Economic evaluations of specialist inpatient care for people with psychosis using data from electronic health records

Koeser, Leonardo

Awarding institution:
King's College London

Economic evaluations of specialist inpatient care for people with psychosis using data from electronic health records

Leonardo A Koeser

Submitted for the award of Doctor of Philosophy
in Health Economics

October 2019

King's College London

Abstract

In order to improve the use of limited health care resources, there is interest in assessing the value for money of treatments for psychotic disorders, a group of serious mental illnesses. However, most studies have assessed the value for money of medications, psychological therapies or community services and relatively little is known about the value of specialist inpatient care. Data that is routinely collected in electronic health records can be the basis for generating relatively inexpensive and timely evidence to support policy relevant questions.

Thus, the aim of this thesis was to assess the value for money of specialist inpatient care for people with psychosis using data from electronic health records. My objectives were (1) To identify approaches to handle unmeasured confounding and measurement error; (2) To conduct an economic evaluation of admission to child and adolescent inpatient care compared to admission to adult wards for young people with psychosis (Analysis 1) and (3) to conduct an economic evaluation of referral to inpatient rehabilitation compared to usual care for adults with persistent forms of psychosis (Analysis 2).

All analyses are based on data derived from the South London and Maudsley Biomedical Research Centre (BRC) clinical records interactive search (CRIS) database. In addition to three approaches to handling confounding that are well-known in health economics, I identify the front-door adjustment as an approach relevant to Analysis 2. I distinguish between four types of measurement error assumptions with respect to non-outcome variables and discuss five potential different strategies to support or enable these measurement error assumptions.

The results of Analysis 1 suggest that a regression discontinuity design is not suitable to compare the impact of admission to child and adolescent and adult ward. The results of Analysis 2 suggest the costs of inpatient rehabilitation are not offset by substantial savings in other service use and there was little evidence to suggest that patients benefit clinically from referral to inpatient rehabilitation.

Acknowledgements

I would like to thank all of those who have made a substantive contributed to this thesis, my friends, my brother, my mother, and the patients without whom this research could not have been possible.

Contents

List of figures

List of tables

List of abbreviations

ATE	Average treatment effect
ATT	Average treatment effect on the treatment
ATU	Average treatment effect on the untreated
BRC	Biomedical Research Centre
CAMHS	Child and Adolescent Mental Health Service
CI	Confidence interval
CPRD	Clinical Practice Research Database
CQC	Care Quality Commission
CRIS	Clinical Records and Interactive Search database
EHR	Electronic health records
HES	Hospital Episode Statistics
HoNOS	Health of the Nation Outcome Scale
ICD	International Classification of Disease
IPCW	Inverse Probability of Censoring Weights
IPTW	Inverse Probability of Treatment Weights
IR	Inpatient rehabilitation
MeSH	Medical Subject Headings
MHSDS	Mental Health Services Data Set
NICE	National Institute of Health and Care Excellence
NLP	Natural Language Processing
PICU	Psychiatric Intensive Care Unit
QALY	Quality-adjusted Life Year
RCT	Randomized Controlled Trial
RDD	Regression discontinuity design
SD	Standard Deviation
SLaM	South London and Maudsley
UK	United Kingdom
WSC	Within study comparison

Chapter 1 Introduction

1.1 Chapter overview

The aim of this thesis was to assess the value for money of specialist forms of inpatient care for people with psychosis. In Section 1.2, I begin by explaining the clinical context to this analysis before provide an introduction to the health economic evaluations with particular reference to the role of electronic health records (Section 1.3). Following these, I given an overview of existing studies assessing the value for money of treatments for psychosis (Section 1.4). I then discuss the clinical and policy significance of assessing the value for money of specialist inpatient care for two specific patient groups, young people (Section 1.5) and people with persistent forms of psychosis (Section 1.6), critically evaluating the existing health economic evidence. This leads on to an examination of the potential approaches to conducting an economic evaluation of inpatient services for these patient groups (Section 1.7). Finally, I will describe the aims, objectives and structure of this thesis (Section 1.8) and state my contributions to it (Section 1.9).

1.2 Psychosis

1.2.1 Features of psychosis

Psychosis is an abnormal condition of the mind characterised by gross impairments of reality, chiefly due to the occurrence of delusions and/or hallucinations without insight (Arciniegas, 2015). For example, a common delusion is the firm belief that other people are trying to harm the person experiencing psychosis despite all evidence to the contrary and a common form of hallucinations is hearing voices that nobody else hears (Bromley et al., 2015). Periods in which psychosis occurs are referred to as psychotic episodes. Psychotic experiences are increasingly conceptualised as lying on a multidimensional spectrum and only distressing and/or disruptive cases are characterised as psychotic illnesses or psychotic conditions (Guloksuz and Os, 2018). Apart from being present in individuals without functional impairment, psychosis can be a feature in neurological and neurodevelopmental

conditions, such as epilepsy and stroke, but the focus of this thesis will be psychiatric conditions that can include psychosis as a symptom (NICE, 2014a). Some of the more prominent types of psychiatric conditions that include psychosis are schizophrenia, schizoaffective disorder and some forms of bipolar disorder. For simplicity and as is customary in the literature, I will use the term 'psychosis' as an umbrella term for these psychiatric conditions hereafter rather than the symptom of psychosis (Hayes and Kyriakopoulos, 2018).

The causes of psychotic disorders are unclear, but research suggests that gene-environment interactions play a role in predisposing individuals to psychosis (van Os et al., 2008). Psychotic episodes can be precipitated by physical, environmental and emotional stressors as well as substance use such as the consumption of cannabis (Griswold et al., 2015). The incidence of psychosis in England has been estimated to be approximately 32 per 100,000 person years and prevalence rates are thought to be around 0.5% of the UK population (Kirkbride et al., 2012; Singleton et al., 2003).

The impact of psychosis on individuals is highly heterogeneous but is often accompanied by deterioration in social and occupational functioning as well as vulnerability to risks and risk taking behaviour (Reed, 2008). In addition, people with psychosis suffer from high rates of physical comorbidity and mortality rates are nearly twice as high as the general population (Dutta et al., 2012). Beyond the people with psychosis themselves, the illness has a substantial impact on their family and carers (Reed, 2008). The impact of psychosis is exacerbated by the fact that this condition is heavily stigmatised because this can, for example, delay help-seeking and lead to social exclusion (Gronholm et al., 2017).

1.2.2 Treatments for psychosis

A range of terms that have been used to refer to people receiving mental health care, such as service user, patient, attendee, survivor, recipient, client or consumer, and preferences between these terms vary (Simmons et al., 2010). In this thesis, I will use the term patient because of the convenience of being able to use the word inpatient to refer to people receiving care while staying in hospital.

In terms of individual-level treatments, the mainstay of care for people with psychosis are antipsychotic drugs or, when mood disorders is present, drugs classed as mood stabilisers (NICE, 2014a). These can be used both during acute periods of symptoms exacerbation and to prevent the recurrence of symptoms. Many patients find antipsychotics useful and they are almost universally prescribed to people with psychotic illnesses who access services in the UK, but they only alleviate symptoms, can have severe side-effects and response to them is both highly heterogenous and difficult to predict (Bentall et al., 2000). Beyond this, other types of pharmacotherapy, psychological and psychosocial interventions or interventions to improve physical health may be offered to patients or the carer but this does not happen routinely (NICE, 2014a).

At a service-level, the care of people with psychosis in the United Kingdom has undergone radical changes since the 1980s as a result of the downsizing and closure of mental health hospitals in favour of care in the community (Burns, 2006). One of the key drivers of this process, known as deinstitutionalisation, has been the desire to avoid the high costs of providing care in an inpatient setting (Knapp et al., 2010). Nonetheless, inpatient care remains indispensable, particularly during psychotic episodes. In fact, if patients are thought to be at risk to themselves or others, they may be hospitalised against their will by health professionals under the Mental Health Act (Burns, 2006). Thus, hospitalisations are often used as an outcome measure in psychosis research (Burns, 2007). Inpatient care is still the most costly health care element in the treatment of psychosis and, with continuing reductions in hospital beds over the last decade or so, some have argued that the deinstitutionalisation process has gone too far (Mangalore and Knapp, 2007; Tyrer et al., 2017). Given the costs associated with hospitalisation, it is important that evidence to identify types of inpatient care that are good value for money is available to support the future development of inpatient care.

1.3 Health economic evaluation

1.3.1 Role of health economic evaluation

Economic evaluation is defined as, “the comparative analysis of alternative courses of action in terms of both the costs and consequences” (Drummond, 2015). Health economic evaluations are a commonly used tool to explore the cost-effectiveness, that is, the value for money, of health care services in the UK and many other high-income countries. The aim of health economic evaluations is to aid health care decision making relating to the allocation of scarce resources, using a systematic and transparent approach. In systems where the individual receives free health care, or does not pay the full cost of the care provided, there is an ethical imperative to assess the cost-effectiveness in addition to considering their effectiveness (Dowie, 2004). This is because pooled resources are limited in such systems such that investments into one form of care imply that benefits are necessarily forgone by other individuals. If the value for money of an intervention were to be disregarded, the allocation of scarce health care resources may not generate the greatest health care benefits to the community.

In order to maximise the benefits generated from scarce resources, the status of economic evaluations has increased in prominence in UK health care decision making over the last 25 years. In particular, economic evaluations have become an integral part of guidance produced by the National Institute of Health and Care Excellence (NICE), an organisation that provides national guidance and advice to improve health and social care in England and Wales. This includes the NICE guidance for psychosis and schizophrenia (NICE, 2014a) and the NICE guidance for bipolar disorder (NICE, 2014b).

At the same time, one should, however, note some of the limitations of economic evaluations (Drummond et al., 2005): (1) In all cases they remains a tool to inform decision making, a starting point to inform discussions about priorities and trade-offs, not a substitute for decision making; (2) Economic evaluation typically does not incorporate all societal values that may be relevant to make a particular decision, for example, consideration of social justice; (3) Economic evaluations commonly embody certain kinds of assumptions and value judgements, such as the idea the decision makers goal is to maximise average outcomes or that if money is not invested in one intervention they will be used for

the next best productive alternative. These assumptions and value judgements may not be warranted in all cases; (4) Economic evaluations are in themselves costly to undertake and are not warranted to inform all health care decisions.

1.3.2 Vehicles for economic evaluations

Three vehicles are commonly employed for conducting economic evaluations: randomised controlled trials (RCTs), observational studies and economic decision models (Baltussen et al., 1999). In RCTs, participants are assigned to different treatment regimens by chance. By contrast, in observational studies, the treatment is assigned on a basis other than randomisation (Faria et al., 2015). I will discuss the relative merits of these two vehicles for economic evaluation in Section 1.7. A limitation that RCTs and observational studies have in common when considering their use in decision making is that they provide estimates of value for money based on data from a specific group of patients in a particular environment. However, typically more than one study or piece of evidence is available to inform a particular decision (Sculpher et al., 2006). This may be because studies estimating the same quantity are available and/or because quantities estimated in single studies do not correspond to the quantities that are directly relevant to inform the decision question of interest (Spiegelhalter et al., 2004). For example, it may be that an RCT or an observational study assesses costs and benefits of an intervention of a limited amount of time but other research suggests that the benefits of an intervention may extended beyond the end of follow-up. To come to a decision that takes into account these multiple pieces of information, some approach to integrating or synthesising the evidence is necessary.

One approach to do so is to develop a decision analytic model, the third major vehicle for economic evaluations. Decision model involve defining mathematical relationships to approximate the impact of the set of treatment options under evaluation and populated using data from any source of evidence that is thought to be of sufficient quality, including individual patient data from RCTs or observational studies, aggregate data and expert opinion (Briggs et al., 2006; Soares et al., 2018). For example, a simple decision model might involve a three-step process of the following kind: (i) a systematic review and meta-analysis of RCTs examining the comparative effect of all relevant interventions on some

intermediate outcome, e.g. relapse rates; (ii) estimating the relationship between the intermediate outcome and endpoints that are of relevance to decision making using observational data, e.g. the mean cost of care for people who do not relapse and those who do relapse; (iii) linking these two elements while incorporating the uncertainty in all inputs and, if relevant, the model structure, to estimate expected cost-effectiveness of treatment strategies while quantifying the decision making uncertainty.

It has been proposed that, compared to other approaches to evidence synthesis for decision making, decision modelling increases the transparency, consistency of decision making by making judgments and uncertainties explicit as well as by facilitating the condensation of large or complex bodies of knowledge into a simplified format (Buxton et al., 1997). However, the extrapolation, selection, simplification and combination process that their creation can involve could make it more difficult to judge to what extent the evidence can be trusted (Ghabri et al., 2018) The dilemma that modelling presents is that complex models are more likely to realistically approximate the decision problem, but are also less transparent, more difficult to understand, more computationally intensive and more prone to errors in their construction (Briggs et al., 2006). Another risk with modelling is that the process of quantification can give the veneer of scientific integrity thereby discouraging further empirical work (OHE, 1997).

If and when decisions or recommendations are made by national entities, such as NICE, guidelines typically stipulate that decision modelling is the required or preferred approach to economic evaluation and evidence synthesis (EUnetHTA, 2015). Indeed, there is a more general consensus that decision modelling can be a valuable tool to aid decision making (Kuntz et al., 2013). However, decision modelling remains a tool and by implication more suited to inform some decisions than others. The extent to which decision models should be used and how they are used has been a source of debate (Barbui and Lintas, 2006; Briggs et al., 2006; Buxton et al., 1997; Grutters et al., 2019; OHE, 1997; Sainfort et al., 2013). In the economic evaluation of service interventions, the type of intervention that will be the focus of this thesis, the use of decision modelling appears less extensive and prominent (Meacock, 2018; Sutton et al., 2018). Systematic reviews by NICE (2014a) and Jin et al. (2020) suggest that only two service interventions, early intervention services and supported employment programmes have been evaluated using a decision analytic approach. In addition, Jin et al.

(2020) show that although 73 decision models were published between 2009 and 2018, none of these assessed the value of a service-level intervention and the 73 decision models only cover a narrow range of interventions. This suggests that the role of decision modelling in informing decisions about the care of people with psychosis is relatively limited in general.

Due to time and resource constraints, constructing a decision model synthesising the new evidence well as all existing evidence using decision models, was considered beyond the scope of this thesis. Instead, I conduct economic evaluations based on single studies and, where appropriate, make recommendations for decision making based on a narrative synthesis of the evidence. More specifically, I conducted economic evaluations based on single observational studies rather than RCTs. I discuss the rationale for this in Section 1.7.

1.3.3 Data sources for economic evaluations

Regardless of whether one undertakes an economic evaluation based on a single study, that is an RCT or an observational study, or an economic evaluation based on a decision model, one can make use of primary data, secondary data or a mixture of the two (Hox, J.J. and Boeije, H.R., 2005). Primary data is data collected for research purposes, often as part of RCTs. Secondary data is data originally collected for non-research purposes that can be reused for research purposes. Often this data is collected routinely in a naturalistic setting. Of particularly relevance to health economic evaluations is data in electronic health records (EHR). The terms EHR or electronic patient record have been used in a number of different ways, but commonly refer to a "longitudinal collection of electronic health information about individual patients and populations" (Gunter and Terry, 2005). Evidence collected in a naturalistic setting is sometimes referred to as 'real world data' and data from EHRs is also frequently considered to be an example of 'big data', which has become a buzzword for complex and large-scale datasets (Collins, 2016; Makady et al., 2017). Frequently, two or more independent datasets are combined for research purposes by means of data linkages. These linkages can be between different types of electronic health records (e.g. between primary care and secondary care data), or between electronic health records and other databases (e.g. mortality records). Some databases of routinely collected data which are

frequently used for medical research in the United Kingdom are Hospital Episode Statistics (HES), Clinical Practice Research Data link (CPRD) and The Health Improvement Network (THIN) (Stewart and Davis, 2016).

Although RCTs are still the most common and preferred source of data for economic evaluations, in parallel with the broader 'big data' movement, the status, capabilities and interest in the use of EHRs and other forms of routinely collected data for economic evaluations has increased over the last decade or so (Faria et al., 2015). For example, in 2007, a task force by the International Society for Pharmacoeconomics and Outcomes Research, published a guidance on the use of such data in coverage and reimbursement decisions, that is what health care should be paid for by the statutory provider and how much should be paid for it (Garrison et al., 2007). In 2015, the NICE Decision Support Unit issued recommendations to improve the quality and transparency of evidence from 'real world data' in technology appraisal and a review of use of such evidence in NICE decision making was published in the subsequent year (Bell et al., 2016; Faria et al., 2015). Similarly, in 2018, the Institute for Clinical and Economic Review produced guidance for the use of real-world data in drug coverage decisions (Pearson et al., 2018).

Due to resource constraints, in this thesis, I conducted economic evaluations based exclusively on secondary data derived from electronic health records.

1.4 Economic evaluations of treatments for psychosis

1.4.1 Overview

A large number of economic evaluations of treatments for psychosis have been carried out, using a variety of study designs and data sources, described later in this section (NICE, 2014a; Zhou et al., 2018). Figure 1, created using the bibliographic software VOSviewer, gives a graphical overview of interventions that have been the focus of existing economic evaluations in the field of psychosis (van Eck and Waltman, 2009). It illustrates the focus of the current literature by identifying intervention-related author keywords used in studies in the database Scopus containing terms related to both economic evaluations ("economic

evaluation" or "cost-effectiveness" or "cost-benefit analysis" or "cost analysis" or "cost-utility analysis" or "cost consequence" or "cost reduction" or "cost offset" or "pharmacoeconomic") and psychosis ("schizophrenia" or "schizoaffective disorder" or "bipolar disorder" or "psychosis" or "psychotic" or "schizophreniform"). I only included English-language articles published between 1 January 1995 and 15th March 2019 and used default options in VOSviewer. For clarity, I merged equivalent terms into the same category. For example, I combined studies using the author keyword 'antipsychotic agents' with those using the keyword 'antipsychotics'. Figure 1 graphically represents three different pieces of information about the included studies: (i) the size of the dots reflects the frequency that the relevant keywords occurred in the literature, with larger dots representing increased frequency; (ii) the distance between keywords reflects the likelihood that two keywords co-occur within a study, with shorter distances representing greater co-occurrence; and (iii) the colours reflect the average normalised citation rate, defined as the average number of citations in documents with a keyword divided by the average number of citations of all included documents published in the same year, with lighter colours representing a higher citation rates (van Eck and Waltman, 2009).

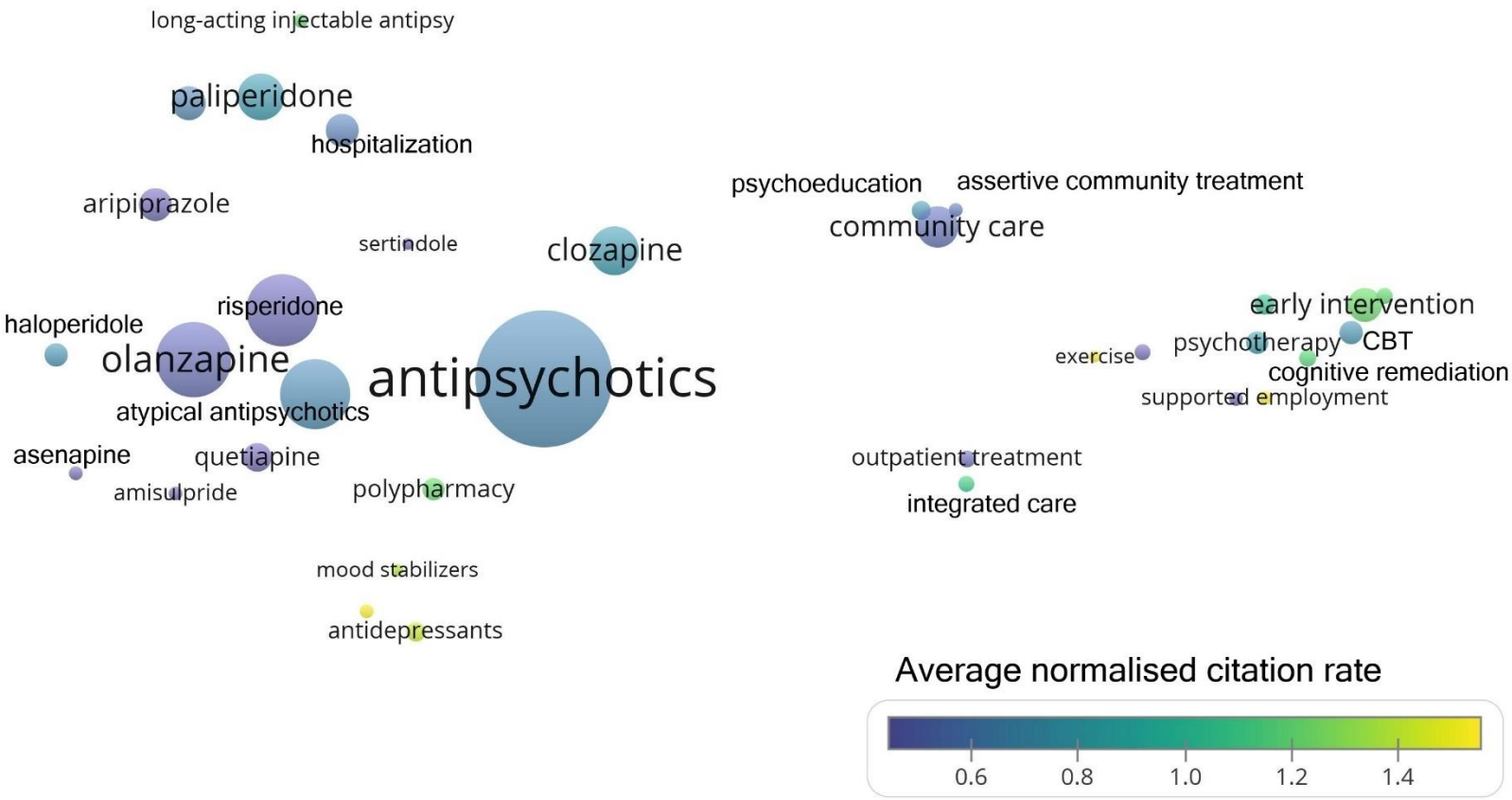


Figure 1: Treatment-related author keywords in economic evaluations for psychosis

Author keywords do not systematically contain the names of interventions evaluated in the study and the simple search strategy employed imperfectly discriminates between economic evaluations for psychosis and other studies. However, Figure 1 provides a useful way to summarise the large number of economic evaluations available. For example, the cluster of circles on the left hand side of the Figure suggests that much more attention has been devoted to the economic evaluation of various antipsychotic medications or classes of antipsychotic medications and, to a lesser extent, other psychiatric drugs, compared to non-pharmacological treatments which are clustered on the right-hand side. Inpatient care does not feature except for the keyword 'hospitalization' at the top left. On closer inspection, the vast majority of studies associated with this keyword, used it because hospitalisation was an outcome measure, rather than an intervention under evaluation.

