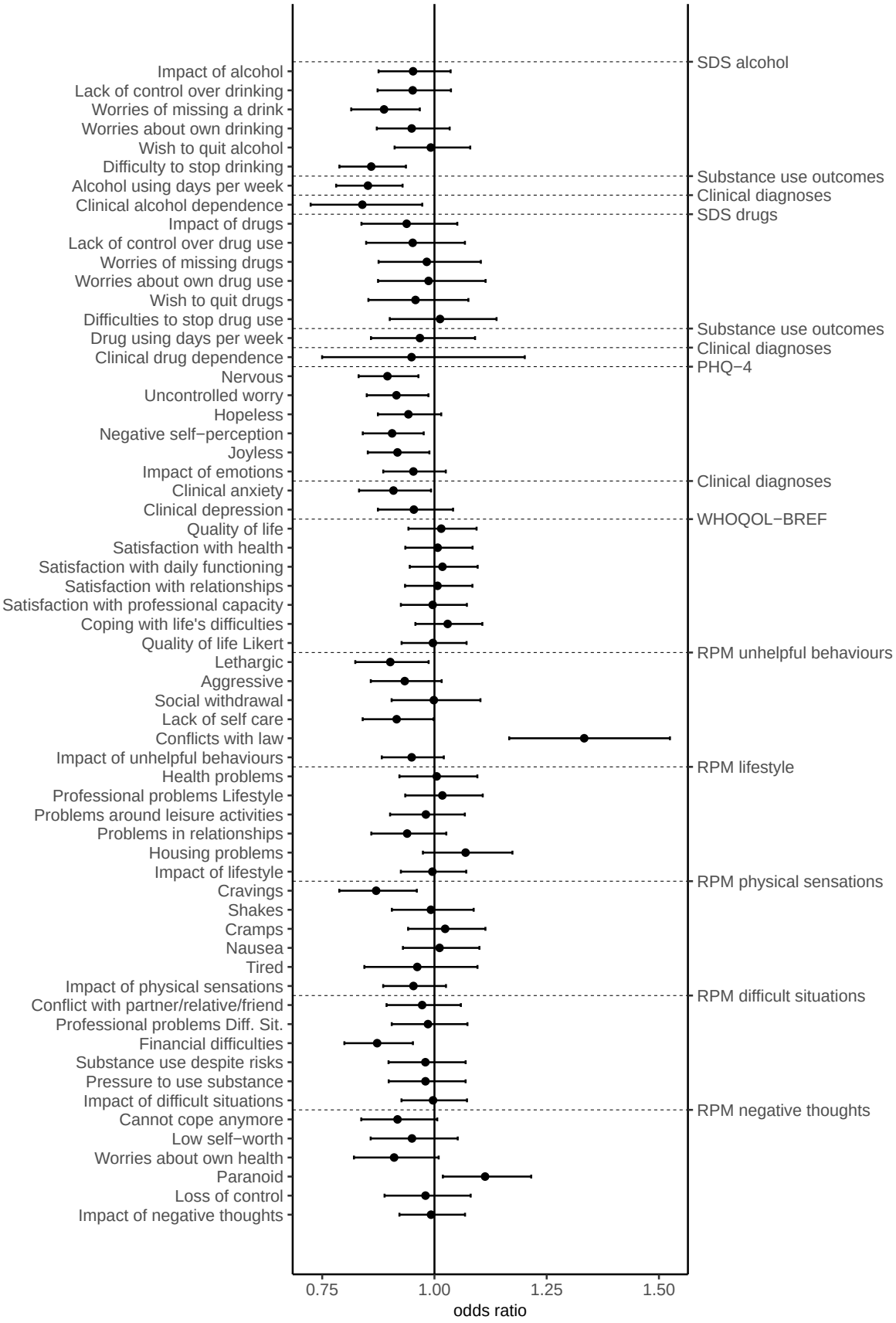

Figure 3.5. Association of participant pre-engagement clinical complexity measured on item and aggregate levels with retention. Data are odds ratios of occupying higher categories on items or fulfilling diagnostic core criteria for a specific psychiatric disorder by participants with a post-engagement assessment vs. those occupying lower categories or not fulfilling criteria. They are shown with their 95% confidence intervals, the severity of dependence scales, the PHQ-4 and the WHOQOL-BREF with affiliated items.

3.5.3 Association of post-engagement outcomes with gender

I next investigated whether the post-engagement outcomes of the DI were associated with self-reported user characteristics. I examined aggregated and item-level outcomes and used gender as an example of a particular characteristic of interest both as a main effect describing the effect of gender on general answering patterns independent of assessment time point but also its statistical interaction with pre- or post-engagement status.

Regarding the main effect of gender, I found significant associations between female gender and increased anxiety at the prospect of missing an opportunity to use a drug (OR = 1.54, p = 7.81 $\times 10^{-5}$) or alcohol (OR = 1.38, p = 4.66 $\times 10^{-5}$), difficulty of stopping drinking (OR = 1.29, p = 0.001), frequency of uncontrolled worry (OR = 1.25, p = 0.001), hopelessness (OR = 1.35, p = 2.28 $\times 10^{-5}$), joylessness (OR = 1.17, p = 0.025), impact of physical sensations (OR = 1.17, p = 0.021), negative thoughts (OR = 1.32, p = 5.25 $\times 10^{-5}$), difficult situations (OR = 1.23, p = 0.003) and emotions (OR = 1.23, p = 0.003), and decreased satisfaction with professional capacity (OR = 0.80, p = 0.001), independent of BFO engagement (see **Fig. 3.6** and **Fig. 3.7**). Female gender was also significantly associated with slightly increased quality of life (OR = 1.14, p = 0.049).

Significant interaction effects between assessment time and gender were found on items on the number of alcohol- and drug-using days per week ($OR_{alcohol}$ = 1.51, p = 0.0002, OR_{drugs} = 1.88, p = 2.77 $\times 10^{-5}$, see **Fig. 3.8**), worries about alcohol use (OR = 1.38, p = 0.004), wish to stop using substances ($OR_{alcohol}$ = 1.33, p = 0.010, OR_{drugs} = 1.47, p = 0.010), experienced lack of control over drug use (OR = 1.45, p = 0.013), quality of life (OR = 0.82, p = 0.032), nervousness (OR = 1.28, p = 0.011) and hopelessness (OR = 0.79, p = 0.015), suggesting slower reduction in recovery progression for women while engaged with BFO.

In addition to examining gender effects, I also looked at the association of assessment time point (pre- or post-BFO-engagement) and responses to questionnaires independently of gender in order to identify differences in symptom severity between the two assessment time points. I found significant associations between assessment time point and responses to 28 out of a total of 29 items (range: 0.86 for intensity of worry about alcohol use - 2.37 for quality of life), and, in line with that, all four cutoff-based measures ($OR_{anxiety}$ = 0.48, p = 9.46 $\times 10^{-26}$, $OR_{depression}$ = 0.52, p = 9.01 $\times 10^{-21}$, $OR_{alcohol\ dependence}$ = 0.64, p = 6.97 $\times 10^{-5}$, $OR_{drug\ dependence}$ = 0.42, p = 4.52 $\times 10^{-8}$).

I was also interested in whether the data held any information about clinically meaningful change that may take place between assessment time points. I found 17% and 38% of participants to report zero alcohol- and drug-using days per week post-BFO-engagement when they had reported at least one

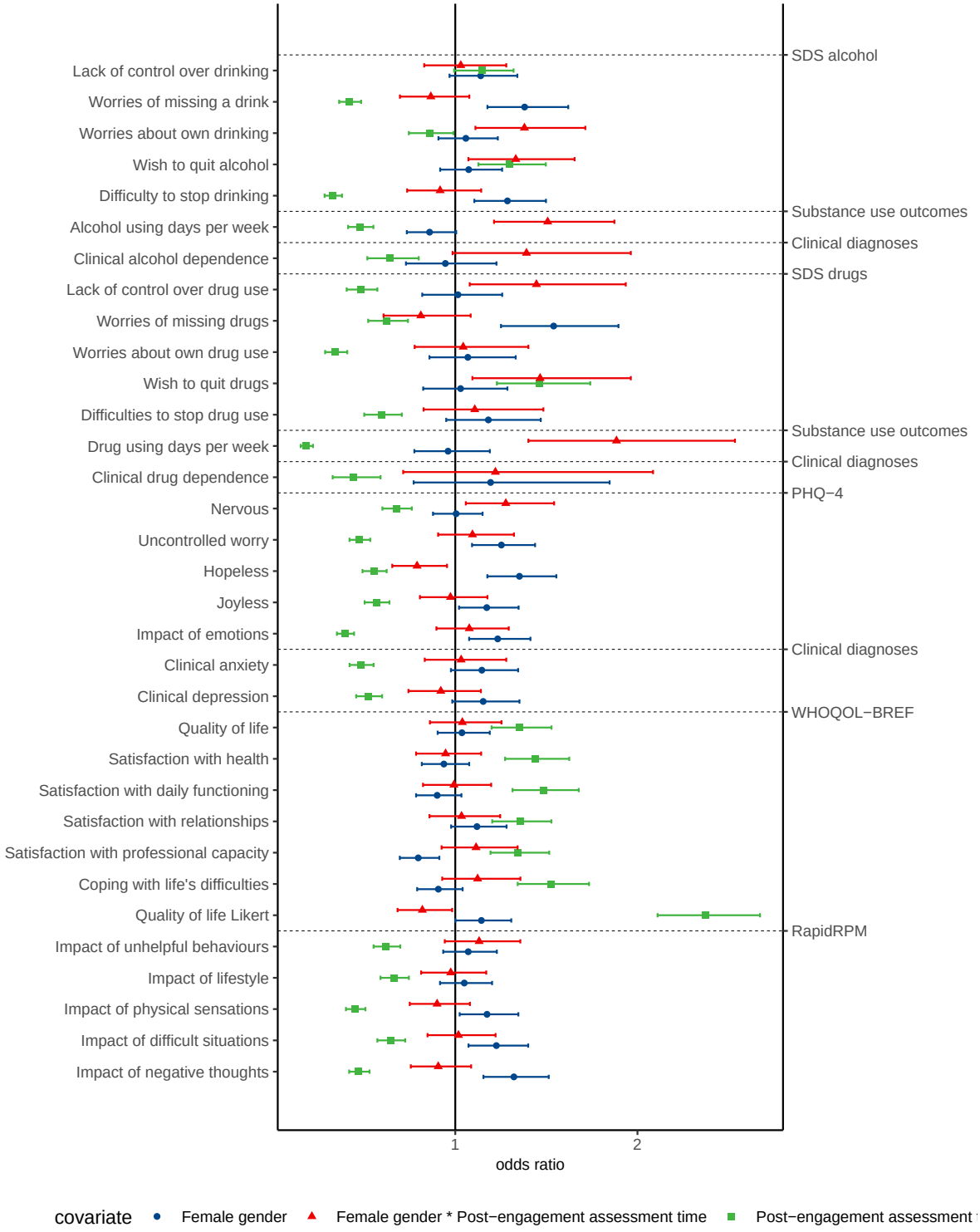

Figure 3.6. Associations of participant recovery progression measured on item and aggregate levels with assessment time pre- and post- BFO engagement, and gender. Odds ratios (reference groups: pre-engagement, male, and pre-engagement male) are shown with their 95% confidence intervals. The PHQ-4 and the WHOQOL-BREF are shown with affiliated items.

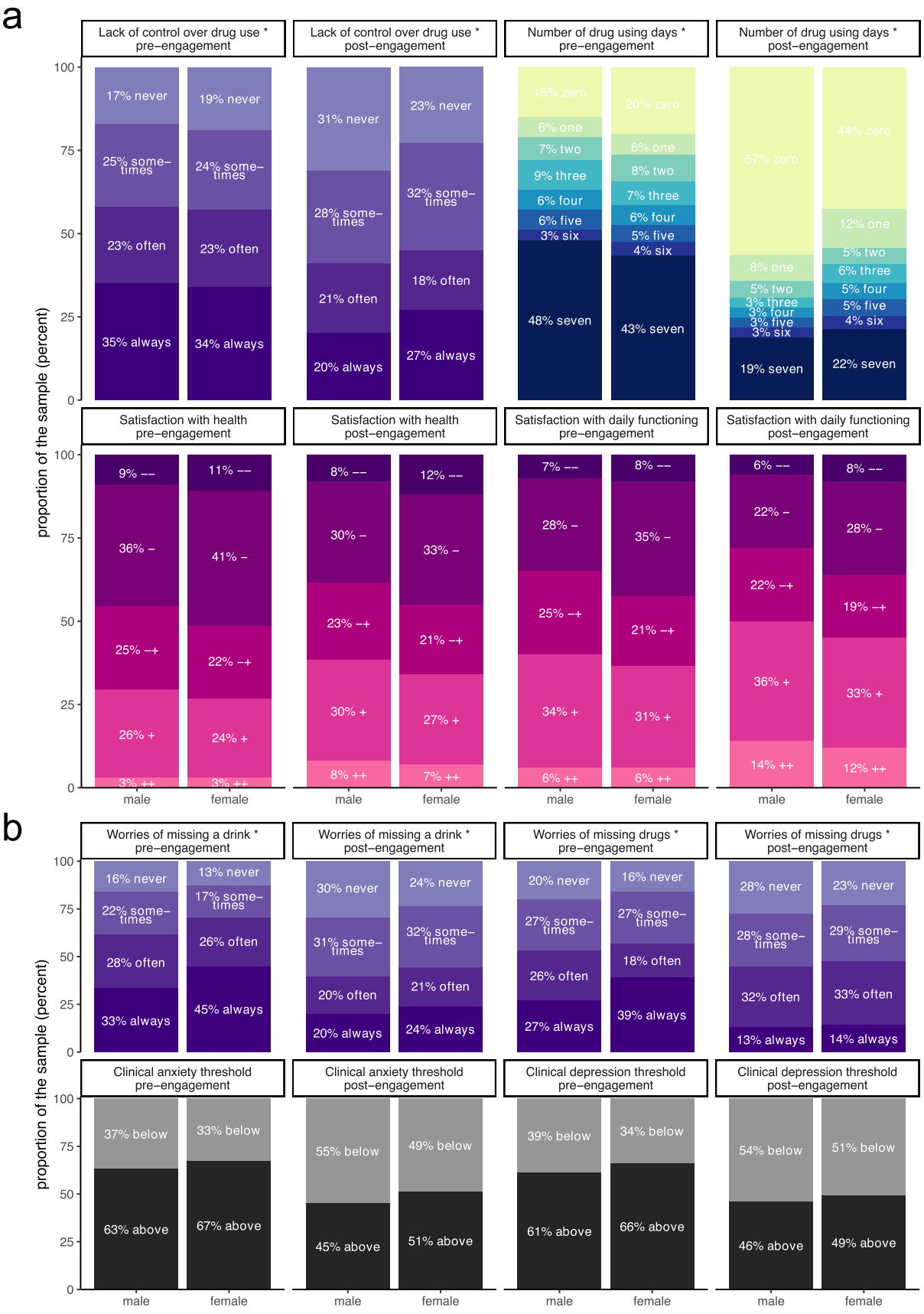

Figure 3.7. Subgroup questionnaire response illustration. Stacked bar plots showing pre- and post-BFO engagement differences in responses between men and women on a selection of questionnaire items and in terms of exceeding the cutoff for anxiety and depression. Each color block reflects a different categorical level for the questionnaire item displayed. An asterisk indicates statistical significance of the interaction between assessment time and gender for (a), and statistical significance of the difference between men and women for (b), suggested by regression models.

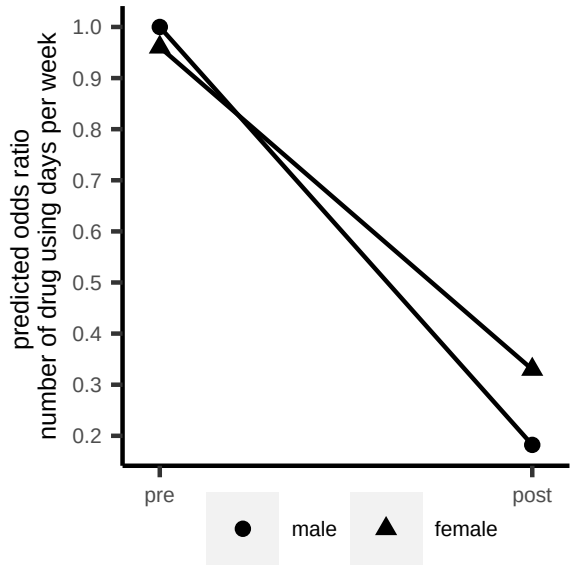

Figure 3.8. Effect of the interaction of time point and gender on drug using days per week. Shown are predicted differences in odds of using drugs on more rather than less days during a typical week. The reference group is men participants at baseline (OR = 1).

such day pre-BFO-engagement. Opposite behaviour (zero alcohol and drug using days pre-, but not post-BFO-engagement) was reported by 5% and 2%, respectively. 24%, 24%, 10% and 12% of participants met cutoffs for anxiety, depression, alcohol, and drug dependence before engagement with BFO, but not afterwards, while for 7%, 8%, 6%, and 4%, respectively, the opposite was true.

Odds ratios in **Fig. 3.6**, for example for weekly drug-using days can be interpreted as such: While across assessment times women do not differ from men regarding their odds of using drugs on more rather than less days, odds are decreased post- compared to pre-engagement regardless of gender. The impact of gender depends on the assessment time such that the odds of using drugs on more days is increased post-, but not pre-engagement for women compared to men.

In conclusion, I found that female substance users reported more severe psychiatric symptoms which largely do not reduce as readily as those of men following use of BFO. Similar analyses based on ethnicity were not possible with this data due to the low numbers of non-white participants.

3.5.4 Outcome prediction

Given these insights, I next set out to evaluate the possibility of building a risk prediction model to stratify users from their pre-engagement response profile on the initial assessment battery. Therefore, I examined whether a participant's retention and post-engagement outcomes can be predicted by training a random forest-based prediction model. **Fig. 3.9** shows acceptable predictive capability for a number

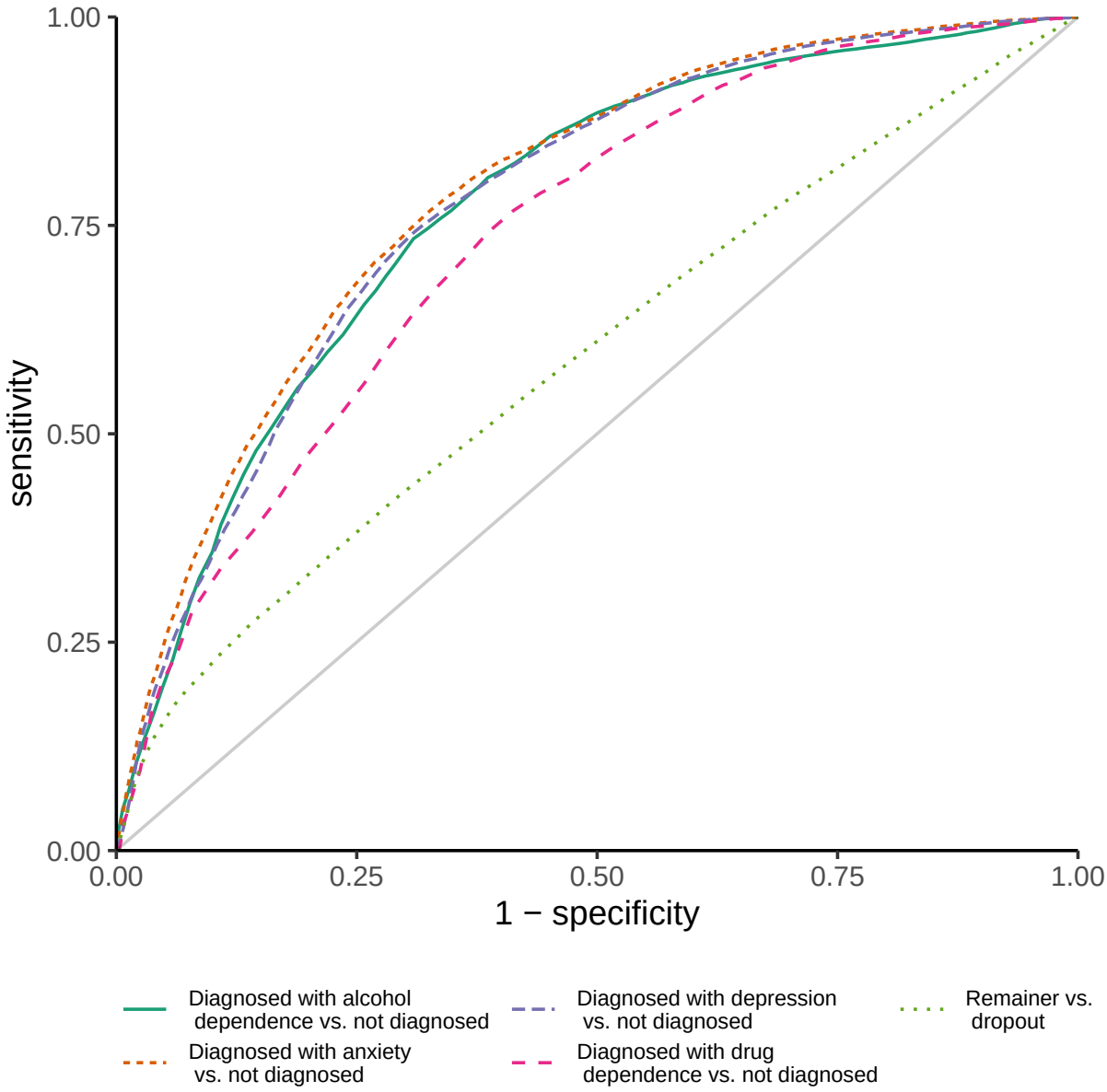

Figure 3.9. Averaged receiver operating characteristic curves for binary outcomes. The curve shows the averaged performance of random forests in 100 validation datasets. $AUC_{retention}$ = 0.59, $AUC_{anxiety}$ = 0.78, $AUC_{depression}$ = 0.79, $AUC_{alcohol\ dependence}$ = 0.77, $AUC_{drug\ dependence}$ = 0.74.

of post-engagement clinical outcomes (area under the receiver operating characteristic curve, abbreviated as AUC = 0.74-0.79), but poor predictive capability for retention (AUC = 0.59). This indicates that outcome heterogeneity of the DI can be partially explained by initial user characteristics (see **Fig. 3.10** for performance on individual level items).

I explored further by using Accumulated Local Effects plots (ALE) to examine how the general functioning related variables ranked most important in **Table 3.3** drive the predictions of meeting the anxiety screener cutoff post-engagement. Examples of effects of answers on these drivers are presented in **Fig. 3.11** and **Fig. 3.12**. These showed, for example, that post-engagement anxiety was increased

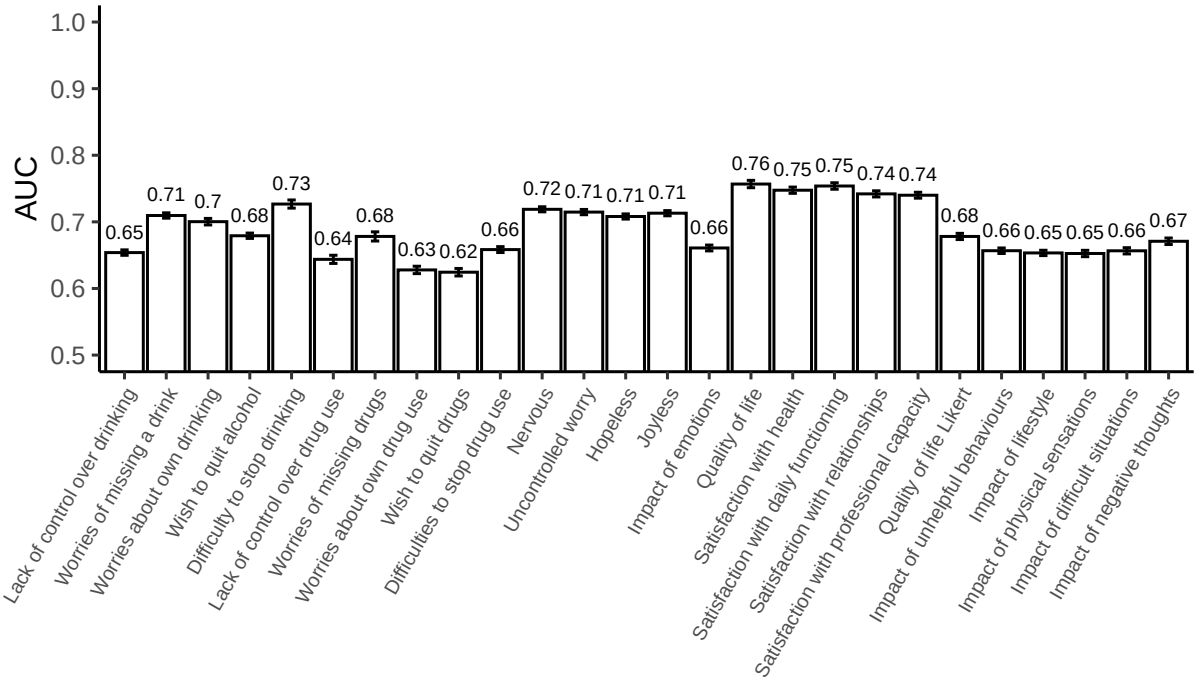

Figure 3.10. Areas under the curves for ordinal outcomes. Bars show the averaged performance of random forests in 100 validation datasets with a 95% confidence interval.

against a backdrop of joylessness, a greater impact of emotions and negative self-perception prior to BFO engagement.

I next decided to examine associations between pre-engagement characteristics of participants and prediction failure, which are illustrated in **Fig. 3.13**. I found significant associations between accessing BFO from a correctional SUD service and repeatedly wrong predictions of anxiety, depression and drug dependence ($OR_{anxiety}$ = 1.60, p = 0.0003, $OR_{depression}$ = 1.52, p = 0.001, $OR_{drug\ dependence}$ = 1.57, p = 0.016). Individuals in correctional, compared to those in community services have hence increased odds of being repeatedly misclassified by my models. Associations between repeatedly wrong predictions of alcohol and drug dependence and the intention to address both alcohol and drug use with BFO ($OR_{alcohol\ dependence}$ = 1.54, p = 0.005, $OR_{drug\ dependence}$ = 0.37, p = 6.59 $\times 10^{-9}$) as well as undershooting the cutoff for alcohol and drug dependence before engagement with BFO ($OR_{alcohol\ dependence}$ = 0.39, p = 5.51 $\times 10^{-9}$, $OR_{drug\ dependence}$ = 0.50, p = 0.005) were also significant. Pre-engagement anxiety was significantly associated with repeated misclassification on post-engagement anxiety and depression ($OR_{anxiety}$ = 1.37, p = 0.021, $OR_{depression}$ = 1.64, p = 0.0002). Repeated misclassification on post-engagement anxiety was significantly associated with female gender (OR = 1.33, p = 0.003), and misclassification on alcohol dependence was associated with participants undershooting the cutoff for depression pre-engagement (OR = 0.75, p = 0.029). Overall, I found no evidence of an association of prediction inaccuracy with ethnicity.