These impressions of the scientific landscape are supported by previous systematic reviews of economic evaluations in psychosis. In terms of pharmacological interventions, Achilla and McCrone (2013) identified 28 economic evaluations of long-acting/extended-release antipsychotics in schizophrenia, Von Scheele et al. (2014) identified 83 economic decision models evaluating the cost-effectiveness of antipsychotics in schizophrenia, and a more recent review by Zhou et al. (2018), which used different inclusion criteria, identified 79 such studies. In contrast, reviews of non-pharmacological interventions have yielded a much more limited number of studies. A recent systematic review of economic evaluations of early intervention services for psychosis identifies 11 studies (Aceituno et al., 2019). A systematic search performed for the NICE guideline for psychosis and schizophrenia yielded 24 economic evaluations of non-pharmacological treatments almost half of which focused on early intervention and intensive care management (NICE, 2014a). The review for the NICE guideline for bipolar disorder identified eight non-pharmacological studies (NICE, 2014b). A review by Thomas and Rickwood (2013) identified six studies assessing the impact of residential mental health services on costs or service use.

In terms of study design, increasingly sophisticated economic decision models, often sponsored by pharmaceutical companies, are by far the most common vehicle for economic evaluations of pharmacological treatments. With few exceptions, such as Frey et al. (2014), these decision models are primarily populated using data from RCTs rather than

observational data. Non-pharmacological interventions appear to be most frequently evaluated within RCTs (NICE, 2014a).

1.4.2 Economic evaluation of inpatient care

Despite the prominence and cost of inpatient care in the care of people with psychosis noted above, these existing systematic reviews also highlight the lack of economic evidence focusing on inpatient care. No systematic review that I am aware of focuses on economic evaluations of inpatient care, NICE guidelines only discuss non-inpatient alternatives to general acute hospitalisations (NICE, 2014a). By contrast, evidence on the cost-effectiveness of specialist inpatient care is relatively limited. Exceptions include Howard et al. (2010) and Trevillion et al. (2019). By implication, resource allocation decisions around the provision of specialist inpatient care are instead likely to be mainly influenced by a combination of clinical judgments, institutional legacies and budget pressures which appears suboptimal. In fact, research on the clinical and economic outcomes of inpatient units is among the top ten research priorities in schizophrenia according to a deliberative exercise which brought patients, carers and clinicians together (Lloyd and White, 2011).

As alternatives to inpatient care for people with psychosis increase (Howard et al., 2010; Slade et al., 2010), there is a growing need for evidence of the effectiveness and cost-effectiveness of specialist inpatient care to support referral decisions made by clinicians and to support future development of inpatient services for people with psychosis. This dissertation attempts to begin to fill this gap by exploring the value for money of specialist inpatient care for young people, namely of child and adolescent mental health service (CAMHS) wards, specialist inpatient care for people with persistent forms of psychosis, namely rehabilitation wards. In the next two Sections (1.5 and 1.6), I describe these two types of inpatient care in more depth.

1.5 Inpatient care for young people

1.5.1 Health care decision problem and its significance

In the UK and many other countries, inpatient care for underage patients with psychosis is generally provided in specialised psychiatric facilities for young people rather than general adult wards (Richardson, 2010). Such facilities are run by staff trained specifically in the care of young people. Therefore, CAMHS ward staff are thought to have a better understanding of the needs of young people and the skills to meet them than mental health professionals who care for adults. In addition, the environment in CAMHS wards is believed to be more therapeutic, for example because severely disturbed older patients may represent a safety risk to young patients on adult wards and because some interventions such as educational support and family therapy can be provided more easily in a CAMHS environment (Park et al., 2011). For these reasons, it is well-accepted and codified in various guidelines that placing underage patients in adult wards is undesirable, potentially detrimental and should only be considered in two exceptional circumstances: (1) in response to a crisis situation, usually when CAMHS beds are already occupied; or (2) if it allows access to a specialist service that is most suited to the young person given their circumstances (Healthy London Partnership, 2016; Mental Welfare Commission, 2012; Richardson, 2010). In fact, admitting a person between 16 and 18 to adult wards is classed as a reportable incident in England (NHS England, 2014).

In practice, however, admission of a young person to a local CAMHS inpatient ward is not always an option as a result of a lack of beds. Instead, patients under the age of 18 are sometimes admitted to adult wards or to specialist child and adolescent facilities far away from their home which can be detrimental to recovery due to discontinuities in care and loss of social support networks. Between April 2017 and March 2018, for example, in England, 246 underage patients were treated for an average of 13 days on adult mental health wards and in 2017 (NHS Digital, 2018), 1,039 underage patients received mental health treatment outside of their local area (Owen, 2018). These persistent issues regularly attract media attention (Buchanan, 2014; Campbell, 2016; Naysmith, 2018). In addition, some parts of the UK, such as Manchester, have chosen to extend CAMHS care to people up

to the age of 25 because the discontinuity in care provision at 18 is widely regarded as problematic (Singh et al., 2017).

Understanding the trade-offs that decision makers face with respect to the cost-effectiveness of alternative effectiveness of alternative inpatient facility options for young people in both of these contexts is valuable to decision makers for a number of reasons. Late adolescence and early adulthood is known to be crucial for people's social, emotional and personal development (Lamb et al., 2008). In addition, inpatient stays in late adolescence and early adulthood are often the first experience of such care for people with psychosis (Kessler et al., 2007). Given that a stay on a psychiatric ward can be a difficult experience, this may shape their relationship with mental health services as well as outcomes in the long-term (Rose et al., 2015). Finally, it is known that, the average cost of CAMHS inpatient care per bed day, i.e. £716, is substantially higher than a bed day on a general adult psychiatric ward (£420) (NHS improvement, 2018).

1.5.2 Review of economic evidence for child and adolescent inpatient care for young people compare to any other care

To identify previous research on the impact of different types of inpatient care modalities on young people with psychosis, undertook a non-systematic review up to 1 July 2020. I particularly drew on existing reviews and databases of the literature in the area. To be more specific, I examined studies include in systematic reviews of economic evaluations in child and adolescent populations with psychosis or with psychiatric conditions more generally (Beecham, 2014; Kilian et al., 2010; Knapp et al., 2016; NICE, 2013), a regularly updated, near-comprehensive database of paediatric economic evaluation, the Paediatric Economic Evaluation Database (Ungar and Santos, 2003), a scoping review by Murcott (2016) on people between 16 and 25 admitted to adult psychiatric wards as well as a non-systematic review by Fusar-Poli (2019) on the effect of integrating mental health services for 12-25 year olds. In addition, I undertook an informal search of the academic literature using Google scholar. My aim was to identify studies that assessed either the impact of different forms of

inpatient care for adolescents with psychosis or the impact of policies to expand the provision of CAMHS beds for young adults beyond the age of 18 regardless of their mental health diagnosis. In either case, my interest lay in studies examining the impact on health care service use or cost thereof.

None of the existing literature reviews included research fulfilling the inclusion criteria specified above. In my informal literature search, however, I identified a recent study by Maxwell et al. (2019) that was relevant to the decision problem described in the previous section (1.5.1). This retrospective cohort study used routinely collected health care data to investigate the impact of implementing a unified youth mental health service for people aged between 14 and 25 in Norfolk, UK. Specifically, the authors compared a 12 month period spanning 2010 and 2011 in which care services were provided according to the traditional CAMHS (10 – 17 years) and adult (18+ years) model described in the previous section (1.5.1) and a 12 month period spanning 2014 to 2015 during which the youth service model caring for 14 to 25 year olds had become established. The outcome measures were the number of referrals, the proportion of accepted referrals and the average number of post-referral service contacts across the service by age at referral.

Maxwell et al. (2019) find that referrals for 14 to 25 year old increased considerably from approximately 7,500 to 12,500 between the two periods of interest. The number of accepted referrals remained approximately the same for 18 to 25 year olds such that the acceptance rate for this cohort fell from 95% to 75%. For adolescents referred between the ages of 14 and 17, on the other hand, both referrals and number of accepted referrals more than doubled. The percentage of accepted referrals decreased from 78% to 59% for this age group. The total number of service contacts increased from approximately 60,000 to 80,000 but the average number of contacts per referral fell for those below the age of 17 from about 11 to 8, slightly increased increase for those aged 18 to 20 at referral from about 5 to 7.5 and, with about 7 to 8 contacts per referral, remained more or less unchanged for those aged 21 to 25 at referral. In short, the contact pattern became more equal across the age of referral. The authors conclude that the provision of services has become more equitable in terms of access to services.

As the authors acknowledge, it is uncertain to what extent these estimates are influenced by random variation over time, changes in data quality, systematic developments in the

provision of care services in Norfolk and to the effects of the service restructuring itself. The authors do not measure health outcomes, cost of care or different forms of service use. They also do not assess the effect of the service restructuring specifically on patients with psychosis or the effects of changes in inpatient care in particular. Thus, lessons that one can draw from this study for the resource allocation decisions discussed in the previous section (1.5.1) are limited.

Given the non-systematic nature of my review, I cannot be confident that this review was exhaustive. However, all of the existing literature reviews and databases suggested that the number of economic evaluations of treatment for psychosis in young people in general is very limited. In fact, as a result of this lack of evidence the NICE (2013) guidance for young people with psychosis refers to the adult guideline for economic evidence while cautioning the reader from generalising findings too readily because of different treatment pathways and resource use. Similarly, Fusar-Poli (2019) specifically notes that economic evaluations of integrating mental health care for young people are lacking.

1.6 Inpatient care for people with enduring psychosis

1.6.1 Health care decision problem and its significance

In the UK and many other countries, inpatient care for adults with psychosis is generally provided in acute psychiatric wards. The environment in acute wards, which has been compared to a "pressure cooker" because of the demand on clinicians to discharge patients as soon as possible, is not thought to be truly recovery oriented (Craig, 2016). The main hospital-based alternative to acute wards for people with severed and enduring forms of psychosis is psychiatric inpatient rehabilitation (referred to from here on as 'inpatient rehabilitation') a specialised, long-term from of care (Killaspy, 2014). Common problems among the patient group for which inpatient rehabilitation is judged to be potentially appropriate include failure to respond to multiple trials of antipsychotics, behavioural problems, severe negative symptoms, cognitive impairments, poor physical health and severe functional impairments (Killaspy, 2009). Inpatient rehabilitation is believed to allow for a more holistic and tailored care than is commonly available in acute wards, providing

the staff time and resources to enable the development of relationships and to maximise patients' biopsychosocial functioning (Bunyan et al., 2017; Lavelle et al., 2012). More specifically,goals of inpatient rehabilitation may include giving clinical staff the opportunity to carry out assessments, optimise antipsychotic and other medication regimes (particularly the administration of the antipsychotic clozapine), providing psychotherapy for relapse prevention, psychoeducation and insight activities, engaging the patient with occupational therapy and social inclusion work, providing guidance on healthy living, and/or planning for a successful community discharge.

Economic evaluations of inpatient rehabilitation are of high policy relevance because almost all mental health trusts in England provide inpatient rehabilitation and, at around 2,100 NHS and 2,300 independent sector rehabilitation beds, the Care Quality Commission (CQC) (2018) has estimated that annual expenditure on this form of care to be more than £500 million (Killaspy et al., 2013). Given the economic constraints facing mental health care services in the UK, this investment in inpatient rehabilitation is currently at risk (Killaspy, 2019). Since length of stay in inpatient rehabilitation wards often exceed one year, this form of care is not only a major investment for the NHS but also a disruption in the life of the patient (Killaspy et al., 2016). In fact, some have raised a concern that lengthy hospitalisations on rehabilitation wards may be a form of 'reinstitutionalisation' (Edwards et al., 2016). On the other hand, inpatient rehabilitation may potentially allow the most disabled patients with psychosis to recover key domains of functioning and lead to long-run savings in health care costs after a high, up-front investment (Liu et al., 2011). Recognising the importance of rehabilitation for adults with complex psychosis, NICE has commissioned the development of a guidance on this topic which is due to be published in 2020 (NICE, 2018).

1.6.2 Systematic review of economic evidence for psychiatric inpatient rehabilitation compared to any other care for people with psychosis

1.6.2.1 Methods

Given the importance of this health care decision problem shown in the previous section (1.6.1) and that, to my knowledge, no review of the economic evidence existed at the time, I undertook a systematic review of such evidence. No standard or consensus definition of systematic reviews exists but a systematic review of such definitions by Martinic et al. (2019) concludes that systematic review typically involves (i) a research question, (ii) sources that were searched with a reproducible strategy, (iii) inclusion and exclusion criteria, (iv) selection methods, (v) critical appraisal and reports of study quality, (vi) information about data analysis and synthesis. I conducted a review to identify studies that inform the decision problem described in the previous section (1.6.1) that meets these criteria and will therefore refer to it as a systematic review. More specifically, my research question was "In people with psychosis, what is impact of inpatient rehabilitation compared to any other form of care on cost and/or service use and any measure of benefit to patient reported alongside (if any) according to any quantitative research design?" I did not undertake a Cochrane review.

I included studies that assessed the impact of mental health inpatient rehabilitation on any type of service use and/or costs compared to any other intervention and were published in English since 1990. I excluded studies in which there was evidence that less than 50% of the patient population had a diagnosis of psychosis and studies set in a low- or middle-income country according World Bank classification (World Bank, 2019). I included all quantitative study designs. Put differently, the population (P) were mental health patients, at least 50% of whom were required to have had the diagnosis of psychosis, the intervention (I) was mental health inpatient rehabilitation in a high-income country, the comparator (C) was any other form of care in a high-income country, and the outcomes (O) of interest were the impact of inpatient rehabilitation on any type of service use and/or costs along with any measure of patient benefit (or just service use and/or costs) using any quantitative study design (S). By necessity, I excluded studies for which the full text was not accessible. For

clarity, I also excluded studies that reported preliminary results of another study that met the inclusion criteria and I removed duplicate studies.

I searched three databases: Embase, PsychInfo and Medline. The search strategy was made up of four elements:

(i) search terms to identify studies in the field of psychosis which I derived from the systematic review for the NICE (2014a) guidance for psychosis in adults. This is an extensive list of more than 30 phrasings to refer to people with psychosis or schizophrenia, including variants of core terms (e.g. psychosis* and psychotic*), variants of more general terms for this patient group (e.g. (chronic or severe) adjacent to the world mental) and symptoms specific to this patient group (e.g. neuroleptic malignant syndrome) as well as their associated terms (e.g. Medical Subject Headings heading (MeSH) terms in MEDLINE)

(ii) Search terms to identify health economic evaluation developed by NHS EED (Glanville et al., 2009). This filter has been shown to perform well, achieving over 99% sensitivity against a gold standard. Briefly, it is made up of more than 10 variants of terms which containing the words ‘economic’, ‘costs’ and their associated terms (e.g. MeSH terms in MEDLINE) are typically used in the context of health economic evaluation (e.g. pharmacoeconomics) and aims to reduces false positives by excluding occurrences which are likely to refer to other fields of research (e.g. studies in which the word ‘costs’ is used near the word ‘metabolic’).

(iii) Search terms related to service use, namely hospitalization, readmission, hospital admission, health care utili*, health service util*, health care use, health care service use, health service use as well as associated terms (e.g. MeSH terms in MEDLINE)

(iv) terms related to inpatient rehabilitation, namely (inpatient or hospital or unit or ward) and (rehabilitation or recovery or "high dependency"). These terms were informed by my clinical collaborators.

To be included studies needed to have keywords comprised by (i), (iv) and either (iii) or (ii) or both (iii) and (ii). I first removed duplicate titles from my search, then screened title and abstracts of studies with help of the software Rayyan and finally checked whether the papers fulfilled the inclusion and exclusion criteria by examining the entire paper (Ouzzani et al., 2016). I ran the final search on 4 October 2019. In addition, I examined the list of references in studies fulfilling the inclusion criteria and studies that cited studies fulfilling the inclusion criteria to identify papers that may have been missed by the aforementioned search terms, that is, I traced citations forwards and backwards. There was no second reviewer. I also extracted the quantitative data without a second reviewer using Microsoft Word 2010. In expectation that this review would soon be superseded by the systematic review undertaken as part of the NICE guidance on care for people with complex forms of psychosis, I judged the reliability of this approach to be sufficient (NICE, 2018). I use means and difference in means as the principal summary measures. The studies were too heterogeneous in their methods and the type of results they reported to meaningfully attempt a quantitative synthesis of their findings. I also did not perform any additional, e.g. subgroup, analyses. I assessed the risk of bias of studies at the study level using a narrative discussion. There appeared to be insufficient basis to attempt to assess risk of bias across studies.

1.6.2.2 Study characteristics and findings

The study flow-chart for the systematic review in Figure 2 shows that I identified nine studies that fulfilled my inclusion criteria. Table 1 provides an overview of the study designs and results of each study. To summarize, of these nine studies, seven were observational studies and two were randomized controlled trials. Three studies were set in the UK and two of the seven observational studies were prospective analyses. Six of the seven observational studies compare outcomes up to the point of admission to inpatient rehabilitation with outcomes following discharge from inpatient rehabilitation. In other words, they undertook uncontrolled, within-patient comparisons. This means that they compared inpatient rehabilitation with the care received prior to inpatient rehabilitation. In contrast to all other studies, Tarasenko et al. (2013), assess the impact of converting an

inpatient rehabilitation to an acute ward at the level of a health care provider rather than the average effect of inpatient rehabilitation at the individual level. Again, this comparison is uncontrolled. In addition to a within-patient comparison, Tsoutsoulis et al. (2018) compare post-discharge rehospitalisation rates in patients discharged from inpatient rehabilitation to matched controls who did not receive such care. In the two RCTs, on the other hand, inpatient rehabilitation is compared to assertive community treatment. In addition, the study by Nordentoft et al. (2010) features a third treatment arm which consists of standard care. To summarise, there appears to be some relevant variation in the comparators. The length of follow-up ranged from the duration of the rehabilitation stay only to two years after discharge from inpatient rehabilitation. Only the studies by Bunyan et al. (2016) and Tarasenko et al. (2013) measured the impact of inpatient rehabilitation on costs. Tarasenko et al. (2013) only measured the impact of the conversion of the ward on per diem cost of inpatient care. Conversely, Bunyan et al. (2016) only account for the cost of non-rehabilitative inpatient care. Neither study formally combined costs and outcomes in their assessment.

There is a marked difference in the conclusions drawn by the RCTs and the observational studies. The two RCTs suggest that inpatient rehabilitation leads to both higher service use than the control groups over the study follow-up and that it is less or equally clinically effective (Lafave et al., 1996; Nordentoft et al., 2010). In contrast, the authors of the observational studies conclude that their results support the provision of inpatient rehabilitation. In the uncontrolled comparisons, service use and health care costs are lower after discharge from inpatient rehabilitation than prior to admission to inpatient rehabilitation. Likewise, in the case of Tarasenko et al. (2013), service use and costs were lower prior to the conversion of the inpatient rehabilitation ward. Compared to matched controls, Tsoutsoulis et al. (2018) estimated that hospitalisation rates were not substantially different in the inpatient rehabilitation group. Given the diverging results of the observational studies and the randomized trials, and, as discussed further in Section 1.7, the fact that these two study designs often have quite different strengths and weaknesses, I will appraise the RCTs by Lafave et al. (1996) and Nordentoft et al. (2010) separately from the observational studies.