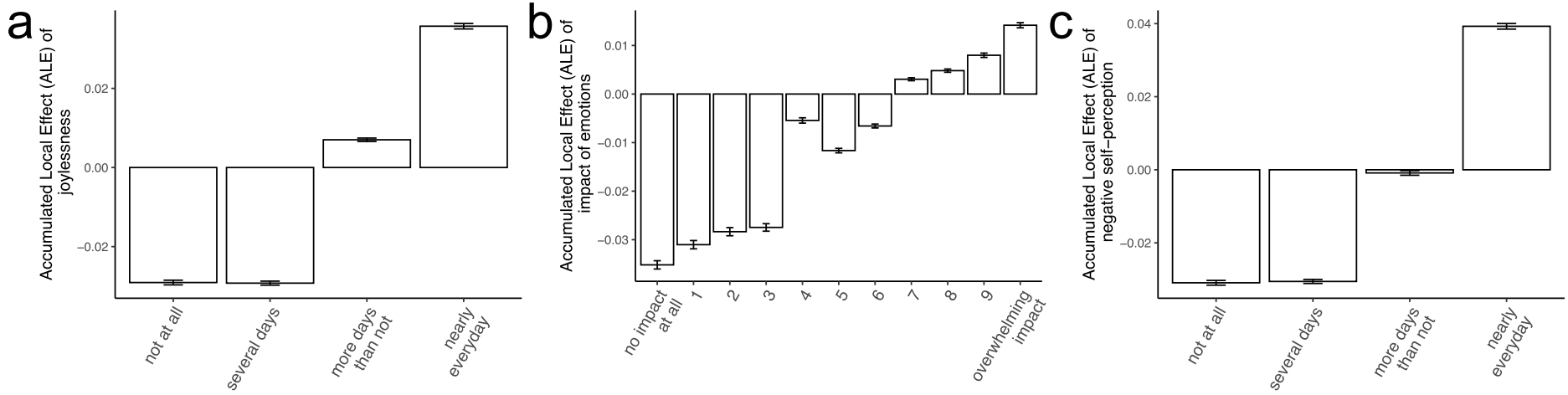

Figure 3.11. Accumulated Local Effects of values on three features on the prediction of anxiety. The predicted probability of post-engagement anxiety increased with pre-engagement levels of (a) joylessness, (b) impact of emotions and (c) negative self-perception.

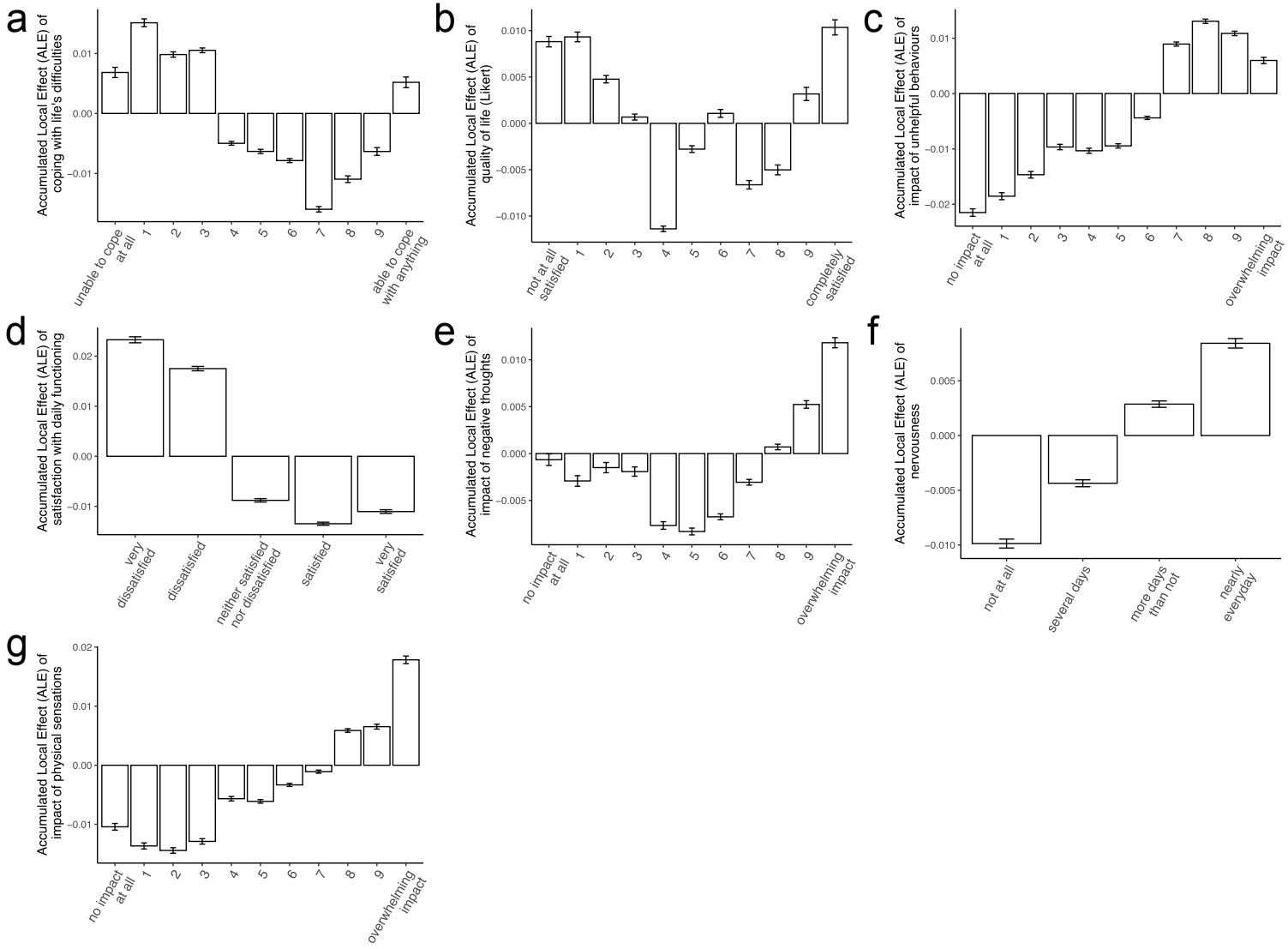

Figure 3.12. Accumulated Local Effects of values on seven features on the prediction of anxiety. Shown are features ranking 4th to 10th based on aggregated importance across general functioning related outcomes.

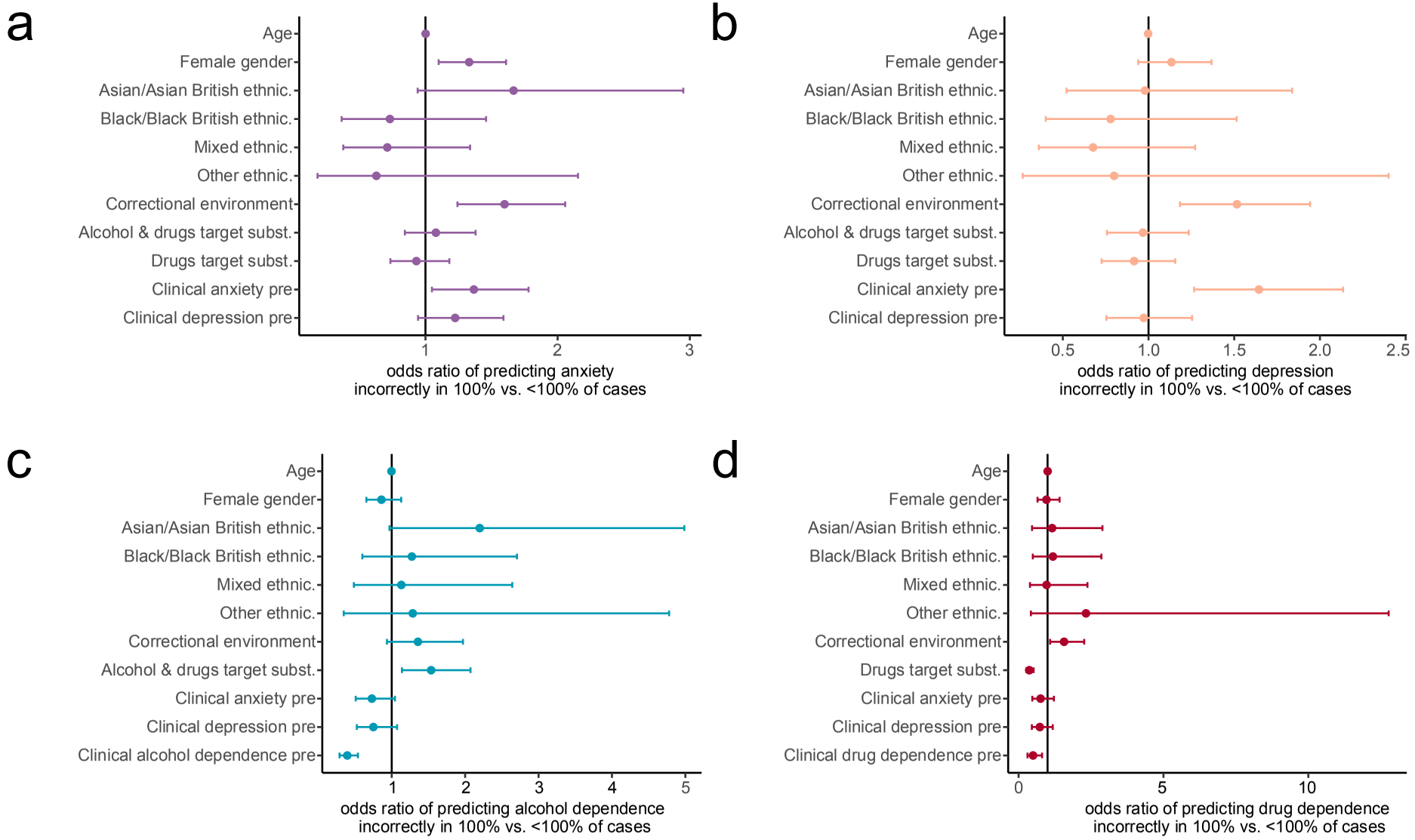

Figure 3.13. Associations of repeated misclassification post-engagement with pre-engagement participant characteristics. Reference groups are male gender, white ethnicity, community environment, alcohol as a target substance for associations with the misclassification of post-engagement (a) anxiety, (b) depression, and (c) alcohol dependence, and alcohol and drugs as target substances for associations with the misclassification of post-engagement (d) drug dependence. ORs are shown with their 95% confidence intervals.

Table 3.3. Pre-engagement drivers of prediction of post-engagement outcomes with random forests, ranked by importance. Ranks are aggregated across general functioning, alcohol dependence and drug dependence related outcomes, respectively.

Rank	Predictor of general functioning	Predictor of alcohol dependence severity	Predictor of drug dependence severity
1	Joyless	Lack of control over drinking	Difficulty to stop drug use
2	Impact of emotions	Worries of missing a drink	Worries about own drug use
3	Negative self-perception	Worries about own drinking	Lack of control over drug use
4	Coping with life's difficulties	Difficulty to stop drinking	Wish to quit drugs
5	Quality of life Likert	Clinical alcohol dependence	Impact of drugs
6	Impact of unhelpful behaviours	Wish to quit alcohol	Negative self-perception
7	Satisfaction with daily functioning	Impact of alcohol	Joyless
8	Impact of negative thoughts	Impact of physical sensations	Impact of negative thoughts
9	Nervous	Satisfaction with health	Impact of physical sensations
10	Impact of physical sensations	Negative self-perception	Coping with life's difficulties

I have therefore identified that it is possible to construct predictive models of DI outcomes from an initially available first-contact symptom profile conducting a retrospective analysis of historical data. Predictive performance was reduced for certain user groups suggesting user heterogeneity in response to the DI and was in line with previous analyses linking the associations of DI outcome with participant pre-engagement characteristics.

3.6 Discussion

My study was a retrospective exploration of large volumes of real-world, user-reported data from a digital health intervention for SUD, and the first of its kind to use a heterogeneous, high-acuity sample. I used the example of BFO where 93% of registrants initially screened positive for dependence on a variety of substances, including Class A drugs such as crack, cocaine and heroin. 72% were dual diagnosis registrants who additionally screened positive for anxiety or depression.

My analysis characterised real-world BFO registrant sociodemographics, mental health problems, and their heterogeneity. My analysis also reported real-world attrition rates and factors potentially implicated in attrition, such as the availability of structured face-to-face support for DI users as it is provided

by correctional treatment services commissioning BFO. I found these services to have lower attrition rates than their counterparts in the community. Lower attrition rates for DIs which provide structured human support is supported by previous research (Linardon, Cuijpers, et al. 2019).

I also described changes in real-world reported mental health problems, quality of life and substance use of men and women registrants after engaging with BFO, and differences in reporting between men and women. These suggested an experience of SUD specific to women, possibly including greater and more persistent clinical complexity which plays out on a symptom level across the psychiatric diagnostic categories of SUD, anxiety, and depressive disorder.

These gender differences are supported by research suggesting substance using women being more likely to experience financial difficulties, parental absence in childhood, having a family history of depression or an alcohol abusing family member or spouse (Khan, Okuda, et al. 2013; Khan, Secades-Villa, et al. 2013; Lewis, Hoffman, and Nixon 2014; Lewis and Nixon 2014; Lopez-Quintero et al. 2011; National Drug Treatment Monitoring System, Office for Health Improvement & Disparities 2023). Conclusions from research on differences between men and women abusing substances are generally complicated by the fact that many studies exclude individuals diagnosed with psychiatric comorbidities which are highly prevalent in women substance users.

My finding on gender differences in a population of registrants for a DI for SUD is of particular interest given the relatively high proportion of women users for both BFO (43%, 47% in community environments) and other substance abuse focused DIs (Bell et al. 2020; Livingston et al. 2021; VanDeMark et al. 2010), relative to gender ratios in individuals presenting to in-person SUD treatment (National Drug Treatment Monitoring System, Office for Health Improvement & Disparities 2023), and the potential DIs therefore present for providing treatment for substance-using women. Supporting women with their specific challenges within gender-specific DI content, for example on psychiatric co-morbidities, trauma and intimate partner relationships is therefore likely to be clinically beneficial (Saraiya et al. 2020; Sugarman, Meyer, Reilly, and Greenfield 2020; Sugarman, Meyer, Reilly, Rauch, et al. 2021). This approach aligns with one of the strategic objectives of the WHO Global strategy on digital health; to promote health equity and gender equality (World Health Organization 2021).

In the study presented in this chapter, I also built prediction models of registrant last-contact SUD symptom severity which was mostly driven by first-contact symptom severity. This is not surprising, as past indicators of symptom severity often inform future mental health status (Garriga et al. 2022). The transdiagnostic character of the mental health problems BFO registrants were facing was reflected in prediction drivers describing symptoms of SUD, anxiety and depression. Registrant heterogeneity was reflected in greater odds of being misclassified by the prediction models for some registrant groups,

for example simultaneous alcohol and drug users, prisoners and registrants undershooting the cutoff for alcohol or drug dependence at baseline. I speculate that this may be due to the possibility of compensating abstinence from one substance with use of the respective other substance (simultaneous alcohol and drug users, registrants undershooting the cutoff for alcohol or drug dependence at baseline), small user group size and a sufficiently different implementation model of BFO in correctional settings (prisoners), and advanced recovery progression (registrants undershooting the cutoff for alcohol or drug dependence at baseline who may use BFO to stay sober).

As I was not able to predict whether a second use event, a preliminary indicator of behavioural engagement, which may in turn be an important indicator of patient benefit, had taken place, it seems precipitate to offer the BFO programme to select addiction service clients.

The implications of my findings on gender differences for precision medicine are limited because I have examined sources of user heterogeneity between user groups rather than individuals. The evidence I produced on the feasibility of prediction of DI outcomes sufficiently motivates follow-up RCTs on the effect of DI user-stratified support systems with potential human elements.

The main limitation of my study's findings about factors in last-assessment DI outcomes is that they were dependent on the availability of such an assessment, and the assumption that meaningful levels of engagement had taken place in between the first- and last-contact assessment. The number of registrants who had completed at least one assessment update was found to be low and may be improved by offering human support to users at risk of levels of engagement too limited to support behaviour change. Evaluating the feasibility of identifying such people from routinely collected data from BFO is the aim of the subsequent chapter. I identified only minimal baseline differences between users who were retained and those who dropped out after first use, but there may be differences on variables unmeasured, for example BFO programme acceptability or benefit which may have been higher in registrants who completed an assessment update. This limitation is shared by similar real-world studies, who encounter similar rates of early disengagement (Ramos et al. 2021).

Further, my definition of user engagement is not based on frequently used indicators of behavioural engagement found in log data, such as the frequency of use, time spent, and number of completions of DI modules (Bell et al. 2020; Ramos et al. 2021). Instead, I define engagement as coming back at least once to complete a second BFO assessment. While my definition of user engagement disregards the potential benefit that registrants may have experienced who have engaged with BFO CBT-based content after their first-contact assessment but did not update the assessment,I believe that my definition is more tolerant of individual preferences of engagement intensity (Chien et al. 2020; Fleming et al. 2018), for which evidence is emerging. My definition also aligns with recent conceptualisations of

engagement which highlights affective and cognitive components of engagement beyond the traditionally measured behavioural investment (O'Brien et al. 2020; Torous, Michalak, and O'Brien 2020).

3.7 Conclusion

As the deployment of DIs continues to proliferate, and users of these interventions support the generation of complex data sets, it becomes incumbent on DI developers to ensure that the data they collect from users are utilised in innovative ways, in order to optimise clinical benefits. This study uses basic service interaction data which may be available to researchers at a minimum, and focuses on associating user heterogeneity with specific static user characteristics. Developers and commissioners of digital health interventions may adapt this analysis to better understand outcomes and the way these are impacted by user heterogeneity in order to evolve their products to deliver more effective services.

3.8 Outlook

These concluding remarks were not included in the published article, and represent a retrospective reflection on the conducted research.

The preceding analysis served as an exploration of BFO data initially available to us, and simultaneously demonstrates that such exploration is important in DI research. I was interested in real-world BFO registrant heterogeneity, and how it influenced digital mental health use and outcomes. I identified several factors, for example registrant gender, and BFO delivery mode, which appear to be associated with BFO outcomes. Note that I did not find these factors to be predictive of outcomes when variables related to baseline substance use and general mental health problems were included in prediction models. These variables were ranked higher in importance for predictions than gender or deployment environment. In the case of gender, this may be due to these variables being related to gender, and being able to represent the information contained in the variable gender in the prediction model. The greatest limitation of this study appears to be the definition of endpoints which reliably indicate user benefit.

Further, I identified statistical and machine learning models with which I could model such real-world data. I did so being precise about type and structure of such data (categorical data characterised by high correlations), and - through the inclusion of symptom items as well as diagnostic screening outcomes into my models - allowed for more detailed, transdiagnostic patterns to be detected, whose im-

portance is increasingly emphasised in mental health outcomes research (Eaton et al. 2023). A possible way to refine these models and explore their value for clinical practice would be evaluating the performance of a simpler prediction model built on only the 10 most important predictors of an outcome (see Chapter 5).

Through the analysis presented in this chapter, I also identified what additional data is needed in order to make more precise statements about BFO outcomes and use. The BFO dataset I had available, for example, does not contain information about whether a last-contact assessment was initiated by a BFO user themselves, or prompted by BFO. This information would have strengthened the validity of our analysis, since voluntarily and mandatorily completed updates may differ systematically from each other: Voluntarily completed updates may be more likely to be completed by users who felt particularly accomplished, or frustrated about their progress at the time of reporting. The absence of this information in the BFO dataset illustrates the typical real-world-data-specific challenge of not knowing the reason why a particular data point - for example an assessment update - is present or absent in a dataset.

Further, in order to make more precise statements about BFO outcomes and use, additional longitudinal data on behavioural engagement of registrants with BFO, for example between assessment time points, is needed, which were available to me at time of conduct of the study reported in Chapter 4 of this thesis. As these data are indispensable to construct better real-world indicators of registrant recovery and benefit from BFO, random forest models of outcomes built in Chapter 3 are best regarded as ML models used for understanding circumscribed aspects of BFO engagement rather than predicting mental health status, for example progression of recovery from SUD, along with Chen et al. (2023) who recognise the value ML models unfit for practical prediction can have for the generation of hypotheses about factors influencing mental health. Specifically, our models may be seen as a reiteration of the importance of pre-treatment symptom severity for recovery from a mental health condition which has been found in previous research, too (Delgadillo, Rubel, and Barkham 2020; Flygare et al. 2020; Garriga et al. 2022; Paul et al. 2019; Van Breda et al. 2018). I further explore use cases of ML prediction models of real-world data from digital mental health such as those collected within the BFO DI in Chapter 4 and Chapter 5, probing ML's proposed role as a companion methodology to traditional statistics in mental health research along the way.

Chapter 4

On the difficulty of predicting engagement with digital interventions for substance use disorders

As engagement with digital interventions (DIs) for mental health often drops sharply in the first week of use (Bricker et al. 2023), earliest possible prediction of to-be-expected engagement levels is desirable to target human support at individuals at risk of low engagement. This chapter presents a case study evaluating whether it is possible to predict if registrants with the "Breaking Free Online" (BFO, see Chapter 2) DI are at risk of low engagement, using data available at the earliest possible time point, in our case, before registrants were able to access cognitive behavioural therapy (CBT) based content. Specifically, registrants had registered with BFO in community addiction services, where structured face-to-face support is not as readily available as in correctional addiction services commissioning BFO. Programme use may be mostly self-directed and engagement more likely to be low, which is why early prediction of engagement levels has a high priority. The content of this chapter was presented as a full conference paper at the conference of the European Federation for Medical Informatics 2023 as "Günther, F., Yau, C., Elison-Davies, S., & Wong, D. (2023). On the difficulty of predicting engagement with digital interventions for substance use disorders. In Hägglund, M., Blusi, M., Bonacina, S., Nilsson, S., Madsen, I. C., Pelayo, S., Moen, A., Benis, A., Lindsköld, L., & Gallos, P. (Eds.), Caring is sharing – exploiting the value in data for health and innovation (pp. 967-971). IOS Press. doi: 10.3233/SHTI230319". In this chapter, I present the accepted paper, with some small changes made to the original text at relevant points, for example the insertion of references supporting my claims.

4.1 Abstract

Self-guided digital interventions may be an important instrument in treating substance use disorder. However, most digital mental health interventions suffer from early and frequent user dropout. Early prediction of engagement would allow identification of individuals whose engagement with digital interventions may be too limited to support behaviour change, and subsequently offer them greater support. To investigate this, I used standard machine learning models to predict different metrics of real-world engagement with a digital cognitive behavioural therapy intervention widely available in UK addiction services. Our set of predictors consisted of baseline data from routinely-collected standardised psychometric measures, and variables derived from them. Areas under the ROC curve, and correlations between predicted and observed values indicated that baseline clinical data does not contain sufficient information about individual patterns of engagement.

4.2 Introduction

DIs for people with substance use disorders (SUDs) are digitised equivalents of traditional face-to-face therapies such as CBT. They are used to complement or temporarily replace equivalent in-person interventions. With DIs being more scalable and 24/7 accessible, they may represent an important instrument in treating SUDs.

To derive improved mental health outcomes via a DI, users need to engage with DI content to a sufficient degree (Gan et al. 2021). However, maintaining user engagement has been a consistent problem for DIs for mental health (Torous, Nicholas, et al. 2018). Early and accurate prediction of level of DI engagement could allow users at high risk of poor engagement to be identified. This could potentially be used to target additional support. Prediction of low engagement on the basis of data collected early into the user journey, if feasible at first user contact with a DI, would make targeted additional support especially effective as dropout after first use is a common phenomenon. On the contrary, if low engagement or dropout is only identified after it already happened, it may be harder to re-engage users in person, with push notifications or emails, or the preventive tailoring of DI content to low engagers may not take effect anymore.

However, it is not clear if prediction is at all possible using such data, since real-world engagement may depend on multiple factors (Baumel and Kane 2018) that may not be reflected in a one-off clinical assessment before user engagement. However, engagement is a prerequisite for beneficial user outcomes, and hence an important prediction target, maybe even more so than clinical outcomes of DI

use which have been targeted with varying success in previous prediction studies (Marinova, Rogers, and MacBeth 2022).

4.2.1 Related Work

Prediction studies in digital mental health have in the past focused on the prediction of mental health outcomes measured at a predefined time point instead of engagement. However, controlled study designs, and the integration of structured human support through care coordinators and therapists into DIs examined in these studies, which may not always be available in the real world, obstruct conclusions about performance of prediction models in the real world where disengagement from self-directed DIs may get in the way of beneficial outcomes (Flygare et al. 2020; Lenhard et al. 2018; Wallert et al. 2022). If real-world data is collected, data analysis does not account for (early) user disengagement (Hornstein et al. 2021; Marinova, Rogers, and MacBeth 2022; Ramos et al. 2021). The result of this analytical approach is that predictions only apply to users who show some level of engagement, which cannot be readily assumed. This, in turn, makes predictions unactionable in practice. In addition, reported predictive performances range around AUC = 0.70 which does not yet warrant human clinical intervention for dropout “prevention” to take effect.

An example for such studies is a study presented by Bricker et al. (2023) who successfully predicted dropout from smoking cessation apps in the second week after download from the number of logins in the first week. However, with this approach, it is not possible to preventively target assistance at the majority of users dropping out early because at the time of them dropping out, not enough data has been collected to make accurate predictions (ibid.).

The aim of this study is therefore to assess whether engagement with the BFO programme can be predicted using data routinely collected at users’ first interaction with the programme.

4.3 Methods

4.3.1 Source of data

Data were routinely collected from users of BFO enrolled between July 2016 and October 2022. Enrolment took place in 513 different community-based addiction services in the UK. In those community-based addiction services, BFO is delivered as a mostly self-directed programme. BFO is a modular

digital CBT programme for SUDs, which for the past decade has been widely available to clients of community addiction services in the UK. Ethical approval for collection, storage and use of data accumulating from routine use of BFO by clients in participating treatment services was obtained from an NHS Research Ethics Committee (London - South East, 16 May 2012 and 22 May 2017, references 12/LO/0076 and 12/LO/0287).

The BFO programme features six modules; each module is split into one part psychoeducation and one complementary part, practice, applying what was learned in psychoeducation to one's own life. These subparts are subsequently referred to as "strategies", specifically, information strategies and action strategies.

Users are required to complete a baseline assessment so that modules can later be recommended to them. The baseline assessment includes four validated questionnaires designed to measure different aspects of SUDs: (i) the Severity of Dependence Scale (SDS), (ii) the Patient Health Questionnaire 4, (iii) the World Health Organization Quality of Life measure (items 1, 2, 17, 18, and 20) and (iv) the Recovery Progression Measure (Elison, Davies, and Ward 2016; Gossop et al. 1995; Kroenke et al. 2009; Skevington, Lotfy, and O'Connell 2004). Responses to questions were recorded on 2-, 4-, 5- and 11-point Likert scales, depending on the questionnaire. In addition, the baseline assessment also recorded user age, gender, ethnicity, abused substances, substance-using days in the preceding week and the user's target for substance-free days per week.

Users struggling with both alcohol and drug abuse are required to answer questions on their alcohol and drug dependence separately. For my study I only used answers to questions on the substance labeled as the primary dependence. If users had not indicated which substance was their primary one, primacy was determined through the highest total SDS score, or if these were equal, through a greater number of substance-using days per week. If these were again equal for both substances, primacy was chosen at random.

In addition to the assessment questionnaire data, dates of user assessments as well as module completion data, specifically the number of completions for the psychoeducation and practice part of each programme module and the date of its most recent completion, were available for analysis.

4.3.2 Predictors

The feature set I used for the development of a prediction model of BFO engagement comprises all 62 items corresponding to every question in the baseline assessment and a set of derived variables. I derived the following variables: (1) baseline abstinence defined as zero substance-using days per week,

(2) the number of days from registration to first assessment completion, (3) the number of clinical complexity inducing factors present for a user (counting in the presence of financial difficulties, cravings, difficulties with physical health, at work, or with housing) and (4) cutoff-based variables on anxiety, depression and substance dependence.

4.3.3 Outcomes

I used 9 derived variables as continuous outcome measures. Each outcome measured a different aspect of user engagement, as follows: (1) the number of days from the first to the last use event, subsequently referred to as the number of accessed days, (2) the number of strategies completed, (3) the number of information strategies completed, (4) the number of action strategies completed, (5) the number of use events (all assessments + strategies completed), (6) the use rate (number of use events / number of accessed days), (7) the percentage of days actively engaged (with the number of days on which an assessment was completed - which empirically fall together with known days of module completion in 67% percent of cases - regarded as active engagement), (8) the median intermission length in days (with days on which no active engagement was registered described as intermission days) and (9) the mean absolute deviation (MAD) intermission length. Log-transformation was applied to all these continuous outcomes due to skewness and excess zeros. In addition, I used the completion of 8 or more strategies as a binary outcome variable. I used this threshold because 8 sessions was the dose of talking therapy commonly received by patients completing a course of treatment through the NHS in England (Population Health, Clinical Audit and Specialist Care Team, NHS Digital 2022) .

4.3.4 Statistical analysis and missing data

I predicted the 9 continuous outcomes and 1 binary outcome independently, using random forests and the XGBoost algorithm with 10-fold cross validation. Stratification was applied to the target variable, with numeric strata being binned into quartiles. Both algorithms were used out-of-the-box without hyperparameter tuning. In the case of XGBoost, all discrete features were one-hot-encoded. The average area under the receiver operating curve was used as a measure of predictive performance for binary engagement outcome variables. Correlations between the observed and predicted values served as an assessment of predictive performance for the continuous engagement outcome variables. The average root mean squared error (RMSE) was used to compare predictive performance between random forests and the XGBoost algorithm. I removed data from users who had >80% data missing on

Table 4.1. User characteristics at baseline.

Characteristics	Statistic/Label	Value
Age in years	mean (SD, range)	40.1 (11.7, 18 - 84)
Gender	Female	47.1% (10,745)
	Male	52.5% (11,967)
	Other	0.3% (79)
Ethnicity	White	93% (21,207)
	Asian / Asian British	1.9% (426)
	Black / Black British	1.7% (382)
	Mixed	2.7% (626)
	Other	0.7% (150)
Primary substances	Alcohol	63.8% (14,533)
	Cocaine	11.7% (2,659)
	Marijuana	7.9% (1,810)
	Heroin	5.5% (1,248)
	Crack	3.5% (805)
	Other (46 other substances)	7.6% (1,741)
Substance dependence (SDS sum score equal to or larger than 3)	Yes	92.6% (20,494)
	No	7.4% (1,631)
Anxiety (sum of first two PHQ-4 items equal to or larger than 3)	Yes	69% (15,593)
	No	31% (7,007)
Depression (sum of last two PHQ-4 items equal to or larger than 3)	Yes	66.5% (15,023)
	No	33.5% (7,577)
Substance-using days in the past week	modes	0 days: 24.5%, 7 days: 38.1%

their baseline assessment (n = 706) as multiple imputation would be difficult for these users. For the remainder of the data, I opted for a complete case analysis as only 5% of these cases had incomplete data (except for intermission length related outcomes which were only available for those updating their assessment at least once), and < 4% of cells were missing in total. All analyses were conducted using R (version 4.2.1), and code is available at https://github.com/franziskagunther/predict-engagement.

4.4 Results

I removed users who were younger than 18 (n = 82) or older than 89 years (n = 3). I also excluded users reporting alcohol consumption of more than 100 standard units of alcohol on a typical day (n = 88), and those reporting a goal of increasing their substance consumption (n = 1314, possibly due to erroneous user interpretation of item as the desired number of substance-consuming instead of substance-free days). Finally, I excluded users whose reports of daily drug consumption was deemed to be clinically infeasible (n = 4). The final dataset contained data from 22,796 users. **Table 4.1** summarises their baseline characteristics.

Figure 4.1. Correlations between variables used for prediction modelling.

I first examined individual feature-outcome correlations (outcomes are located in rows/columns from MAD intermission length up to Accessed days; see also Predictors and Outcomes sections above) and found these to be low (see **Fig. 4.1**, Pearson correlation used for continuous variables, Cramer's V for nominal variables, polychoric correlation for ordinal variables, and square root of R^2 for continuous - discrete variable combinations). Cross-feature correlations (between variables detailed in the Predictors section) were high, instead. Finally, I used XGBoost and random forests to see if the combination of predictors could predict outcomes, but predictive performance was poor in all cases.

Because the random forest showed slightly better out-of-the-box performance than the XGBoost algorithm across outcomes with regards to RMSE and AUC, I report results for it here. I obtained an average AUC of 0.57 [CI: 0.56-0.58] for the prediction of completing n=8 or more modules. Model performance did not improve when other values of n were tested (AUC for the prediction of completing one or more modules: 0.54 [CI: 0.54-0.55]).

Predictive performance for continuous outcomes was similarly low and correlations between observed and predicted outcomes ranged between 0.03 and 0.13.

4.5 Discussion

Prediction of real-world engagement in self-guided DIs for mental health, including those addressing SUD, could contribute to addressing one of the field's biggest problems; early and frequent dropout (Baumel, Edan, and Kane 2019; Baumel and Kane 2018; Fleming et al. 2018; Pratap et al. 2020). Many DIs routinely administer assessments on users' clinical characteristics before providing access to DI content (Carl et al. 2020; Mehta et al. 2021), in order to recommend content and to establish a baseline for evaluation of symptom change. Such assessments, in theory, represent routinely obtained sets of predictors of possibly non-beneficial engagement at the earliest possible time point.

I conducted a prediction study with real-world data from the (mostly) self-guided BFO programme in which all users, regardless of their actual pattern of engagement with the system, were included in my analysis. This decision was made to allow for a wide range of potentially beneficial engagement patterns instead of restricting beneficial engagement to a prescribed pattern of use. Despite using modern prediction modelling methods (Borisov et al. 2021), I was unable to accurately predict a range of engagement metrics from baseline assessment data. While the prediction of exact continuous engagement metrics was expected to be difficult, this suggests that it is not possible to predict who will engage more with the BFO programme from clinical information at first access.

Multiple unmeasured, and potentially difficult-to-measure, factors implicated in user engagement may make prediction of user engagement challenging. Predicting user engagement is likely more difficult in individuals struggling with substance use whose often unstructured lifestyle (Davies et al. 2015), and multiple areas of clinical complexity, likely interfere with engagement. Given the lack of prediction accuracy in this study, triaging new addiction service clients for BFO use on the basis of their baseline assessment data may exclude individuals who may engage and possibly benefit from the programme if they were introduced to it.

This research has some limitations. First, my metrics of engagement were behavioural, and do not reflect possible cognitive or emotional involvement of users with the BFO programme (O'Brien et al. 2020). Further, engagement does not equate to benefit, which may be achieved after minimal engagement (Schleider et al. 2021). However, by including continuous engagement variables, I have attempted to reflect that beneficial engagement can have individually different outlooks and that the binarisation of it on the basis of often arbitrarily defined "minimal engagement" often ignores this. In

its account for the number of temporally distributed user events, my approach also agrees with recent conceptualisations of engagement as “continuing to come back” to a DI (Torous, Michalak, and O’Brien 2020). My attempt of allowing for a variety of different engagement patterns also resulted in intentionally not removing engagement outliers, which may have additionally complicated prediction.

This study focused on a single DI and the results may not be generalisable to other DIs. Examination of other DIs is desirable but challenging due to limited access to commercially-sensitive data sets for independent researchers.

4.6 Conclusion

Early prediction of engagement, and intervention before dropout are desirable in digital mental health. My case study of prediction modelling of engagement in digital CBT for substance dependence suggests that information beyond clinical baseline characteristics is necessary to achieve accurate predictions.

4.7 Outlook

These concluding remarks were not included in the published article, and represent a retrospective reflection on the conducted research.

Early disengagement is a problem for most digital interventions for mental health (Pratap et al. 2020). This is especially true for those interventions targeting highly debilitating mental disorders such as substance dependence (Cross et al. 2022), and those interventions without considerable levels of practitioner involvement integrated (for example through integration of chat communication between users and dedicated practitioners which monitor and accompany intervention use into the DI user interface). As practitioner workload may often make entertainment of such close monitoring impossible (Chekroud et al. 2021), interventions such as BFO, which target SUDs and do not implement such chat communications, but enable practitioner support with BFO as needed in face-to-face sessions if an individual makes use of those, have a right to exist in the digital mental health space.

Early disengagement rates observed for most DIs, but, unsurprisingly, most often for self-guided DIs reveal the adherence bias in those previous studies which assess mental health outcomes of only those individuals which have shown substantial behavioural engagement with a DI, and on the basis of these measured outcomes make conclusions on user benefit from this DI. In the study conducted in

this chapter, therefore, I emphasise the importance of engagement with the DI whose data was available to us, BFO, over clinical outcome measurements, and choose multiple endpoints reflecting different aspects of engagement.

Importantly, the data I had available allowed a behavioural, quantitative definition of engagement. Even though these endpoints describing engagement likely differ from hard-to-operationalise true patient benefit, too, they prevent us from measuring engagement exclusively on a binary scale, and therefore, from making unrealistic assumptions about how beneficial user engagement, and more generally, benefit from a DI looks like, and when benefit should be assessed. Conducting this study, I have come to believe that beneficial user engagement with a DI may look different from user engagement with social media and commercial apps, whose purely quantitative approach to desired engagement often influences the choice of endpoints in studies on digital mental health interventions.

In the study presented in this chapter, I have attempted to predict BFO engagement (as I defined it), using data available at the earliest possible time point; before BFO registrants had access to CBT-based content. The rationale for this was that face-to-face support or content tailoring may be targeted at those users who are predicted to engage too little - a judgement that could be made by practitioners presented with expected engagement rates for a particular patient - before these individuals start the BFO programme.

I tested a variety of different user engagement measures, reflecting different aspects of user engagement, but could not establish acceptable predictive performance with state-of-the-art machine learning models, using detailed baseline information about BFO registrant substance use and general mental health. This result is in line with previous research on engagement with DIs, for example a recent study in which future engagement with a DI for eating disorders could only be predicted when a predictor set with baseline information was enriched with past engagement data (Linardon, Fuller-Tyszkiewicz, et al. 2022).

On the basis of my findings, I speculate that engagement may depend on a multitude of patient- and DI-related factors interacting with each other. Patient-related factors may, for example, include digital literacy, attitude towards digital health, or the ability to form and sustain habits. Importantly, it may not be feasible to collect this additional information on patients due to limits to how much self-report data can be collected in real-world situations: Registrants may not be willing to fill out lengthy questionnaire batteries before DI use, and may cease engagement before even accessing DI content. DI related factors may have to do with its usability and perceived usefulness.

The most promising direction of research building on my findings, therefore, would, from my perspective, be a detailed investigation into how people, who are found to benefit from a digital intervention,

use this intervention. Ideally, commonalities between these users could then be identified, and made measurable to inform endpoints of future studies on digital mental health interventions. More details of a potential study which may realise these objectives is presented in Chapter 6.

Chapter 5

The effect of multicollinearity on reliability of local feature attribution for mental health outcome predictions

5.1 Introduction

In Chapter 3 and Chapter 4, I have developed machine learning (ML) prediction models of mental health related outcomes. Specifically, these were outcomes relating to a digital intervention (DI) for substance dependence; engagement with this DI in Chapter 4, and clinical assessment related outcomes after provision of access to the DI content in Chapter 3. These prediction models are just two examples of ML models in mental health outcomes research, which have increased in popularity in recent years. Their superior performance in other research domains, and related liberty of assumptions about the distribution of the modelled data make them attractive to researchers who are interested in accurate prediction of mental health outcomes in order to optimise distribution of limited therapeutic resources (Chekroud et al. 2021).

Further, in the wake of the increasing digitisation of mental health care, large datasets have become available for secondary use. These seem - as per their volume, fast and inexpensive growth, and high ecological validity - to account better for the complexity and heterogeneity of the experience of mental illness than previously used small experimental samples, and are less expensively collected than data in large clinical trials or longitudinal cohort studies (Koppe, Meyer-Lindenberg, and Durstewitz 2021). ML models may be fit to process such data due to their ability to aggregate small, non-linear effects - common in the domain of mental healthcare research - across large feature sets, and to detect and

leverage interactions between variables to predict an outcome. The introduction of ML models does not only mark a methodological shift from traditional statistics to ML, but also has a potential effect on clinical practice by shifting researchers' previous focus on hypothesis testing as the main goal of an analysis to the prediction of outcomes.

The incorporation of ML models into the methodological toolsets of mental health researchers, however, has some disadvantages as well, and especially pertinent is the following: Most ML models are considered "black boxes". This means that how they arrive at a prediction or decision is much less transparent to engineers building the model, and mental health professionals and patients impacted by the model's decisions, than the decision process of, for example, a classical general linear model based method: Many general linear model based methods are arguably more transparent through their model coefficients.

As ML models have gained popularity in other domains, and are suggested for high-stakes legal, educational, and health-related environments, it has been deemed important that model decisions are made transparent to stakeholders. Otherwise, human oversight of real-world utility and also hazardousness of these models is limited in high-risk situations. Situations in which high-risk decisions have to be taken will likely also occur in mental healthcare. Practically, non-transparency of ML models may result in reservations from clinicians and patients towards the use of these models (Abrams 2023; Wang, Kaushal, and Khullar 2020). In this vein, lack of model transparency has been suggested as a barrier towards real-world implementation of artificial intelligence in healthcare (He et al. 2019). Methodology to make ML models more transparent to stakeholders of model utility has, therefore, been continuously developed alongside the models themselves.

This methodology is also often referred to contributing to the "explainability" of the ML models. "Explainability" is - next to "transparency" and "interpretability" - one of the various terms in the scientific literature relating to a person's understanding of the reasoning or decision process of a given model. Definitions of each of these terms vary between researchers, and different terms can refer to the same concept. Below, I refer to the definitions in the literature which have influenced my choice for the term "explainability", and how I have defined it eventually.

One contribution to the discourse around terminology that influenced my choice was a paper by Lipton (2016) who distinguishes between transparency and post-hoc interpretability. With transparency, they refer to the inherent property of a model to allow humans with our limited cognitive capacity to calculate model outputs from inputs and parameters and to understand components of the model intuitively. Another criterion for a transparent model according to Lipton (ibid.) is that it allows researchers to retrace in detail how the model algorithm arrived at a solution. With post-hoc interpretability, Lipton (ibid.)

refer to human-understandable explanations or communications of the behaviour of models which are not transparent. Markus, Kors, and Rijnbeek (2021) built on these definitions, characterising a model as explainable if it is either inherently transparent as defined by Lipton (2016), or post-hoc interpretable (ibid.), under the condition that these post-hoc explanations also reflect model behaviour accurately. I adopt the definition by Markus, Kors, and Rijnbeek (2021).

Definitions of terms in the scientific community, and definitions laid out by regulators have recently been found to differ fundamentally from each other (Gyevnar, Ferguson, and Schafer 2023): Regulators may regard explainability less as a goal to be achieved than rather as a "means that is needed to promote a range of very different values" related to human rights (ibid.). This can be exemplified by the phrasing of the European Union's GDPR laws, stating that a data subject has the right to "express his or her point of view and to contest the decision" which is "based solely on automated processing", and to obtain "meaningful information about the logic involved" (European Parliament and Council of the European Union 2016). The proposed AI Act, in which requirements for such "meaningful information" are laid out, emphasises algorithmic accountability, contestability of decisions, and human empowerment in the interaction with models (European Commission 2021), which go beyond most definitions of explainability proposed in the scientific literature.

Below, I provide a brief introduction to the topic of ML for reference. I focus on supervised learning, which is commonly used for prediction modelling. After that, I introduce a class of explainability methods which is often used to explain ML prediction models.

5.2 Machine learning overview

Machine learning is a field of artificial intelligence focused on the development of algorithms and models that enable computers to learn patterns and make predictions or decisions without being explicitly programmed. The key idea is to allow computers to learn from data and improve their performance over time.

ML algorithms and models identify patterns and relationships within the data, allowing the system to generalise and make predictions on new, unseen data. ML models often undergo iterative training processes, where they adjust their parameters to improve performance.

Supervised learning is a subcategory of ML where the algorithm is trained on, or learns from a labeled dataset. This labeled dataset consists of observed input data - with individual variables within that input data often called features - which is paired with corresponding observed output labels. An output

is the variable in the dataset that researchers are interested in predicting. An input (variable) is a variable that is known and that may contain some information about the value of the output. The goal for the algorithm is to learn a mapping function relating inputs to outputs. Predictions on new, unseen datapoints with the supervised ML model are possible after training.

Prediction models vary in their explainability. A linear regression model, for example, is transparent in terms of the coefficients which explain the mapping from model inputs to output. A coefficient for a particular feature, or input variable, commonly represents the change in the output for a one-unit change in this feature or input. The mapping from inputs to output is much more complex in, for example, a neural network model, for which such coefficients are not available. Therefore, neural networks are considered less explainable.

Explainability of an ML model can be defined on the level of individual predictions (resulting in so-called local explanations), and overall model behaviour (global explanations). Local explanations aim to explain why a prediction model made a particular prediction for a specific data point from the available dataset. Global explanations, in turn, aim to summarise model behaviour across the whole dataset.

In the following, I introduce a class of explanations called “feature attribution”.

5.3 Feature attribution

One kind of explainability which is frequently used by ML engineers in practice is feature attribution. Feature attribution attaches importance to every feature in the feature set according to how important it is for the model behaviour as a whole - called global feature attribution - or for an individual prediction - called local feature attribution. The concept of importance is differently interpreted by different researchers. Local feature attributions refer to individual feature values of an individual data point for which a prediction was made, global ones to features (or input variables).

Feature attribution is especially useful for mental health contexts because explanations themselves are easy to communicate to stakeholders like clinicians, and - when their global variant is used - can sometimes take over the prominent role model coefficients play in classical statistics - not in their specific interpretation, but as entities generally referring to feature importance.

5.3.1 Feature attribution in mental health outcomes research

Feature attribution is increasingly used in mental health outcomes research, attributing importance in feature sets used for the prediction of digital and face-to-face treatment outcomes with ML models (Delgadillo, Rubel, and Barkham 2020; Flygare et al. 2020; Garriga et al. 2022; Koutsouleris et al. 2016; Lenhard et al. 2018; Paul et al. 2019; Van Breda et al. 2018). I identified three use cases of feature attribution in this domain, that I associate with a stakeholder group each. I would like to emphasise that these use cases need to be realised in concert for overall benefit from feature attribution as a post-hoc analysis method.

First, feature attribution may be helpful for model developers debugging a model, and gauging potential model bias. Since datasets in mental healthcare are not yet large enough to prevent ML model overfitting (Koppe, Meyer-Lindenberg, and Durstewitz 2021), feature attribution may help to understand whether a model focuses on unanticipated artifacts that should not drive prediction, for example study site identifiers or proxies for those. Further, feature attribution may help to detect model discrimination of patient subgroups ahead of deployment: It would for example be useful to detect model focus on protected patient characteristics which, when these characteristics are used for clinical decision making, results in patient discrimination (Obermeyer et al. 2019). This is relevant when model decisions are supposed to inform treatment and resource allocation.

Secondly, global and local feature attribution may be used to make general ML model behaviour and individual model predictions more transparent for practitioners who are often tasked with contributing to human oversight of the model after deployment. Note in this context that explanations can also induce a false sense of trust in model accuracy, which is especially fatal when these explanations do not reflect true model behaviour (Eiband et al. 2019; Ghassemi et al. 2018; Kroll 2018; Lockey et al. 2021; Poursabzi-Sangdeh et al. 2021). Further, feature attributions may also effect misplaced causal reasoning: A therapist may for example hold the false belief that focusing on a patient's symptom which was attributed much importance for the prediction of a clinical outcome during therapy sessions may change this patient's prognosis.

Thirdly, given appropriate model performance and (external) validation, feature attribution may help researchers understand mental health and illness better by suggesting yet unknown variable relationships. This, in turn, could lead to the generation of hypotheses if these relationships are unknown as yet, or the support of existing hypotheses about factors evoking, sustaining, and resolving mental health problems, a potential capability of this methodology that was also proposed for feature selection (Chen et al. 2023). In this vein, global feature attributions have, for example, suggested personality

traits to be implicated into the characterisation of subcategories of major depressive disorder because their importance was highlighted in the prediction of treatment response clusters of major depressive disorder patients (Paul et al. 2019).

Many feature attribution methods have been introduced in the scientific literature. In the remainder of this chapter, I focus on local feature attribution methods which may be more important for the case in which prediction models are deployed in practice and an explanation for a specific prediction is sought. Below, I talk more in detail about some popular local feature attribution methods.

5.3.2 Local feature attribution methods

Local feature attribution methods in the literature can broadly be separated into two classes; perturbation- and gradient-based methods. Perturbation-based methods attribute importance based on the effect of feature value perturbations on the prediction. This means that the prediction model is probed on unobserved, perturbed data points which deviate to varying extent, i.e. on a specific subset of features, from the data point for which an explanation is requested. The deviation of the prediction for an unobserved, perturbed data point from the original prediction is an indicator for the importance of the feature values whose value was perturbed. What makes many of these perturbation-based methods useful in practice is that they are model agnostic, i.e. applicable to any chosen prediction model.

Gradient-based methods can be thought of as computing the machine learning model analogues of linear regression coefficients, i.e. the derivative of a model output with respect to its inputs, which is commonly processed further for better interpretation, for example by smoothing. As gradient-based methods can only be applied to models which are differentiable, which many models used for categorical data from mental health contexts are not, these methods will not be a focus of this chapter. In the following, the mechanics of two popular post-hoc perturbation-based feature attribution methods, LIME and Shapley value based feature attribution, are described.

LIME

Local Interpretable Model-agnostic Explanations (LIME) approximate the behaviour of the prediction model for an individual prediction with a transparent surrogate model, for example a regularised linear regression (Ribeiro, Singh, and Guestrin 2016). The functioning of this method is illustrated in **Fig. 5.1**. The surrogate model is built on artificial data resulting from small perturbations of the feature values of the original data point, and predictions output when the prediction model to be explained is fed the input variables including the perturbed feature values. Perturbations are generated with a kernel, and

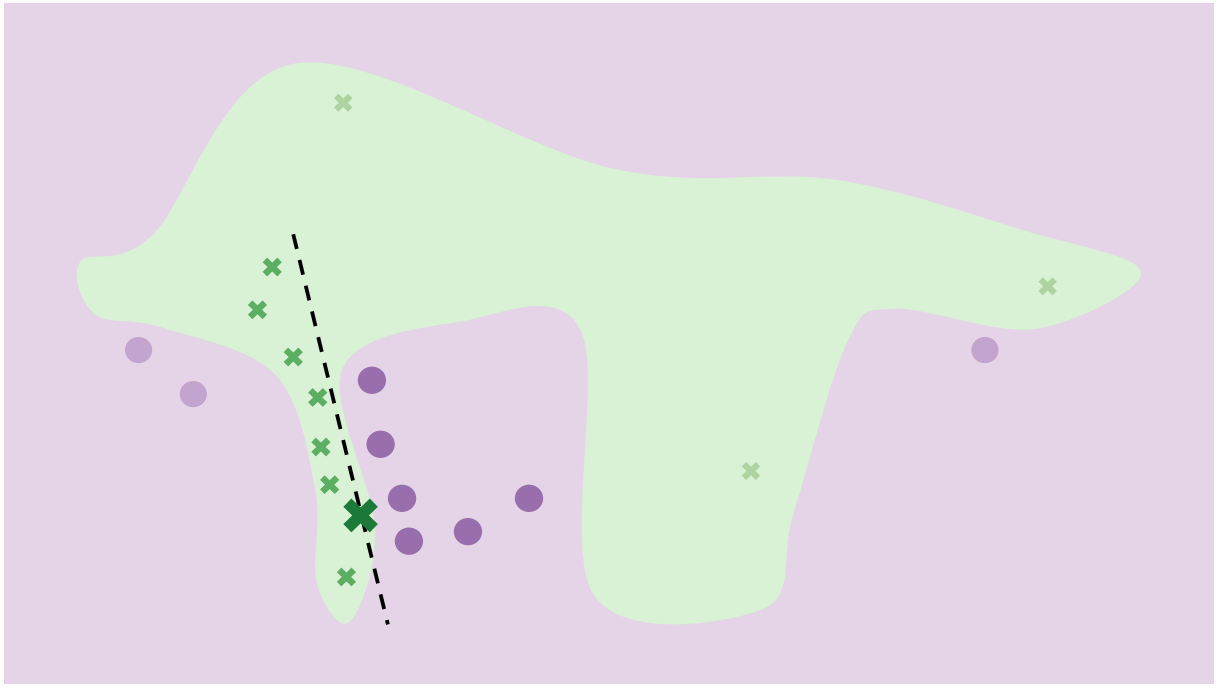

Figure 5.1. Basic concept of LIME approximation of local model behaviour with a transparent surrogate model. The coloured areas correspond to decision regions for a complex binary classification model. The bold green cross represents the data point of interest. Light green crosses and light violet points correspond to artificial data around the instance of interest. The dashed line represents a simple linear model fitted to the artificial data. The simple linear model describes local behaviour of the black-box model around the data point of interest.

the transparent surrogate model weights new data points according to their proximity to the original. If for example a regularised linear regression surrogate model is used, then, importance is attributed to a feature value depending on the magnitude of the regression coefficient for that feature in the surrogate model built from the artificial data set.

Shapley value based methods

A range of feature attribution methods, including the popular so-called SHAP (Lundberg and Lee 2017), have been proposed which are based on the Shapley value. The Shapley value is a cooperative game theoretic concept which was adapted for feature attribution (Shapley 1953; Strumbelj and Kononenko 2010; Štrumbelj and Kononenko 2014). Originally, it was designed to fairly distribute the joint return of a group of players in a game according to the contribution of each player. In feature attribution, the Shapley value can describe the contribution of a feature value to an individual prediction in the sense of its marginal contribution to all possible subsets of feature values for this particular data point. With contribution, I mean the magnitude of the average difference between the prediction for the original data point and the prediction for any subset of the feature values (feature values out of this subset are replaced by feature values of another random data point in the dataset).

The Shapley value has a couple of desirable theoretical properties. The Shapley values for all feature values, for example, sum up to the difference between the prediction for the individual data point, and the average prediction for the data set. Also, Shapley values can be aggregated across data points to obtain a global estimate of feature importance (Covert, Lundberg, and Lee 2020).

Determining the marginal contribution of a feature value to all possible subsets of feature values is computationally expensive. Most of the time, a Monte Carlo procedure, sampling subsets, is used. Researchers often refer to so-called SHAP (or KernelSHAP) when Monte Carlo samples are chosen by means of a kernel (Lundberg and Lee 2017).

Despite their potential relevance for the trust stakeholders can place in ML model predictions, research has found that feature attribution methods can be unreliable. Below, I detail in which ways this may be the case.

5.3.3 Feature attribution reliability

Human evaluation of the plausibility of feature attributions in terms of whether it "makes sense" that an ML model relies on a certain feature to make a prediction may be necessary, but is not sufficient to judge the quality of feature attributions. These attributions, by design, only approximate model behaviour (Rudin 2019). Several teams of researchers have published evidence that popular feature attribution methods may, inherently, not present reliable information about model reasoning: They have for example been found to be vulnerable to adversarial attacks which can hide classifier dependence on a feature from feature attribution methods (Slack et al. 2020), and sensitive to ML model hyperparameters such as random seeds (Bansal, Agarwal, and Nguyen 2020) as well as infinitesimal perturbations to input features (Agarwal et al. 2022).