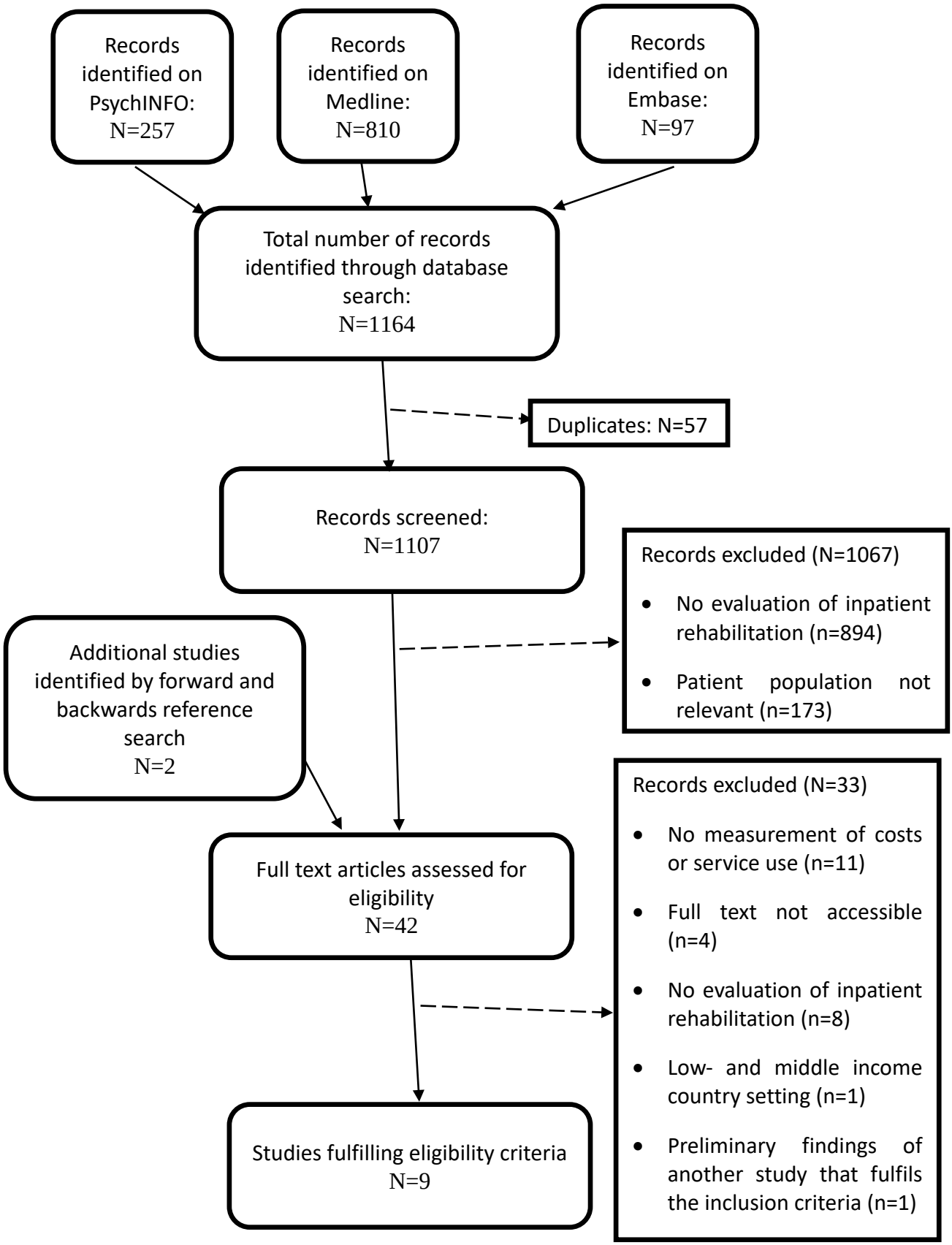


Figure 2 Study flow-chart for systematic review of evaluations of inpatient rehabilitation

First author (Publication year), Country	Study type; Comparison	Patients (Number of sites)	Main clinical effects	Service use or cost thereof
Bruseker and O'Halloran (1999), Australia	Prospective cohort study; within-patient (pre-post admission)	76 (1)	Comparing admission scores with 1-year follow-up scores, Brief Psychiatric Rating Scale scores improved by 18%, Scale for the Assessment of Negative Symptoms scores by 32%, and Quality of Life Scale scores by 19%.	Among people who had an admission in the year prior to admission to inpatient rehabilitation, both number of admissions were lower (35 vs. 98) and number of bed days (277 vs. 1050) in the year after discharge compared to the year before admission to IR
Bunyan et al (2016), United Kingdom	Retrospective cohort study; within-patient (pre-post admission)	22 (1)	Comparing period prior to admission with period after discharge, higher engagement in regular daytime activities (14% vs. 53% patients) and lower number of patients with record of one or more major risk incident (86% vs. 50%)	Average length of inpatient rehabilitation: 701 days; Lower in average cost of care (£66,000 vs. £18,000) and average bed days (380 vs. 111) when comparing two years before admission to IR with two years after discharge
Lafave et al.	Randomized	65 (1)	Patients in the assertive	Average length of inpatient stay over one-year

First author (Publication year), Country	Study type; Comparison	Patients (Number of sites)	Main clinical effects	Service use or cost thereof
(1996), United States	controlled trial; inpatient rehabilitation vs. assertive community treatment		community treatment group score higher on all dimensions of a subjective quality of life measure of the Quality of Life Interview, except health, and higher on an objective measure of quality of life. Brief psychiatric rating scale scores were comparable in the two groups. Ratings on the Client Satisfaction questionnaire were at the same levels across the group. Patients in the assertive community treatment group felt more in control of their lives according to ratings on the Environmental Index	follow-up: 39 days (assertive community treatment group), 256 days (inpatient rehabilitation group. Proportion of patients living in a community setting higher throughout follow-up. At one-year follow-up, more patients lived independently (50% vs. 20%) and more lived in a supervised setting (50% vs. 35%) and fewer were hospitalized (0% vs. 45%) in the assertive outreach compared to the inpatient rehabilitation group. Over a two-year follow-up the percentage of clients living in community settings was found to be consistently higher in those randomized to assertive community treatment

First author (Publication year), Country	Study type; Comparison	Patients (Number of sites)	Main clinical effects	Service use or cost thereof
MacPherson et al. (2017), United Kingdom	Prospective cohort study; within-patient (pre-post admission)	43 (1)	Minimal improvements in Health of the Nation Outcome Scale ratings; improvements in Camberwell Assessment of Need domains related to social functioning; Slightly higher fraction in employment (data not shown)	Average length of inpatient rehabilitation: 380 days; Higher share of patients living in own tenancy prior to compared to after admission (44% vs. 74%)
Nordentoft et al. (2010), Denmark	Randomized controlled trial; hospital-based rehabilitation vs. standard care vs. assertive community treatment	94 (2)	Difference of -0.34 (95% Confidence Interval -1.2 to 0.5) on the Scale for Assessment of Positive Symptoms and a difference of -0.45 (95% Confidence Interval -1.4 to 0.5) on the Scale for Assessment of Negative Symptoms in favour of those receiving assertive	Those randomized to hospital-based rehabilitation had highest number of bed days and number of days in supported housing during almost all points of a five-year follow-up whereas those hospitalized to assertive community had the lowest whereas those receiving standard care were in between (approximately 550 vs. 290 vs. 400 days over 5 years). With respect to number of days in supported housing the ranking was the

First author (Publication year), Country	Study type; Comparison	Patients (Number of sites)	Main clinical effects	Service use or cost thereof
			community treatment compared to those receiving hospital care	same (approximately 420 vs. 220 vs. 240 days over 5 years)
Petrie and Mountain (2009), United Kingdom	Retrospective cohort study; within-patient (pre-post admission)	35 (1)	Number of Mental Health Act uses lower (2.1 vs. 0.5) in the 2 years prior to admission compared to 2 years after discharge	Average length of inpatient rehabilitation: 668 days; Lower in the average number of bed days (478 vs. 116) and lower number of admissions (2.5 vs. 1.2) in the 2 years prior to admission to inpatient rehabilitation compared to 2 years after discharge
Sullivan et al. (1991), United States	Retrospective cohort study; within-patient (pre-post admission)	Not stated (1)	75% of relatives agreed or strongly agreed that the program made significant improvements to their lives or the looked after person; Post-discharge, quality of life is stated to have improved over the past year	Reduction in time in hospital by 70% when comparing the two years prior to admission with the two year after admission
Tarasenko et al. (2013),	Retrospective cohort study;	Not stated (1)	Higher use of restraints or seclusion incidents in the	Cost per diem costs on ward ($391 vs. $503/510); Lower discharge rates seven months and two

First author (Publication year), Country	Study type; Comparison	Patients (Number of sites)	Main clinical effects	Service use or cost thereof
Canada	within-system (pre-post transformation of an IR ward to an acute ward)		transition period and after closure of inpatient rehabilitation ward compared to period prior to closure	years after closure of IR ward compared to three years prior to closure (0.162 vs. 0.125 vs. 0.192); Higher use of crisis centre in the year after IR closure and two years after compared to two years prior to closure (850 vs. 1100 vs. 700 days); Fewer discharges to less restrictive settings by community rehabilitation programme in the 3 years after closure compared to 5 years before closure (9 vs. 5 cases per year)
Tsoutsoulis et al. (2018), Australia	Retrospective cohort study; Within-patient (pre-post admission) and with matched controls	504 (1)	Not applicable	Average length of inpatient rehabilitation: 125 days; Lower rate of admission (100% vs. 33%), lower number of admissions (1.5 vs. 0.6) and longer time to admission (110 vs. 152 days) when comparing 12-month period prior to inpatient rehabilitation with 12-month period after discharge; Higher rate of admission (33% vs. 31%), higher number of admission (0.6 vs. 0.5)

First author (Publication year), Country	Study type; Comparison	Patients (Number of sites)	Main clinical effects	Service use or cost thereof
				and short time to admission (152 vs. 171 days) when comparing 12-months after discharge with matched controls

Table 1 Overview of evaluations of inpatient rehabilitation

1.6.2.3 Appraisal of the randomized trials

A strength common to the identified RCTs is that due to random allocation of treatment to trial participants, it is expected that both observed and unobserved factors are balanced out between groups at baseline, making it more likely that the findings are internally valid. Both RCTs also provide a clear description of the treatments that are being compared which aids interpretation of the findings. Given that in both trials at least one of the treatment groups is smaller than 30, a common weakness of the RCTs is that the results may be affected by chance bias (Torgerson and Torgerson, 2003). In addition, measurement of the benefits of different treatments to the patient is limited, the studies do not include measures that allow for the cost-effectiveness of inpatient rehabilitation to be compared to investments elsewhere in the health care sector, such as quality adjusted life year, nor is the impact of inpatient rehabilitation on the family and carers of patients considered.

A strength that is specific to the trial by Nordentoft et al. (2010) is that with a length of follow-up of five years, it is much more likely to capture an adequate proportion of the potential benefits of inpatient rehabilitation which assess outcomes over a follow-up period of at most two years after discharge for inpatient rehabilitation. The robustness of its findings is also strengthened by blinding of interviewers to treatment allocation and use of registry data to reduce loss to follow-up. However, the results of the study by Nordentoft et al. (2010) are unlikely to be of high relevance to decision making in the United Kingdom because the authors assesses the value of inpatient rehabilitation for people with first-episode psychosis as opposed to people with enduring forms of psychosis, that is for a patient group which may have a lower ability to benefit from an intensive long-term treatment like inpatient rehabilitation.

The patients in Lafave et al. (1996), on the other hand, appear to be more comparable to the population accessing inpatient rehabilitation in the UK. However, the length of follow-up, two years for the proportion of patients living in community setting and one year for all other outcome measures, is limited. In addition, the authors do not base their analysis on patients who were randomized but the patients who accepted their assigned treatment. Since rates of acceptance were relatively low and differed between the treatment arms

(53% for those randomized to inpatient rehabilitation and 72% for those assigned to assertive community treatment) this may have led to bias in the analysis.

1.6.2.4 Appraisal of the observation studies

One of the strengths of the identified observational studies is that three of them are set in the UK and therefore more likely to be generalizable to UK decision making. In addition, all three studies set in the UK have been published within the last 10 years which further adds to their generalizability. (2018)However, the observational studies also have some considerable limitations:

Confounding: One of the main weaknesses of the observational studies is that they all implicitly assume that there is no time-variant confounding in their before-and-after analyses. In other words, it is assumed that service use and patients' clinical status would have remained unchanged had patients not been admitted to inpatient rehabilitation (or had the inpatient rehabilitation ward not been transformed in Tarasenko et al. (2013)). Referrals for inpatient rehabilitation are typically made following a period during which the patient has been particularly unwell and after one or more hospitalisations. Given the episodic and fluctuating nature of psychosis, it is possible that outcomes would have improved regardless of inpatient rehabilitation care, thus potentially overestimating the relative effectiveness and cost-effectiveness of inpatient rehabilitation. Similarly, the analysis by Tarasenko et al. (2013) at the organisation level may be affected by temporary fluctuations, long-term trends or changes in care provision unrelated to the conversion of the inpatient rehabilitation ward.

The fact that hospitalisation rates do not differ between the inpatient rehabilitation group and a matched cohort in Tsoutsoulis et al. (2018), illustrates the impact that different assumptions with respect to confounding can have in the observational studies . Yet, the implicit assumptions underlying the matched analysis in Tsoutsoulis et al. (2018) are potentially less plausible than those in the before-and-after analysis because the authors only match for the patients' gender, age, and primary diagnosis. Not only is this set of variables too limited to plausibly account for confounding in this context more generally, but the authors report that there is an observed imbalance between the groups in terms of

prior service use which they do not match for. It is known that prior service use is one of the strongest predictors of future service use in psychosis so the matched comparison is likely to be biased against inpatient rehabilitation (Sernyak and Rosenheck, 2003). In fact, contrary to their initial motivation for identifying a control group stated in their introduction (i.e. to reduce potential bias in the before-and-after designs), in their discussion section, Tsoutsoulis et al. (2018) themselves do not appear to give credence to the assumptions needed to interpret the comparison between IR and matched controls as a causal effect. Instead, they interpret equal rates of hospitalisation over the follow-up as showing that hospitalisation rates in the inpatient rehabilitation group return to "normative levels" after receiving rehabilitation care, i.e. in a descriptive fashion.

Generalizability of evidence: An implicit assumption that is likely to be violated in many of the identified observational studies, is that the samples used in the analyses are representative. Bunyan et al. (2016) excluded patients who were admitted for less than 6 weeks from the analysis and only included patients who were discharged rather than all those admitted within the study period, MacPherson et al. (2017) excluded patients with a length of stay of less than 3 months, Petrie and Mountain (2009) excluded patients who were not discharged by the end of the study period and, in Bruseker and O'Halloran (1999), 50% of patients dropped out of the study. At least two of the observational studies appeared motivated by cuts to the funding to inpatient rehabilitation services or the risk thereof (Bunyan et al., 2016; Tarasenko et al., 2013). This may have led to a bias across studies in favour of inpatient rehabilitation if this meant that studies suggesting a positive effect of inpatient rehabilitation were more likely to be published.

Scope of measurement: Even if one is willing to disregard these weaknesses, the extent to which findings are informative to current decision making in the UK appears limited. One reason for this is the restricted scope of measurement: (i) Five out of seven studies do not incorporate the length and cost of stay on inpatient rehabilitation wards themselves in their evaluations. While in some studies it is possible to retrospectively adjust the analysis for this omission, given the extensive length of stays and the fact that at approximately £350 per bed day, the unit cost of inpatient rehabilitation in England is similar to that of staying on a general acute ward, this is a significant weakness of (CQC, 2018; NHS improvement, 2018); (ii) With follow-ups of at most two years, the length over which benefits and costs are

measured of the included studies is relatively limited for a chronic condition like psychosis. (iii) As with the RCTs, the measurements of benefits of treatment are limited and difficult to compare across studies.

Relevance of estimated quantities: Another reason that limits the usefulness of the identified observational studies to current decision-making is that the relevance of the estimated quantities is unclear or reduced for a number of reasons: (1) The treatments are not well-characterised, that is it is unclear to what inpatient rehabilitation and the comparator treatments entail (VanderWeele and Hernan, 2013); (2) In the case of the within-patient before-and-after analyses, the implicit comparator to inpatient rehabilitation is the mix of care received prior to inpatient rehabilitation admission. However, this may be a suboptimal comparator because the care received in the period prior to inpatient rehabilitation may not have included realistic alternatives to inpatient rehabilitation, such as care by community rehabilitation teams while staying in 24-hour supported housing; (3) A conversion of inpatient rehabilitation wards to acute wards without expansion of relevant community services as considered by Tarasenko et al. (2013), is unlikely to be an appropriate policy alternative because it does not replace inpatient rehabilitation by a service that caters to the particular needs of people with persistent and complex forms of psychosis; (4) Outcomes are not measured at regular time-points from the point at which the decision maker chooses a treatment option. Instead, most studies, focus on follow-up after discharge. For example, a study might measure outcomes one year after discharge from inpatient rehabilitation rather than one year after admission or referral to inpatient rehabilitation. This implies that the period between admission and the measurement of outcomes will depend on the length of inpatient rehabilitation stay which differs for every patient in the study. I would argue that the estimated parameters, do not directly correspond to what decision makers are likely to be interested in, namely what consequence a current action has over a given time period rather than a variable time period. (5) As illustrated in Figure 3, the pathway to inpatient rehabilitation begins at the point of referral not at the point of admission to inpatient rehabilitation. In contrast to acute psychiatric care, where, in most cases, only hours or a few days elapse between referral and admission, in the context of inpatient rehabilitation, the distinction between referral and admission is meaningful because, to my knowledge, it is more common for long

periods of time to pass between referral and admission to inpatient rehabilitation or exit from the pathway to inpatient rehabilitation. Yet, most of the observational studies start follow-up at the point of admission and only include patients who have been admitted. This has two implications: First, since patients most commonly stay on an acute psychiatric ward while waiting for assessment or a rehabilitation bed to become available, potentially costly delays may be omitted from the economic evaluation. Second, the characteristics of the patients that are ultimately admitted to inpatient rehabilitation is likely to be different from patients who are referred to inpatient rehabilitation because they have both undergone a selection process and because time has passed between referral and admission to inpatient rehabilitation. Thus, the effect of admitting patients to inpatient rehabilitation compared to not admitting them does not correspond to the effect of referring compared to not referring them in a hypothetical world in which it is possible to determine whether a patient is suitable for inpatient rehabilitation at referral and admit them immediately to inpatient rehabilitation without investing additional resources in service provision. Put differently, I would argue a more policy-relevant question appears to be the effectiveness and cost-effectiveness of referring patients to inpatient rehabilitation rather than the cost-effectiveness of admitting patients to inpatient rehabilitation because, if rehabilitation wards were closed, patients would not be referred to begin with.

1.6.2.5 Limitations of the systematic review

As indicated above, when undertaking this review it was clear that it would soon be superseded by the systematic review undertaken as part of the NICE guidance on care for people with complex forms of psychosis (NICE, 2018). Therefore, I judged it to be a better use of time to limit myself to the core features of this review and invest more attention on the aspects of this thesis that would be more likely to have a potential to reach a broader audience. However, this also means that the results of this review should be regarded with some caution. It is possible that studies in the nonpublished literature have been overlooked by the choice of search strategy, however, given the low quality of research identified in the published literature it appears unlikely that the main conclusions of this review would have seen major change. Evidence suggests that, typically, about 5-10% more

studies would have been identified if a second reviewer had participated in the screening of studies (Stoll et al., 2019; Waffenschmidt et al., 2019). Similarly, mistakes in the data extraction process are more likely to have occurred. I also did not attempt to rate the quality of the evidence using a formal, quantitative approach such as ROBINS-I or undertake a meta-analysis (Sterne et al., 2016). I deliberately limited this review to studies containing economic evidence but this excluded evaluations that only considered the clinical benefit of inpatient rehabilitation. Thus, this review provides only an incomplete appraisal of the evidence supporting inpatient rehabilitation. By the same token, qualitative research may have provided a theoretical basis for how inpatient rehabilitation brings about change. Finally, I was unable to assess the risk of bias across studies.

1.6.2.6 Conclusion

Previous observational studies unanimously suggest that their evidence supports the provision of inpatient rehabilitation. However, these findings are likely to be significantly biased in favour of inpatient rehabilitation because these studies typically exclude the cost of inpatient rehabilitation, use before-and-after analyses and because they are based on selective samples. By contrast, the two RCTs identified suggest that inpatient rehabilitation is expensive and does not appear to yield meaningful improvements in patient outcomes but the settings and time horizons of these trials are very different from the decision making problem encountered in the UK so it is unclear whether these findings can be generalised.

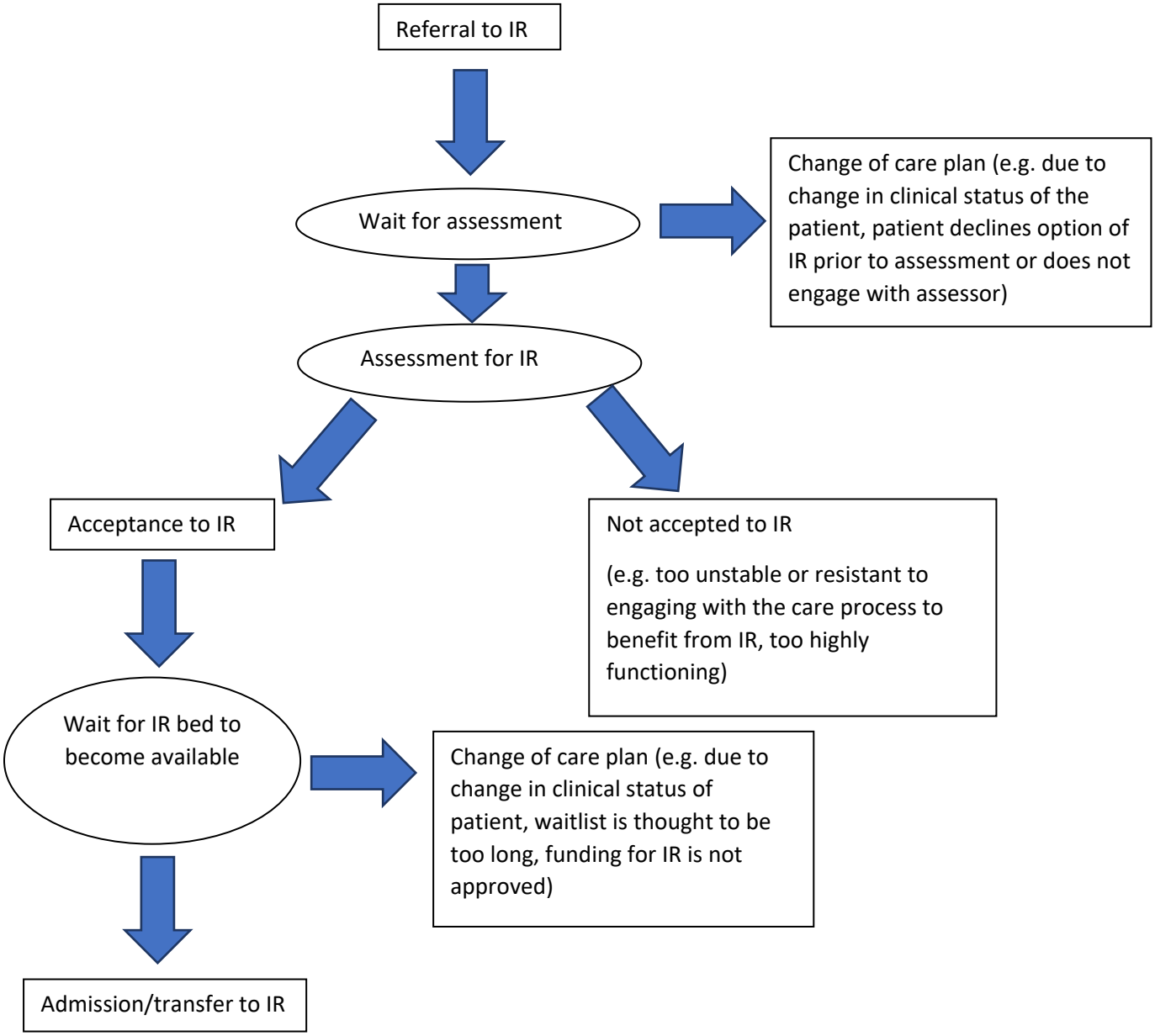


Figure 3 Schematic diagram of pathway to inpatient rehabilitation (IR)

1.7 Approaches to the economic evaluation of inpatient care

The importance of the choice between the main alternative inpatient options in the UK for young people with psychosis (adult versus specialised wards for children and adolescents) and adults with persistent forms of psychosis (inpatient rehabilitation versus usual care) coupled with the limited available evidence, calls for further research to support resource allocation decision making. As indicated above, conducting an RCTs could be one possible way to generate new evidence for value for money of inpatient care for people with psychosis. The key advantage of RCTs is that, due to random allocation of treatment to trial participants, estimates are likely to have high internal validity. Furthermore, RCTs are both conceptually simple, require minimal assumptions and their methodology is both familiar to and well-accepted by decision makers (Buxton et al., 1997; Higgins et al., 2008).