Also, doubts were raised about whether feature attributions are faithful to the underlying ML model, with faithfulness meaning that feature attributions reflect the ground truth about which features are used by the model for decision making. For real-world data, and situations in which the interior of the model is inaccessible, for example due to API-only access to the model output, such ground truth about feature importance is typically unavailable. Some researchers have attempted to circumvent this problem by inducing a ground truth about feature importance in real-world data: Zhou et al. (2021), for example, found feature attribution methods to fail to detect induced reliance of neural networks on spurious artifacts. However, inducing a ground truth to probe feature attribution methods on whether they will recover this ground truth is not always possible in data science practice.

Krishna et al. (2022) have recently circumvented this problem, proposing that feature attributions for an individual prediction may be more believable when two different local feature attribution methods agree with each other on the attributions. Importantly, they also found that disagreement between feature attributions from different methods is frequent in data science practice, and that data scientists do not know how to deal with this problem effectively. They term this the “disagreement problem”, and examine its severity in several datasets, among them two tabular datasets with 7 and 20 features, respectively, which bear some similarity to tabular datasets collected for mental health outcomes research which are commonly categorised by categorical features due to patient self-reports. In addition to that, Markus, Fridgeirsson, et al. 2023 have recently shown that disagreement of global feature attribution methods is prevalent for prediction models of outcomes relevant for clinical medicine made with electronic health record data, and examined the influence of some dataset characteristics on the magnitude of disagreement, using semi-natural datasets.

Little research exists to date on how the level of feature set multicollinearity, i.e. the strength of correlations between features used by the prediction model, influence aforementioned disagreement between feature attribution methods. What is known is that, under multicollinearity, Shapley value attributions, for example, may not be a good representation of feature values’ importance: Feature values out of the subset, clipped to feature values in the subset from a random other instance, may not respect feature correlations, resulting in the prediction model being evaluated on unrealistic instances. Adversarial attacks on global Shapley value based explanations such as the one from Dimanov et al. (2020) are feasible because a single feature’s information content can be completely represented by other features, which is possible because features are correlated with each other. Further, Frye, Mijolla, et al. (2021) demonstrate in simulated and empirical data that tight correlation between two features can result in negative global Shapley values for one of the correlated features (and would therefore indicate - counterintuitively - that using this feature would be detrimental to the performance of the model). Moreover, experiments from Aas, Jullum, and Løland (2021) on simulated data suggest that Shapley value based feature attributions are suboptimal under multicollinearity because the method evaluates the prediction model on unrepresentative examples.

Research in this domain, generally, struggles with defining a ground truth of feature importance against which to evaluate the attributions, meaning which features are truly important to a black-box model. Several alternatives for Shapley value based local feature attribution methods have been proposed for prediction on multicollinear data in the literature (Frye, Mijolla, et al. 2021; Olsen et al. 2022), some of them tackling specific reasons for correlation, for example, causal ancestry (Frye, Rowat, and Feige 2020). However, many of them are expensive to compute, or can be applied to data of specific types only.

The question of how multicollinearity influences feature attribution method disagreement is of particular relevance for mental health outcomes research: Multicollinearity as an influence factor on the reliability of feature attributions is of interest because it is frequently present in mental healthcare related datasets, e.g. through the inclusion of self-report questionnaires, which continue to constitute the primary measurement instrument in mental healthcare. Generally, research considers many interacting levels of analysis for mental health, which encourages multicollinearity: A high-level outcome such as treatment success can be influenced by a multitude of distal and proximal factors in a patient's present and past which may be related to each other, for example, childhood adversity, strategies to cope with stress, and current depressive symptoms.

The feature set I have used in Chapter 3 to predict maintenance/attainment of sub-threshold mental health problems after provision of access to the content of a DI is a good example for a feature set in mental health outcomes research: I built my ML models with data from an assessment battery that was completed by registrants with the digital substance dependence intervention BFO before access was provided to this intervention's CBT-based content. Having a baseline assessment battery like this one available as a predictor set is frequently the case in mental health outcomes research (Carl et al. 2020; Mehta et al. 2021), also for non-digital interventions. These assessment batteries often consist of sociodemographic variables, and questionnaire items. In Chapter 3, I employ global feature attribution methods inbuilt to the random forest model I used for prediction, and visualised averaged local effects with plots by Apley and Zhu (2020) to help understand which features are associated with outcomes most and what averaged effects a value on a particular feature had on the predicted probability of a certain outcome. Now, I am interested in how valid conclusions from such post-hoc feature attribution analyses with regards to feature importance really are when features are as highly correlated as they were in my analyses.

Hence, with the application of prediction modelling of BFO data, and more broadly, mental health outcomes research using ML prediction models, in mind, I aim to explore the reliability of feature attribution methods when multicollinearity is present, with a particular focus on feature attribution method disagreement.

5.4 Objectives

I have described the problem of local feature attribution method unreliability for ML models, and how it is of relevance to the prediction of mental health outcomes with data characterised by multicollinearity. In the following, I present a study in which I aim to a.) explore the effect of feature set multicollinear-

ity on the degree of reliability of local feature attribution explanations of predictions made in synthetic data with reasonable similarity to datasets collected in mental healthcare, and b.) to illustrate the degree of feature attribution method disagreement for predictions made in semi-natural, and natural mental health related datasets with considerable multicollinearity present. I opt for using synthetic, semi-natural, and natural datasets to also be able to evaluate explanations against a ground truth.

To understand the effect of feature set multicollinearity on feature attribution reliability, I varied the degree of feature set multicollinearity in a synthetic dataset and observed disagreement of two popular local feature attribution methods on individual predictions (Krishna et al. 2022). I also observed disagreement of local feature attribution methods with the ground truth of a transparent prediction model to understand how well local feature attribution methods approximate a comparably easy-to-explain model under varying degrees of feature set multicollinearity.

To explore how relevant Krishna et al. (ibid.)'s disagreement problem is for mental health outcomes research, I matched the synthetic data closely to real-world data from mental healthcare contexts by simulating self-report questionnaire data. I also contrasted the effect of varying feature set multicollinearity on explanation reliability with the effects I obtain when modifying other characteristics of my synthetic data. As my synthetic datasets can only approximate the complexity of real-world data from mental healthcare contexts which are likely characterised by certain unknowable variables, I also quantify disagreement on semi-natural and natural datasets with multicollinear features.

One cause of the multicollinearity in datasets used in mental healthcare research whose effect I aim to quantify is the inclusion of data resulting from responses to self-report questionnaires into feature sets. Each of these questionnaires comprises multiple questions, called items, which are - through optimisation of internal consistency, a quality criterion for psychological questionnaires - designed to correlate highly with each other.

Questionnaires used in mental health outcomes research are ideally designed to refer to distinct mental health related phenomena. Since these phenomena are, in reality, often conflated on some level - owing to the known challenges of the field of psychiatry to establish non-overlapping classes of pathology - yet another layer of multicollinearity is introduced. In my models in Chapter 3, both these layers are present: To respect the transdiagnostic character of many mental health problems, and the relevance of symptoms rather than underlying disorders for mental healthcare practice (Abi-Dargham et al. 2023; Borsboom, Cramer, and Kalis 2019), I included individual items instead of sum scores of questionnaires into the feature sets for this study, resulting in correlations between items from the same questionnaire, and correlations between items from different questionnaires. With the use case of ML models in mental health outcomes research in mind, I aim to account for both layers of multicollinearity

Dataset	Prediction model	Feature attribution method reliability measure
synthetic	elastic net	disagreement of methods with model coefficients
	random forest	disagreement amongst methods
	XGBoost	
semi-natural	random forest	
natural	random forest	

Figure 5.2. Overview over study methodology.

- on an item, and questionnaire level - in my simulations.

As the two papers from Krishna et al. (2022) and Markus, Fridgeirsson, et al. (2023) are closely and importantly related to the research presented in this chapter, I will make explicit reference to them in the Results section.

5.5 Methodology

I start this section by reporting my approach to quantifying the impact of varying the degree of multicollinearity in synthetic data on feature attribution method reliability for predictions made on this dataset, the primary aim of this study. I detail how synthetic data was simulated to represent data from a mental health questionnaire, and how I varied the degree of multicollinearity in it. I then describe which prediction models and feature attribution method implementations I considered, and which aspects of feature attribution method reliability I considered.

After this, I describe the changes I made to my initial sampling model to vary the number of items per questionnaire, and the number of non-predictive features in my synthetic datasets. The rationale behind these steps was to put the effect of varying the degree of feature set multicollinearity into context by comparing this effect to the effect of varying other dataset characteristics.

At last, I describe how I quantify feature attribution method disagreement in semi-natural, and fully natural data.

Key elements of my methodology are summarised in **Fig. 5.2**.

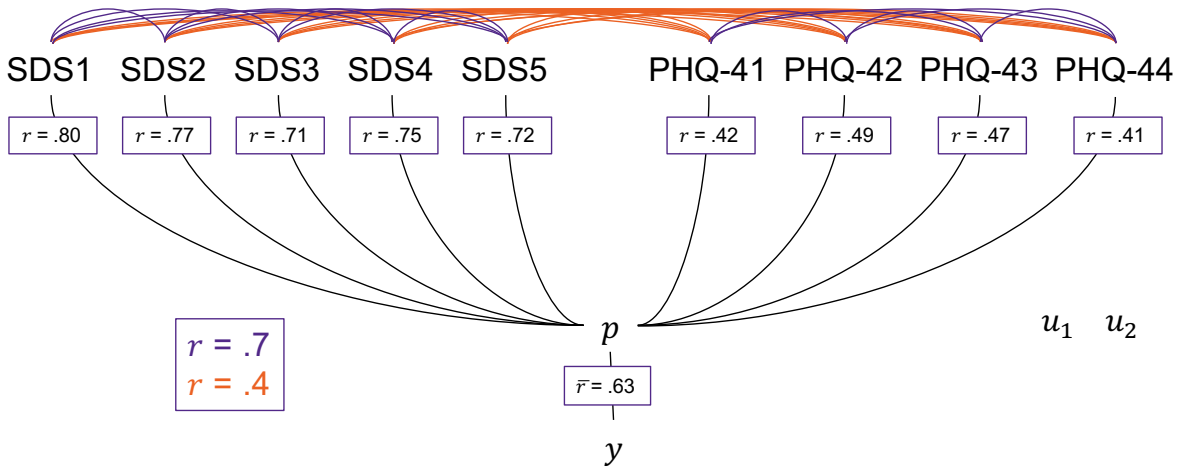

Figure 5.3. Initial sampling model with five SDS items and four PHQ-4 items. p denotes a general psychopathology factor, u_1 and u_2 two features non-predictive of the outcome, and y the outcome. Correlations between SDS items and p are randomly sampled from $r \in \{0.70, 0.71, 0.72, 0.73, 0.74, 0.75, 0.76, 0.77, 0.78, 0.79, 0.80\}$, correlations between PHQ-4 items and p from $r \in \{0.40, 0.41, 0.42, 0.43, 0.44, 0.45, 0.46, 0.47, 0.48, 0.49, 0.50\}$. Through joint and conditional sampling procedures, empirical correlations are slightly lower than shown.

5.5.1 Experiments on synthetic data

I start by providing details of my sampling model, which I recycle with modifications in all experiments using synthetic and semi-natural data.

Sampling model and data simulation

In the following, I present my initial sampling model for the experiments varying degrees of multicollinearity. From this initial sampling model I simulate data, and deviate strategically to represent different degrees of multicollinearity in the feature set by modifying the respective parameters. It is illustrated in **Fig. 5.3**.

This initial sampling model represents a deliberately simple scenario in which I account for both the layers of multicollinearity caused by questionnaire design (correlations of items from the same questionnaire) and questionnaire construct overlap (correlations of items from different questionnaires) present in many tabular datasets available for mental healthcare research. Specifically, I simulate responses to a total of nine items belonging to two self-report questionnaires, the Patient Health Questionnaire 4 (PHQ-4) (Kroenke et al. 2009), a screener for anxiety and depression related symptoms, and the Severity of Dependence Scale (SDS) (Gossop et al. 1995), querying the presence of symptoms related to substance use disorder. I also simulate the latent variable p which underlies the item variables to a varying extent, and two control variables u_1 and u_2 which are unrelated to the outcome y, and are subsequently referred to as “non-predictive features”.

Questionnaire items SDS1-SDS5 as well as PHQ-41-PHQ-44, p, u_1 and u_2 are considered to be on an ordinal scale with four response categories (0 - 3, 0 representing no impairment, and 3 representing significant impairment). Data on these variables are sampled from a joint distribution with the "genOrdCat" function available in the R package "simstudy", using correlation matrices describing variable relations exemplified in **Fig. 5.3**. Through conditional sampling, I obtain a binary outcome y in whose sampling model only p is included with a sampling coefficient of b = 5. I assume a prevalence of 0.2 to conform with typical clinically relevant outcomes in mental healthcare research which may be often rare, as is for example the event of experiencing a mental health crisis in the next month used as a prediction target in Garriga et al. (2022).

One element in my sampling model which causes feature set multicollinearity is the correlation between items from the same questionnaire. I am interested in the effects of multicollinearity on feature attribution reliability, and hence vary the correlation coefficients of items within the same questionnaire such that any $r \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$ within one variation.

For each variation of the initial sampling model I undertake, initially in terms of varying levels of multicollinearity, I simulate 15 times 10.000 data points from the respective newly initialised sampling model, and build a new prediction model on each of the 15 repetitions.

Frequency counts for one variation (10.000 data points) are shown in **Fig. 5.4**. Below, I provide an extended rationale for my initial sampling model.

Extended rationale for my sampling model In my sampling model, I allowed items from two different questionnaires to covary, but to a lesser extent than items from the same questionnaire. In this way, I account for the frequent co-occurrence of the mental health problems that each of the individual questionnaires assesses, while also acknowledging that certain symptoms, for example those related to substance abuse, which are collated in the SDS, co-occur more often than others.

I decide against relating items of each questionnaires to variables relating to the diagnostic categories most closely associated with these questionnaires, such as substance use disorder in the case of the SDS or major depressive disorder in the case of the last two PHQ-4 items. These disorders often present alongside each other in clinical practice, and their scientific validity has been subjected to criticism. Instead, I relate all questionnaire items to a general psychopathology factor p. p itself is inspired by the general psychopathology factor p proposed in psychiatric research to explain covariation in psychopathological symptoms (Pettersson et al. 2020).

In certain cohorts, we can expect some psychiatric phenomena which are assessed with questionnaires, to indicate greater case severity than others, and therefore greater degrees of psychopathol-

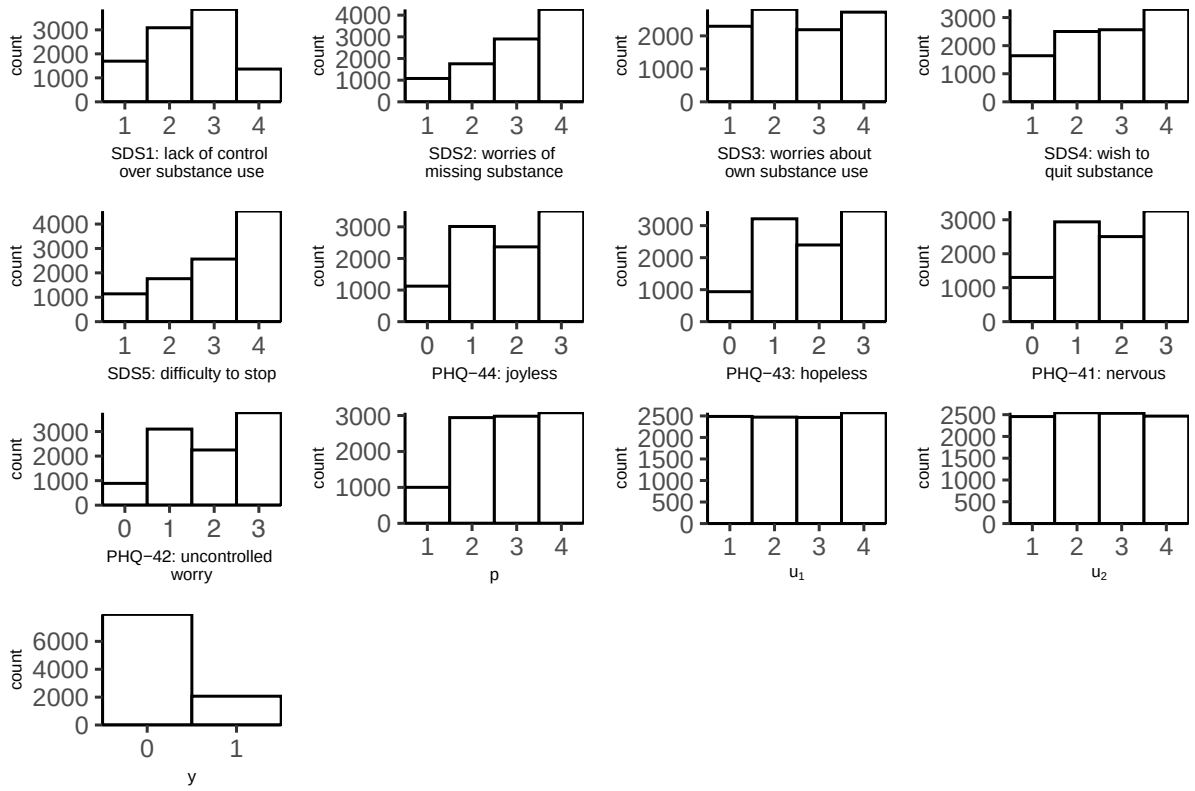

Figure 5.4. Frequency counts of simulated data with five SDS items, four PHQ-4 items, two non-predictive items, the general psychopathology factor p, and y.

ogy. For this simulation, I hold substance using individuals such as registrants with the digital substance dependence programme BFO in mind as respondents to questionnaires. For these individuals, we can expect SDS items to relate more strongly with the general psychopathology factor p than PHQ-4 items.

My choice of a common ordinal scale for questionnaire items, p, and u_1 and u_2 rules out potential effects of feature scale on feature attributions (Strobl et al. 2007) that may obscure the effect of multicollinearity.

Prediction models

The prediction models I built took as an input the simulated data for the questionnaire items SDS1-SDS5 and PHQ-41-PHQ44 as well as non-predictive features u_1 and u_2. They were used to predict the binary outcome y.

I experimented with three different prediction models which are frequently used in mental health outcomes prediction: an elastic net, implemented with the R package “glmnet”, a random forest (R package “randomForest”), and an XGBoost (R package “xgboost”). I opted for tree-based models instead

of neural networks as recent research has found that they continue to outperform neural networks on tabular datasets (Borisov et al. 2021). In contrast to the tree-based prediction models which often perform well out-of-the-box, I decided to do hyper-parameter training for the elastic net over a grid of $\alpha \in \{0.01, 0.1, 1\}$ as the elastic net penalty (α = 1 being a pure lasso regression and α = 0 a pure ridge regression) and λ as a sequence of 100 equally spaced numbers between 0.0001 and 1 as the overall strength of penalty.

Feature attribution method implementations

I explain individual predictions from my models with two different post-hoc perturbation-based feature attribution methods. One of them is based on the Shapley value, specifically, I used a Monte Carlo sampling based approximation to it using the R package “fastshap”. The other one is the LIME implementation of the R package “lime”.

Measurement of feature attribution method reliability

The prediction models I use are located on a continuum of explainability, with the elastic net, thanks to its coefficients available after model fit, being closer to the transparent, glass-box end of that continuum. Consequently, I decide to operationalise feature attribution method reliability (1) as the disagreement amongst post-hoc feature attribution methods for my random forest and XGBoost models (inherently less explainable), and (2) as the disagreement of post-hoc feature attribution methods with available model coefficients for my elastic net models (inherently more explainable, ground truth available). In the Results section, I denote (1) as disagreement, and (2) as faithfulness.

I adopt my operationalisation of disagreement for any two feature attributions (whether local feature attributions through LIME or Shapley values, or essentially global feature attributions through elastic net coefficients) from Krishna et al. (2022). Specifically, I adopt their metrics rank agreement, sign agreement, signed rank agreement, and pairwise rank agreement, which are defined below. I also use Spearman rank correlations, which Krishna et al. (ibid.) use as well.

The metrics rank agreement, sign agreement, and signed rank agreement depend on the parameter k, the number of top ranked features in a feature attribution that is of interest. k = 5 would for example result in rank agreement, sign agreement, and signed rank agreement of the five features ranked highest by the two feature attributions LIME and Shapley values which I use in this study. Due to my comparably small number of features, I compute these metrics (1) for k = 11, or all features in the feature set

if their number does not equal 11, and (2) for k = 5 as this is a typical number of features looked at in practice to quickly gauge whether two explanations disagree with each other.

Krishna et al. (2022) propose to let the metrics Spearman rank correlation and pairwise rank agreement depend on a parameter $F = \{f_1, f_2...\}$, which contains a set of (not necessarily top-ranked) features for which agreement of feature attributions is of particular interest. I let F contain the whole feature set.

Rank agreement This metric quantifies the fraction of features that are not only common between the sets of top k features of two feature attribution methods, but also have the same rank.

The formal definition of this metric, RankAgreement(A_a, A_b, k), for two feature attributions A_a and A_b and a chosen k is:

$$\frac{|\bigcup_{s \in S} \{s \mid s \in \text{top\_feat}(A_a, k) \wedge s \in \text{top\_feat}(A_b, k) \wedge \text{rank}(A_a, s) = \text{rank}(A_b, s)\}|}{k} \tag{5.1}$$

S is the complete feature set. top_feat(A, k) gives back the set of top k features for the attribution A with regards to the size of the feature importance values. rank$()$ gives back the position or rank of the feature s according to the attribution A.

Sign agreement This metric quantifies the fraction of features that are not only common between the sets of top k features of two feature attribution methods, but also have the same sign.

SignAgreement(A_a, A_b, k) is defined as follows:

$$\frac{|\bigcup_{s \in S} \{s \mid s \in \text{top\_feat}(A_a, k) \wedge s \in \text{top\_feat}(A_b, k) \wedge \text{sign}(A_a, s) = \text{sign}(A_b, s)\}|}{k} \tag{5.2}$$

where sign$()$ gives back the sign of the feature s according to the attribution A.

Signed rank agreement This metric quantifies the fraction of features that are not only common between the sets of top k features of two feature attribution methods, but also have the same rank and sign.

This is how SignedRankAgreement(A_a, A_b, k) is defined:

$$\frac{|\bigcup_{s \in S} \{s \mid s \in \text{top\_feat}(A_a, k) \wedge s \in \text{top\_feat}(A_b, k) \wedge \text{rank}(A_a, s) = \text{rank}(A_b, s) \wedge \text{sign}(A_a, s) = \text{sign}(A_b, s)\}|}{k} \tag{5.3}$$

Pairwise rank agreement This metric quantifies the degree to which the relative ordering of every pair of features in F is the same for both attributions.

PairwiseRankAgreement(A_a, A_b, F) is defined as:

$$\frac{\sum_{i,j \text{ for } i<j} \mathbb{1}\left[\text{RelativeRanking}(A_a, f_i, f_j) = \text{RelativeRanking}(A_b, f_i, f_j)\right]}{\binom{|F|}{2}} \tag{5.4}$$

where RelativeRanking(A, f_i, f_j) is an indicator function giving back 1 if feature f_i is more important than feature f_j according to explanation A, and 0 otherwise.

These metrics are presented in my Results section as aggregations over 15 repetitions of data simulation for one particular variation of the degree of multicollinearity present in the synthetic data.

Varying the number of items per questionnaire

I want to contrast the effect of varying multicollinearity in a feature set on feature attribution method reliability with the effect of varying other potentially influential characteristics of a dataset. One of them is the number of items per questionnaire. To do this, I almost completely recyle the procedure of my experiments varying degrees of multicollinearity outlined above, with some small modifications described below.

Instead of varying the correlation coefficients of items within a questionnaire, I hold the correlation of items from the same questionnaire fixed at r = 0.7, and vary the number of SDS items and PHQ-4 items I simulate. Specifically, my initial sampling model in **Fig. 5.3** is characterised by a slight imbalance (four PHQ-4 items and five SDS items) of the number of items per questionnaire.

My modification consists of stipulating (1) a scenario with four SDS and four PHQ-4 items, the "balanced" scenario, and (2) a scenario with six SDS and two PHQ-4 items, the "unbalanced" scenario. Since both scenarios have the same total number of included questionnaire items, I can rule out the influence of the total number of features, which seems to have a negative effect on disagreement when larger (Markus, Fridgeirsson, et al. 2023). For (1), I did not simulate data for the last SDS item SDS5. For (2), I simulated data for the first SDS item SDS1 twice, and did not simulate data for the last two PHQ-4 items PHQ-43 and PHQ-44.

Varying the number of non-predictive features

I also conduct experiments varying the number of non-predictive features (u_1 and u_2 in my initial sampling model in **Fig. 5.3**). Again, I almost completely re-use the procedure of my experiments varying degrees of multicollinearity.

However, I hold the correlation of items from the same questionnaire fixed at r = 0.7. I also changed the number of non-predictive features added to the feature set to 0, 5 and 10.

5.5.2 Illustration of feature attribution method reliability in semi-natural data

To obtain semi-natural data, and illustrate the degree of feature attribution method reliability in it, my modifications to the procedure laid out in the section describing experiments varying degrees of multicollinearity in synthetic data are minimal. They affect (1) the sampling model, (2) the prediction model used on the data, and (3) the measure of feature attribution method reliability.

With regards to (1), I create semi-natural data using an empirical correlation matrix from a natural dataset to describe the correlation of data on simulated SDS and PHQ-4 items in **Fig. 5.3**. The empirical correlation matrix was obtained from a natural dataset including responses of registrants with the digital substance dependence programme BFO to an assessment battery administered to them before engagement with the CBT-based content. This assessment battery also contained SDS and PHQ-4. The dataset is identical with the one used for engagement prediction in Chapter 4 of this thesis.

With regards to (2), I use a random forest model only in order to make results obtained in semi-natural data comparable to those obtained on natural data (see next section). This means, that for (3), I only report feature attribution method disagreement between LIME and Shapley value based feature attribution.

5.5.3 Illustration of feature attribution method disagreement in natural data

To showcase the extent of feature attribution method reliability in a fully natural dataset, I again made use of the dataset used for engagement prediction in Chapter 4 of this thesis. I used data from BFO registrants who completed at least one programme module before their last recorded post-engagement assessment. This leaves me with a total of 1.385 registrants. I use their pre-engagement assessment data (67 features) - excluding the number of days from registration to first assessment completion - to predict self-reported abstinence (zero days of consuming primary substance) in the week before the last recorded assessment update. No feature describing engagement after the first-contact assessment completion was used in this prediction model.

My model selection process included a random forest model and an XGBoost model which may be chosen over regression-based models when prediction models are built on high-dimensional, natural data such as the one I used. As the random forest model performed better than the XGBoost model within 10 repeats of 10-fold cross validation on 80% of the BFO dataset set aside for training, I proceeded with it. I then trained one random forest on this subset of 80%, and tested its performance on the remaining 20% of the data which had not been used during model selection.

5.6 Results

I first report my findings from experiments on synthetic data, and then go on to illustrate feature attribution method disagreement in semi-natural, and natural datasets.

5.6.1 Experiments on synthetic data

I begin with results of my experiments varying multicollinearity in synthetic datasets.

Experiments varying multicollinearity

I begin reporting disagreement amongst feature attribution methods in data with varying degrees of multicollinearity. Disagreement was quantified as reported in the Methodology section “Measurement of feature attribution method reliability”. Specifically, top 5 and top 11 rank, sign, and signed rank agreement, as well as pairwise rank agreement, and the Spearman rank correlation coefficient (all features considered) between LIME and Shapley value based feature attributions for individual predictions of

XGBoost and random forest models are reported. The variation of multicollinearity consisted of the variation of the correlation between simulated items from the same questionnaire, for example, the correlation between items simulated to represent responses to the PHQ-4. Correlations were varied between 0.5 and 0.9.