However, on average, randomisation only leads to balance between the groups at baseline. If there are unobserved factors that influence dropout from the RCT or treatment switching then the internal validity of trials is reduced (Latimer et al., 2014; Szymczynska et al., 2017). In addition, even if RCTs are internally valid, they may not generalise to other settings which they are meant to inform, resulting in a lack of external validity (generalisability of the results to other settings or contexts). One reason for this is that, in an RCT, treatments may be delivered in a way and in an environment that differs from how these treatments would be implemented in routine practice and thus may not be relevant to the decision making context (Baltussen et al., 1999). Secondly, the treatment effect in those taking part in trials may differ from that in the wider patient population in unobserved ways. For example, in the context of psychosis, participants in RCTs may be different from non-participants because there is evidence to suggest that clinicians have a tendency to suggest patients for participation in RCTs that are less unwell (Patel et al., 2017). Moreover, some patients with psychosis lack the capacity to give informed consent and suicidal patients are typically excluded from RCTs on ethical grounds (Wilson and Stanley, 2006). In an RCT set in inpatient rehabilitation wards conducted by Killaspy et al. (2017), for example, 87% of mental health trusts in the UK were willing to participate in the study. Among the patients in these trusts, 13% lacked capacity to give informed consent to participate in the study, 10% were unavailable at the time of research interview and 16% declined to participate, leaving only 62% being included in the study. In the case of the evaluation of CAMHS services, if patients

are underage, parental consent may be required for participation in an RCT which could create a barrier to enrolment into a study.

A further difficulty with RCTs relates to issues around feasibility. For example, it is unclear whether randomising underage patients to adult and CAMHS inpatient care would be ethically feasible because, as discussed above, there does not appear to be clinical equipoise. The extension of CAMHS inpatient care to patients above the age of 18, on the other hand, may be practically difficult to implement in an RCT because this would involve a significant restructuring of service provision. A prospective observational study could avoid these issues around feasibility but would inherit all other weaknesses of RCTs and loose the strengths RCT gain through means of randomization.

When they are feasible, RCTs tend to be time-consuming and expensive. For example, in 2005 prices, RCTs funded by the Health Technology Assessment (HTA) program cost approximately £1.3 million on average (Raftery et al., 2015). To save costs, the length of follow-up is often relatively short, RCTs may not compare all relevant treatment options and/or sample sizes may be low. This may mean that they do not capture all relevant costs and benefits of the treatment or that parameter estimates have high variance. Low sample sizes are particularly problematic in economic evaluations because cost data is typically more variable and non-normally distributed than clinical outcome measures, thus larger sample sizes are generally required to obtain precise estimates of differences in costs as compared to differences in effects (Mihaylova et al., 2011; Petrou and Gray, 2011). Concerns regarding the cost of conducting an RCT appear particularly significant with respect to the economic evaluation of inpatient rehabilitation. The median length of stay on rehabilitation wards has been previously reported to be more than 16 months (Killaspy et al., 2016). Moreover, the benefits of IR are thought to mostly accrue after discharge from hospital and inpatient care is the most expensive treatment class in schizophrenia. Therefore, an RCT with a follow-up length of several years would be required for inpatient rehabilitation to have a realistic chance of being shown to be cost-effective.

Since inpatient rehabilitation and CAMHS inpatient care are already provided in routine clinical practice, another option is to carry out economic evaluation based on observational data. If routinely collected data from sources such as EHRs is used, observational studies can be undertaken at a far lower cost than RCTs. Observational studies based on routinely

collected data can also act as a precursor to an RCT. For example, they can be used to get an idea of the variance in the outcome variable and, with due caution, the approximate magnitude of the treatment effect. Since consent is on an opt-out basis, samples drawn from EHRs are more likely to be representative and thus generalizable, loss to follow-up may be lower or less problematic than in RCTs, and studies based on EHRs can lower the administrative burden to study participants (Franklin et al., 2017). Specifically in relation to service use data, three benefits of EHRs compared to conventional approaches are: (i) they have the potential to yield more precise service use and cost estimates because, unlike self-report service use questionnaires commonly used in RCTs, they do not suffer from recall issues and often contain a more specific description of the service that was used (e.g. health care resources (HRG) codes) allowing more accurate costing; (ii) EHRs enable the analyst to more easily test whether alternative decisions with respect to the cost perspective impact the results of the economic evaluations (Asaria et al., 2016); and (iii) in the long-run, the research cost in terms of cleaning and costing the service use data are reduced because the same code for data processing can be reused for multiple studies based on the same database.

Typically, the greatest concern when conducting observational studies is that their internal validity may be compromised, i.e. estimated treatment effect may be biased, because variables that have both an effect on the treatment decision and the outcome are unmeasured, a phenomenon that can be referred to as unmeasured confounding (Hernán and Robins, 2018). Observational studies not based on routinely collected data and prospective observational studies share the disadvantage of RCTs in that they can be similarly costly, untimely and potentially limited in terms of the scope of measurement or comparison. The principle disadvantage of routinely collected data is that it is not recorded for use in research but for clinical or administrative purposes. In general, this means that variables that are of importance to the analysis may be completely missing or the recording quality, recording styles and incentives to record different kinds of data can vary across fields, clinicians, services and time. Of course, there is also uncertainty in the validity the more traditional self-report resource-use measurement but the validity of data in large databases can be more difficult to judge because consumers of research may be less familiar with the particular database used for a research study and because it is more difficult to

build in quality checks in the data collection process (Thorn et al., 2013). Measurement error resulting from data quality issues can lead to distortions in the estimation of the quantity of interest and greater uncertainty in the estimates (Buonaccorsi, 2010). In addition, in contrast to typical RCTs or prospective cohort studies, patients' clinical status in EHRs is often not monitored at systematic time intervals (Kreif et al., 2018). In terms of service use data, the main disadvantages of EHRs compared to self-report data are: (i) one cannot easily distinguish between patients with missing service use data and patients who did not access a service; (ii) any given set of EHRs is likely to collect reliable data only on a limited subset of services and over a limited geographic area. Thus, they cannot replace self-report methods when trying to capture societal costs such as unpaid care; and (iii) whereas the cost of working with self-report data is well understood and under the control of the researcher, this is not necessarily the case with EHRs (Franklin et al., 2017). For these reasons, understanding the quality of the data and harmonising variables can be one of the most time-consuming aspects of conducting economic evaluations based on data from EHRs. On a more practical level, issues around information governance can be challenging and/or time-consuming when working with EHRs. A less apparent difference between RCTs and observational studies is that since RCTs are more established, methodological rigor is often higher than in observational studies (Dhiman et al., 2020; Kreif et al., 2013b; Velentgas et al., 2013). Problems with the design and analysis of observational studies are thus more likely, and some have argued that these differences compared to RCTs can be as important as the lack of randomization (Hernán et al., 2008; Rubin, 2008).

Ultimately, the credibility of results obtained by observational studies compared to RCTs depends on the specific application at hand because, for each study, one needs to weigh beliefs about the potential impact of all of the aforementioned sources of bias (e.g. the potential bias due to unmeasured confounding in an observational study against the bias due to a potentially small sample size and unrepresentative sample in an RCT) (Imai et al., 2008). To empirically inform such an exercise, a stream of research has emerged which makes use of a research design referred to as within study comparison (WSC), which involves the comparison of the treatment effect estimates from RCTs with estimates from observational studies that share the same target population (Wong and Steiner, 2018). As explained by Wong et al. (2018) there are at least five difficulties in interpreting WSCs: (1)

Particularly in early WSCs, RCTs and their corresponding observational study differed in ways other than assignment of treatment (e.g. the way that the outcomes were measured) and (2) different causal quantities or estimands were targeted by the different types of studies; (3) Although RCTs typically serve as the benchmark for evaluation the performance of observational studies (e.g. due to differential attrition), they may not provide valid treatment effects themselves; (4) There is a lack of consensus as to what metrics to use to evaluate the performance of observational studies; (5) It is unclear to what extent the results from WSCs generalise to different contexts (e.g. different populations, treatments and study designs) and the pool of existing WSCs is relatively limited to date, particularly in the context of health care treatments. With these caveats in mind, the authors of recent reviews of WSCs that attempt to reduce the impact of threats to their validity indicate that, under certain circumstances, well-conducted observational studies can achieve a high level of internal validity (Chaplin et al., 2018; Cook et al., 2008; Weidmann and Miratrix, 2020). These circumstances include studies using regression discontinuity designs and studies matching on rich covariate information within the same geographic area, the types of studies that make up this thesis and will be described in more detail in the subsequent chapters. Distinct from the empirical evidence on the credibility of observational studies compared to RCTs, it is also worth considering the relative perception of these two sources of evidence by decision makers since this affects their relative potential to affect decision making. The current manual for developing NICE guideline as well as related NICE publications state that RCTs are the preferred source of evidence to assess the effectiveness of an intervention, that observational studies may also be used for this purpose and that observational studies should be routinely included in a systematic review of the evidence (Bell et al., 2016). This methodological guidance is currently under development. The preference for RCTs, where feasible, for estimates of effectiveness is likely to remain but the role of evidence from observational studies should be clarified (NICE, 2020). However, not all decision makers share these views, particularly in the context of local decision making (Cairney and Oliver, 2017).

To summarise, as common in the evaluation of service-level intervention, practical considerations of implementation, timeliness and high costs, as well as concerns around external validity and length of follow-up make observational studies a relevant alternative

to RCTs for the decision problem that are of interest in this thesis despite the potential risks around their internal validity (Meacock, 2018; Sutton et al., 2018)

1.8 Aim, objectives and structure of the thesis

Given these considerations, the aim of my thesis is to assess the value for money of specialist inpatient care for people using specialist inpatient care for people with psychosis based on single studies and data derived from electronic health records. My specific objectives were:

(1) To identify approaches to handle unmeasured confounding and measurement error
(2) To conduct an economic evaluation of admission to CAMHS inpatient care compared to admission to adult wards for young people with psychosis (Analysis 1)
(3) To conduct an economic evaluation of referral to inpatient rehabilitation compared to usual care for adults with persistent forms of psychosis (Analysis 2)

The structure of the thesis parallels with these three objections, that is in Chapter 2 I address objective 1, in Chapter 3 I address objective 2, in Chapter 4 I address objective 3. In the final chapter (Chapter 5), I conclude with an overarching discussion of this thesis, including the main findings of the thesis, its implication for policy making, strength and limitations, and potential areas for future research.

1.9 Personal contribution to the thesis

The work in this thesis was funded by internal grant money from King's Health Economic at King's College London and from the Maudsley Biomedical Research Centre (BRC). It was not earmarked for a specific topic. I, the candidate, therefore, took the lead in shaping the direction of this PhD. I conceived the idea behind Analysis 1 and developed Analysis 2 based on the suggestion of David O'Flynn. I explored the database with advice from my clinical collaborators, Alexander Tulloch, David O'Flynn, Matthieu Crews, Shahzad Alikhan, Tom Craig and Johnny Downs, my supervisor Richard Hayes and many other people working at the BRC. I wrote the scripts to extract the data and manually coded all but a small

proportion of the data from free text notes. I made use of natural language processing applications previously developed by researchers at the BRC and, where noted, of versions of data previously 'cleaned' by Alexander Tulloch. Alexander Tulloch and I also jointly worked on costing CRIS service use. I was responsible for developing the analysis plans for all analyses. The clinical collaborators mentioned above provided advice on clinical and policy aspects of the two evaluations. I carried out all statistical analyses and interpreted the findings. My supervisors provided feedback on some of the chapters and some of the analyses. To make the distinction between work predominately undertaken in collaboration or views formed through discussion and the work for which I was largely responsible clearer, in the following chapters, I will use the first personal singular and plural pronouns, i.e. 'I' and 'we', as I judged appropriate.

Chapter 2 Handling unmeasured confounding and measurement error

2.1 Introduction

In the previous chapter, I identified unmeasured confounding and measurement errors as two potential threats to the validity of economic evaluations using observational data. The aim of this chapter is to identify potential strategies to mitigate these threats. To provide some context, in Section 2.2, I briefly describe the available data sources to set the context which led me to investigate these methodological issues. I then justify my focus on unmeasured confounding, outline common approaches to handling this methodological issue in previous health economic evaluations and describe a novel alternative (Section 2.3). In Section 2.4, I discuss my rationale for examining methods to handle measurement errors, review potential measurement error assumptions, and discuss strategies to support and enable these assumptions. In preparation for the analyses in the following two chapters, in each case, I will outline which approaches I used for the evaluation of inpatient care for young people (Chapter 3) and the evaluation of inpatient rehabilitation (Chapter 4).

2.2 Data sources

2.2.1 Clinical records interactive search (CRIS) database

My main source of data was the South London and Maudsley (SLaM) Clinical Record Interactive Search (CRIS) database at the Maudsley Biomedical Research Centre (BRC) which was set up in 2008 (Stewart et al., 2009). The SLaM NHS Foundation Trust is a provider of secondary mental health care and CRIS contains a de-identified version of all but a small part of SLaM's system of electronic health records (Perera et al., 2016). The SLaM catchment area covers four administrative units, known as boroughs, in the south of London, United Kingdom (see Figure 4). It includes both urban areas of high deprivation (mainly in the boroughs of Lambeth, Lewisham, Southwark) and more affluent suburban locations (mainly in the borough of Croydon). SLaM also accepts some referrals from outside of its catchment area.

The use of CRIS is subject to a security and governance framework and the responsible oversight committee approved the use of CRIS for the two economic evaluations making up this PhD (project #13-090 and #15-004) (Stewart et al., 2009). While the CRIS database has been used in more than 150 published studies, the data quality has not been comprehensively investigated and no data dictionary exists. Therefore, in Appendix A, I note what is known about the quality of variables that will be relevant for the evaluations in the next two chapters and undertake some additional assessments of data quality.

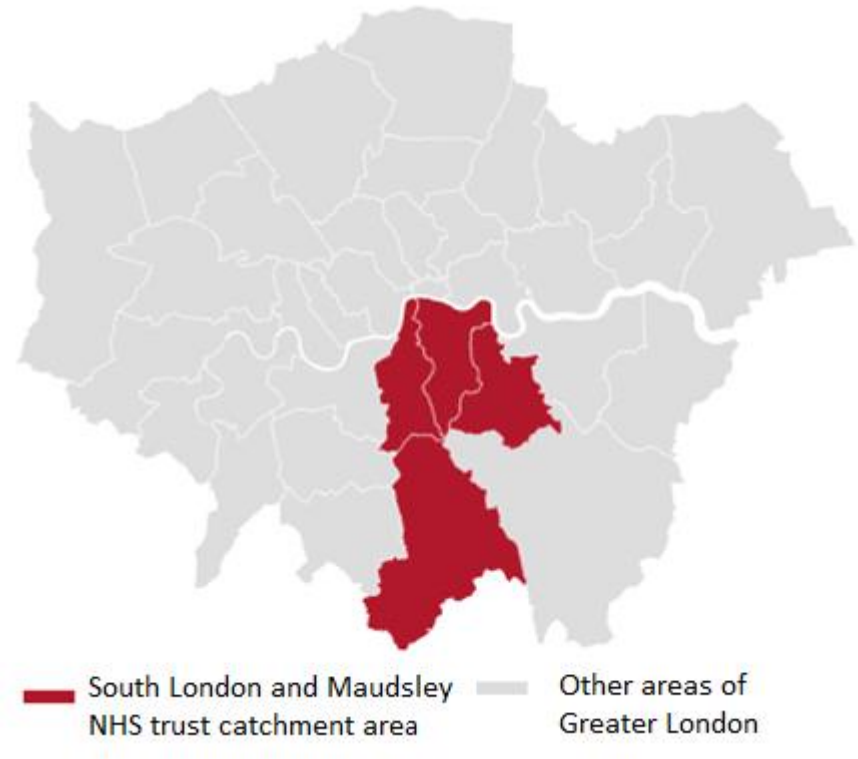


Figure 4 Map of areas covered by the CRIS database within Greater London

2.2.2 Data extraction

One of the key ways in which CRIS differs from more commonly used databases of routinely collected data that are used for health economic research, such as Hospital Episode Statistics (HES) or Clinical Practice Research Datalink (CPRD), is that it allows the researcher to access two types of data: (a) conventional data from 'structured fields', meaning data which is readily available for use in research because it is stored in a known, simple and,

typically, numerical format (e.g. age, ethnicity and service contacts); and (b) more uncommonly, the vast majority of SLaM's unstructured clinical notes. These clinical notes usually contain more extensive and nuanced information in text form (Stewart and Davis, 2016). A bespoke algorithm that masks patient identifiers in these clinical notes in the vast majority of cases (e.g. by replacing patient names in the clinical notes by 'ZZZZZZ') ensures that clinical notes are available for research while protecting the anonymity of the patient to a sufficiently reliable degree (Fernandes et al., 2013). In order to derive data from these unstructured clinical notes that is amenable for quantitative research, there are three potential approaches:

(i) Keyword or key phrase searching – Searches can be carried out to identify whether certain keywords or key phrases of interest are contained within a patient's clinical notes. A crude approach might, for example, be to determine the number of clinical notes containing the term 'psychosis', alternative terms (e.g. relevant diagnostic codes) and misspellings thereof, prior to the study baseline date and consider everyone who has at least one document containing these terms to have a diagnosis of psychosis at baseline.

(ii) Manual coding – The researcher can read the clinical records and manually code or validate the information of interest. Continuing with the above example, instead of only identifying the presence or absence of psychosis-related keywords, one could read the clinical notes going back in time starting from the baseline date and consider the context that the words were used in. For example, if a clinical note stated that "ZZZZZZ does not have a diagnosis of psychosis" then, in contrast to approach (i), one would code the patient as not having a diagnosis of psychosis. In addition, this allows the researcher to determine the temporality of the diagnosis, that is whether the diagnosis was still valid at baseline rather than whether the patient ever had a diagnosis of psychosis

(iii) Natural language processing (NLP) – NLP involves the application of computational techniques to analyse and synthesise information from the free text in CRIS (Stevenson, 2010). For example, if we consider the same, relatively simple example sentence ("ZZZZZZ does not have a diagnosis of psychosis") used

in (ii) the NLP developer may write an application which analyses this sentence based on grammatical or linguistic rules. Thus, the NLP application may be able to recognize that the possessive verb ('have') in this sentence relates to the object ('a diagnosis of psychosis') and that this possessive verb is negated ('does not') in relation to the subject of the sentence ('ZZZZZZZ'), the patient that this clinical note relates to. Thus, like approach (ii), NLP applications can, in principle, correctly recognize the above sentence as not being a diagnosis of psychosis. In contrast to approach (ii), however, once the application is developed, this approach can be scaled to evaluate the entire set of clinical notes in CRIS. Some NLP approaches are based on such grammatical or linguistic rules but most existing NLP applications for CRIS use machine learning approaches or approaches hybrid versions, that is, those combining the two approaches. At its most basic, machine learning approaches involve two steps: (1) the manual coding of a set of 'gold standard' annotations of the concept of interest; (2) use of algorithms that weigh linguistic 'features' of the text data in this 'gold standard' set to optimize predictive performance of classifications in the remaining documents. For example, if a sentence containing the word psychosis, also contains a reference to a family member, say the word 'father', the algorithm might be less likely to classify this sentence as an instance in which the patient was diagnosed with psychosis because in the 'gold standard' annotations sets, such sentences often describe the father as having psychosis. Jackson et al. (2017) and Patel et al. (2015) provide further detail on the operation of some the main sets of NLP applications that have been developed for CRIS.

For this thesis, I extracted both structured data, NLP data and data used for manual coding of terms from CRIS by writing queries in the Structured Query Language (SQL) version of the database.

2.2.3 Data linkages

Apart from allowing access to clinical notes, CRIS has also been linked to several other databases using deterministic linkage, which involves matching records using an identical

identifier or set of identifiers (Perera et al., 2016). Of relevance to the analyses in this thesis were existing linkages with HES and the Zaponex Treatment Access System (ZTAS).