I found that, surprisingly, for predictions of an XGBoost, agreement between LIME and the Shapley value based feature attribution method increases slightly with stronger inter-item correlations within questionnaires on several metrics; top 5 signed rank agreement, and top 11 rank, sign and signed rank agreement.

Conversely, for predictions of a random forest model, I observe an opposite trend; slightly decreasing top 5 sign agreement, top 11 rank agreement, and pairwise rank agreement with stronger inter-item correlations.

These findings are illustrated in **Fig. 5.5**.

I also looked at faithfulness of feature attribution methods to the underlying model when degrees of multicollinearity were varied between r = 0.5 and r = 0.9. In this case, the model was an elastic net. Specifically, I explained the elastic net predictions with LIME and Shapley value based feature attributions, and quantified disagreement of these attributions with the elastic net coefficients. More detail is provided in the Methodology section “Measurement of feature attribution method reliability”.

I observed that when k = 5, feature attribution rank agreement with regression coefficients even increases slightly with stronger correlations between items from the same questionnaire. When k = 11, looking at signed rank agreement, the strength of correlations between items from the same questionnaire does not seem to have an effect. Variability of the Spearman rank correlations becomes slightly stronger with stronger inter-item correlations, i.e. agreement is more likely to be unusually high or low when correlations between items from the same questionnaire are high. Spearman rank correlations seems to be slightly more likely to be negative with increasing multicollinearity for LIME on this metric. I illustrate this in **Fig. 5.6**.

Experiments varying the number of items per questionnaire

To put the effect of multicollinearity I observed into context, I modified other characteristics of my initial sampling model, such as the level of imbalance of the number of items per questionnaire. Note that I simulate responses to two questionnaires, and vary the number of items for both of them. I regard the case in which I have simulated data for four SDS items and four PHQ-4 items as “balanced”, and the case in which I simulate data for six SDS items and two PHQ-4 items as “imbalanced”. Details of this

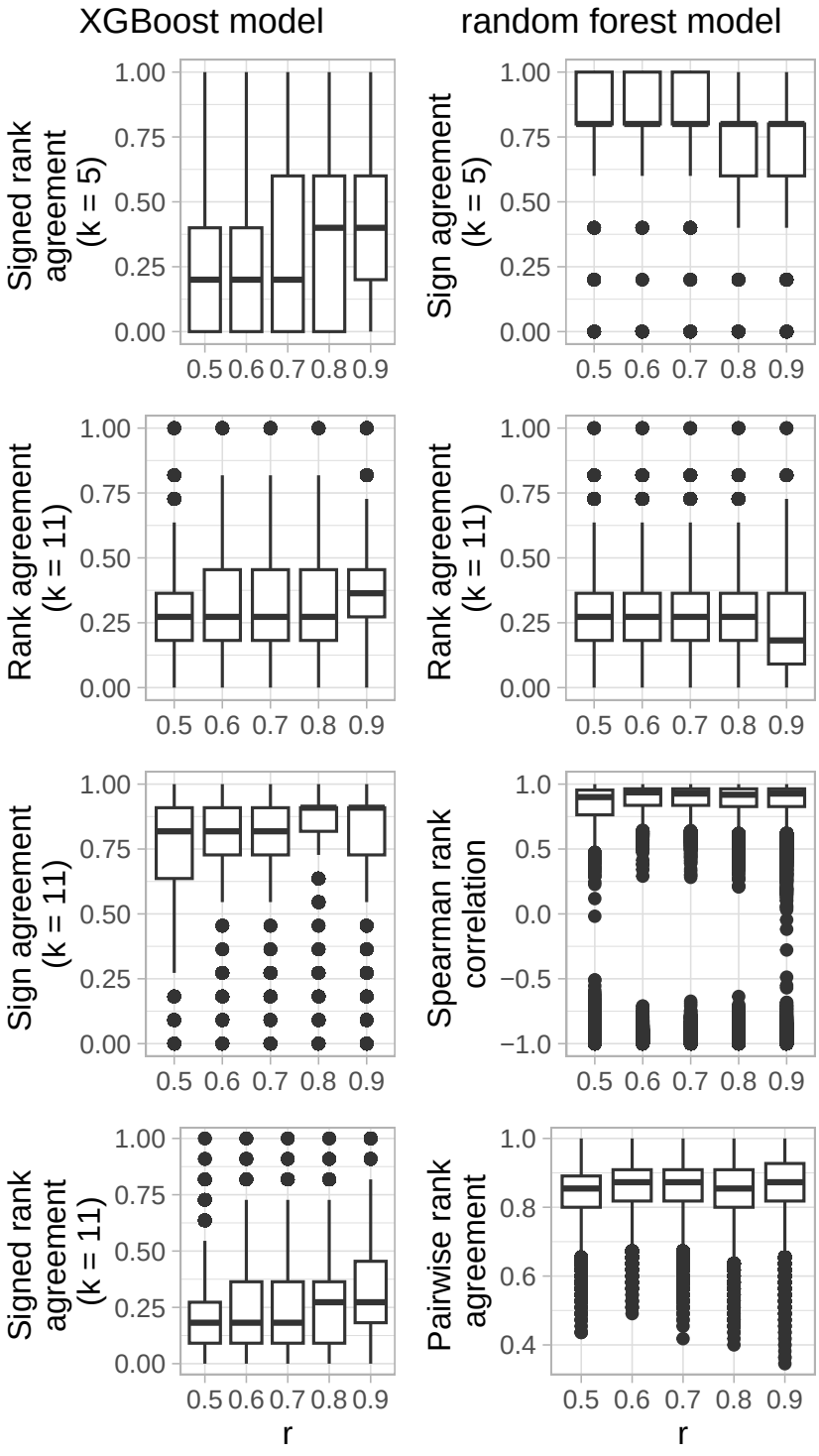

Figure 5.5. Distribution of disagreement between LIME and Shapley values for predictions of an XGBoost model (left) and a random forest model (right) built on simulated data. Data are characterised by varying strengths of correlations between items from the same questionnaire (x axis). The model predicted a simulated hypothetical outcome. Values closer to 1 indicate more agreement.

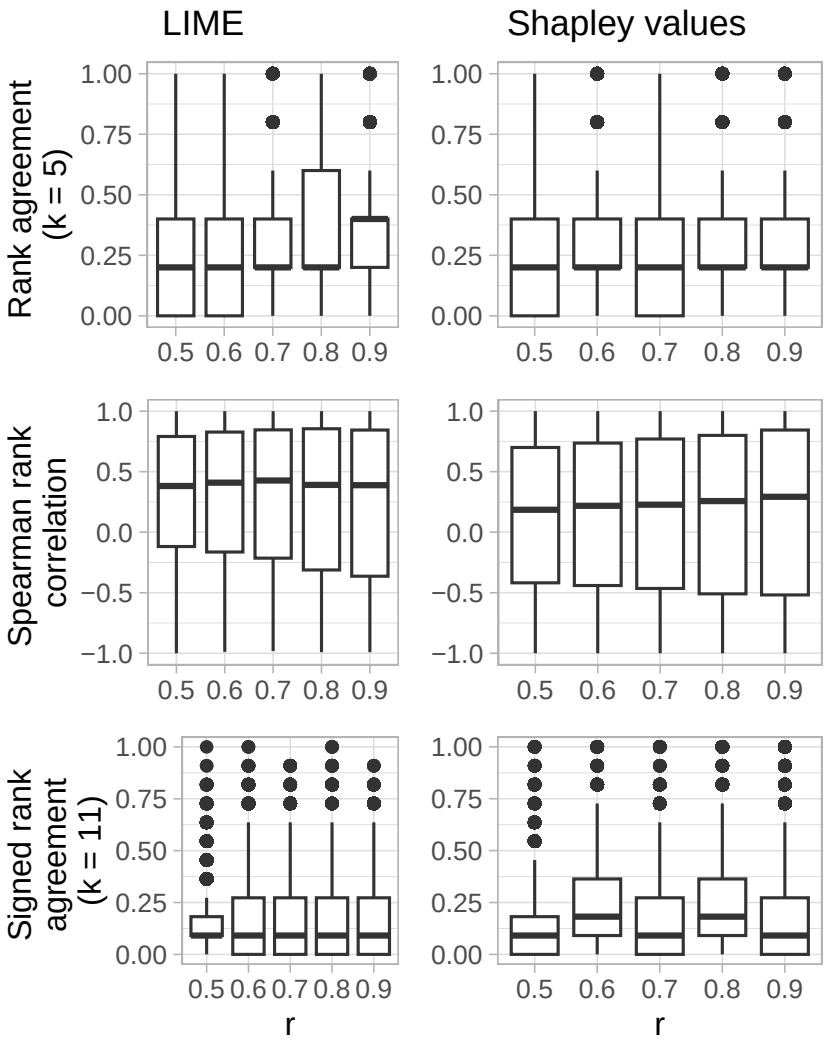

Figure 5.6. Distribution of disagreement between LIME (left) / Shapley values (right) with coefficients of an elastic net built on simulated data. Data are characterised by varying strengths of correlations between items from the same questionnaire (x axis). The model predicted a simulated hypothetical outcome. Values closer to 1 indicate more agreement.

modification of my initial sampling procedure are laid out in the Methodology section “Varying the number of items per questionnaire”. Again, I observed feature attribution disagreement and faithfulness as defined in the Methodology section “Measurement of feature attribution method reliability”.

Balancing the number of items per questionnaire has, across prediction models (XGBoost and random forest models) and select metrics (top 5 signed rank agreement, top 10 rank and signed rank agreement), a small positive effect on feature attribution method agreement. This is illustrated in **Fig. 5.7**. There are also instances, in which this does not hold, for example for top 10 sign agreement and an XGBoost model.

I next report faithfulness of LIME and Shapley value based feature attributions to underlying elastic net models built on simulated data in which I varied imbalance of the number of items per questionnaire.

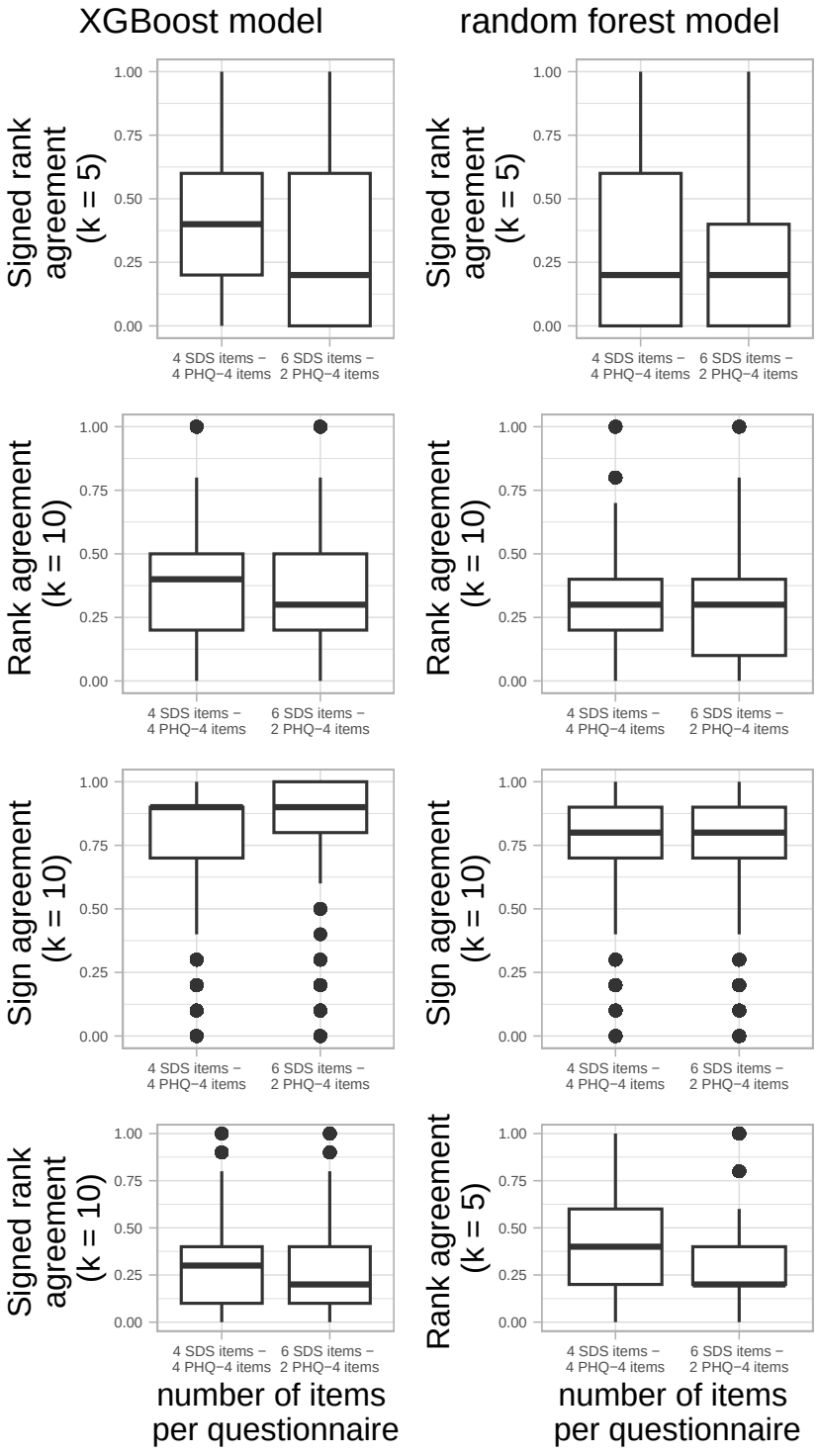

Figure 5.7. Distributions of disagreement between LIME and Shapley values for predictions of an XG-Boost model (left) and a random forest model (right) built on simulated data. Data are characterised by varying levels of imbalance of the number of items per questionnaire (x axis). The model predicted a simulated hypothetical outcome. Values closer to 1 indicate more agreement.

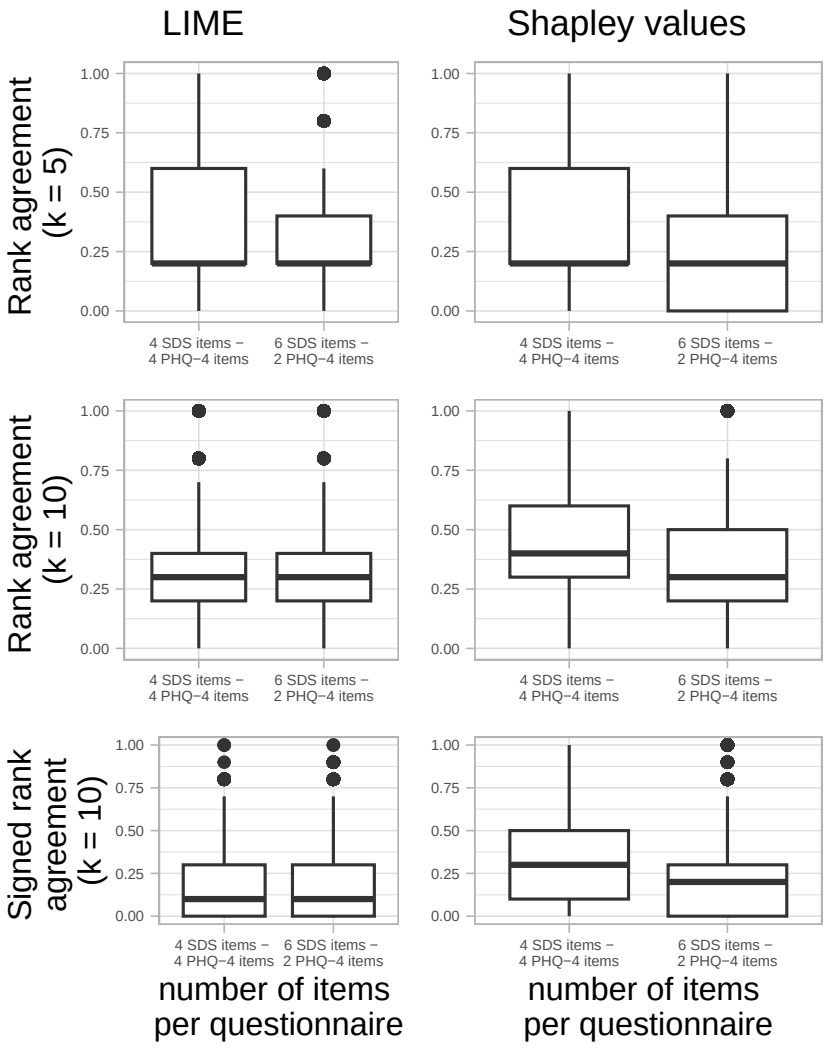

Figure 5.8. Distributions of disagreement between LIME (left) / Shapley values (right) with coefficients of an elastic net built on simulated data. Data are characterised by varying levels of imbalance of the number of items per questionnaire (x axis). The model predicted a simulated hypothetical outcome. Values closer to 1 indicate more agreement.

Across these feature attribution methods and for some metrics (top 5 and 10 rank agreement, top 10 signed rank agreement), balance of the number of items per questionnaire has a small positive effect on faithfulness of feature attribution methods to coefficients of an elastic net. This seems to be especially true for the Shapley value. I visualise this in **Fig. 5.8**.

Experiments varying the number of non-predictive features

Additionally, I have also varied the number of non-predictive features in the feature set, and observed the effect on feature attribution method disagreement and faithfulness. I have experimented with adding 0, 2 (the original u_1 and u_2), 5, and 10 non-predictive features to the feature set. Details of this modification of my initial sampling procedure are laid out in the Methodology section “Varying the number of non-predictive features”.

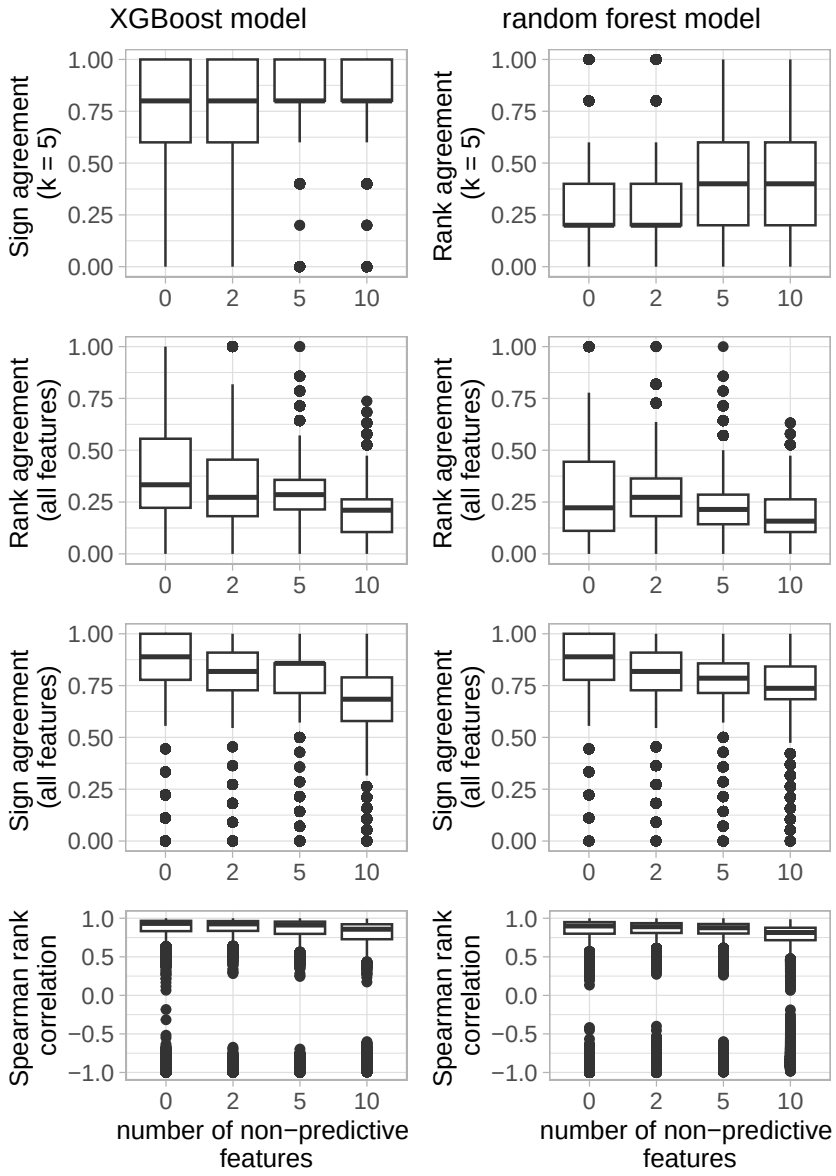

Figure 5.9. Distributions of disagreement between LIME and Shapley values for predictions of an XGBoost model (left) and a random forest model (right) built on simulated data. Data are characterised by varying numbers of non-predictive features added to the feature set (x axis). The model predicted a simulated hypothetical outcome. Values closer to 1 indicate more agreement.

Adding non-predictive features to the feature set had comparable effects on disagreement between feature attribution methods for an XGBoost and random forest model. It seemed to have a positive effect on feature attribution method agreement on select metrics when only the top 5 features are considered. When taking all features into account, however, I observe a negative effect across metrics. I illustrate this in **Fig. 5.9**.

Effects of including non-predictive features into the feature set on faithfulness of feature attribution methods to the underlying elastic net model are less clear. Including more non-predictive features seems to have a negative effect on feature attribution faithfulness of LIME to elastic net model

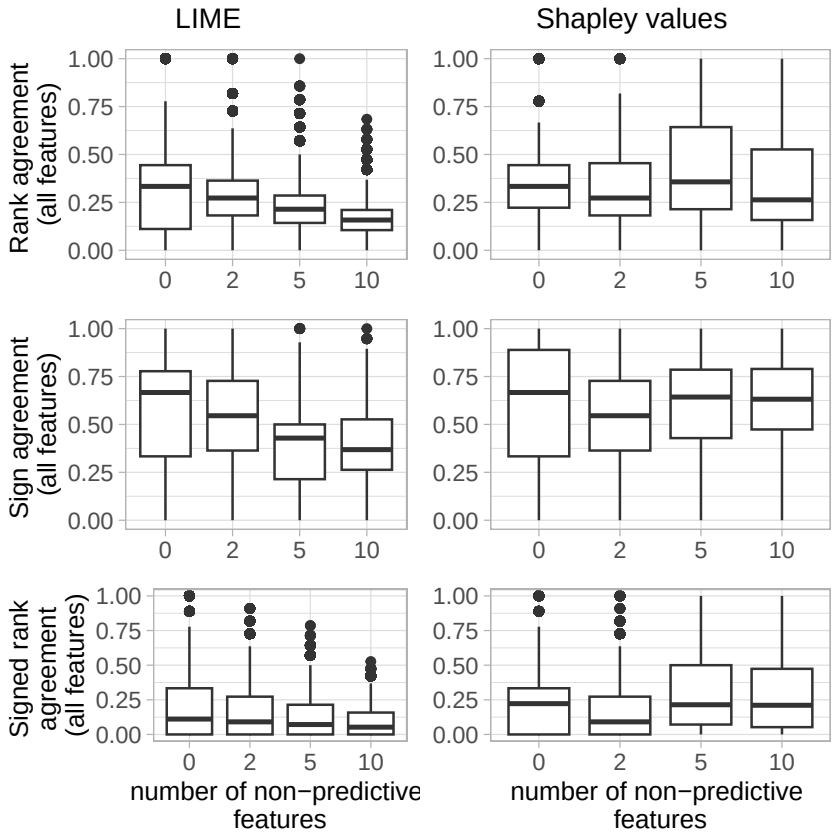

Figure 5.10. Distributions of disagreement between LIME (left) / Shapley values (right) with coefficients of an elastic net built on simulated data. Data are characterised by varying numbers of non-predictive features added to the feature set (x axis). The model predicted a simulated hypothetical outcome. Values closer to 1 indicate more agreement.

Table 5.1. Average F1 score over 15 repetitions of the simulation variation, varying the strength of correlations between items from the same questionnaire. Numbers in round brackets are standard deviations.

r	XGBoost	random forest	elastic net
0.5	0.906 (0.005)	0.903 (0.005)	0.906 (0.004)
0.6	0.898 (0.005)	0.896 (0.005)	0.898 (0.005)
0.7	0.892 (0.005)	0.892 (0.005)	0.893 (0.005)
0.8	0.889 (0.006)	0.890 (0.006)	0.891 (0.006)
0.9	0.889 (0.005)	0.890 (0.005)	0.891 (0.005)

coefficients on select metrics, for example max k rank agreement, sign agreement and signed rank agreement. This is illustrated in **Fig. 5.10**.

Note that I found no marked differences in accuracy in terms of the F1 score for the models I built on simulated data, as reported in **Table 5.1**, **Table 5.2** and **Table 5.3**.

Table 5.2. Average F1 score over 15 repetitions of the simulation variation, varying the level of imbalance of the number of items per questionnaire. Numbers in round brackets are standard deviations.

imbalance	XGBoost	random forest	elastic net
4 SDS items - 4 PHQ-4 items	0.890 (0.005)	0.890 (0.005)	0.891 (0.005)
6 SDS items - 2 PHQ-4 items	0.895 (0.005)	0.895 (0.005)	0.895 (0.005)

Table 5.3. Average F1 score over 15 repetitions of the simulation variation, varying the number of non-predictive features added to the dataset. Numbers in round brackets are standard deviations.

non-predictive features	XGBoost	random forest	elastic net
0	0.893 (0.004)	0.892 (0.004)	0.894 (0.004)
2	0.892 (0.005)	0.892 (0.005)	0.893 (0.005)
5	0.891 (0.004)	0.892 (0.004)	0.892 (0.004)
10	0.892 (0.005)	0.893 (0.005)	0.894 (0.005)

5.6.2 Illustration of feature attribution method reliability in semi-natural data

Further, I aim to illustrate LIME and Shapley value based feature attribution reliability in semi-natural data, generated using an empirical correlation matrix to describe feature set correlations in otherwise synthetic data. I focus on disagreement of these feature attribution methods with each other.

The average F1 score I obtained over 15 random forest models built on these semi-natural datasets, was F1 = 0.898 (SD: 0.003).

As shown in **Table 5.4**, I observe agreement amongst feature attribution methods to be relatively high for predictions on my semi-natural dataset, but only compared to levels of agreement obtained in fully natural datasets (Krishna et al. 2022; Markus, Fridgeirsson, et al. 2023). An explanation for this may be the small number of features used to build the prediction model. Agreement is, however, generally still much lower than what would be desired to confidently rely on a particular attribution.

Table 5.4. Feature attribution method disagreement between LIME and Shapley values for a random forest model built with semi-natural data. Modelled data was simulated using the correlation matrix obtained from a real-world BFO dataset. The model predicted a simulated hypothetical outcome. Values closer to 1 indicate more agreement.

Agreement metric	Value
Rank agreement (k = 5)	0.32
Sign agreement (k = 5)	0.75
Signed rank agreement (k = 5)	0.25
Rank agreement (k = 11)	0.27
Sign agreement (k = 11)	0.75
Signed rank agreement (k = 11)	0.19
Spearman rank correlation	0.64
Pairwise rank agreement	0.83

Table 5.5. Feature attribution method disagreement between LIME and Shapley values for a random forest model built with natural user data from BFO. The predicted outcome was post-engagement abstinence. Values closer to 1 indicate more agreement.

Agreement metric	Value
Rank agreement (k = 5)	0.36
Sign agreement (k = 5)	0.35
Signed rank agreement (k = 5)	0.19
Rank agreement (k = 30)	0.09
Sign agreement (k = 30)	0.44
Signed rank agreement (k = 30)	0.05
Spearman rank correlation	0.17
Pairwise rank agreement	0.71

5.6.3 Illustration of feature attribution method disagreement in natural data

Finally, I illustrate the reliability of LIME and Shapley value based feature attributions for predictions made with models built on natural data from users of the digital substance dependence intervention BFO. Specifically, a random forest model was built to predict abstinence from BFO users' primary substance after having engaged with the intervention. Predictions were made with baseline data available before access to the intervention's CBT-based content. I report disagreement of LIME and Shapley value based feature attribution.

Prediction of post-engagement abstinence with a random forest was possible with good performance (accuracy = 0.745, AUC = 0.811, F1 score = 0.706, precision = 0.771, recall = 0.651).

As reported in **Table 5.5**, I observe lower agreement rates between LIME and Shapley value based feature attribution methods than Krishna et al. (2022) on some metrics (sign agreement, Spearman rank correlation). Their experimental results for a random forest model on the German Credit dataset (Hofmann 1994) with 20 features and $k \in \{7, 20\}$ are the results from other research teams that are most comparable to ours. However, lower agreement is expected in datasets with a larger number of features (Krishna et al. 2022; Markus, Fridgeirsson, et al. 2023), and my feature set was almost 3.5 times the size of the one Krishna et al. (2022) used.

To illustrate what disagreement between LIME and Shapley value based feature attribution may look like for a prediction of post-engagement abstinence on an individual datapoint, I provide two examples of disagreement for the predictions for two BFO users in **Fig. 5.11** and **Fig. 5.12**. Shown are the five responses to questionnaire items that users had provided at baseline and that are ranked highest as per their LIME or Shapley value, indicating greatest importance of these responses for the respective predictions.

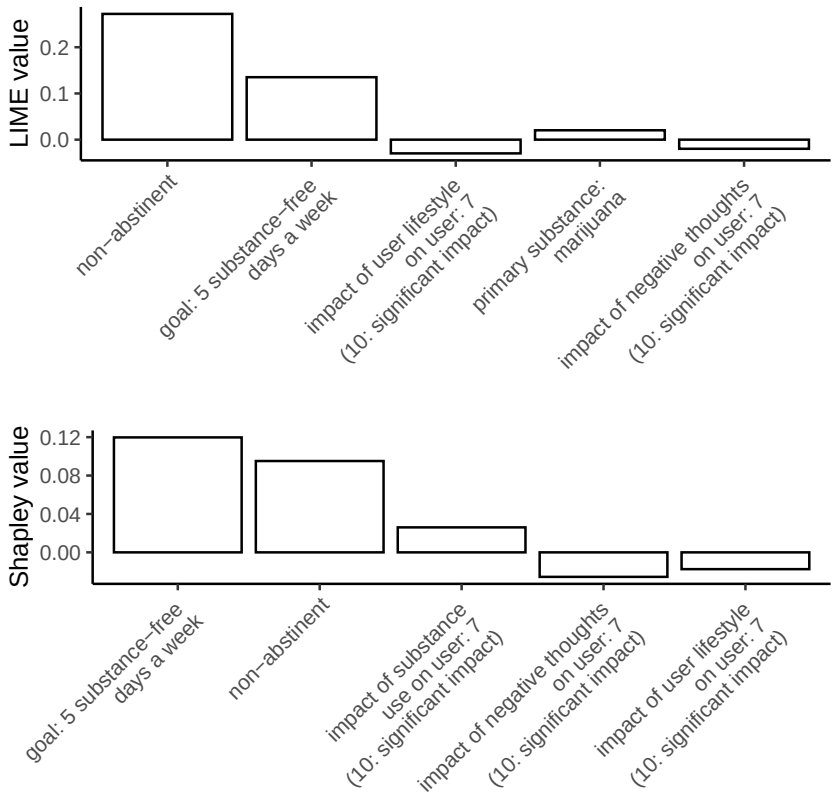

Figure 5.11. Illustration of disagreement on LIME and Shapley value based feature attributions for a BFO user predicted and observed to be non-abstinent post-engagement. The x axis shows the top 5 feature values specified by this BFO at the baseline assessment pre-engagement, ordered by importance.

The random forest model correctly predicts non-abstinence post-engagement for the feature attributions for the BFO user in **Fig. 5.11**. I observe that for the feature values which are ranked both into the most important five ones by LIME and Shapley value based feature attribution, signs are the same. Sign agreement is therefore expected to be reasonable for this data point. In contrast, rank agreement and pairwise rank agreement are expected to be suboptimal because the rank order as well as the order of pairs of feature values is different. Further, while the fact that this BFO user's primary substance is marijuana is included in the LIME explanation, there is no reference to this fact in the Shapley value explanation.

A second example where disagreement between LIME and Shapley value based feature value attributions is particularly strong is presented in **Fig. 5.12**: The model correctly predicts abstinence for the user whose baseline responses are ranked in importance by LIME and Shapley values. While the same feature values are included in both explanations (though in different rank orders), their signs are opposite of each other, which makes the decision process of the model behind this prediction for this BFO user (post-engagement abstinence) very intransparent.

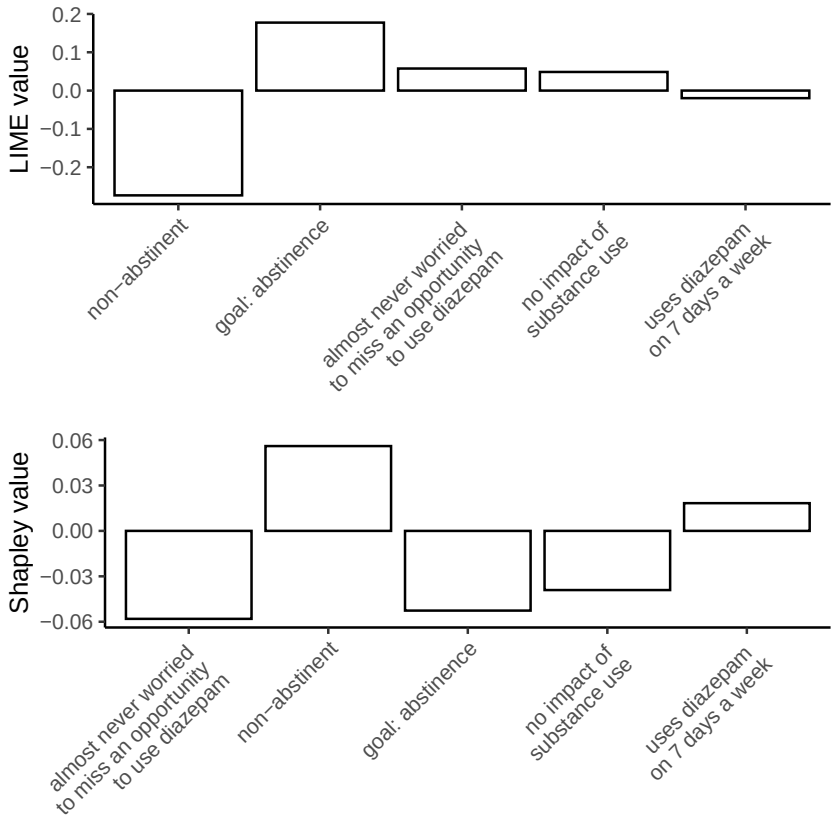

Figure 5.12. Illustration of disagreement on LIME and Shapley value based feature attributions for a BFO user predicted and observed to be abstinent post-engagement. The x axis shows the top 5 feature values specified by this BFO at the baseline assessment pre-engagement, ordered by importance.

5.6.4 Discussion

To explore the effect of multicollinearity on local feature attribution method reliability for mental health research, I have built prediction models on simulated questionnaire data with varying degrees of multicollinearity present, and observed reliability of prediction explanations in terms of explanation method disagreement as inspired by Krishna et al. (2022) and faithfulness to the underlying model. My concept of multicollinearity was inspired by datasets mainly consisting of categorical (questionnaire) data.

I found small effects of multicollinearity on feature attribution method disagreement on select metrics. The direction of the effects depended on the explained prediction model class, that is, whether an XGBoost, or a random forest was explained. Effects are likely more pronounced in datasets with larger feature sets, therefore, the effects found in my very small datasets likely represent a lower limit. Research building on this study would be advised to increase the number of simulated questionnaires or the number of items within them, and examine whether the effect of the total number of features on feature attribution reliability interacts with the effect of the correlations these features have with each

other. It remains to be shown whether the directionality of the trend we observed will stay the same as the size of the feature set grows.

Effects of the degree of feature set multicollinearity on feature attribution faithfulness are even less clear, and, in fact, hardly seem present at all in my experiments. This suggests that feature attribution methods explain simple regression-based models well, regardless of the degree of multicollinearity in the data. Related research finds small effects of multicollinearity on global feature attribution method disagreement, but is not able to disentangle it from the effect of the feature set size (Markus, Fridgeirsson, et al. 2023).

Given that my findings with small simulated datasets rather represent trends than large effects, it is also possible that multicollinarity may not affect perturbation-based feature attribution methods like LIME and Shapley values much at all in datasets whose relatively simple joint distribution of mostly ordinal variables may not leave much opportunity to generate wildly unrealistic data points as in data with mixed data types and continuous, non-normal distributions.

In comparison, the effects of the relative numbers of items per included questionnaire on feature attribution method disagreement and faithfulness seemed more homogeneous across prediction models and metrics than the effect of feature set multicollinearity, favouring balance of the number of items per questionnaire. Concerning faithfulness, Shapley value representation of a simple regression based model seem to be impacted by imbalance of the number of items in the included questionnaires more than LIME. I speculate that this may be a result of the symmetry axiom of the Shapley value, stipulating that features which contribute equally to a prediction should be attributed the same Shapley value. In line with this, irregularities of global Shapley values when differently sized groups of features which are correlated only with their own group members are included into feature sets have been observed by Catav et al. (2021) .

In summary, my experiments with the relative numbers of items per questionnaire suggest that including items from questionnaires which are differently sized may have a detrimental effect on feature attribution method reliability. As there is a real possibility for this to happen in mental health outcomes research, this finding is particularly relevant for researchers, because it is an argument against relying much on these feature attribution methods when the number of items per questionnaire are very different, or, more generally, when the employed feature set has a block correlation structure with differently sized blocks.

Further, I compared the effect of feature set multicollinearity to the effect of the number of non-predictive features in the feature set in order to be better able to gauge its impact compared to the impact of other dataset characteristics. With regards to feature attribution method disagreement, the effect of

the number of non-predictive features in the feature set seems more homogeneous across prediction models, and seems to depend on whether disagreement of some top features or the disagreement of all features is of interest. My results suggest that, if all features are concerned, the effect on disagreement is negative.

With regards to feature attribution method faithfulness to the coefficients of an elastic net model when an increasing number of non-predictive features are included in the feature set, I could only identify an effect on LIME feature attributions. Specifically, this effect seems to be negative.

Importantly, since I increase the total number of features while adding non-predictive features, the effect I observe cannot completely be disentangled from the negative effect of an increasing number of features on feature attribution disagreement that has recently been suggested by Markus, Fridgeirsson, et al. (2023) and seems to empirically be present in Krishna et al. (2022). I suggest future studies to attempt to isolate these both effects. Holding the number of features constant while converting some informative features to non-informative features may not be a viable way to approach this, since predictive performance may lessen with the proportion of uninformative features in a feature set, and the effect of predictive performance and the number of uninformative features in the dataset could not be isolated from each other, either.

In general, the extent of feature attribution method reliability in my simulation experiments with a small number of features seems unsatisfactory considering ML engineers' and researchers' reliance on them in practice. Agreement between methods, and faithfulness to a simple model seem to depend to a certain degree on each of a multitude of factors, amongst them are: the correlation structure of the data (to a greater extent the size of groups of correlated features, to a lesser extent the magnitude of correlation, as suggested by our results), the proportion of features from the feature set tested for reliability, the prediction model, the number of non-predictive predictors included in the feature set, the feature attribution method, and how strictly agreement or faithfulness is defined (for example whether only ranks are of interest, or whether sign and rank are expected to agree between feature attributions).

This is in agreement with findings from Krishna et al. (ibid.) and Markus, Fridgeirsson, et al. (2023). In comparison to these two previous studies which examined the disagreement problem in natural and semi-natural data, the impact which model performance (which I found to be comparable across manipulations) can have on feature attribution method reliability in my simulation experiments is limited. This strengthens the common observation of feature attribution method unreliability that I share with these studies.

One limitation of the study presented in this chapter is that our simulation is only one of many possible ones. The existence of a psychopathology factor underlying responses to questionnaire items which

my sampling model assumes, for example, is not uncontroversial (Stein et al. 2022). Future research may also consider that questionnaires may differ in their internal consistency, which means that items from one questionnaire may be more strongly correlated with each other than items from a different questionnaire. Further, not all mental health outcome prediction models include questionnaire items into their feature sets, but use aggregate measures, which may lessen the influence of a block correlation structure, and eventually multicollinearity in the dataset. I argue that the inclusion of individual questionnaire items is important as they often represent symptoms, and including these allows prediction algorithms to learn transdiagnostic predictive patterns, which are highly prevalent in mental illness and increasingly emphasised in research and clinical practice (Eaton et al. 2023).

I extended my analysis of feature attribution method reliability to semi-natural and natural data, and find disagreement of feature attribution methods comparable to those in Krishna et al. (2022) and Markus, Fridgeirsson, et al. (2023), and in expected magnitude given the number of features I use. Krishna et al. (2022), for example, make predictions with models from different families of supervised ML prediction models on two real-world tabular datasets unrelated to health. Their results on these datasets for a random forest model offer themselves for comparison to a certain extent, however, it is important to bear in mind that their datasets contain only a fraction of the features used to construct my prediction model on real-world BFO data, and therefore, feature attribution method agreement obtained for my prediction model is expected to be lower.

To advance this line of research, future studies may consider to investigate feature attribution method reliability including other feature attribution methods. Especially worthwhile may be the quantification of disagreement between unconditional and conditional Shapley value based feature attribution methods. With the terms “unconditional” and “conditional”, I refer to these methods’ estimation of the distribution of feature values not in a to-be-evaluated subset not conditioning, and conditioning on the feature values in the subset. Conditioning on the feature values in the subset was proposed to counteract the production of unrealistic data points while estimating Shapley values (Aas, Jullum, and Løland 2021).

Recent implementations of such conditional Shapley values have accounted for increasingly complex and more realistic data distributions, and have improved in terms of estimation time (Olsen et al. 2023). Markus, Fridgeirsson, et al. (2023), for example, have found considerable disagreement between unconditional and conditional global Shapley values. Future research may explore whether this extends to local conditional and unconditional Shapley values, and what datapoints for which these methods disagree, may have in common.

Some concluding remarks in terms of feature attribution for ML models in mental healthcare outcomes research are warranted. The reliability of feature attribution (and similarly, feature selection) methods

can be crucial for the evaluation of the clinical usefulness, safety, and fairness of an ML prediction model of mental health outcomes, and the relevance of research building on feature importance analyses. The study presented in this chapter supports accumulating evidence that naive reliance on feature attribution methods in practice may lead to incorrect conclusions about feature importance. Given that many outcome prediction models in mental healthcare may not yet reach levels of performance comparable to those achieved in this study, and prediction accuracy may further impair feature attribution reliability, the warning against naive use becomes even more pertinent.

To note, generally, the disagreement problem may arise from the different computation, and consequently, concept that LIME and Shapley value based feature attributions have of feature importance. Therefore, the possibility exists that feature attributions from either method are reliable in the sense that each reflects a different ground truth about feature importance. However, as long as feature importance stays an ambiguous term, researchers cannot easily determine which method's attributions to believe in, and which method to employ for which dataset and which concept of feature importance. For a naive use of these methods, reliance on them seems unjustified.

This last point is also stressed by a recent study which has suggested that the mere provision of local explanations of ML models may not necessarily support, and may even harm the responsible real-world deployment of ML models and negatively affect clinical practice by promoting over-reliance on model suggestions (Jacobs et al. 2021). A thorough understanding of the empirical and theoretical properties of ML model explainers is therefore needed ahead of use of these explainers, which I hope to have contributed to in this study.

5.7 Conclusion

Feature attribution methods are widely used to explain ML prediction models, increasingly also in mental health outcomes research. My study on the influence of characteristics frequently found in datasets used in mental health, especially multicollinearity, on the reliability of these methods calls the naive use of these methods into question.

Chapter 6

Conclusion

As highlighted in the introduction to this thesis, real-world data has great potential to contribute to our understanding about mental health and illness as well as potential treatment. However, their routine collection also comes with downsides, among them the difficulty to establish causality, the extra effort that needs to be spent on data preparation, missing data, context dependence and related difficulties with reproducibility and replicability, and data reflecting rather than contributing to a clinician's knowledge (Adler et al. 2020; Beaulieu-Jones et al. 2021; Chekroud et al. 2021; Chikersal et al. 2020; Ewbank et al. 2021; Faurholt-Jepsen et al. 2016; Faurholt□Jepsen et al. 2015; Hoogendoorn et al. 2017; Liu and Panagiotakos 2022). The latter results in algorithms not being able to provide knowledge beyond what clinicians already know, or not being able to provide this knowledge earlier than clinicians can provide it. The potential of real-world data to increase inclusivity in mental health research by collecting data from patient populations which are underrepresented in clinical studies, as discussed in Chapter 2, also depends on the inclusivity of the services which routinely collect these data (Liu and Panagiotakos 2022).

In this thesis, I have used real-world data from the digital substance dependence intervention BFO to explore several research questions of interest to the digital mental health research community. As digital solutions are increasingly considered as augmentations of face-to-face mental healthcare, research in this area is important to ensure that interventions are useful for clinical practice.

Conducting research on this particular intervention was of particular interest because little research exists on interventions targeting high acuity populations, and substance use disorder. People registering with BFO exhibit severe impairments in terms of substance abuse, and significant levels of comorbid anxiety and depression. This is not surprising given the recruitment of BFO registrants through face-to-face addiction services in the UK: These services may serve as a first point of contact when substance use is as severe and debilitating for the individual and its surroundings that denial of impairment, or

delusions of control over substance use cannot be maintained anymore, or referral from general practitioners or the criminal justice system takes place.

Specifically, I have achieved the following: Firstly, I have described the heterogeneity of the real-world user base of BFO, and established some associations between BFO user / delivery characteristics, and BFO outcomes in initially available data. Secondly, I have conducted a study suggesting that clinical assessment data available before DI engagement may not contain enough information to predict behavioural engagement. Thirdly, I have found some evidence for unreliability of popular feature attribution methods, used to explain ML prediction models which are hoped to be deployed to clinical benefit in (digital) mental healthcare practice in terms of these methods, being affected by the structure and type of mental health data. (With “structure”, I refer to questionnaires of different size being included in prediction feature sets. With “type’’, I refer to commonly small correlations between features and outcomes in psychology and psychiatry research which introduces a real possibility that features nearly unrelated to the outcome are included in the feature set.) Found unreliability makes it less likely that features have been attributed importance correctly, and that human oversight over ML models is feasible. I have made these findings in one of the largest databases of its kind, coming from a widely available mental health DI.

Similar to many studies conducted with real-world data, and many studies conducted in the digital mental health domain, my studies, especially those conducted in Chapter 3 and Chapter 4 were complicated by the challenge of defining clinically meaningful study endpoints. In my case, the concept of user benefit from a DI could only be approximated with the datasets available to me.

In fact, we do not know much about what beneficial use of a digital intervention looks like to date, and how it may be different from the use of social media and commercial apps. This results in uncertainty about when an assessment of mental health or recovery progression should be scheduled, whether it should be for example based on time passed, or an amount of interaction that has taken place with the DI. Uncertainty also prevails with regards to what the relative validity of patient self-reported benefit, clinician-assessed benefit, and objective engagement metrics would be for evaluating the benefit of a person from a DI. When evaluating benefit from a DI using real-world data, I suggest - at a minimum - to jointly consider longitudinal self-report and behavioural engagement data, for example module completions, and to eventually choose a study endpoint respecting their dependence on each other. Often unavailable to researchers using real-world data from DIs is also data on user motives to continue to engage, disengage, or re-engage. I suggest future research directions to address this problem further along in this chapter.

Our finding in Chapter 5 comments on the safe deployment of ML prediction models in mental healthcare which explainability, if reliable, can be an important tool for. Sustainable deployment of these models in practice is rare to date (Chekroud et al. 2021). Factors contributing to this may be ML prediction models lacking in actionable levels of accuracy, and unreliability of explanations (as suggested by our findings, too) which could make model oversight, and subsequent model improvement through ML engineers, assisted by clinicians, possible.

To note, machine learning is not mandatory for making accurate predictions. Regression-based models may yield similar model performance at times, and are arguably more explainable (Perlis 2013; Shortreed et al. 2023). The clinical value of prediction models for mental healthcare more generally has also been called into question: Model predictions may be accurate, but clinicians may not be able to account for these predictions in their clinical practice, for example due to them only being trained to offer specific treatments and not others which may be recommended by prediction models. Similarly, a recent study on the implementation of an algorithm predicting antidepressant response reported that clinicians most common reaction to predicted non-response was dosage increase (Browning et al. 2021). Other barriers to prediction models' clinical value are self-fulfilling prophecies in which predictions influence treatment delivery, and the under-development of independent clinical judgement and empathy in practitioners (Chekroud et al. 2021).

Further, it remains to be shown what the advent of large language models such as the ones used in the conversational AI ChatGPT, mean for the digital delivery of mental health content, and for the analysis of real-world data accruing from it (Abrams 2023; Lyubomirsky 2023). Large language models are seen by some as a possibility to improve sub-threshold mental health problems such as stress in day-to-day life or low mood due to a chronic illness. Conversational agents which are powered by large language models may, through conversation, ideally encourage self-reflection and empathy, teach strategies for happiness, communication, and social interaction in a hands-on way, or point cognitive distortions out (Lyubomirsky 2023). However, serious concerns have also been raised about human interaction with such agents effecting decreased authenticity or investment into real human connection and relationships. Also, practical reports have emerged in which they behaved in a psychologically harmful way towards humans (Roose 2023).

In the following, I take the opportunity to outline some directions for future research building on the work presented in this thesis.

6.1 Future Work

Even though the analysis of real-world data from BFO has provided me with some interesting insights into digital substance dependence intervention engagement and its users, these data also have limits which I have outlined in the beginning of this chapter. This has led me to believe that research teams should conduct studies using real-world data and more traditional studies using purposefully collected data alternatingly or, at specific times during a process of evidence generation and discovery. Conducting exploratory, data-driven research on BFO made sense at the current time point since little evidence is yet generated on DIs for substance dependence, and few hypotheses about mechanisms of change, outcomes, and engagement have been proposed.

For future work, I propose, however, studies with more control exerted over the data collection process. They are tailored to the BFO programme, however, similar studies may be conducted for other DIs. My propositions respond to the observation that "necessary data around engagement, effect sizes, necessary dose, and duration of effect remain unknown for almost all digital health technologies" (Torous, Bucci, et al., 2021, p. 328).

A particular challenge for digital mental health interventions is the high disengagement rates, and for researchers, the definition of user benefit. Since user engagement as reported in RCTs does not seem to translate readily into real-world engagement rates (Baumel, Edan, and Kane 2019), the conclusions that can be drawn from these studies are limited. With regards to user engagement studied with real-world data, in turn, there is uncertainty about the quantity and quality of user interactions equating to user benefit, and about what healthy engagement with digital technology looks like, in the first place. I propose therefore to invest resources into characterising beneficial engagement, and the many outlooks it may take on, before more inference is drawn from routinely collected data. Evidence from such studies may then inform analysis plans for future real-world data studies and controlled trials.

I propose a study (1), which, depending on its findings, may be followed by a study (2). For study (1), I propose to conduct semi-structured interviews with a diverse sample of people at a treatment service offering BFO. In the following I outline a potential procedure for study (1). A potential study (2) is outlined afterwards.

6.1.1 Proposed study 1

This study would consist of semi-structured interviews with people at a treatment service commissioning BFO. The clientele of this treatment service would, before interviews start, be examined closely in

terms of (a) its sociodemographics, and (b) its substance use severity. This includes for (a) the examination of age, ethnicity, education, social class, and income distributions. For (b), this would involve describing prevalence of poly-drug use and class A drug use (such as crack, cocaine and heroin), and distributions of age at first SUD episode for people seeking help at this service. This would inform the person-specific questions asked in the interview, which may later be used to quantify the association of sociodemographics with benefit.

The most important inclusion criterion of this interview study would be that a participant had been offered BFO at treatment entry, and is willing to talk about their experience. This experience may include continued use, brief use, intermittent use, and refusal of the offer to register. Monetary compensation for participation would likely be imperative. Recruitment would likely have to continue until a reliable number of service users who report benefit from BFO had been interviewed, and overrepresentation of them in the sample would be desirable.

The recruitment pool of this study could also be enlarged by asking clinicians at the treatment service to identify patients who benefitted from using BFO in any way. Clinicians would also be asked to elaborate on how they identified that these patients derived benefit from using BFO. These patients could subsequently be contacted for semi-structured interviews as they are described in the following.

Recruited study participants, in the following called interviewees, would first be asked whether they accepted the offer from staff to use BFO. If they did not, they would be asked to elaborate on their motivation to decline. Prompts may be given to understand what role the following factors play: a lack of understanding what a digital intervention entails, insecurity about own digital literacy, uncertainty about how the use of a DI would affect their access to face-to-face care, technology ownership, and, if an interviewee is not in possession of any device that allow BFO access, travel time to a treatment service or public library.

Interviewees who successfully created an account with BFO would be asked instead about the date they created this account, and about the last date they remembered using BFO. Then they would be asked to tell their story of using it, and if they did discontinue and show little motivation to re-engage with BFO, why they did so. Possible reasons for discontinuation could be arising technical problems, subjective need for face-to-face care, stress and time constraints (interviewers may want to clarify whether these were effected by increased responsibilities in a person's job, or at home, or a mental health crisis, or both), lost hope in recovery after a relapse, or inadequate content for individual difficulties. The contribution of these factors could potentially be discussed with the participant.

Interviewees who successfully created an account with BFO would subsequently be asked whether they benefited from BFO in any way (even if they do not actively use it now). For better analysis, an-

swers would ultimately be recorded in a binary fashion. Nevertheless, interviewees would be asked to elaborate, for example by asking them to detail why they answered Yes or No. Reasons for benefit could be connected to the acquisition of digital literacy, the achievement of engaging with CBT-based content, an improved understanding of substance use triggers, the achievement of sticking to a routine, adaptability to lapse and relapse, mood improvements, establishment of self-care practices, motivation and hope, and resistance to urges. Interviewees should be assisted in identifying which are true for them.

Unless an interviewee's pattern of use can already be described well through their responses to previous questions, interviewees who successfully created an account with BFO and reported benefiting from it would subsequently be asked to describe their overall pattern of and motivation for use. They would also be asked whether they would characterise it as regular (and at which intervals it was regular), or sporadic (and what were motivating factors to use BFO in those moments), and whether there was ever a transition between regular and sporadic use.

In a thematic analysis of interviewees answers, similarities and differences between interviewees who reported benefit from BFO would ideally be established. Similarities may lead to clusters of "beneficiaries" who engaged in a similar fashion, and benefited in similar ways. The percentage of interviewees reporting benefit would be of interest, and the identification of factors driving benefit and discontinuation would be the objective.

6.1.2 Potential study 2

With evidence from study (1), a study (2) could potentially be conducted that may take the form of a controlled trial. Note that similar studies, specifically, RCTs evaluating BFO with varying degrees of human support against standard treatment, are already underway at several treatment services (Elison-Davies, Davies, et al. 2018; Elison-Davies, Pittard, et al. 2023; Quilty et al. 2022). Therefore, study (2) represents a new idea for a controlled trial, building specifically on research conducted for this thesis.