HES is a database of electronic health records covering England and Wales that is widely used for health economic research (Sinha et al., 2013). The linkage with CRIS allows access to data on inpatient service use since 1997, outpatient service use since 2003 and emergency care since 2007 for patients who have a record in CRIS (as well as service use of non-SLaM patients while they were living in the SLaM catchment area). Approximately 90% of SLaM patients have been linked to HES. The remaining 10% have not been linked due to, in approximately equal proportions, failure in record linkage or the patient's choice to opt out from the linkage. For data protection reasons, without further ethical approvals, HES data is only available for people who have had a contact with SLaM while under the age of 18 if they subsequently had a contact with SLaM services after becoming adults.

ZTAS is a database created to monitor prescribing and blood values of users of Zaponex, the most commonly prescribed brand name version of the antipsychotic medication clozapine. The current linkage contains data on SLaM patients prescribed Zaponex between 2007 and 2016 but only if they were not prescribed Zaponex at another NHS trust at some later stage. A separate database exists for monitoring the use of Denzapine, the liquid version of clozapine, but this database is not linked to CRIS.

In addition to the existing linkages described above, my collaborators and I linked some data on private sector bed use, out of area care spells and financial data that we obtained from SLaM to supplement or examine the validity of existing data. Private sector bed use and out of area care spells were linked deterministically using patients' NHS ID by the Clinical Data Service Lead responsible for CRIS. We linked the financial data deterministically ourselves based on service names as described in more detail in Appendix B.2. For simplicity, in the rest of this thesis, I use CRIS as a shorthand to refer to the combination of data sources described in this section.

I also considered making use of linkage to Lambeth DataNet, a clinical database from general practices in Lambeth, one of the four boroughs that SLaM covers. However, no equivalent database exists for the other three boroughs, the quality of this dataset is unclear and a three-way linkage with HES has not yet been approved. For these reasons, I

did not make use of this dataset. Whilst this means no primary care data was available for the economic evaluations presented in this thesis, primary care costs are known to be low compared to the cost of secondary health care services (Jin and Mosweu, 2017). The CRIS-ONS linkage provides an alternative source of all-cause mortality data which can differ from mortality data in CRIS, as well as cause-specific mortality data. Since cause-specific mortality data was not of interest and mortality was unlikely to be very different in absolute terms between the groups under comparison, I also did not make use of this linkage.

2.3 Unmeasured confounding

2.3.1 Rationale for methodological exploration

As noted in Chapter 1, since treatments are not randomly assigned in observational studies, the risk that treatment effects cannot be identified, that is computed from the available data without distortion, due to confounding is a chief concern in observational studies. To design and interpret studies based on observational data, it is therefore good practice to carefully assess what assumptions with respect to confounding are necessary for identification, to consider what extent they are plausible and what their potential direction or magnitude might be (Kreif et al., 2013b).

Whether the assumption that there is no unobserved confounding holds or the extent to which one's analysis deviates from the ideal conditions, cannot be tested directly. However, based on our subject matter knowledge and the limited range of potential confounders that are readily available in CRIS, that is those available via in structured fields, NLP application or keyword search, there appears to be a significant risk of unmeasured confounding. The risk of bias due to unmeasured confounding, is likely to be particularly acute in the case of Analysis 1 – the economic evaluation of inpatient care for young people – for the following reasons: (i) At hospitalization, young people have typically not been in contact with SLaM for a long time or the hospitalization may even be their first contact with mental health services. Thus, the amount of information recorded in clinical notes can be sparse; (ii) The performance of NLP applications has typically been optimized based on a random sample from CRIS. Since these random samples mostly consist of adult patients, it is likely that NLP

application may perform less well in young people. Thus, confounders derived from NLP applications are less likely to be good proxies for the true confounders; (iii) At admission to CAMHS inpatient care, the Children's Global Assessment Scale (CGAS) is typically administered (Shaffer et al., 1983) whereas at admission to an adult ward the Health of the Nation Outcome Scale (HoNOS) is the routine measure of outcome (Wing et al., 1998). Disease severity is therefore not measured using the same standardised instrument in the two treatment groups. Thus, potential differences in disease severity between young people admitted to CAMHS and young people admitted to adult wards cannot be easily accounted for in the analysis. In the case of Analysis 2 – the economic evaluation of inpatient rehabilitation – there are some key patient characteristics that both likely motivate referral to inpatient rehabilitation and have an effect on outcomes which are not readily available in CRIS or only crudely measured (again, see Appendix A for a discussion of data quality). These include patients' social functioning, living conditions, the persistence and complexity of their mental health needs as well as their history of medication use and response to medication.

In the context of the CRIS database, one approach to increase the plausibility of the standard assumption that there is no unmeasured confounding, is to extract more information about patient characteristics using one or more of the three approaches described in Section 2.2.2. This approach leaves the standard 'no unmeasured confounding' assumption qualitatively unchanged but reduces the quantity of unmeasured information. However, the more potential confounders are included in an analysis, the more likely it will be that some observations do not have an identical or similar counterpart in terms of all confounders in the other treatment group. In other words, the more confounders are included the more likely it will be that there will be a lack of overlap and the need to extrapolate (D'Amour et al., 2020; Petersen et al., 2012) Even if there is no lack of overlap, the more variables are included, the less likely it is that a saturated or somewhat saturated model, that is a model contained terms for all interactions of confounder, can be adequately fitted which increases the risk of model misspecification (Harrell, 2015). Moreover, within the context of CRIS, expanding the set of confounders risks resorting to variables that are measured with error. The issues surrounding measurement error will be discussed more extensively in Section 2.4. In short, although adding more measured confounders into the regression model is desirable from the perspective of reducing unmeasured confounding,

this comes at a price, limiting the potential for reducing bias due to confounding. Finally, despite the wealth of information contained in clinical notes in CRIS, the most relevant confounders may be missing from them.

Another approach that has been advocated by some researchers in the casual inference literature have advocated the use of sensitivity analysis for unmeasured confounding (see for example Haneuse et al. (2019) and VanderWeel and Arah (2011)). The aim of such sensitivity analyses is to explore the potential direction, magnitude and uncertainty arising from unmeasured confounding (Beesley et al., 2020). In addition, some have argued that the very process of quantifying and modelling potential bias combats researcher's tendency for overconfidence (Lash et al., 2014). This is an area of research that is still under development and a large number of sometimes conceptually complex approaches have been proposed (Schneeweiss, 2006; Uddin et al., 2016; Zhang et al., 2018). To remain within the scope of this thesis, I will, therefore, limit myself to stating that, broadly speaking, these approaches involve specifying a bias model and its parameters within the analysis of interest (Lash et al., 2014). At its simplest one could conduct a simple sensitivity analysis assigning a fixed value to the parameter that is believed to be biased. At its most complicated and most computationally intensive, one can vary multiple parameters at once while assigning probabilistic distributions to these. According to Lash et al. (2014) there are four potential sources of information to assign values to the bias parameters: an internal validation study or sub-study in which the analyst is able to collect information on the confounders that are not measured in the sample as a whole (e.g. as part of propensity score calibration) (Uddin et al., 2016); use of external validation data; elicitation of expert opinion (Turner et al., 2009); or without reference to subject matter knowledge, e.g. by identifying magnitude of the bias parameter confounding that would explain away the observed treatment effect or calculating the estimated treatment effect over a very broad ranges of potential bias (Haneuse et al., 2019).

In relation to the empirical applications in the next two chapters, it is conceivable that clinicians could, for example, rate patients' clinical status based on the medical notes and a standardised scoring protocol, thus recovering unmeasured confounders (e.g. social functioning in the evaluation of inpatient rehabilitation). This internal validation study could be used to inform the choice of bias parameter. However, such an exercise would imperfect

due to the fact that clinical notes are incomplete, require clinical training that I do not have and be time-intensive. Thus, this approach was beyond the scope of this thesis. As for external data that could inform the bias parameter, unfortunately, I am unaware of any evidence in the literature that could be used for a sensitivity analysis with respect to unmeasured confounding. Eliciting expert opinion was also not a feasible option in either of the empirical applications, because the amount of input from experts required for such an exercise was not available within the scope of this thesis. Moreover, there are some significant methodological challenges to eliciting expert opinions that would need to be overcome such as the risk of overconfidence in experts' opinion or the challenge of how to combine the opinions from multiple experts (Lash et al., 2014; Turner et al., 2009). With respect to sensitivity analyses that assess what amount of bias that would be necessary to change the results, several objections have been noted in the literature (Lash et al., 2014). Most obviously, it does not help the reader to understand whether the value that invalidates the findings of the base case analysis is plausible. In fact, at their simplest, such sensitivity analyses merely yield a transformation of the point estimate of the base case analysis. In addition, analysing multiple sources of bias becomes more difficult and it has been argued that knowing what values would change the findings of a study can bias one's view of the plausibility of such a value (Lash et al., 2014). A broader concern with these sensitivity analyses is that, even in academic research, they are rarely used and, as noted above, they can be conceptually complex (Zhang et al., 2018). Thus, I would argue that it is not self evident that the increased formalism introduced by a quantitative assessment of the impact of unmeasured confounding always adds to understanding given the difficulty of communicating these analyses to both the academic reader and decision makers without statistical training. To summarize, conducting sensitivity analyses for unmeasured confound was either not feasible or not useful within the context of the empirical applications in this thesis.

A third approach is to consider alternatives to the 'no unmeasured confounding' assumption, that is to use qualitatively different approaches. Specifically, in the next two sections, I will first compare the 'no unmeasured confounding' approach with two other well-known classes of approaches, such as instrumental variable approaches, before introducing an identification strategy that is less well-known but is of value in Analysis 2.

Just like the 'no unobserved confounding' assumption, these alternative identifying assumption rely on untestable assumption. Nonetheless, they can be useful when analysing observational data for at least four reasons: (i) Subject matter knowledge or indirect evidence may suggest that an alternative assumption is likely to be more plausible than the conventional 'no unmeasured confounding' assumption. If so, one can obtain parameter estimates that are likely to be closer to the 'true' value by substituting the conventional by the alternative approach (Faria et al., 2015). Moreover, differences between approaches that vary in their plausibility may give some indirect evidence regarding the approximate magnitude of remaining confounding; (ii) If the alternative assumption is believed to be similarly plausible to the 'no unmeasured cofounding' assumption, the alternative assumption can complement the analysis by helping in quantifying the sensitivity of the results to violations of the 'no unmeasured confounding' assumption; (iii) Alternative identifying assumptions may allow the analyst to place bounds on the 'true' parameter of interest. For example, if there is reason to believe that one approach yields estimates that are likely to be biased upwards, and another approach yields estimates that are biased downwards, then the true treatment effect will lie somewhere between the two estimates (Ding and Li, 2019; Glynn and Kashin, 2017); (iv) In other cases, there may be reason to believe that different approaches yield estimates that are biased in the same direction. The relative magnitude of the estimates would then allow the analyst to determine which of the two estimates is more plausible (Glynn and Kashin, 2018).

2.3.2 Common identifying assumption

2.3.2.1 Introduction to directed acyclic graphs

Identification strategies with respect to unmeasured confounding that have been used in health economic evaluations can, to my knowledge, be classified into three broad classes (Faria et al., 2015; Kreif et al., 2013b). To reduce the use of mathematical notation, I will use causal diagrams to illustrate the main assumptions underlying these approaches. Specifically, I will use directed acyclic graphs (DAGs) which are an increasingly popular tool to conceptualise and explore causal problems (Pearl, 2010; Textor et al., 2016).

In short, DAGs consist of two elements: First, stochastic variables, that is, variables that have a random probability distribution. In Figure 5 these are represented by rectangles surrounding the shorthand variable name. Second, DAGs consist of arrows between these stochastic variables which indicate the possible existence of causal effects (Morgan and Winship, 2014). The variable from which an arrow originates is believed to potentially have a causal effect on the variable to which the arrow points to. For example, $A \rightarrow Y$ indicates that variable represented by the shorthand A is believed to potentially have a causal effect on variable represented by the shorthand Y. The arrows do not specify what functional form the relationship between the two variables takes (e.g. linear, quadratic etc.) nor whether the causal effect is positive or negative. Absence of arrows between two variables indicates that it is assumed a causal relationship between them does not exist. From this point onwards in this thesis, I will use A to denote what treatment a patient is assigned to, Y to denote the outcome measure, X to denote measured confounders and U to denote unmeasured confounders.

2.3.2.2 No unobserved confounding

As a starting point, Figure 5(a) represents the 'no unobserved confounding' scenario that I referred to in the more informal discussion in the previous section (Section 2.3.1). Treatment A is assumed to potentially cause the outcome Y (indicated by $A \rightarrow Y$) and, by definition, in a large sample context, the measured confounders X have a causal effect on both A and Y (indicated by $A \leftarrow X \rightarrow Y$). The DAG reflects the 'no unmeasured confounding' assumption by the absence of unmeasured confounders U that have a causal effect on both A and Y, that is the absence of $A \leftarrow U \rightarrow Y$ (Pearl, 2010). In this scenario, the effect of A and Y can be identified by standardising or reweighting X across levels of A, for example, by means of regression or weighting estimators (Hernan and Robins, 2020). In the literature, this identification strategy is also referred to as back-door adjustment, unconfoundedness, exogeneity, selection on observable or conditional independence (Imbens and Wooldridge, 2009; Jones and Rice, 2011; Morgan and Winship, 2014). The 'no unobserved confounding' approach is well-understood and widely used. As discussed above, however, there are reasons to doubt the assumption underlying this approach in both of the empirical applications in this thesis.

2.3.2.3 Fixed effects approaches

One alternative the 'no unobserved confounding' assumption is to replace it with the following three assumptions: (a) unmeasured confounders U do not change over multiple measurements of the same unit of analysis (e.g. patients); (b) the effect on the unmeasured confounders U on the outcome does not change across the measurements; (c) outcomes Y in one period do not affect treatments A in all subsequent periods nor does treatment A in one period affect outcomes Y in all subsequent period (Imai and Kim, 2019).

For simplicity, Figure 5(b) shows a simple two-period scenario (see Imai and Kim (2019) for a DAG extended to three periods). The two periods are indicated by the bracketed numbers 0 and 1 following the shorthand variable names if they are time-variant. For example, $A(0)$ represents the assigned treatment at the first point in time and $A(1)$ the assigned treatments at the subsequent time point. Figure 5(b) is similar to Figure 5(a) with respect to the causal relationships between A, X and Y. However, these causal relationships are assumed to be present both at time 0 and time 1. Thus, visually, whereas there is only one triangular relationship in Figure 5(a), there are two triangular relationships in Figure 5(b). In addition to this, it allows for a causal effect of past treatment on current treatments, that is $A(0) \rightarrow A(1)$, for a causal effect of past confounders on current confounders, that is $X(0) \rightarrow X(1)$ and, crucially, for a time-invariant unmeasured confounder U. If one assesses the effect of treatment in Y, in this scenario, U is essentially controlled for because the causal effects $U \rightarrow Y(0)$ and $U \rightarrow Y(1)$ are assumed to be identical or at least known. Thus, the treatment effect is identifiable. However, as explained above, this approach assumes that there are no lagged effects of treatment on the outcome. In other words, it precludes $A(0) \rightarrow Y(1)$ because otherwise, $A(0)$ both has a causal effect on $A(1)$ and $Y(1)$ which is not accounted for cannot be adjusted for. Similarly, it precludes $Y(0) \rightarrow A(1)$ and $Y(0) \rightarrow Y(1)$. A limitation of approaches of this kind is that sufficient variation in the treatment variable, confounders and/or the outcomes is necessary.

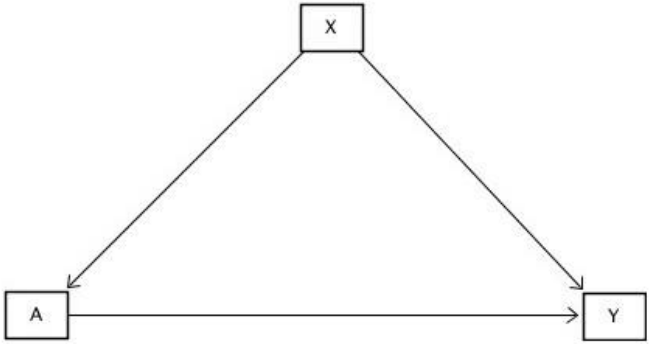
Some versions of these approaches that are based on unit-invariant effects, such as before-and-after and difference-in-difference analyses, are widely used in health economic research and conceptually relatively simple (Faria et al., 2015). Within the context of this thesis, one approach is to use patients as the unit of analysis and assume that patient-specific unmeasured confounders do not change between relevant time points. For

example, one could assume that unmeasured confounders are time-invariant across different hospitalizations of the same patient. Intuitively, this approach removes the patient-specific effects by focusing on changes in outcomes rather than absolute levels. In fact, for comparability with previous literature, I will use this identification strategy in a sensitivity analysis in the evaluation of inpatient rehabilitation. In the evaluation of CAMHS inpatient care, on the other hand, many patients are only admitted to either a CAMHS or an adult ward. This implies that an analysis based on variations between hospitalizations would not be possible without restricting the analysis sample. Similarly, terminal events, such as death, cannot be analysed in a within-patient comparison. In addition, particularly in the case of an unstable condition such as psychosis, it can be difficult to judge over what period patient-specific effects can be considered to be stable and to what extent they are important enough to warrant approaches of the kind shown in Figure 5(b) as opposed to back-door approaches in Figure 5(a) that allow for the analyst to completely adjust for effects of time-variant measured variables.

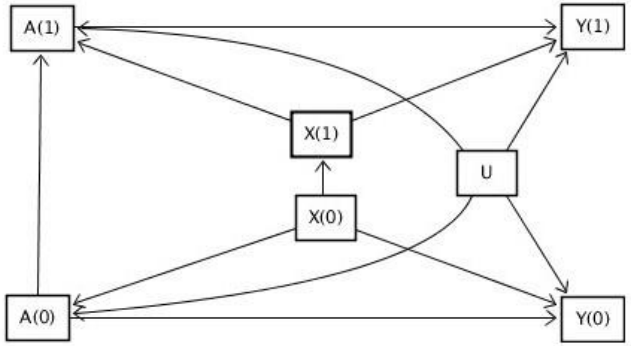
A third potential identification strategy relies on the availability of a variable Z, known as the instrumental variable (IV), that only affects the outcome Y through its effect on the treatment A and that it is independent of the unmeasured confounders U (Kennedy, 2008). In other words, as shown in Figure 5(c), there are no arrows that directly or indirectly connect Z with Y other than via $Z \rightarrow A \rightarrow Y$ and there are now causal effects of the form $Z \leftarrow U \rightarrow A$. Given these assumptions, the treatment effect $A \rightarrow Y$ can be identified because changes in Z will cause a non-deterministic change in A. In turn, one can estimate the expected change in Y given this change on A induced by Z. Intuitively, the function of Z is to approximate the role of a randomization mechanism (Faria et al., 2015). A variant of the IV approach, is when the Z is a deterministic function of some measured confounders X, that is the level of Z is entirely determined by X. This could be represented by adding a the arrow $X \rightarrow Z$ in Figure 5(c). (Oldenburg et al., 2016). This case is known as a regression discontinuity design (RDD) (Lee and Lemieux, 2010). As described in more detail in Chapter 3, I will explore this approach in the evaluation of inpatient care for young people. The potential IV in this analysis will be whether a patient is 18 on the date of hospitalization because, by design, whether a patient is admitted to a CAMHS or adult ward is strongly affected by this, but, in itself, turning 18 is not expected to affect outcomes.

IV and fuzzy RDD approaches are popular among economists and can yield credible estimates of causal effects even if few measured confounders are available (Kennedy, 2008). However, IV methods are conceptually more complex than approaches based on back-door adjustment and unit-invariant fixed effects and, at least the use of RDD approaches, appears to be relatively rare in the context of psychiatric research (Moscoe et al., 2015; Swanson and Hernán, 2018). Although IV methods can yield estimates that are less biased, they can increase the variance of the estimate and parameter estimation tends to be more complex (Boef et al., 2014; O'Keeffe and Baio, 2016). One should also note that, without further assumptions, IV and fuzzy RDD approaches estimate parameters that may not be of direct interest to decision makers (Imbens, 2010). Finally, in many cases it is not possible to find an instrumental variable that strongly determines treatment assignment A. In particular, the assignment to treatments in psychiatry does not appear to be governed often by thresholds as presupposed by the fuzzy RDD design. In the case of the evaluation of inpatient rehabilitation, for example, there did not appear to be any obvious possibility to make use of an instrumental variable or RDD approach.