Endpoints of study (2) should be informed by the knowledge gained in study (1) about user benefit. A draft outline of this trial could involve users of two different treatment services matched by the substance use and sociodemographic profile of its clientele. New service users who are offered BFO would also be asked if they would take part in a controlled trial. This would entail their app use being recorded (timestamped assessments and module completions), and coming in for on-site visits for toxicology screens, a clinical assessment of their SUD, and (in the case of those having access to BFO) binary recorded self-reported benefit from BFO at 3, 6, and 12 months after this first service visit. Findings

from such a study may allow a better understanding of differences in use patterns between those with access to BFO who benefit from it, and those who do not benefit. It may also allow to identify differences in recovery from SUD between control participants, and people with access to BFO who benefit, and those who do not benefit.

6.2 Closing remarks

In conclusion, our use of real-world data from the digital substance dependence intervention BFO as a data source to generate evidence about BFO users and outcomes led me to believe that real-world data studies show promise to complement controlled studies in mental health intervention research. Conducted at the right time point in a succession of studies, they may, when transparency about the data origin and data generation process exists, be used for patient benefit, especially in order to generate hypotheses and, citing a recent comment paper on methodological issues and opportunities of real-world evidence within the European Health Data Space, “ensure the trial’s relevance and representation of real-world clinical scenarios such as patient population or treatment patterns” (Kympouropoulos, 2023, p. 2). In closing, I emphasise the importance of exercising caution in choosing study endpoints and interpreting results.

References

Aas, Kjersti, Jullum, Martin, and Løland, Anders (2021). “Explaining individual predictions when features are dependent: More accurate approximations to Shapley values”. In: *Artificial Intelligence* 298, Article 103502. DOI: https://doi.org/10.1016/j.artint.2021.103502 (cited on pp. 86, 111).

Abi-Dargham, Anissa et al. (2023). “Candidate biomarkers in psychiatric disorders: State of the field”. In: *World Psychiatry* 22.2, pp. 236–262. DOI: https://doi.org/10.1002/wps.21078 (cited on p. 88).

Abrams, Zara (Jan. 2023). “AI is changing every aspect of psychology. Here’s what to watch for.” In: *Monitor on Psychology* 54.5. URL: https://www.apa.org/monitor/2023/07/psychology-embracing-ai (visited on 11/29/2023) (cited on pp. 79, 115).

Adler, Daniel A. et al. (2020). “Predicting early warning signs of psychotic relapse from passive sensing data: An approach using encoder-decoder neural networks”. In: *JMIR mHealth and uHealth* 8.8, Article e19962. DOI: 10.2196/19962 (cited on p. 113).

Agarwal, Chirag et al. (Apr. 2022). “Rethinking stability for attribution-based explanations [Conference presentation]”. In: *International Conference on Learning Representations* (cited on p. 85).

Alexander, Karen, Sanjuan, Pilar, and Terplan, Mishka (2023). “The use of ecological momentary assessment methods with people receiving medication for opioid use

disorder: A systematic review". In: *Current Addiction Reports* 10.3, pp. 366–377. DOI: https://doi.org/10.1007/s40429-023-00492-5 (cited on p. 18).

Ali, Mir M., Teich, Judith L., and Mutter, Ryan (2017). "Reasons for not seeking substance use disorder treatment: Variations by health insurance coverage". In: *The Journal of Behavioral Health Services & Research* 44, pp. 63–74. DOI: https://doi.org/10.1007/s11414-016-9538-3 (cited on p. 39).

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders (text revision)*. 4th ed. (cited on p. 17).

– (2013). *Diagnostic and statistical manual of mental disorders*. 5th ed. URL: https://doi.org/10.1176/appi.books.9780890425596 (cited on p. 16).

Anvari, Farid (Sept. 2023). *Non-interval scale use: Self-report ratings are not on an interval scale*. URL: https://twitter.com/farid_anvari/status/1703127702742384782 (visited on 11/03/2023) (cited on p. 45).

Apley, Daniel W. and Zhu, Jingyu (2020). "Visualizing the effects of predictor variables in black box supervised learning models". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82.4, pp. 1059–1086. DOI: https://doi.org/10.1111/rssb.12377 (cited on pp. 45, 87).

Argyriou, Evangelia et al. (2023). "Individual factors predict substance use treatment course patterns among patients in community-based substance use disorder treatment". In: *PLOS ONE* 18.1, Article e0280407. DOI: 10.1371/journal.pone.0280407 (cited on p. 36).

Bankiewicz, Urszula and Robinson, Chloe (2020). *Health Survey for England 2019*. Tech. rep. NHS Digital. URL: https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/2019 (cited on p. 37).

Bansal, Naman, Agarwal, Chirag, and Nguyen, Anh (June 2020). “SAM: The sensitivity of attribution methods to hyperparameters [Poster session]”. In: *Conference on Computer Vision and Pattern Recognition*. Seattle, Washington, USA (cited on p. 85).

Baumel, Amit, Edan, Stav, and Kane, John M. (2019). “Is there a trial bias impacting user engagement with unguided e-mental health interventions? A systematic comparison of published reports and real-world usage of the same programs”. In: *Translational Behavioral Medicine* 9.6, pp. 1020–1033. DOI: https://doi.org/10.1093/tbm/ibz147 (cited on pp. 39, 74, 116).

Baumel, Amit and Kane, John M. (2018). “Examining predictors of real-world user engagement with self-guided eHealth interventions: Analysis of mobile apps and websites using a novel dataset”. In: *Journal of Medical Internet Research* 20.12, Article e11491. DOI: 10.2196/11491 (cited on pp. 23, 68, 74).

Beaulieu-Jones, Brett K. et al. (2021). “Machine learning for patient risk stratification: Standing on, or looking over, the shoulders of clinicians?” In: *npj Digital Medicine* 4, Article 62. DOI: https://doi.org/10.1038/s41746-021-00426-3 (cited on p. 113).

Beck, A. T. et al. (1993). *Cognitive therapy of substance abuse*. The Guilford Press (cited on p. 29).

Bell, Lauren et al. (2020). “Engagement with a behavior change app for alcohol reduction: Data visualization for longitudinal observational study”. In: *Journal of Medical Internet Research* 22.12, Article e23369. DOI: 10.2196/23369 (cited on pp. 24, 39, 63, 64).

Ben-Zeev, Dror et al. (2017). “CrossCheck: Integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse”. In: *Psy-*

chiatric Rehabilitation Journal 40.3, pp. 266–275. DOI: https://doi.org/10.1037/prj0000243 (cited on p. 15).

Black, Carol (2020). *Review of drugs part one*. Tech. rep. Home Office. URL: https://assets.publishing.service.gov.uk/media/5f02e08ce90e075c5128f311/2SummaryPhaseOne+foreword200219.pdf (cited on pp. 20, 39).

– (2021). *Review of drugs part two: Prevention, treatment, and recovery*. Tech. rep. Home Office. URL: https://www.gov.uk/government/publications/review-of-drugs-phase-two-report/review-of-drugs-part-two-prevention-treatment-and-recovery (cited on p. 20).

Borisov, Vadim et al. (2021). "Deep neural networks and tabular data: A survey". In: *IEEE Transactions on Neural Networks and Learning Systems*. DOI: 10.1109/TNNLS.2022.3229161 (cited on pp. 74, 93).

Borsboom, Denny, Cramer, Angélique O. J., and Kalis, Annemarie (2019). "Brain disorders? Not really: Why network structures block reductionism in psychopathology research". In: *Behavioral and Brain Sciences* 42, Article e2. DOI: https://doi.org/10.1017/S0140525X17002266 (cited on p. 88).

Boumparis, Nikolaos et al. (2017). "Internet interventions for adult illicit substance users: A meta-analysis". In: *Addiction* 112.9, pp. 1521–1532. DOI: https://doi.org/10.1111/add.13819 (cited on pp. 20, 22, 23).

Bricker, Jonathan et al. (2023). "Can a single variable predict early dropout from digital health interventions? Comparison of predictive models from two large randomized trials". In: *Journal of Medical Internet Research* 25, Article e43629. DOI: 10.2196/43629 (cited on pp. 67, 69).

Browne, Teri et al. (2016). “Barriers and facilitators to substance use treatment in the rural south: A qualitative study”. In: *The Journal of Rural Health* 32.1, pp. 92–101. DOI: https://doi.org/10.1111/jrh.12129 (cited on p. 39).

Browning, Michael et al. (2021). “The clinical effectiveness of using a predictive algorithm to guide antidepressant treatment in primary care (PReDicT): An open-label, randomised controlled trial”. In: *Neuropsychopharmacology* 46, pp. 1307–1314. DOI: https://doi.org/10.1038/s41386-021-00981-z (cited on p. 115).

Burgess-Hull, Albert and Epstein, David H. (2021). “Ambulatory assessment methods to examine momentary state-based predictors of opioid use behaviors”. In: *Current Addiction Reports* 8.1, pp. 122–135. DOI: https://doi.org/10.1007/s40429-020-00351-7 (cited on p. 17).

Burton, Robyn et al. (2016). *The public health burden of alcohol and the effectiveness and cost-effectiveness of alcohol control policies. An evidence review*. Tech. rep. Public Health England. URL: https://assets.publishing.service.gov.uk/media/5b6c5703ed915d3119112af6/alcohol_public_health_burden_evidence_review_update_2018.pdf (cited on p. 20).

Campbell, Aimee N.C. et al. (2014). “Internet-delivered treatment for substance abuse: A multisite randomized controlled trial”. In: *American Journal of Psychiatry* 171.6, pp. 683–690. DOI: https://doi.org/10.1176/appi.ajp.2014.13081055 (cited on p. 23).

Carl, Jenna R. et al. (2020). “Efficacy of digital cognitive behavioral therapy for moderate-to-severe symptoms of generalized anxiety disorder: A randomized controlled trial”. In: *Depression and Anxiety* 37.12, pp. 1168–1178. DOI: https://doi.org/10.1002/da.23079 (cited on pp. 74, 87).

Catav, Amnon et al. (July 2021). “Marginal contribution feature importance - an axiomatic approach for explaining data [Poster session]”. In: *International Conference on Machine Learning* (cited on p. 109).

Chaple, Michael et al. (2014). “Feasibility of a computerized intervention for offenders with substance use disorders: A research note”. In: *Journal of Experimental Criminology* 10.1, pp. 105–127. DOI: https://doi.org/10.1007/s11292-013-9187-y (cited on p. 23).

Chapman, Sarah C. E. and Horne, Rob (2013). “Medication nonadherence and psychiatry”. In: *Current Opinion in Psychiatry* 26.5, pp. 446–452. DOI: 10.1097/YCO.0b013e3283642da4 (cited on p. 13).

Chekroud, Adam M. et al. (2021). “The promise of machine learning in predicting treatment outcomes in psychiatry”. In: *World Psychiatry* 20.2, pp. 154–170. DOI: https://doi.org/10.1002/wps.20882 (cited on pp. 16, 75, 78, 113, 115).

Chen, Ji et al. (2023). “Leveraging machine learning for gaining neurobiological and nosological insights in psychiatric research”. In: *Biological Psychiatry* 93.1, pp. 18–28. DOI: https://doi.org/10.1016/j.biopsych.2022.07.025 (cited on pp. 66, 82).

Chien, Isabel et al. (2020). “A machine learning approach to understanding patterns of engagement with internet-delivered mental health interventions”. In: *JAMA Network Open* 3.7, Article e2010791. DOI: 10.1001/jamanetworkopen.2020.10791 (cited on pp. 23, 24, 39, 64).

Chikersal, Prerna et al. (Apr. 2020). “Understanding client support strategies to improve clinical outcomes in an online mental health intervention [Conference presentation]”. In: *Conference on Human Factors in Computing Systems*. Honolulu, Hawaii, USA (cited on p. 113).

Corrigan-Curay, Jacqueline, Sacks, Leonard, and Woodcock, Janet (2018). “Real-world evidence and real-world data for evaluating drug safety and effectiveness”. In: *JAMA* 320.9, pp. 867–868. DOI: 10.1001/jama.2018.10136 (cited on p. 13).

Covert, Ian C., Lundberg, Scott, and Lee, Su-In (Aug. 2020). “Understanding global feature contributions with additive importance measures [Poster session]”. In: *International Conference on Neural Information Processing Systems* (cited on p. 85).

Cross, Shane P. et al. (2022). “Factors associated with treatment uptake, completion, and subsequent symptom improvement in a national digital mental health service”. In: *Internet Interventions* 27, Article 100506. DOI: https://doi.org/10.1016/j.invent.2022.100506 (cited on pp. 23, 75).

Davies, Glyn et al. (2015). “The role of lifestyle in perpetuating substance use disorder: The Lifestyle Balance Model”. In: *Substance Abuse Treatment, Prevention, and Policy* 10.1, Article 2. DOI: https://doi.org/10.1186/1747-597X-10-2 (cited on pp. 6, 28, 41, 74).

DeAngelis, Tori (Jan. 2021). “Can real-world data lead to better interventions?” In: *Monitor on Psychology* 52.6. URL: https://www.apa.org/monitor/2021/09/news-real-world-data (visited on 11/29/2023) (cited on pp. 13, 14).

Degenhardt, Louisa et al. (2017). “Estimating treatment coverage for people with substance use disorders: An analysis of data from the World Mental Health Surveys”. In: *World Psychiatry* 16.3, pp. 299–307. DOI: https://doi.org/10.1002/wps.20457 (cited on p. 39).

Delgadillo, Jaime, Rubel, Julian, and Barkham, Michael (2020). “Towards personalized allocation of patients to therapists”. In: *Journal of Consulting and Clinical Psychology* 88.9, pp. 799–808. DOI: https://doi.org/10.1037/ccp0000507 (cited on pp. 16, 66, 82).

Diehl, Alexander et al. (2007). “Alcoholism in women: Is it different in onset and outcome compared to men?” In: *European Archives of Psychiatry and Clinical Neuroscience* 257.6, pp. 344–351. DOI: https://doi.org/10.1007/s00406-007-0737-z (cited on p. 37).

Dimanov, Botty et al. (Aug. 2020). “You shouldn’t trust me: Learning models which conceal unfairness from multiple explanation methods [Conference presentation]”. In: *European Conference on Artificial Intelligence*. Santiago de Compostela, Spain (cited on p. 86).

Dugdale, Stephanie, Elison, Sarah, et al. (2016). “Using the transtheoretical model to explore the impact of peer mentoring on peer mentors’ own recovery from substance misuse”. In: *Journal of Groups in Addiction & Recovery* 11.3, pp. 166–181. DOI: https://doi.org/10.1080/1556035X.2016.1177769 (cited on pp. 27, 41).

– (2017). “A qualitative study investigating the continued adoption of Breaking Free Online accross a national substance misuse organisation: Theoretical conceptualisation of staff perceptions”. In: *The Journal of Behavioral Health Services & Research* 44, pp. 89–101. DOI: https://doi.org/10.1007/s11414-016-9512-0 (cited on pp. 27, 41).

Dugdale, Stephanie, Ward, Jonathan, et al. (2016). “Using the behavior change technique taxonomy v1 to conceptualize the clinical content of Breaking Free Online: A computer-assisted therapy program for substance use disorders”. In: *Substance Abuse Treatment, Prevention, and Policy* 11.1, Article 26. DOI: https://doi.org/10.1186/s13011-016-0069-y (cited on pp. 29, 42).

Eaton, Nicholas R. et al. (2023). “A review of approaches and models in psychopathology conceptualization research”. In: *Nature Reviews Psychology* 2, pp. 622–636. DOI: https://doi.org/10.1038/s44159-023-00218-4 (cited on pp. 66, 111).

Eiband, Malin et al. (Aug. 2019). “The impact of placebic explanations on trust in intelligent systems [Poster session]”. In: *Conference on Human Factors in Computing Systems*. Glasgow, UK (cited on p. 82).

Elison, Sarah, Davies, Glyn, and Ward, Jonathan (2015). “Effectiveness of computer-assisted therapy for substance dependence using Breaking Free Online: Subgroup analyses of a heterogeneous sample of service users”. In: *JMIR Mental Health* 2.2, Article e13. DOI: 10.2196/mental.4355 (cited on pp. 32, 33).

– (2016). “Initial development and psychometric properties of a new measure of substance use disorder “Recovery Progression”: The Recovery Progression Measure (RPM)”. In: *Substance Use & Misuse* 51.9, pp. 1195–1206. DOI: https://doi.org/10.3109/10826084.2016.1161052 (cited on pp. 6, 28, 43, 70).

Elison, Sarah, Dugdale, Stephanie, et al. (2017). “The Rapid Recovery Progression Measure: A brief assessment of biopsychosocial functioning during substance use disorder recovery”. In: *Substance Use & Misuse* 52.9, pp. 1154–1163. DOI: https://doi.org/10.1080/10826084.2017.1299183 (cited on pp. 28, 43).

Elison, Sarah, Humphreys, Lloyd, et al. (2014). “A pilot outcomes evaluation for computer assisted therapy for substance misuse – an evaluation of Breaking Free Online”. In: *Journal of Substance Use* 19.4, pp. 313–318. DOI: https://doi.org/10.3109/14659891.2013.804605 (cited on pp. 27, 41).

Elison, Sarah, Jones, Andrew, et al. (2017). “Examining effectiveness of tailorable computer-assisted therapy programmes for substance misuse: Programme usage and clinical outcomes data from Breaking Free Online”. In: *Addictive Behaviors* 74, pp. 140–147. DOI: https://doi.org/10.1016/j.addbeh.2017.05.028 (cited on pp. 28, 32–34).

Elison, Sarah, Ward, Jonathan, et al. (2017). “Feasibility of a UK community-based, eTherapy mental health service in Greater Manchester: Repeated-measures and between-groups study of ‘Living Life to the Full Interactive’, ‘Sleepio’ and ‘Breaking Free Online’ at ‘Self Help Services’”. In: *BMJ Open* 7.7, Article e016392. DOI: 10.1136/bmjopen-2017-016392 (cited on pp. 32–34, 40).

Elison-Davies, Sarah, Davies, Glyn, et al. (2018). “Protocol for a randomized controlled trial of the Breaking Free Online Health and Justice program for substance misuse in prison settings”. In: *Health & Justice* 6.1, Article 20. DOI: https://doi.org/10.1186/s40352-018-0078-1 (cited on pp. 34, 118).

Elison-Davies, Sarah, Märtens, Kaspar, et al. (2021). “Associations between baseline opioid use disorder severity, mental health and biopsychosocial functioning, with clinical responses to computer-assisted therapy treatment”. In: *The American Journal of Drug and Alcohol Abuse* 47.3, pp. 360–372. DOI: https://doi.org/10.1080/00952990.2020.1861618 (cited on pp. 32–35, 40).

Elison-Davies, Sarah, Pittard, Lauren, et al. (2023). “Examining outcomes for service users accessing the Breaking Free Online computer-assisted therapy program for substance use disorders via a ‘telehealth’ approach: Protocol for a two arm, parallel group randomized controlled trial”. In: *Addiction Science & Clinical Practice* 18.1, Article 39. DOI: https://doi.org/10.1186/s13722-023-00391-0 (cited on pp. 34, 118).

Elison-Davies, Sarah, Wardell, Jeffrey D., et al. (2021). “Examining correlates of cannabis users’ engagement with a digital intervention for substance use disorder: An observational study of clients in UK services delivering Breaking Free Online”. In: *Journal of Substance Abuse Treatment* 123, Article 108261. DOI: https://doi.org/10.1016/j.jsat.2020.108261 (cited on pp. 32–34, 40).

Epstein, David H., Tyburski, Matthew, Craig, Ian M., et al. (2014). “Real-time tracking of neighborhood surroundings and mood in urban drug misusers: Application of a new method to study behavior in its geographical context”. In: *Drug and Alcohol Dependence* 134, pp. 22–29. DOI: https://doi.org/10.1016/j.drugalcdep.2013.09.007 (cited on p. 18).

Epstein, David H., Tyburski, Matthew, Kowalczyk, William J., et al. (2020). “Prediction of stress and drug craving ninety minutes in the future with passively collected GPS data”. In: *npj Digital Medicine* 3, Article 26. DOI: https://doi.org/10.1038/s41746-020-0234-6 (cited on p. 18).

Epstein, Edward S. (1969). “A scoring system for probability forecasts of ranked categories”. In: *Journal of Applied Meteorology* 8.6, pp. 985–987. DOI: https://doi.org/10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2 (cited on p. 45).

European Commission (2021). *Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts*. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206 (visited on 11/15/2023) (cited on p. 80).

European Monitoring Centre for Drugs and Drug Addiction (2020). *European Drug Report 2020. Trends and developments*. Tech. rep. Publications Office of the European Union. URL: https://www.emcdda.europa.eu/publications/edr/trends-developments/2020_en (cited on p. 37).

European Parliament and Council of the European Union (2016). *General Data Protection Regulation*. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj (visited on 03/18/2024) (cited on p. 80).

Ewbank, M. P. et al. (2021). “Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning

approach to automatic coding of session transcripts". In: *Psychotherapy Research* 31.3, pp. 300–312. DOI: https://doi.org/10.1080/10503307.2020.1788740 (cited on p. 113).

Faurholt-Jepsen, Maria et al. (2016). "Behavioral activities collected through smartphones and the association with illness activity in bipolar disorder: Smartphone data in bipolar disorder". In: *International Journal of Methods in Psychiatric Research* 25.4, pp. 309–323. DOI: https://doi.org/10.1002/mpr.1502 (cited on p. 113).

Faurholt□Jepsen, Maria et al. (2015). "Smartphone data as an electronic biomarker of illness activity in bipolar disorder". In: *Bipolar Disorders* 17.7, pp. 715–728. DOI: https://doi.org/10.1111/bdi.12332 (cited on p. 113).

Firth, Josh A. et al. (2020). "Using a real-world network to model localized COVID-19 control strategies". In: *Nature Medicine* 26, pp. 1616–1622. DOI: https://doi.org/10.1038/s41591-020-1036-8 (cited on p. 14).

Fleming, Theresa et al. (2018). "Beyond the trial: Systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety". In: *Journal of Medical Internet Research* 20.6, Article e199. DOI: 10.2196/jmir.9275 (cited on pp. 39, 64, 74).

Flygare, Oskar et al. (2020). "Predictors of remission from body dysmorphic disorder after internet-delivered cognitive behavior therapy: A machine learning approach". In: *BMC Psychiatry* 20.1, Article 247. DOI: https://doi.org/10.1186/s12888-020-02655-4 (cited on pp. 66, 69, 82).

Frye, Christopher, Mijolla, Damien de, et al. (Mar. 2021). "Shapley explainability on the data manifold [Poster session]". In: *International Conference on Learning Representation* (cited on p. 86).

Frye, Christopher, Rowat, Colin, and Feige, Ilya (Sept. 2020). “Asymmetric Shapley values: Incorporating causal knowledge into model-agnostic explainability [Poster session]”. In: *Conference on Neural Information Processing Systems* (cited on p. 86).

Gan, Daniel Z. Q. et al. (2021). “Effect of engagement with digital interventions on mental health outcomes: A systematic review and meta-analysis”. In: *Frontiers in Digital Health* 3, Article 764079. DOI: https://doi.org/10.3389/fdgth.2021.764079 (cited on p. 68).

Garriga, Roger et al. (2022). “Machine learning model to predict mental health crises from electronic health records”. In: *Nature Medicine* 28, pp. 1240–1248. DOI: https://doi.org/10.1038/s41591-022-01811-5 (cited on pp. 15, 16, 63, 66, 82, 91).

Gehrmann, Julia et al. (2023). “What prevents us from reusing medical real-world data in research”. In: *Scientific Data* 10, Article 459. DOI: https://doi.org/10.1038/s41597-023-02361-2 (cited on p. 14).

Ghassemi, Marzyeh et al. (2018). “ClinicalVis: Supporting clinical task-focused design evaluation”. URL: https://doi.org/10.48550/arXiv.1810.05798 (cited on p. 82).

Goldberg, Simon B., Lam, Sin U, et al. (2022). “Mobile phone-based interventions for mental health: A systematic meta-review of 14 meta-analyses of randomized controlled trials”. In: *PLOS Digital Health* 1.1, Article e0000002. DOI: https://doi.org/10.1371/journal.pdig.0000002 (cited on pp. 20–22).

Goldberg, Simon B., Sun, Shufang, et al. (2023). “Selecting and describing control conditions in mobile health randomized controlled trials: A proposed typology”. In: *npj Digital Medicine* 6, Article 181. DOI: https://doi.org/10.1038/s41746-023-00923-7 (cited on p. 22).

Gossop, Michael et al. (1995). “The Severity of Dependence Scale (SDS): Psychometric properties of the SDS in English and Australian samples of heroin, cocaine

and amphetamine users". In: *Addiction* 90.5, pp. 607–614. DOI: https://doi.org/10.1046/j.1360-0443.1995.9056072.x (cited on pp. 6, 43, 70, 90).

Gyevnar, Balint, Ferguson, Nick, and Schafer, Burkhard (Sept. 2023). "Bridging the transparency gap: What can explainable AI learn from the AI Act? [Conference presentation]". In: *European Conference on Artificial Intelligence*. Krakow, Poland (cited on p. 80).

Hammarlund, Rebecca A. et al. (2018). "Review of the effects of self-stigma and perceived social stigma on the treatment-seeking decisions of individuals with drug- and alcohol-use disorders". In: *Substance Abuse and Rehabilitation* 9, pp. 115–136. DOI: 10.2147/SAR.S183256 (cited on p. 39).

He, Jianxing et al. (2019). "The practical implementation of artificial intelligence technologies in medicine". In: *Nature Medicine* 25, pp. 30–36. DOI: https://doi.org/10.1038/s41591-018-0307-0 (cited on p. 79).

Henson, Philip et al. (2021). "Anomaly detection to predict relapse risk in schizophrenia". In: *Translational Psychiatry* 11, Article 28. DOI: https://doi.org/10.1038/s41398-020-01123-7 (cited on p. 15).

Hirk, Rainer, Hornik, Kurt, and Vana, Laura (2020). "mvord: An R package for fitting multivariate ordinal regression models". In: *Journal of Statistical Software* 93.4, pp. 1–41. DOI: https://doi.org/10.18637/jss.v093.i04 (cited on pp. 43, 44, 46, 48).

Hofmann, Hans (1994). *Statlog (German Credit Data)*. DOI: 10.24432/C5NC77. URL: https://archive.ics.uci.edu/dataset/144 (visited on 11/29/2023) (cited on p. 106).

Hoogendoorn, Mark et al. (2017). "Predicting social anxiety treatment outcome based on therapeutic email conversations". In: *IEEE Journal of Biomedical and Health Informatics* 21.5, pp. 1449–1459. DOI: 10.1109/JBHI.2016.2601123 (cited on p. 113).

Hornstein, Silvan et al. (2021). “Predicting therapy outcome in a digital mental health intervention for depression and anxiety: A machine learning approach”. In: *DIGITAL HEALTH* 7. DOI: https://doi.org/10.1177/20552076211060659 (cited on p. 69).

Jacobs, Maia et al. (2021). “How machine-learning recommendations influence clinician treatment selections: The example of antidepressant selection”. In: *Translational Psychiatry* 11, Article 108. DOI: https://doi.org/10.1038/s41398-021-01224-x (cited on p. 112).

Jennings, Katie (Apr. 2023). “Pear Therapeutics files for bankruptcy as CEO blames shortfalls on insurers”. In: *Forbes*. URL: https://www.forbes.com/sites/katiejennings/2023/04/07/pear-therapeutics-files-for-bankruptcy-as-ceo-blames-shortfalls-on-insurers/ (visited on 11/29/2023) (cited on p. 23).

Kelley, Sean W. and Gillan, Claire M. (2022). “Using language in social media posts to study the network dynamics of depression longitudinally”. In: *Nature Communications* 13, Article 870. DOI: https://doi.org/10.1038/s41467-022-28513-3 (cited on p. 16).