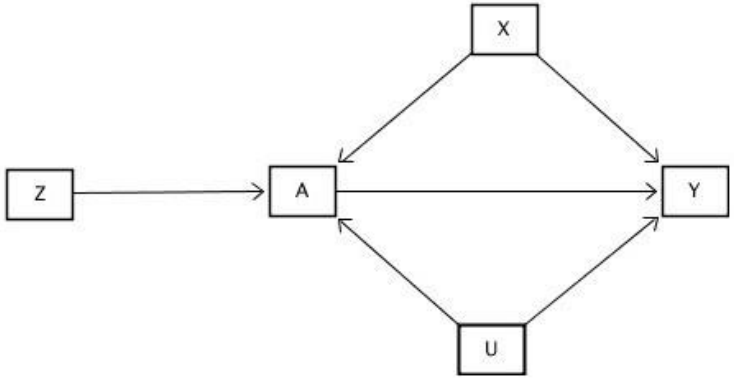
(a) 'No unobserved confounding'



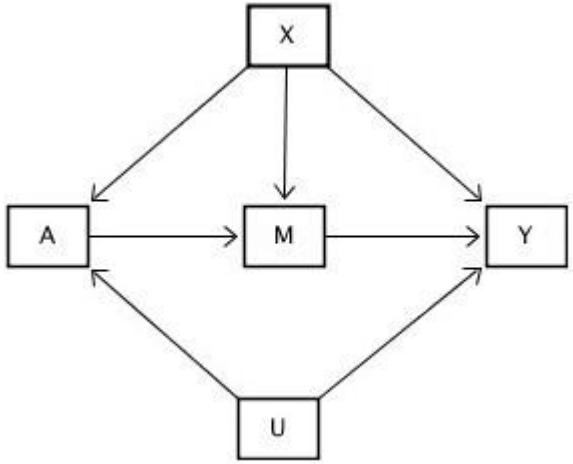
(b) Fixed effects approaches



(c) Instrumental variable approach



(d) Front-door adjustment



$\boldsymbol{A}$ =treatment variable; $\boldsymbol{X}$ =measured confounders; $\boldsymbol{Y}$ =outcome; $\boldsymbol{U}$ =unmeasured confounders; $\boldsymbol{Z}$ =instrumental variable; $\boldsymbol{M}$ =mediator; where applicable, the numbers in brackets correspond to the time period/hospitalizations at which variables are measured; Directed Acyclic graphs were drawn using the software DAGitty (Textor et al., 2016)

Figure 5 Causal diagrams illustrating identifying assumptions with respect to unmeasured confounding

2.3.3 The front-door adjustment

Other strategies to handle unmeasured confounding have been proposed in the broader casual inference literature (Glynn and Gerring, 2013; Lipsitch et al., 2010; Zhang et al., 2018). Figure 5(d) illustrates one of these, known as the front-door adjustment approach, that is of relevance to Analysis 2 in thesis. Like the instrumental variable approach in Figure 5(c), the front-door adjustment approach sidesteps the fact that U is a confounder with respect to the effect of A on Y by identifying a variable in the causal mechanism that is not confounded. However, it is not based on a variable that is upstream of A, that is on a variable that has a causal effect on A. Instead, it is based on a mediator variable, denoted by M, that is downstream with respect to A, that is a variable on which A has a causal effect. In addition to assuming that M is unconfounded, the front-door approach assumes that A has no direct effect on Y, that is the effect of A on Y is entirely mediated by M. As shown in Figure 5(d), the reason that the treatment effect is identifiable despite the presence of unmeasured confounder U, is that U is only a confounder with respect to the total effect $A \rightarrow Y$. If the total effect $A \rightarrow Y$ is decomposed into $A \rightarrow M$ and $M \rightarrow Y$, however, U does not lead to bias because it is not a have a causal effect on both A and M nor a causal effect on both M and Y. These identifying assumptions are equivalent to those invoked when estimating the complier average causal effects (CACE), meaning the effect of the treatment if all patients who were assigned to it were also receiving it (DiazOrdaz et al., 2018). However, in the case of the front-door approach, the target parameter remains the average effect of A on Y, that is assignment to treatment, not the effect of M on Y, that is receipt of treatment, as in CACE. Consequently, estimation of the target parameter using the front-door approach differs from estimation of CACE.

A special case of the front-door adjustment is when A represents treatment assignment, M is a measure of compliance or treatment receipt and there can only be non-compliance in the treatment group, that is non-compliance is one-sided. Typically, this is because the control group does not have access to the treatment. Under one-sided non-compliance, the value of M is constant in the control group because nobody receives the treatment in the control group so the effect of M on Y can only be estimated in the treatment group. This implies that data is need only from people assigned to treatment, that is from the treated

arm, for front-door adjustment in this scenario (Glynn and Kashin, 2018). As discussed in more detail in Chapter 4, we will encounter this one-sided non-compliance scenario in the evaluation of inpatient rehabilitation. I identify two mediators of the effect of referral to inpatient rehabilitation: whether the referral to inpatient rehabilitation is accepted or declined and whether patients are transferred to an inpatient rehabilitation ward following acceptance of their referral or whether they are removed from the waiting list. These mediators are one-sided mediators of non-'compliance' because patients who are not referred to inpatient rehabilitation necessarily cannot have their referral accepted nor be transferred to inpatient rehabilitation.

The strengths and limitations of the front-door approach are similar to those of instrumental variables methods. As indicated above, an advantage of the front-door approach under the one-sided non-compliance scenario is that, if its assumptions are satisfied, treatment effects are identifiable even when neither pre-treatment data nor data from individuals who have not been assigned to treatment are available (Glynn and Kashin, 2018). The price for this is that, since estimation is based on a smaller sample, the variance of estimates will be higher and/or the analysis will rely more on the correct model specification than other approaches. This is particularly the case when there are substantial imbalances in the proportion of patients who comply and those who do not comply. In addition, this approach does not appear to be well-known in the health economic literature nor do there appear to be many empirical applications in the broader causal inference literature (Faria et al., 2015; Glynn and Kashin, 2018; Kreif et al., 2013b). Therefore, consumers of research are likely to be less familiar with the front-door approach compared to the approaches described in the previous section. This is likely to make it more difficult to communicate the assumptions underlying the front-door approach clearly enough so that readers can judge their plausibility. Parameter estimation can be also be more complex than many other approaches because both the joint uncertainty of the effect of A on M and the effect of M on Y needs to be estimated. Glynn and Kashin (2017) use a bootstrap approach but a seemingly unrelated regression (SUR) approach may be a simpler alternative estimator that is also well-known in the context of health economic evaluations (Willan et al., 2004). In addition, it is likely that in many cases a mediator that approximately satisfies the assumptions show in Figure 5(d) is not available. For example, in the evaluation of CAMHS

inpatient care there are no clear candidates for a mediator M. Finally, in the one-sided non-compliance scenario, for example, without further assumptions, the front-door adjustment only yields average treatment effect on the treated, meaning the effect of treatment on those assigned to it (Glynn and Kashin, 2018). Since this is the target parameter in Analysis 2, this is not a limitation within the context of this application. To summarize, I identified four potential identification strategies. In Analysis 1, I will use an instrumental variable approach, specifically and RDD approach. In Analysis 2, I will use the other three approaches.

2.4 Measurement error

2.4.1 Rationale for the methodological exploration

Measurement errors are the difference between measured values and their true quantities. It can be shown that, unless one is willing to make further assumptions or one has additional data, the target parameter cannot be identified in presence of measurement error and/or that the estimates will be more imprecise (Hernán and Cole, 2009; VanderWeele and Hernán, 2012). Apart from simple cases and when the variance of the measurement error is small relative to the variance of the variable, the direction and magnitude of bias due to measurement error can be difficult to predict even when measurement error is random (Buonaccorsi, 2010; Kennedy, 2008). I would also argue that measurement error can make it more difficult to assess the credibility of study findings. For example, the addition of a confounder that is measured with error in a statistical model is likely to reduce the bias of the parameter of interest although to a lesser extent than if it was measured without error. By statistical criteria, therefore, even the inclusion of poor proxy confounders is recommended (Kennedy, 2008). However, if the additional confounder is measured with significant error, an intuitive judgement about the remaining amount of unmeasured confounding becomes more difficult because it is unclear to what degree the effect of the mismeasured confounder has been adjusted for.

As discussed in more detail in Appendix A, many variables that are readily available in CRIS, that is data from structured fields, NLP application outputs or keyword/keyphrase search

results, are either known to be measured with potentially non-negligible error or their accuracy is unknown. To be more specific, in the evaluation of inpatient care for young people with psychosis, measurement error in service use and particularly measurement error in the baseline diagnosis were of concern. In the evaluation of inpatient rehabilitation, in addition to these two variables, measurement error in the treatment variable, symptoms derived from clinical notes and whether patients had a history of taking the antipsychotic drug clozapine use were potentially problematic. However, since CRIS allows access to free text clinical notes, as described in Section 2.2.2, one can either develop and/or improve NLP applications or manually code variables of interest to obtain variables with lower measurement error. In other words, in contrast to most existing economic evaluations based on routinely collected data, it is possible to quantify, reduce and/or adjust for some of the bias due to measurement error. Developing and/or improving NLP applications is, however, often time-intensive and can be technically complex. Similarly, reading CRIS records to manually coding all relevant variables is not practically feasible and any approach other than manually coding a random subsample, such as stratified sampling which I describe further below, increases the conceptual and potentially the technical complexity of the analysis. To summarize, there are three competing demands arise when handling measurement error using CRIS data: (1) risk of biased and/or imprecise estimates; (2) conceptual and/or technical complexity of the analysis; (3) demands on researchers' time. Given these competing demands, an informed choice of among statistically principled approaches to handling measurement error appeared warranted in this thesis.

Like in the discussion of approaches to handling unmeasured confounding, I will first describe common identification assumptions with respect to measurement errors (Section 2.4.2). In section 2.4.3, I will then discuss the strengths and limitations existing strategies to support or enable the identifying assumptions in face of the three competing demands listed at the end of the preceding paragraph. In contrast to other common statistical issues, such as missing data or unmeasured confounding, the handling of measurement errors appears to have received relatively little attention in the health economic evaluation literature or the medical literature more broadly (Brakenhoff et al., 2018; Marschner, 2006; Smith et al., 2018). This may partially be explained by the fact that the opportunities for adjusting for measurement error tend to be more limited when working with databases that

do not allow access to free text clinical notes. Therefore, it appeared more appropriate to consider practices in published CRIS studies rather than the economic evaluation literature as a reference point for the discussion of approaches to handling of measurement error. To this end, I reviewed all published papers listed on the online compendium of CRIS studies (https://www.maudsleybrc.nihr.ac.uk/facilities/clinical-record-interactive-search-cris/cris-publications; last accessed 16 August 2019) that had causal inference or descriptive analysis as their primary goal, and made use of variables in their analyses that could have been or were manually verified using the free text. I excluded conference abstracts and studies that made use the same dataset as another published CRIS study.

2.4.2 Common measurement error assumptions

To sharpen intuition, I will first describe common measurement error assumptions informally. For clarity, I will also limit myself to discussion of different types of measurement error assumptions with respect to non-outcome variables, i.e. the treatment variable, confounders and eligibility criteria. With respect to outcome variables, throughout this thesis, I will simply assume that measurement errors are random, have a zero mean and are additive to the regression equation. Under this common assumption measurement error in the outcomes, estimates remain unbiased but their precision decreases (Cameron and Trivedi, 2005). Intuitively, this is because errors 'balance out' on average and simply become part of the regression error term but there is more 'noise' in the data so it is more difficult to determine the true effect.

With respect to non-outcome variables, on the other hand, I will consider four possibilities which are summarised in *Table 2*:

> *Assumption (a)*: one can assume that the readily available CRIS variable is measured without error. For example, one can extract gender the patient is according to their clinical record and use this as a confounder in the analysis. By definition, this approach requires no time for manual coding. It is conceptually and technically simple. However, it represents the strongest assumption, precisely the assumption that, as discussed in the previous section (2.4.1), is in doubt in many cases in CRIS.

Assumption (b): one can read through a subset of the medical records of the analysis to check whether the readily available variable of interest is measured correctly, i.e. internally validate some data points, an make use of this validated data where available. For example, one could read through the medical record of a patient to determine whether there is evidence that they were prescribed clozapine before the date of interest in a subset of the patients that comprise the sample. In the analysis one would then make use of the manually validated data where it was validated and use the readily available data when it is not available. Assumption (b) must be weaker than Assumption (a) because, on average, the manual validated data is known to have a lower measurement error than readily available data by virtue of the manually checking process. This comes at the cost that manually checking is more time intensive than just using readily available data. Also, as discussed in the next section (2.4.3) depending on how the observations which are manually validated are sampled, it can be somewhat or significantly more technically complex to implement.

Assumption (c): One can assume that the manually validated version variable of interest is correct in the subset in which it was validated as in assumption (b) but relax the assumption that the data that has not been manually validated has no measurement error by assuming that the relationship between the mismeasured variable and the validated variable is the same in the subsample where both are available and the rest of the sample (where only the readily available version is measured). For example, one could read through medical notes to determine whether someone has had a history of having been prescribed clozapine. One can then assume that in the rest of the sample one can accurately estimate the probability that someone has had a history of clozapine given the number of the mentions the word clozapine in the free text (the ready available data). This assumption is weaker than assumption (b) because it does not assume that when a variable has not been validated, the readily available version is correct. However, to be able to estimate a stable model of the relationship between the readily available and the manually validated version of the variable, manual coding will be necessary. Conceptually, I would argue that this is the most complex assumption because it

both requires the choice of a suitable strategy to sample the subset of variables that are manually validated and an assessment of whether the model between the readily available and the manually validate variable is robust.

Assumption (d): Finally, one can assume that the manually validated data has no measurement error and read all relevant patient records to obtain this data for the analysis. This is the weakest assumption because all data has been verified. However, it is also the most time consuming to implement for the same reason. Conceptually it is as simple as Assumption (a) because all data comes from one source and it has a simple mathematical representation.

Following this intuitive discussion, I proceed to a more formal proceeding to a mathematical definition. As in Section 2.3.2, I will denote outcomes by Y but collectively refer to non-outcome variables as V, that is to confounders, treatment variables and/or indicator variables reflecting whether an observation fulfils the eligibility criteria for the study. For simplicity, as above, I consider a binary treatment variable A. I use a superscript asterisk to denote the readily available version of a variables (e.g.V^*), a superscript plus sign to denote the manually coded versions of a variable, typically referred to as internal validation data (e.g. V^+), whereas variable without superscript denote the true measurement (e.g. V). By definition, the measurement error of the readily available data is at least as large in absolute terms as the measurement error in the internal validation data, that is $|Y^+ - Y| \leq |Y^* - Y|$ and $|V^+ - V| \leq |V^* - V|$. S_V represents a binary indicator equal to 1 if V is manually coded in an observation and equal to 0 otherwise. I use $\mathbb{P}(\,)$ to denote a probability function.

Given this notation we can define the aforementioned assumptions as follows:

Assumption (a): $V = V^*$ meaning the readily available version of the non-outcome variable (V^*) is the same as the true version of the non-outcome variable (V). This assumption makes no reliance on manual validation, such that is makes no assumption about S_V. Given the aforementioned properties of the manually coded version V^+ relative to V^*, it implies that $V = V^+$ is also assumed to hold

Assumption (b): $V = S_V \cdot V^+ + (1 - S_V) \cdot V^*$ where $0 < P(S_V = 1) < 1$, meaning that the readily available version of the non-outcome variable has no measurement error ($V = V^*$) when the variable is not validated ($S_V = 0$) as in Assumption (a), but

when it is manually validated ($S_V = 1$), one only assumes that the manually validated version has no error ($V = V^+$). It can be seen that if $\mathbb{P}(S_V = 1) = 0$, i.e. no observation was manually validated this assumption would be equivalent to Assumption (a). This assumption implies that $S_V = \mathbb{I}(V^* \neq V^+)$ where $\mathbb{I}(\,)$ is an indicator function. In other words, if the readily available non-outcome variable is not the same as the manually coded version, the manually coded version is sampled because otherwise the non-outcome variable will be measured with error and the effect cannot be identified. Again, given the properties of the manually coded version V^+, it implies that $V = V^+$ is also assumed to hold

Assumption (c): It should be noted that this assumption appears to take many different forms in the literature (Guolo, 2008). Here, I represent a fairly general versions described by Blackwell et al. (2017). Again, one assumes that $V = V^+$. However, unlike Assumption (b) we relax the relationship between the V^* and V by assuming that the difference between these two versions of the variable can be bridged by a function $f(\,)$ of the form $V^* = f(V^+, Y, \eta)$ with parameters η that can be consistently estimated from the data. Also, unlike Assumption (b), we assume that $\mathbb{P}(S_V = 1|V^+, Y) = \mathbb{P}(S_V = 1|V^{+\prime}, Y)$, meaning that, conditional on other variables, the manually coded variable is sampled independently of its realised value. This is to ensure that the relationship between V^+ and V^* is not distorted by the process with which V^+ is sampled. As in the case of Assumption (b), Assumption (c) is only a meaningful option if some but not all of the observations have been manually validated, i.e. $0 < P(S_V = 1) < 1$

Assumption (d): $V = V^+$ meaning the manually validated version of the non-outcome variable (V^+) is the same as the true version of the non-outcome variable (V). This assumptions implies that $\mathbb{P}(S_V = 1) = 1$, i.e. that all observations have been manually validated. As described above, Assumptions (a)-(c) also assume that $V = V^+$ but add additional assumptions whereas in Assumption (d) it the only assumption that is made.

In my review of CRIS studies, I identified 81 papers that fulfilled the aforementioned inclusion criteria (see Figure 6 for the study flow-chart). Measurement error assumptions were rarely explicitly stated so this summary shown is based on my interpretation of

existing analyses. With respect to the outcome, like in this thesis all studies appear to have assumed that additive, zero mean errors. With respect to non-outcome variables, Figure 7 shows that assumptions (a) was the most common, that is analyses that did not make use of manually validated data in mismeasured variables. However, 33% of studies also manually validated all observations in one or more variable, that is $\mathbb{P}(S_V = 1) = 1$ (Assumption (d)) and 7% manually validated as subset of observations used in their analysis, that is $0< P(S_V = 1) < 1$ (Assumption (b)). No existing CRIS study partially estimated the measurement error model (Assumption (c)).

As described in more detail in Appendices C and D, with respect to the non-outcome variables, I also did not use a measurement error model but, depending on the variable, invoked one of the three previously used assumptions (see Figure 7), i.e. I made use of Assumptions (a), (b) and (d). More specifically, in Analysis 1 I decided to use Assumption (d) with respect to diagnosis data because other assumptions did not appear plausible and Assumption (a) with respect to all other non-outcome variables. In Analysis 2, I invoked Assumption (d) for symptom variables, Assumption (b) for the diagnosis and clozapine history variables and Assumption (a) with respect to all other non-outcome variables based on my assessment of the data quality (see Appendix A).

	Assumption (a)	Assumption (b)	Assumption (c)	Assumption (d)
Description	The readily available version of the variable is measured without error	If the readily variable has not been manually validated then it is measured without error, otherwise the manually validated version is without error	The manually validated version of the variable is measured without error, the relationship between the manually validated and the readily available variable can be estimated consistently and manual validation of the variable does not depend on the value that the manually validated variable takes.	The manually validated version of the variable is measured without error
Mathematical definition	$V = V^*$	$V = S_V \cdot V^+ + (1 - S_V) \cdot V^*$ where $0 < P(S_V = 1) < 1$	$V = V^+$ and $\mathbb{P}(S_V = 1 \mid V^+, Y) = \mathbb{P}(S_V = 1 \mid V^{+\prime}, Y)$ and for a function $f()$ of the form $V^* = f(V^+, Y, \eta)$, parameters η can be consistently estimated where $0 < P(S_V = 1) < 1$	$V = V^+$
Strength of assumption	Strongest	Weaker than assumption (a)	At least as weak as assumption (b) given same $\mathbb{P}(S_V = 1)$	Weakest
Manual coding burden	Lowest	Moderate/High	Moderate/High	Highest
Conceptual complexity	Low	Moderate/High	Highest	Low

$\mathbb{P}()$ *= probability function; uppercase asterisk = readily available version of variable; uppercase plus-sign = manually coded version of variable; uppercase dash = alternative realization of the same variable,* $\boldsymbol{Y}$ *= outcome;* $\boldsymbol{V}$ *= non-outcome variable;* $\boldsymbol{S_V}$ *= indicator variable for whether the* $\boldsymbol{V}$ *was manually coded for an observation;*

Table 2 Overview of common measurement error assumptions with respect to non-outcome variables

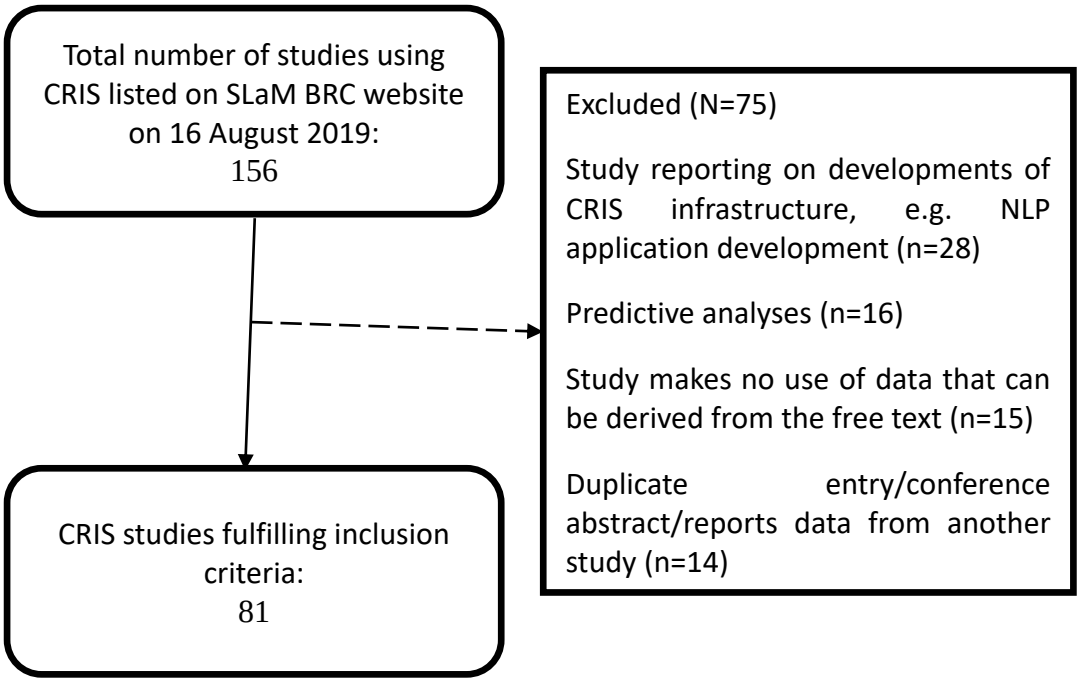
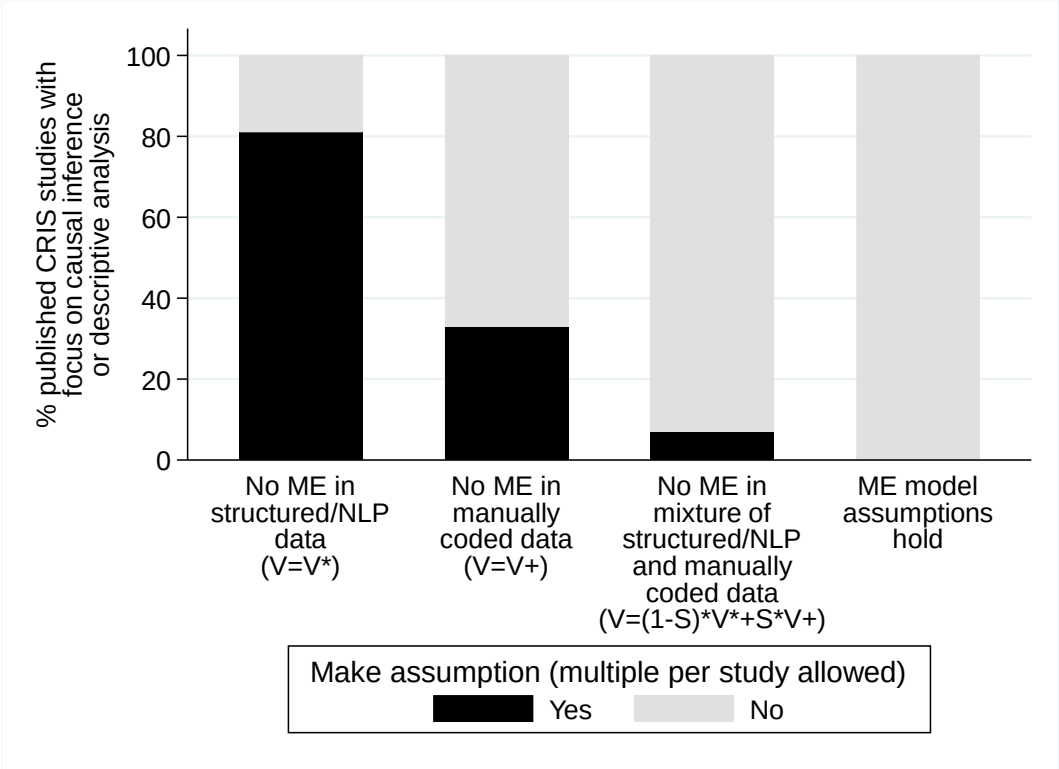


Figure 6 Flow-chart for review of published SLaM CRIS studies

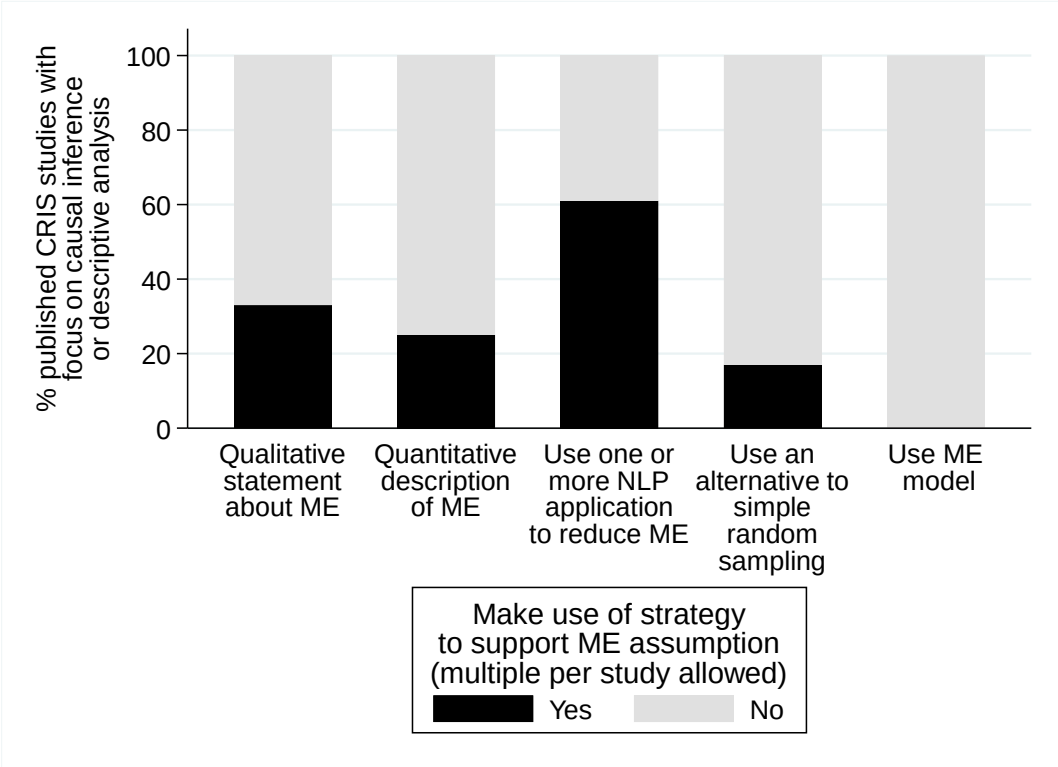


ME: measurement error; NLP: natural language processing

Figure 7 Identifying assumptions with respect to measurement errors in non-outcome variables used in existing SLaM CRIS studies

2.4.3 Strategies to support or enable measurement error assumptions

I identified five types of strategies to support or enable the measurement error assumptions described in the previous section: (i) qualitative discussions of the impact of measurement error; (ii) quantitative description of the impact of measurement error; (iii) approaches to reduce measurement errors in V^* and/or Y^*; (iv) alternatives to simple random sampling strategies and (v) estimators for measurement error models. These strategies are complementary to each other but only a limited number may be appropriate to a specific analysis. For clarity, I will limit my discussion to what I judged to be best practices in published CRIS studies. Figure 8 provides an overview of how frequently each of these strategies have been used in existing CRIS studies.



ME: measurement error; NLP: natural language processing

Figure 8 Strategies to support or enable measurement error assumptions in existing CRIS studies

2.4.3.1 Qualitative bias analysis

The first approach - qualitative discussion based on subject matter knowledge - can help inform stakeholders to what extent the measurement error assumptions discussed in the preceding section are credible and in what way deviations from them may have an impact on the study findings. Qualitative statements ultimately underpin all quantitative evidence, and they are an accessible and potentially time-efficient approach to communicate the impact of measurement error. However, qualitative approaches frequently fail in presenting an unbiased picture of measurement errors, for example, because of limits to human ability to reason under uncertainty (Lash et al., 2009). They have a less rigorous theoretical grounding and can be less precise, comprehensive and transparent in characterising the impact of measurement error (Neumann et al., 2016). Among published CRIS studies, 33% make a statement related to the quality of the data or measurement error (see Figure 8). In both of the following studies, I make use of this approach.

2.4.3.2 Measurement error quantification

The purpose of the second strategy – quantitative description of the impact of measurement error – is to overcome some of the disadvantages of a purely qualitative approach. Judging from the published CRIS literature, the generation of descriptive evidence involves a four-step process: (1) drawing a sample from that is representative of the observations that constitute the analysis of interest or a sample that can be reweighed to become representative; (2) manually coding variables that are used in the analysis, that is sampling V^+ and/or Y^+, within this subsample; (3) calculation of one or more performance metric that comprehensively summarise the impact of measurement error on some variable that is of direct interest to the analysis and quantifying the sampling uncertainty of the performance metric; (4) making a qualitative judgement as to what extent measurement error threatens the validity of the study based on the estimated performance metric.

The strength of this approach is that it is conceptually simple, technically straightforward to implement and appears to be well accepted in the NLP literature (Velupillai et al., 2018). However, a key limitation of approach is that it only quantifies the difference between V^+ and V^*, and/or between Y^+ and Y^*, rather than the impact on the parameter(s) of interest.

It is, of course, possible to run the analysis of interest in the subsample for which both V^* (or Y^*) and V^+ (or Y^+) are available and assess how the estimate of the parameter of interest based on V^* (or Y^*) compares to estimates using V^+ (or Y^+). However, this does not yield an estimate of the sampling uncertainty of the bias which would require a procedure that similar in complexity to methods for adjusting for measurement error. Finally, the theoretical basis for determining the size of the validation subsample and the basis on which measurement errors of a given magnitude are judged to be negligible appears unclear.

Among the published CRIS studies, 25% made use of this approach (Figure 8). The two most frequently used performance statistics were positive predictive value (also referred to as precision), that is number of true positive over the sum of true positives and false positives, and sensitivity (also referred to as recall), that is the number of true positives over the sum of true positives and false negatives when comparing V^+ and V^*. Within the context of the two economic evaluations making up this thesis, I did not make use of this approach because I considered targeted sampling approaches, discussed in Section 2.4.3.4, to be more attractive strategies.

2.4.3.3 Reducing measurement error in proxy variables

Broadly speaking, the third strategy – approaches to reduce measurement errors in V^* or Y^* - involves processing existing data (e.g. data derived from the free text) and/or improving the prediction model for V^* or Y^* based on subject matter knowledge and empirical examination of the causes for measurement error. At its simplest, informal approaches to 'cleaning' the data are often used to reduce measurement error in health economic research but, as described in section 2.2.2, in the case of CRIS the possibility of developing or modifying NLP applications potentially allows the analyst to reduce measurement error in a more sophisticated fashion. However, as noted above, there are technical barriers to the development of NLP applications, particularly for researchers with background in health economics. In addition, for a given amount of time invested in developing strategies to , the added benefit of developing NLP applications or other approaches to reduce measurement error over a simple keyword/keyphrase search in terms of reduction in measurement error

may be uncertain or low. This is particularly the case when complex concepts, such as those involving temporality or ambiguous terminology, are in question. Finally, when the concept of interest is rare, devising NLP or non-NLP strategies to reduces measurement error in V^* or Y^* may be more time-intensive than manually coding all or a substantive number of observations.

Among published CRIS studies, 61 % make use of one or more NLP application but only few studies develop a new application or another approach to reduce measurement error in proxy variables (Figure 8). I will take a similar approach, that is make use of existing NLP applications but not develop these further (Patel et al., 2015). In the evaluation of inpatient rehabilitation, we were interested in expanding the set of measured variables by including an indicator as to whether patients were self-neglecting or not. However, using targeted approach to manual sampling as described in the next section (Section 2.4.3.4) appeared to be a more time-efficient approach than developing a new NLP application.

2.4.3.4 Alternatives to simple random sampling

The basic principle behind the fourth class of strategies – alternatives to simple random sampling - is to target the manual coding of variables that will lead to the lowest bias and/or variance of the parameter of interest for a given amount of manual coding. Within this class, two types of approaches can be distinguished: those which enable the analyst to manually code a greater proportion of the analysis sample by reducing its size and those that enable the analyst to prioritise observations for manual coding when not all observations are manually coded.

Matching is one of the potential sampling procedures that aims at reducing the analysis sample so that a greater proportion can be manually coded. It typically entails three steps: (i) defining a metric that reflects the 'similarity' between two units of analysis, the so-called distance measure or matching metric, or the defining subgroups that are thought to be sufficiently 'similar' in terms of measured confounders; (ii) pairing of observations based on the measure of closeness or subgroups defined in (i); (iii) checking the quality of balance and overlap in the resulting sample and, potentially, repeat steps (i) and (ii) to optimise its desired properties, whether by hand or in an automated fashion (Stuart, 2010). Matching

increases the overlap and balance in measured confounders between the treatment and the control group. This reduces the reliance on correct specification of the regression model. Thus, matching can yield estimates with lower bias and mean squared error compared to simple random sampling of an equally sized population and manual coding an equal number of observations (Stuart and Ialongo, 2010). The use of matching as a strategy to enable the analyst to manually code a greater proportion of the sample does not preclude the use of matching to address observed confounding, the purpose for which it is more commonly used in common in health economics (Kreif et al., 2016). On the contrary, the matching procedure would typically remain unchanged whether matching is used as a strategy to support reduction of measurement error or only to reduce observed confounding, and thus yield the same benefits in terms of improved overlap and balance of confounders. However, conventionally the improvements in overlap and balance of confounders are the only motivation for choosing a matching approach and the resulting reduction in sample size is an undesirable side effect of the approach. By contrast, if matching is used to support manual coding, the reduction in sample size is the primary purpose, but the improvements in overlap and balance of confounders is what makes matching superior to simple random sampling (Stuart and Ialongo, 2010). In other words, in the latter case, both effects of matching are valued.

An alternative to matching is stratified random sampling (Lohr, 2009). In stratified random sampling, subpopulations, referred to as strata, are created based on V^* and/or Y^*. Then a certain proportion of observations is sampled at random from each stratum. Matching or stratifying sampling based on pre-treatment variables that are not assumed to be free of measurement error can reduce bias but there are two cases in which this can increase bias: (1) when the measurement error is an instrumental variable, which leads to what is referred as Z-bias; (2) when the measurement error is a collider variable, that is a pre-treatment variable that is associated with both treatment and outcome but only because it is affected by causes of the treatment and causes of the outcome, which leads to what is referred to as M-bias (Schneeweiss, 2018). In practice, simulation studies suggest that the risk of large biases due Z- or M-bias is likely to be low in practice (Liu et al., 2012; Myers et al., 2011)

The advantage of stratified random sampling compared to matching is that, subject to appropriate reweighting of the sample, it allows sampling based on the outcome variable.

This is particularly valuable when the primary outcome is as 'rare event' such as mortality or, individuals with very high service use because then the between-strata variance explains the majority or all of the overall variance in the outcome (Lohr, 2009). In addition, if the target parameter is the average treatment effect on the treated (ATT) or the average treatment effect on the untreated (ATU), matching only allows the analyst to decrease the size of one of the two treatment groups, that is the control group in the case of the ATT and the treated group in the case of the ATU. By contrast, stratified sampling, can help reduce bias more generally (McNamee, 2005). However, there appears to be only relatively limited research with respect to calculating 'optimal' sampling fractions across strata and, within mainstream statistical software, routines to assist in optimal stratified sampling appear to be limited (McNamee, 2005; Reilly, 1996). Stratified sampling is also more likely to require incorporation of sampling weights in the statistical model than matching. This can increase the complexity of the analysis considerably when combined with other approaches such as bootstrapping (Deb et al., 2017). Finally, unlike stratified random sampling, matching is a well-accepted and well-developed approach in the health economics literature (Deb et al., 2017; Diamond and Sekhon, 2012; Kreif et al., 2016). A disadvantage of both matching, stratified random sampling and simple random sampling is, of course, that, while they enable the analysist to make weaker measurement error assumptions, this comes at the cost of increasing the variance of the estimate of due to the reduction in the size of the analysis size.

The second alternative to simple random sampling is based on the recognition the expected change in the estimate of the parameter of interest or the change in the uncertainty surrounding it is not constant across observations. There appear to be at least three reasons for this: (i) the expected probability that a variable is measured with error is not constant across all observations. For example, if there are hundreds of documents containing the term 'clozapine' prior to baseline, it is almost certain that patients were prescribed the antipsychotic clozapine at this point so there is little value in manually validating a representative sample of such observations; (ii) the impact of a given amount of measurement error on the parameter estimate is not constant for all observations. For example, measurement error in a patient's diagnosis will have a greater impact on cost estimates if he/she has particularly high cost of service use because, being an outlier, they

will affect regression coefficients to a greater extent; (iii) When a measurement error model is estimated and it is correctly specified, manual validation will not reduce bias but targeted sampling could reduce the variance of the estimate. For example, if one of the treatment groups is substantially smaller than the other, then manual coding an additional observation the smaller treatment group it likely to decrease the variance to of the parameter of interest more than sampling an additional observation in the larger treatment group. Given reason (i), one way to targeting observations could be to be to develop subgroups by V^* (or Y^*) based on prior knowledge or after some simple random sampling of data and continue manually coding only those strata in which there was a significant amount of difference between V^* and V^+ (or between Y^* and Y^+). Given reason (ii), I propose that another way to target observations could be to calculate the change in the estimate of the parameter of interest if for a given observation $V^* \neq V^+$ (or $Y^* \neq Y^+$). Specifically, if V is binary one could estimate how the parameter of interest would change if for each observation V^* was instead equal to $|V^* - 1|$, that is if the binary variable was 0 instead of 1 or 1 instead of 0. If V^* is not binary, one could estimate the effect of replacing V^* by its predicted value using all other variables. If the analysis of interest is a linear regression and the interest lies in manual verification of whether an observation either fulfils the eligibility criteria or whether the outcome is measured with error, then DFBETA is a readily computable statistic that allows the analyst to rank observation based on the impact of measurement error on the parameter of interest (Belsley et al., 2005). I am unclear about a formal approach to target observations based on reason (iii). An advantage of this second class of sampling strategy compared to the first is that it avoids the reduction in the analysis sample. However, its potential to reduce bias due to measurement error is much more limited. While it appears likely that in other disciplines such as auditing these approaches have been formalised, I am unaware of this literature, so the theoretical basis of this approach is relatively weak.

Among the reviewed CRIS studies, 10% used a matching approach and 7% of existing studies sampled observations based on the expected probability of measurement errors (Figure 8). None of the studies used a stratified random sampling approach or sampled observations based on the impact of a measurement error on the parameter estimate. Given the large number of potential controls in the evaluation (n=12,828), in the evaluation of inpatient rehabilitation, I used a matching approach to reduce the size of the control group to a

number (n=337) for which manual coding of the variables such as evidence of self-neglect was feasible within the time constraints of this thesis. Further, I manually validated patients' diagnosis based on the predicted probability of measurement error. Finally, in both evaluations, I manually validated the service use/costs of a small number of observations based on the potential impact of measurement error on the parameter of interest.

2.4.3.5 Measurement error models

To implement measurement assumptions that partially estimate the measurement error models, an appropriate estimator needs to be chosen. Popular approaches in the literature include the simulation and extrapolation approach, Bayesian methods, regression calibration, multiple imputation and multiple overimputation (Blackwell et al., 2015; Buonaccorsi, 2010; Carroll et al., 2006; Yi, 2017). Without going into detail with respect to the principles behind each of these approaches, the key advantages of these is that they enable the analyst to invoke a weaker measurement error assumption (what is referred to as Assumption (c) in Section 2.4.2) and may allow the analysis to account for the uncertainty due to measurement error. Multiple imputation is the most attractive option among those listed above because it is a widely used and well-accepted approach in the health economic evaluation literature (Gabrio et al., 2017). It reconceptualises the problem of measurement error as a problem of missing data, that is as a problem of missing V^+ and/or Y^+. The disadvantage of these approaches is, however, that they can be computationally intensive, there may be difficult in making them compatible with other statistical approaches, such as matching or weighting and they can rely heavily on model specification (Blackwell et al., 2015). Finally, while there has been significant research on measurement error models with respect to error in the outcomes Y, the treatment variable A and the measured confounders X, there has been little research with on measurement error models for eligibility criteria (Yi, 2017). Within a multiple imputation framework, for example, measurement error in the eligibility criterion could be conceptualised as having missing data in a binary variable that interacts with all confounders and the treatment variable. However, interaction terms pose technical challenges in multiple imputation and inference for those not fulfilling the inclusion criteria is not of interest it is unclear to what extent the insights from the missing

data literature can be transferred to a measurement error (Bartlett et al., 2015). Bayesian approaches may be more suitable in this context but pose problems in terms of computational intensity and conceptual complexity (Keogh and Bartlett, 2019)

Since none of the existing CRIS analyses invoked the relevant identifying assumption none made use of a measurement error model (Figure 8). While I considered the use of multiple imputation to reduce the amount of manual coding of variables, I did not make use of it in either of the analyses in the next two chapters. In the evaluation of inpatient care for young people, the reason for this that the main variables with measurement error, whether patients had a diagnosis of psychosis, was an eligibility criterion which would have led to the complications described above. In other words, it was unclear whether multiple imputation could be used to impute whether a patient had a diagnosis because this has, to my knowledge, not been adequately explored in the methodological literature. In the evaluation of inpatient rehabilitation, this was because a multiple imputation would have further added to the computational intensity and conceptual complexity of the analysis. More specifically, use of multiple imputation (or another approach to estimating the measurement error model) would have added to the already significant computational burden resulting from using a bootstrap and boosted regression approach. Further, it was unclear to what extent multiple imputation is compatible with approaches such as boosted regression. Another reason for not using a measurement error model in the evaluation of inpatient rehabilitation is that, while it would have increased the sample size available for the base case analysis of this study, it was unlikely that it would have meaningfully decreased bias because measurement error remaining after manual coding was low. At the same time, I would argue that in the base case analysis of Analysis 2, potential bias due to confounding is a more critical issue than sampling uncertainty. Thus, using a measurement error model is unlikely to have meaningfully changed the interpretation of the findings.

2.5 Conclusion

In this chapter, I have shown that unmeasured confounding and measurement errors are significant concerns given the nature of the CRIS data. Yet, by exploring existing and non-standard methodologies, I have identified approaches that can help in alleviating these

concerns in some cases. To address unmeasured confounding, I will explore the use of a regression discontinuity design in the evaluation of inpatient care for young people (Chapter 3) and, among others, the front-door adjustment for the evaluation of inpatient rehabilitation (Chapter 4). In terms of measurement error, I judged the diagnosis variable in the evaluation of inpatient care for young people and the treatment variable in the evaluation of inpatient rehabilitation to be too central to the analysis to be even partially based on data from structured fields and/or NLP outputs. However, using matching as an approach to reduce the size of the control group in the evaluation of inpatient rehabilitation and targeting observations for manual observations based on the predicted probability of measurement error or their influence on the analysis appeared to be valuable approach in both studies.