Keyes, Katherine M. et al. (2010). “Telescoping and gender differences in alcohol dependence: New evidence from two national surveys”. In: *American Journal of Psychiatry* 167.8, pp. 969–976. DOI: https://doi.org/10.1176/appi.ajp.2009.09081161 (cited on p. 37).

Khan, Sharaf, Okuda, Mayumi, et al. (2013). “Gender differences in lifetime alcohol dependence: Results from the National Epidemiologic Survey on Alcohol and Related Conditions”. In: *Alcoholism: Clinical and Experimental Research* 37.1, pp. 1696–1705. DOI: https://doi.org/10.1111/acer.12158 (cited on pp. 37, 63).

Khan, Sharaf S., Secades-Villa, Roberto, et al. (2013). “Gender differences in cannabis use disorders: Results from the National Epidemiologic Survey of Alcohol and Re-

lated Conditions". In: *Drug and Alcohol Dependence* 130.1-3, pp. 101–108. DOI: https://doi.org/10.1016/j.drugalcdep.2012.10.015 (cited on pp. 37, 63).

Koppe, Georgia, Meyer-Lindenberg, Andreas, and Durstewitz, Daniel (2021). "Deep learning for small and big data in psychiatry". In: *Neuropsychopharmacology* 46, pp. 176–190. DOI: https://doi.org/10.1038/s41386-020-0767-z (cited on pp. 78, 82).

Koutsouleris, Nikolaos et al. (2016). "Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: A machine learning approach". In: *The Lancet Psychiatry* 3.10, pp. 935–946. DOI: https://doi.org/10.1016/S2215-0366(16)30171-7 (cited on pp. 16, 82).

Krishna, Satyapriya et al. (2022). "The disagreement problem in explainable machine learning: A practitioner's perspective". URL: https://arxiv.org/abs/2202.01602 (visited on 11/29/2023) (cited on pp. 86, 88, 89, 93, 94, 105, 106, 108, 110, 111).

Kroenke, Kurt et al. (2009). "An ultra-brief screening scale for anxiety and depression: The PHQ–4". In: *Psychosomatics* 50.6, pp. 613–621. DOI: https://doi.org/10.1016/S0033-3182(09)70864-3 (cited on pp. 6, 43, 70, 90).

Kroll, Joshua A. (2018). "The fallacy of inscrutability". In: *Philosophical Transactions of the Royal Society A* 376.2133, Article 20180084. DOI: https://doi.org/10.1098/rsta.2018.0084 (cited on p. 82).

Kympouropoulos, Stelios (2023). "Real world evidence: Methodological issues and opportunities from the European Health Data Space". In: *BMC Medical Research Methodology* 23, Article 185. DOI: https://doi.org/10.1186/s12874-023-02014-3 (cited on pp. 36, 119).

Lappan, Sara N., Brown, Andrew W., and Hendricks, Peter S. (2020). "Dropout rates of in-person psychosocial substance use disorder treatments: A systematic review

and meta-analysis". In: *Addiction* 115.2, pp. 201–217. DOI: https://doi.org/10.1111/add.14793 (cited on p. 36).

Lenhard, Fabian et al. (2018). "Prediction of outcome in internet-delivered cognitive behaviour therapy for paediatric obsessive-compulsive disorder: A machine learning approach". In: *International Journal of Methods in Psychiatric Research* 27.1, Article e1576. DOI: https://doi.org/10.1002/mpr.1576 (cited on pp. 69, 82).

Lewis, Ben, Hoffman, Lauren A., and Nixon, Sara Jo (2014). "Sex differences in drug use among polysubstance users". In: *Drug and Alcohol Dependence* 145, pp. 127–133. DOI: https://doi.org/10.1016/j.drugalcdep.2014.10.003 (cited on pp. 37, 63).

Lewis, Ben and Nixon, Sara Jo (2014). "Characterizing gender differences in treatment seekers". In: *Alcoholism: Clinical and Experimental Research* 38.1, pp. 275–284. DOI: https://doi.org/10.1111/acer.12228 (cited on pp. 37, 63).

Liddell, Torrin M. and Kruschke, John K. (2018). "Analyzing ordinal data with metric models: What could possibly go wrong?" In: *Journal of Experimental Social Psychology* 79, pp. 328–348. DOI: https://doi.org/10.1016/j.jesp.2018.08.009 (cited on p. 45).

Linardon, Jake, Cuijpers, Pim, et al. (2019). "The efficacy of app-supported smartphone interventions for mental health problems: A meta-analysis of randomized controlled trials". In: *World Psychiatry* 18.3, pp. 325–336. DOI: https://doi.org/10.1002/wps.20673 (cited on p. 63).

Linardon, Jake, Fuller-Tyszkiewicz, Matthew, et al. (2022). "An exploratory application of machine learning methods to optimize prediction of responsiveness to digital interventions for eating disorder symptoms". In: *International Journal of Eating Dis-*

orders 55.6, pp. 845–850. DOI: https://doi.org/10.1002/eat.23733 (cited on p. 76).

Lipton, Zachary C. (June 2016). “The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery”. In: New York, NY, United States. DOI: https://doi.org/10.1145/3236386.3241340 (cited on pp. 79, 80).

Liu, Fang and Panagiotakos, Demosthenes (2022). “Real-world data: A brief review of the methods, applications, challenges and opportunities”. In: *BMC Medical Research Methodology* 22, Article 287. DOI: https://doi.org/10.1186/s12874-022-01768-6 (cited on p. 113).

Livingston, Nicholas A. et al. (2021). “Differential alcohol treatment response by gender following use of VetChange”. In: *Drug and Alcohol Dependence* 221, Article 108552. DOI: https://doi.org/10.1016/j.drugalcdep.2021.108552 (cited on p. 63).

Lockey, Steven et al. (May 2021). “A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions [Conference presentation]”. In: *Hawaii International Conference on System Sciences*. Maui, Hawaii, USA (cited on p. 82).

Lopez-Quintero, Catalina et al. (2011). “Probability and predictors of transition from first use to dependence on nicotine, alcohol, cannabis, and cocaine: Results of the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC)”. In: *Drug and Alcohol Dependence* 115.1-2, pp. 120–130. DOI: https://doi.org/10.1016/j.drugalcdep.2010.11.004 (cited on p. 63).

Lundberg, Scott M. and Lee, S.-I. (May 2017). “A unified approach to interpreting model predictions [Conference presentation]”. In: *International Conference on Neural Infor-*

mation Processing Systems. Long Beach, California, USA. DOI: https://www.youtube.com/watch?v=wjd1G5bu_TY (cited on pp. 84, 85).

Lyubomirsky, Sonja (2023). “Everything everywhere all at once”. In: *AI Anthology*. Ed. by Eric Horvitz. Microsoft Unlocked. URL: https://unlocked.microsoft.com/ai-anthology/sonja-lyubomirsky (visited on 11/29/2023) (cited on p. 115).

Marinova, Nushka, Rogers, Tim, and MacBeth, Angus (2022). “Predictors of adolescent engagement and outcomes - a cross-sectional study using the togetherall (formerly Big White Wall) digital mental health platform”. In: *Journal of Affective Disorders* 311, pp. 284–293. DOI: https://doi.org/10.1016/j.jad.2022.05.058 (cited on p. 69).

Markus, Aniek, Fridgeirsson, Egill, et al. (July 2023). “Understanding the size of the feature importance disagreement problem in real-world data [Poster session]”. In: *3rd Workshop on Interpretable Machine Learning in Healthcare at the International Conference on Machine Learning*. Honolulu, Hawaii, USA (cited on pp. 86, 89, 96, 105, 106, 109–111).

Markus, Aniek F., Kors, Jan A., and Rijnbeek, Peter R. (2021). “The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies”. In: *Journal of Biomedical Informatics* 113, Article 103655. DOI: https://doi.org/10.1016/j.jbi.2020.103655 (cited on p. 80).

Marlatt, G. A., Bowen, S., et al. (2010). “Mindfulness-based relapse prevention for substance abusers: Therapist training and therapeutic relationships”. In: *Mindfulness and the therapeutic relationship*. Ed. by S. Hick, T. Bien, and Z. Segal. The Guilford Press (cited on p. 29).

Marlatt, G. A. and D. M. Donovan, eds. (2005). *Relapse prevention: Maintenance strategies in the treatment of addictive behaviors*. 2nd ed. The Guilford Press (cited on p. 29).

Marsch, Lisa A. (2021). “Digital health data-driven approaches to understand human behavior”. In: *Neuropsychopharmacology* 46, pp. 191–196. DOI: https://doi.org/10.1038/s41386-020-0761-5 (cited on pp. 14, 15, 39).

Marsch, Lisa A. et al. (2014). “Web-based behavioral treatment for substance use disorders as a partial replacement of standard methadone maintenance treatment”. In: *Journal of Substance Abuse Treatment* 46.1, pp. 43–51. DOI: https://doi.org/10.1016/j.jsat.2013.08.012 (cited on p. 23).

McHugh, R. Kathryn et al. (2018). “Sex and gender differences in substance use disorders”. In: *Clinical Psychology Review* 66, pp. 12–23. DOI: https://doi.org/10.1016/j.cpr.2017.10.012 (cited on p. 45).

Mehta, Ashish et al. (2021). “Acceptability and effectiveness of artificial intelligence therapy for anxiety and depression (Youper): Longitudinal observational study”. In: *Journal of Medical Internet Research* 23.6, Article e26771. DOI: 10.2196/26771 (cited on pp. 74, 87).

Michie, Susan et al. (2013). “The Behavior Change Technique Taxonomy (v1) of 93 hierarchically clustered techniques: Building an international consensus for the reporting of behavior change interventions”. In: *Annals of Behavioral Medicine* 46.1, pp. 81–95. DOI: https://doi.org/10.1007/s12160-013-9486-6 (cited on pp. 6, 29).

Miller, Sam, El-Bahrawy, Abeer, et al. (2020). “Predicting drug demand with Wikipedia views: Evidence from darknet markets. [Conference presentation]”. In: *The Web Conference*. Taipei, Taiwan (cited on p. 19).

Miller, William R. and Rose, Gary S. (2015). “Motivational interviewing and decisional balance: Contrasting responses to client ambivalence”. In: *Behavioural and Cognitive Psychotherapy* 43.2, pp. 129–141. DOI: 10.1017/S1352465813000878 (cited on p. 29).

National Drug Treatment Monitoring System, Office for Health Improvement & Disparities (Nov. 2023). *ViewIt*. URL: https://www.ndtms.net/ViewIt/Adult (visited on 11/23/2021) (cited on pp. 37, 48, 63).

NHS Wales Informatics Service (2019). *Treatment data - substance misuse in Wales 2018-19*. Tech. rep. Welsh Government. URL: https://www.gov.wales/sites/default/files/publications/2019-10/treatment-data-substance-misuse-in-wales-2018-19.pdf (cited on p. 48).

O’Brien, Heather L. et al. (2020). “Beyond clicks and downloads: A call for a more comprehensive approach to measuring mobile-health app engagement”. In: *BJPsych Open* 6.5, Article e86. DOI: https://doi.org/10.1192/bjo.2020.72 (cited on pp. 65, 74).

Obermeyer, Ziad et al. (2019). “Dissecting racial bias in an algorithm used to manage the health of populations”. In: *Science* 366, pp. 447–453. DOI: 10.1126/science.aax2342 (cited on p. 82).

Office for Health Improvement and Disparities (2021). *Sex ratios at birth in the United Kingdom, 2015 to 2019: Report*. Tech. rep. Department of Health and Social Care. URL: https://www.gov.uk/government/statistics/sex-ratios-at-birth-in-the-united-kingdom-2015-to-2019/sex-ratios-at-birth-in-the-united-kingdom-2015-to-2019-report (cited on p. 50).

Office for National Statistics (Dec. 2021). *Population estimates by ethnic group and religion, England and Wales: 2019*. URL: https://www.ons.gov.uk/peoplepopul

ationandcommunity/populationandmigration/populationestimates/articles/populationestimatesbyethnicgroupandreligionenglandandwales/2019#:~:text=2.-,Ethnicity%20in%20England%20and%20Wales,points%20since%20the%202011%20Census (visited on 03/17/2024) (cited on p. 50).

Olsen, Lars H. B. et al. (2022). “Using Shapley values and variational autoencoders to explain predictive models with dependent mixed features”. In: *Journal of Machine Learning Research* 23, pp. 1–51. URL: http://jmlr.org/papers/v23/21-1413.html (cited on p. 86).

Olsen, Lars Henry Berge et al. (2023). “A comparative study of methods for estimating conditional Shapley values and when to use them”. URL: https://arxiv.org/pdf/2305.09536.pdf (visited on 11/29/2023) (cited on p. 111).

Panlilio, Leigh V., Stull, Samuel W., Bertz, Jeremiah W., et al. (2021). “Beyond abstinence and relapse II: Momentary relationships between stress, craving, and lapse within clusters of patients with similar patterns of drug use”. In: *Psychopharmacology* 238, pp. 1513–1529. DOI: https://doi.org/10.1007/s00213-021-05782-2 (cited on pp. 18, 19).

Panlilio, Leigh V., Stull, Samuel W., Kowalczyk, William J., et al. (2019). “Stress, craving and mood as predictors of early dropout from opioid agonist therapy”. In: *Drug and Alcohol Dependence* 202, pp. 200–208. DOI: https://doi.org/10.1016/j.drugalcdep.2019.05.026 (cited on p. 18).

Paul, Riya et al. (2019). “Treatment response classes in major depressive disorder identified by model-based clustering and validated by clinical prediction models”. In: *Translational Psychiatry* 9, Article 187. DOI: https://doi.org/10.1038/s41398-019-0524-4 (cited on pp. 16, 66, 82, 83).

Pelissier, Bernadette, Jones, Nicole, and Cadigan, Timothy (2007). “Drug treatment aftercare in the criminal justice system: A systematic review”. In: *Journal of Substance Abuse Treatment* 32.3, pp. 311–320. DOI: https://doi.org/10.1016/j.jsat.2006.09.007 (cited on p. 39).

Perlis, Roy H. (2013). “A clinical risk stratification tool for predicting treatment resistance in major depressive disorder”. In: *Biological Psychiatry* 74.1, pp. 7–14. DOI: https://doi.org/10.1016/j.biopsych.2012.12.007 (cited on p. 115).

Pettersson, Erik et al. (2020). “The general factor of psychopathology: A comparison with the general factor of intelligence with respect to magnitude and predictive validity”. In: *World Psychiatry* 19.2, pp. 206–213. DOI: https://doi.org/10.1002/wps.20763 (cited on p. 91).

Phillips, Karran A., Epstein, David H., and Preston, Kenzie L. (2013). “Daily temporal patterns of heroin and cocaine use and craving: Relationship with business hours regardless of actual employment status”. In: *Addictive Behaviors* 38.10, pp. 2485–2491. DOI: https://doi.org/10.1016/j.addbeh.2013.05.010 (cited on p. 39).

Population Health, Clinical Audit and Specialist Care Team, NHS Digital (2022). *Psychological therapies, annual report on the use of IAPT services, 2021-22*. Tech. rep. 12. NHS Digital. URL: https://digital.nhs.uk/data-and-information/publications/statistical/psychological-therapies-annual-reports-on-the-use-of-iapt-services/annual-report-2021-22 (cited on p. 71).

Poursabzi-Sangdeh, Forough et al. (Oct. 2021). “Manipulating and measuring model interpretability [Conference presentation]”. In: *Conference on Human Factors in Computing Systems*. DOI: https://www.youtube.com/watch?v=0CYTLkQOV2E (cited on p. 82).

Pratap, Abhishek et al. (2020). “Indicators of retention in remote digital health studies: A cross-study evaluation of 100,000 participants”. In: *npj Digital Medicine* 3, Article 21. DOI: https://doi.org/10.1038/s41746-020-0224-8 (cited on pp. 23, 74, 75).

Quilty, Lena C. et al. (2022). “Peer support and online cognitive behavioural therapy for substance use concerns: Protocol for a randomised controlled trial”. In: *BMJ Open* 12.12, Article e064360. DOI: 10.1136/bmjopen-2022-064360 (cited on pp. 34, 118).

Ramos, Lucas A. et al. (2021). “Predicting success of a digital self-help intervention for alcohol and substance use with machine learning”. In: *Frontiers in Psychology* 12, Article 734633. DOI: https://doi.org/10.3389/fpsyg.2021.734633 (cited on pp. 24, 40, 64, 69).

Ribeiro, Marco Tulio, Singh, Sameer, and Guestrin, Carlos (2016). “”Why should I trust you?”: Explaining the predictions of any classifier [Conference presentation]”. In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, California, USA (cited on p. 83).

Riper, Heleen et al. (2018). “Effectiveness and treatment moderators of internet interventions for adult problem drinking: An individual patient data meta-analysis of 19 randomised controlled trials”. In: *PLOS Medicine* 15.12, Article e1002714. DOI: https://doi.org/10.1371/journal.pmed.1002714 (cited on pp. 20, 22, 23).

Roose, Kevin (Feb. 2023). “A conversation with Bing’s chatbot left me deeply unsettled”. In: *New York Times*. URL: https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html (visited on 11/29/2023) (cited on p. 115).

Rudin, Cynthia (2019). “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nature Machine Intelli-*

gence 1, pp. 206–215. DOI: https://doi.org/10.1038/s42256-019-0048-x (cited on p. 85).

Russell, Cayley et al. (2021). “Identifying the impacts of the COVID-19 pandemic on service access for people who use drugs (PWUD): A national qualitative study”. In: *Journal of Substance Abuse Treatment* 129, Article 108374. DOI: https://doi.org/10.1016/j.jsat.2021.108374 (cited on p. 39).

Santomauro, Damian F et al. (2021). “Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic”. In: *The Lancet* 398.10312, pp. 1700–1712. DOI: https://doi.org/10.1016/S0140-6736(21)02143-7 (cited on p. 12).

Saraiya, Tanya C. et al. (2020). “Perspectives on trauma and the design of a technology-based trauma-informed intervention for women receiving medications for addiction treatment in community-based settings”. In: *Journal of Substance Abuse Treatment* 112, pp. 92–101. DOI: https://doi.org/10.1016/j.jsat.2020.01.011 (cited on p. 63).

Schleider, Jessica L. et al. (2021). “A randomized trial of online single-session interventions for adolescent depression during COVID-19”. In: *Nature Human Behaviour* 6.2, pp. 258–268. DOI: https://doi.org/10.1038/s41562-021-01235-0 (cited on p. 74).

Shapiro, Allison et al. (2021). “Characterizing COVID-19 and influenza illnesses in the real world via person-generated health data”. In: *Patterns* 2.1, Article 100188. DOI: https://doi.org/10.1016/j.patter.2020.100188 (cited on p. 14).

Shapley, L. S. (1953). “A value for n-person games”. In: *Contributions to the theory of games (AM-28), volume II*. Ed. by Harold William Kuhn and Albert William Tucker. Princeton University Press, pp. 307–318 (cited on p. 84).

Shortreed, Susan M. et al. (2023). “Complex modeling with detailed temporal predictors does not improve health records-based suicide risk prediction”. In: *npj Digital Medicine* 6, Article 47. DOI: https://doi.org/10.1038/s41746-023-00772-4 (cited on p. 115).

Skevington, S. M., Lotfy, M., and O’Connell, K. A. (2004). “The World Health Organization’s WHOQOL-BREF quality of life assessment: Psychometric properties and results of the international field trial. A report from the WHOQOL Group”. In: *Quality of Life Research* 13.2, pp. 299–310. DOI: https://doi.org/10.1023/B:QURE.0000018486.91360.00 (cited on pp. 43, 70).

Slack, Dylan et al. (Feb. 2020). “Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods [Conference presentation]”. In: *AAAI/ACM Conference on AI, Ethics, and Society*. New York, New York, USA (cited on p. 85).

Sobell, Linda C. et al. (2001). “Cross-cultural evaluation of two drinking assessment instruments: Alcohol timeline followback and inventory of drinking situations”. In: *Substance Use & Misuse* 36.3, pp. 313–331. DOI: https://doi.org/10.1081/JA-100102628 (cited on p. 22).

Staiger, Petra K. et al. (2020). “Mobile apps to reduce tobacco, alcohol, and illicit drug use: Systematic review of the first decade”. In: *Journal of Medical Internet Research* 22.11, Article e17156. DOI: 10.2196/17156 (cited on pp. 20–22).

Stein, Dan J. et al. (2022). “Psychiatric diagnosis and treatment in the 21st century: Paradigm shifts versus incremental integration”. In: *World Psychiatry* 21.3, pp. 393–414. DOI: https://doi.org/10.1002/wps.20998 (cited on pp. 16, 111).

Stern, Ariel D et al. (2022). “Advancing digital health applications: Priorities for innovation in real-world evidence generation”. In: *The Lancet Digital Health* 4.3, e200–e206. DOI: https://doi.org/10.1016/S2589-7500(21)00292-2 (cited on p. 14).

Strobl, Carolin et al. (2007). “Bias in random forest variable importance measures: Illustrations, sources and a solution”. In: *BMC Bioinformatics* 8.1, Article 25. DOI: https://doi.org/10.1186/1471-2105-8-25 (cited on p. 92).

Strumbelj, Erik and Kononenko, Igor (2010). “An efficient explanation of individual classifications using game theory”. In: *The Journal of Machine Learning Research* 11, pp. 1–18 (cited on p. 84).

Štrumbelj, Erik and Kononenko, Igor (2014). “Explaining prediction models and individual predictions with feature contributions”. In: *Knowledge and Information Systems* 41.3, pp. 647–665. DOI: https://doi.org/10.1007/s10115-013-0679-x (cited on p. 84).

Sugarman, Dawn E., Meyer, Laurel E., Reilly, Meghan E., and Greenfield, Shelly F. (2020). “Feasibility and acceptability of a web-based, gender-specific intervention for women with substance use disorders”. In: *Journal of Women’s Health* 29.5, pp. 636–646. DOI: https://doi.org/10.1089/jwh.2018.7519 (cited on p. 63).

Sugarman, Dawn E., Meyer, Laurel E., Reilly, Meghan E., Rauch, Scott L., et al. (2021). “Exploring technology-based enhancements to inpatient and residential treatment for young adult women with co-occurring substance use”. In: *Journal of Dual Diagnosis* 17.3, pp. 236–247. DOI: https://doi.org/10.1080/15504263.2021.1940412 (cited on p. 63).

Svendsen, Thomas S. et al. (2021). “Securing participant engagement in longitudinal substance use disorder recovery research: A qualitative exploration of key retention factors”. In: *Journal of Psychosocial Rehabilitation and Mental Health* 8.3, pp. 247–259. DOI: https://doi.org/10.1007/s40737-021-00222-y (cited on p. 36).

Titov, Nickolai et al. (2020). “User characteristics and outcomes from a national digital mental health service: An observational study of registrants of the Australian

MindSpot Clinic". In: *The Lancet Digital Health* 2.11, e582–e593. DOI: https://doi.org/10.1016/S2589-7500(20)30224-7 (cited on pp. 40, 45).

Torous, John, Bucci, Sandra, et al. (2021). "The growing field of digital psychiatry: Current evidence and the future of apps, social media, chatbots, and virtual reality". In: *World Psychiatry* 20.3, pp. 318–335. DOI: https://doi.org/10.1002/wps.20883 (cited on pp. 19–21, 116).

Torous, John, Lipschitz, Jessica, et al. (2020). "Dropout rates in clinical trials of smartphone apps for depressive symptoms: A systematic review and meta-analysis". In: *Journal of Affective Disorders* 263, pp. 413–419. DOI: https://doi.org/10.1016/j.jad.2019.11.167 (cited on p. 23).

Torous, John, Michalak, Erin E., and O'Brien, Heather L. (2020). "Digital health and engagement—looking behind the measures and methods". In: *JAMA Network Open* 3.7, Article e2010918. DOI: 10.1001/jamanetworkopen.2020.10918 (cited on pp. 65, 75).

Torous, John, Nicholas, Jennifer, et al. (2018). "Clinical review of user engagement with mental health smartphone apps: Evidence, theory and improvements". In: *Evidence Based Mental Health* 21.3, pp. 116–119. DOI: 10.1136/eb-2018-102891 (cited on pp. 39, 68).

US Food and Drug Administration (2018). *Framework for FDA's real-world evidence program*. URL: https://www.fda.gov/media/120060/download (cited on p. 14).

– (Feb. 2023). *Real-world evidence*. URL: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence (visited on 04/25/2024) (cited on p. 13).

US Substance Abuse and Mental Health Services Administration (2020). *Behavioral health workforce report*. Tech. rep. US Substance Abuse and Mental Health Ser-

vices Administration. URL: https://annapoliscoalition.org/wp-content/uploads/2021/03/behavioral-health-workforce-report-SAMHSA-2.pdf (cited on p. 12).

Van Breda, Ward et al. (2018). “Predicting therapy success for treatment as usual and blended treatment in the domain of depression”. In: *Internet Interventions* 12, pp. 100–104. DOI: https://doi.org/10.1016/j.invent.2017.08.003 (cited on pp. 66, 82).

VanDeMark, Nancy R. et al. (2010). “An exploratory study of engagement in a technology-supported substance abuse intervention”. In: *Substance Abuse Treatment, Prevention, and Policy* 5, p. 10. DOI: https://doi.org/10.1186/1747-597X-5-10 (cited on p. 63).

Vos, Theo et al. (2020). “Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019”. In: *The Lancet* 396.10258, pp. 1204–1222. DOI: https://doi.org/10.1016/S0140-6736(20)30925-9 (cited on pp. 20, 39).

Wallert, John et al. (2022). “Predicting remission after internet-delivered psychotherapy in patients with depression using machine learning and multi-modal data”. In: *Translational Psychiatry* 12, p. 357. DOI: https://doi.org/10.1038/s41398-022-02133-3 (cited on p. 69).

Wang, Fei, Kaushal, Rainu, and Khullar, Dhruv (2020). “Should health care demand interpretable artificial intelligence or accept “black box” medicine?” In: *Annals of Internal Medicine* 172.1, pp. 59–60. DOI: https://doi.org/10.7326/M19-2548 (cited on p. 79).

Ward, Jonathan, Davies, Glyn, et al. (2017). “Achieving digital health sustainability: Breaking Free and CGL”. In: *International Journal of Health Governance* 22.2, pp. 72–

82. DOI: https://doi.org/10.1108/IJHG-07-2016-0037 (cited on pp. 26, 27, 34, 40).

Ward, Jonathan, Elison-Davies, Sarah, et al. (2019). “Clinical and demographic patient characteristics, alcohol treatment goal preference and goal attainment during computer-assisted therapy with Breaking Free Online”. In: *Journal of Substance Use* 24.6, pp. 681–687. DOI: https://doi.org/10.1080/14659891.2019.1651915 (cited on pp. 32–34, 40).

Weisel, Kiona K. et al. (2019). “Standalone smartphone apps for mental health—a systematic review and meta-analysis”. In: *npj Digital Medicine* 2, Article 118. DOI: https://doi.org/10.1038/s41746-019-0188-8 (cited on pp. 20–22).

Whittaker, Robyn et al. (2019). “Mobile phone text messaging and app-based interventions for smoking cessation”. In: *Cochrane Database of Systematic Reviews*. DOI: https://doi.org/10.1002/14651858.CD006611.pub5 (cited on pp. 20–23).

World Health Organisation (2022). *World mental health report: Transforming mental health for all*. Tech. rep. World Health Organization. URL: https://www.who.int/publications/i/item/9789240049338 (cited on p. 12).

World Health Organization (2021). *Global strategy on digital health 2020-2025*. Tech. rep. World Health Organization. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf (cited on p. 63).

Zhou, Yilun et al. (Dec. 2021). “Do feature attribution methods correctly attribute features? [Conference presentation]”. In: *1st Workshop on eXplainable AI approaches for debugging and diagnosis at Conference on Neural Information Processing Systems* (cited on p. 85).