Chapter 3 Economic evaluation of inpatient care for young people (Analysis 1)

3.1 Introduction

In this chapter, I conduct an economic evaluation of inpatient care modalities for young people with psychosis. In Section 1.5.1, I described the broader context for this analysis, its relevance to health care decision making and existing evidence on this topic. To summarize, child and adolescent (CAMHS) inpatient ward provide specialist care for young people who are under the age of 18 at admission. CAMHS inpatient care is more costly but is also believed to be more therapeutic than that provided on adult wards. I identified three reasons for why a comparison between these care models for inpatient care is of interest to decision makers: (a) due to shortages of CAMHS beds, at times, clinicians need to choose between admission to local adult wards or out of area CAMHS wards; (b) some mental health trusts have or are considering to expand provision of CAMHS inpatient care to the age of 25; (c) the effect of inpatient care quality on outcomes is of general interest to decision makers in the UK. It is known that economic evaluation of interventions for children and adolescents are very limited in general and I am unaware of any quantitative study comparing CAMHS and adult inpatient care (Beecham, 2014). I only identified one study assessing the impact of integrating CAMHS and adult care for young people with psychosis but this research had major limitations (Maxwell et al., 2019).

Given the background to this study, it had two objectives: (i) to estimate the average effect of being admitted to an adult ward instead of a CAMHS ward, among those underage patients that were admitted to an adult ward; (ii) to estimate the average effect of admitting all patients aged between 18-25 to CAMHS wards instead of adult wards. I conducted a cost-consequence analysis, that is I present the differences in costs and other outcomes alongside each other without attempting to summarise them into a single measure (Mauskopf et al., 1998).

3.2 Methods

3.2.1 Study design and comparison of interest

I undertook a retrospective cohort study based on routinely collected data using a regression discontinuity design (RDD), a study design explained and motivated in more detail below. I targeted two parameters, known as estimands, in my analysis. Firstly, our interest lay in the average change in outcomes, if patients who were admitted to an adult ward while being under the age of 18, had instead been admitted to a CAMHS ward. This effect is known as the average treatment effect on the untreated (ATU) (Morgan and Winship, 2014). Secondly, we assessed the average impact of shifting the age threshold that separates CAMHS and adult inpatient care from 18 to 25 for all patients. This effect is known as the average treatment effect (ATE). In both cases, we were interested in a hypothetical scenario in which nobody moves away from the study area, that is moving out of the study area was the censoring event. In Appendix C.3, I define these quantities more formally.

3.2.2 Data source and setting

The source of data, described in more detail in Chapter 2, was the South London and Maudsley (SLaM) Biomedical Research Centre (BRC) case register, CRIS (Perera et al., 2016). To summarize, CRIS is an anonymized version of SLaM's electronic health records covering secondary mental health care in four geographical areas in the south of London, UK. It allows both access to information from structured fields, free text clinical notes and data derived from these notes through natural language processing (NLP). In this study, I make use of data from structured fields or data derived by reading free-text clinical notes but not NLP data. I obtained approval for this specific project by the CRIS oversight committee (project number 13-090).

3.2.3 Study population

To be included in the analyses, patients needed to be admitted to a SLaM inpatient ward between 1 April 2008 and 31 March 2018. Further, we only included patients aged 17 or 18 on the date of admission. As discussed further below, we chose this narrow age range to be able to relax the assumptions underlying our analysis. Patients also needed to have a primary working diagnosis of psychosis at admission, that is a diagnosis captured by the ICD-10 codes F1x.5, F2x, F30, F31, F32.3 or F33.3 where F1x.5 and F2x refer to all diagnoses of the same form (e.g. F10.5, F15.5, F20, F25, F29) (WHO, 1992). We considered the working diagnosis at admission to be the more appropriate than the discharge diagnosis for this analysis for two reasons: First, it is less likely to be influenced by whether a patient was first seen by an adult or a CAMHS mental health professional. Second, the complaints presenting at admission are those that motivate admission to both CAMHS and adult inpatient care rather than diagnoses that emerge during the course of the admission. In order to determine the working diagnoses at admission, I reviewed clinical notes in chronological order starting from a week before the date of admission until I considered it possible to make a diagnostic judgement. Finally, from this population we excluded patients who were living outside of SLaM catchment area at the time of admission. SLaM CAMHS inpatient wards are more likely to admit out-of-area residents than adult wards so including out of area patients could have led to an imbalance in patient characteristics between those admitted to adult and CAMHS wards.

3.2.4 Outcome measures

The length of stay in psychiatric inpatient care within one year of admission to the index episode was our primary outcome measure in this study. We chose the length of inpatient stay rather than cost of care as our primary measure because it is unclear whether the differences in unit cost between CAMHS and adult inpatient bed days can be accounted for by factors specific to CAMHS wards (e.g. provision of schooling which is not needed for most adult patients) or the non-specific elements (e.g. differences in staffing levels and staff contacts for all patients psychosis) whose effects is of interest in this study. The total cost of

community and inpatient psychiatric care within one year of the start of the index admission was, however, one of our secondary outcomes. As described in more detail in Appendix B, I used financial data from SLaM to cost this service use, inflating figures to 2018 price levels. Further, I assessed differences in the proportion of people who were detained under the Mental Health Act, that is had a compulsory admission, during parts of their index admission, the length of detention among those who were detained under the Mental Health Act, the length of the index admission and number of community contacts within one year of the start of the index admission.

3.2.5 Regression discontinuity design

One approach to estimating the ATU described above could be to compare underage patients admitted to adult wards with underage patients admitted to CAMHS wards under the 'no unmeasured confounding' assumption. Since, as described in Chapter 1, underage patients are typically only admitted to adult ward when there are bed shortages, one may argue that that this assumption is plausible. However, since age-inappropriate admissions are a relatively rare event, estimates would be very imprecise using this approach based on data from only one trust. Therefore, this identification strategy would not be practical in this study. Similarly, to estimate the ATE described above, one could compare patients aged 18 to 25 admitted to adult wards with those 18 to 25 admitted to CAMHS wards under the 'no unmeasured confounding' assumption. However, not only is the number of 18- to 25-year olds admitted to CAMHS wards also small but, the further away from the threshold of 18, the more likely it will be that no patient is admitted to CAMHS wards. In other words, it is likely that there is a lack of overlap in the age distribution of those admitted to CAMHS and adult wards. Given that research suggests that people with early onset psychosis have a different level of premorbid functioning and different illness trajectories compared to those with late-onset psychosis and that these factors are not well or consistently recorded in CRIS, again, a different identification strategy is required if extrapolation is to be avoided (Golay et al., 2017; Immonen et al., 2017; Petersen et al., 2012).

An alternative that could be used to evaluate the estimand of interest within this context given the age-based division of services (see Section 1.5.1) is to compare the outcomes

between people admitted to CAMHS wards aged 17 or 18 to those admitted to adult ward aged 17 or 18. In technical terms, this approach is known as a locally randomized regression discontinuity design (RDD) (Branson and Mealli, 2018). Its assumption may be weaker than the 'no unobserved confounding' approaches described above insofar as clinical knowledge suggest that there are unlikely to be meaningful variation in average disease severity or their ability to benefit from inpatient care among those at risk of hospitalization across this narrow age range. Unlike the aforementioned approaches based on the 'no unobserved confounding' assumption, this approach does, however, also assume is that the characteristics of patients who are not just at risk of but in fact admitted to inpatient care does not vary by within this age range. In other words, it assumes that admission thresholds are the same for patients aged between 17 or 18. More specifically, we assume that, conditional on the number of previous psychiatric admissions to SLaM, the gender and ethnicity of the patient and that a patient is admitted age 17 or 18 to a ward, the type of ward that they are admitted to (i.e. a CAMHS or adult ward), is independent of their unmeasured characteristics. We chose not to adjust for the psychosis subtype at admission because we believe that the judgements made in relation to these were likely to be influenced by the ward type a patient was admitted to. In contrast to the 'no unobserved confounding' assumption, this assumption has some quantifiable implications which provide us with circumstantial evidence to judge its plausibility (Lee and Lemieux, 2010). First, there should be no substantial differences in hospitalization rates between 17- and 18-year olds. Second, although we adjust for confounders so it is not necessary for this condition to hold, we would also expect that there are no substantial difference in measured confounders in those hospitalized on a CAMHS ward and those hospitalized on an adult ward at the age of 17 or 18.

To draw conclusions in relation to the objectives of this study, when replacing the 'no unobserved confounding' assumption, it is also necessary to make an assumptions about the how the treatment effect varies by age at admission. In term of the first objective of this study, the effect of admitting underage patients to CAMHS instead of adult wards, we would expect that, if anything, this treatment effect would be larger the for patients who are younger than 17 at admission. In other words, we would expect that, if at all, younger patients would be more adversely affected by age-inappropriate care. Thus, we assumed

that, if the treatment effect is not the same for 17 and 18-years olds as for all underage patients admitted to adult wards, then the above comparison at least represent a lower bound to the ATU. By contrast, it appeared plausible to assume that, if anything, the potential benefit of extending CAMHS inpatient care to 25 would diminish for older the patient. Thus, if effects of CAMHS care are not constant across age, our estimates represent an upper bound for the ATE. In Appendix C.4, I specify these assumptions and others made in the analysis more formally. I am unaware of a third alternative to estimate the parameter of interest that would be feasible to implement with the dataset at hand.

3.2.6 Statistical analysis

I used a graphical approach to assess whether patients hospitalized at the ages of 17, were different in terms of the covariates compared to those hospitalized at 18, or whether there was a change in the number of admissions. To complement this, I assessed whether there were statistically significant differences while keeping in mind that this test of model assumptions was likely to have low power, that it could only detect instances in which there is evidence that the RDD assumptions may be violated rather than where there is evidence to support them and that, ultimately, the magnitude of differences not the statistical significance is of relevance (Bilinski and Hatfield, 2018). In the base case analysis, I used a logistic regression approach for binary data, linear regression approach for non-censored non-binary data and censored regression models for non-binary censored data. The censored regression model assumes that conditional on the covariates contained in the model, data is censored randomly. This assumption appears a reasonable base case because it is unclear in what direction service use changes when people move to a different provider. In each case, I used cluster robust standard errors to account for correlation between repeated hospitalisations of the same people.

3.2.7 Sensitivity analyses

To assess the sensitivity of the results to violations of the local independence assumption, I re-analysed the data adding a linear and quadratic term for age at admission to the regression and, separately from this, used the same approach as in the base case analysis but restricted the study population to patients admitted aged between 17.5 and 18.5 years. I assessed the sensitivity of the results with respect to the model specification by using a coarsened exact matching approach and the sensitivity of the results with respect to the choice of estimator by using common alternatives where relevant (e.g. a linear regression instead of a negative binomial regression for modelling count data) (Iacus et al., 2012). Further, I investigated the impact of the approach to handling censored data by conducting a complete case analysis and by assuming that if data was censored, no services were used during the unobserved follow-up. Intuitively, both of these approaches make fairly extreme assumptions but, in absence of evidence to inform more appropriate sensitivity analyses with respect to missing data, the results from these analyses can give some indication regarding the potential uncertainty introduced by missing data.

3.3 Results

3.3.1 Descriptive statistics

Figure 9 shows the flow-chart for this study. I identified 358 patients hospitalized on a psychiatric ward when they were 17 or 18 who fulfilled our inclusion criteria. Of the 358 patients, 229 patients were initially hospitalized on an adult ward and 129 on a CAMHS ward. Overall, approximately 20% were of a non-white ethnicity, about 75% were men, in about 60% of cases it was the first known SLaM admission, and, on average, patients had spent 8 days in psychiatric care in the 6 months prior to the index admission. Figure 10 show the bimonthly averages of measured confounders and the yearly averages by age group. For example, the first dot in the top left figures show that among those patients hospitalized aged between 17 and 17 years and 2 month, approximately 20% were non-white. Many subsequent figures follow a similar format. As shown in Figure 10, in terms of these measured confounders, those hospitalized aged 17 were very similar to those hospitalized aged 18 and

none of the differences were statistically significant. Most patients who were admitted to either ward type had a primary working diagnosis on the schizophrenia spectrum (F2x) (Figure 29). Figure 11(a) shows the percentage of patients admitted to each of the two ward types by age at admission. The figure shows that about 10% of 17 year old patients were admitted to adult ward but only one patient aged 18 at admission was admitted to a CAMHS ward. In other words, almost all admissions to age-inappropriate wards involved underage patients being admitted to adult wards. Among those aged 17 at admission there was no clear tendency for older patients to be more likely to be admitted to an adult wards. Clinical notes suggested that, in about 70% of cases, age-inappropriate admissions were due to CAMHS bed shortages (see Figure 11(b)). Figure 12 shows that hospitalization rates for 18-year olds were more than a third higher than those for 17-year olds. This difference appeared relatively stable across the age range under consideration and was also statistically significant. The percentage of people who moved out of the catchment area over the course of the 1-year follow-up was relatively low in absolute terms (<20%) and comparable in both treatment groups (see Figure 30).

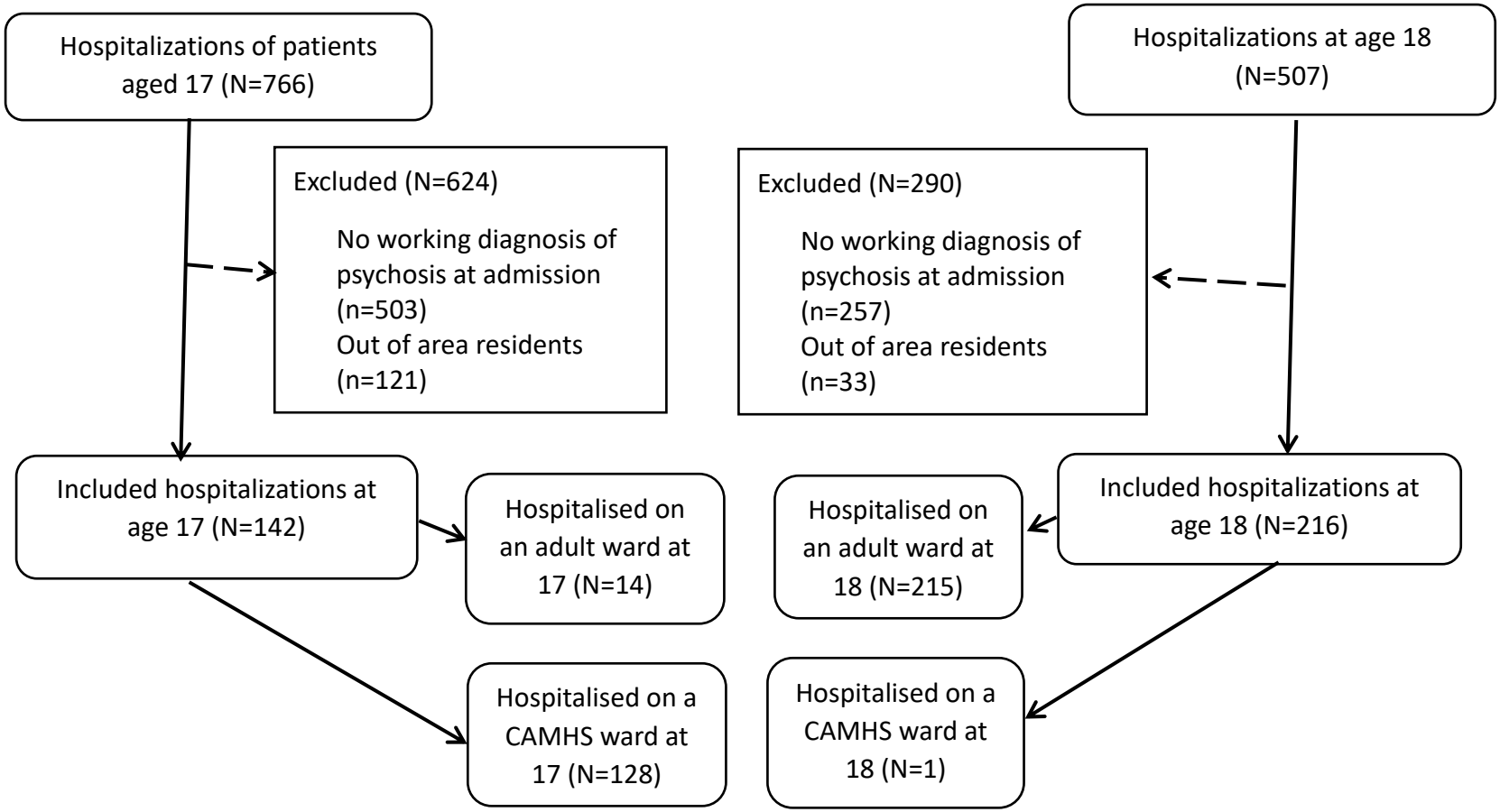


Figure 9 Study flow-chart (Analysis

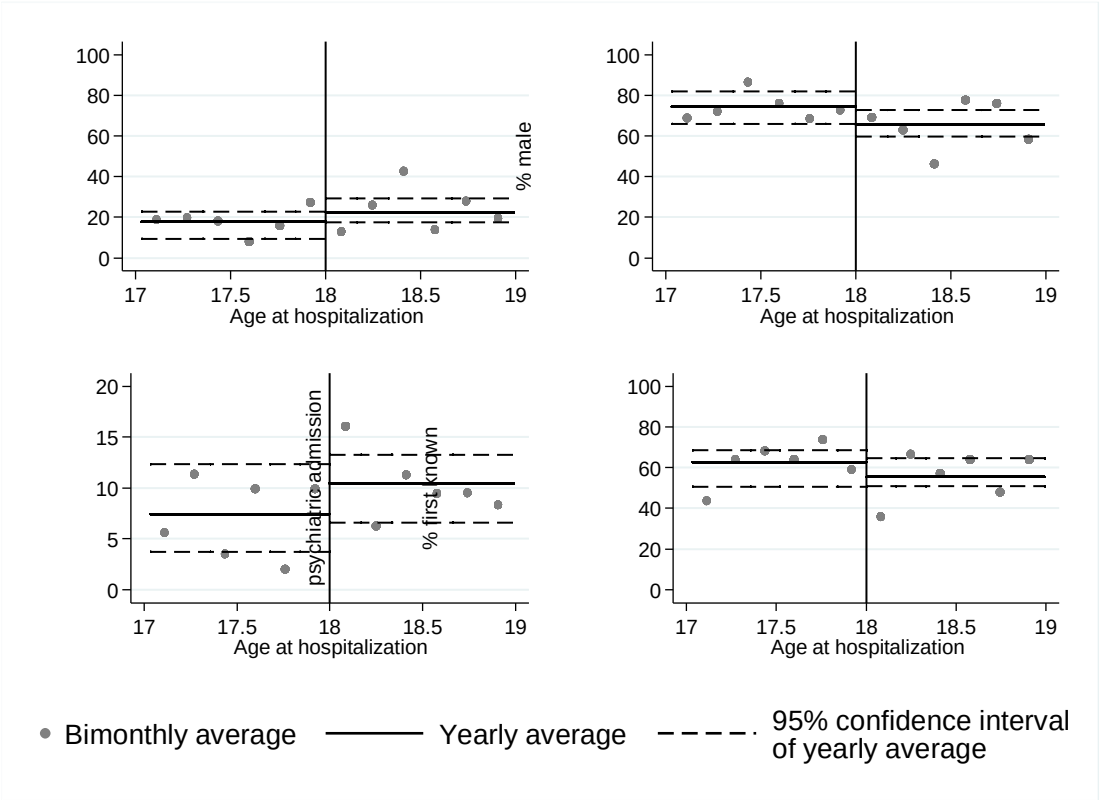


Figure 10 Patient characteristics and their distribution by age of admission

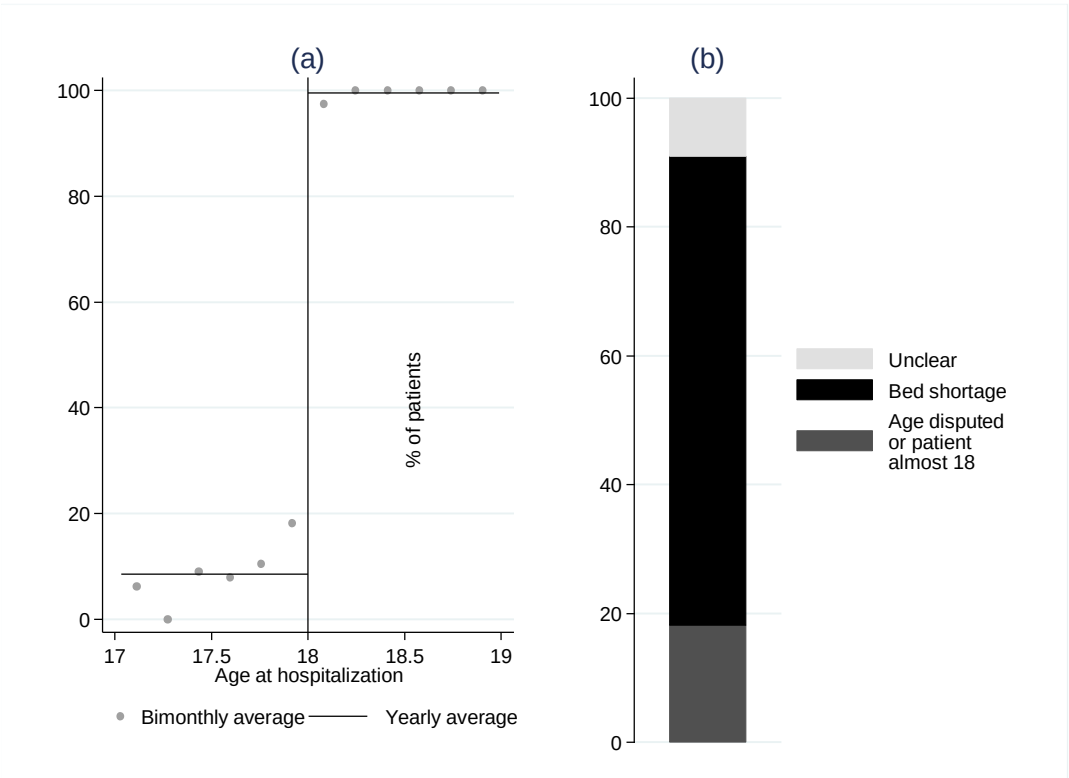


Figure 11 Type of psychiatric ward admitted to by age at admission (a) and reasons for age inappropriate admissions (b)

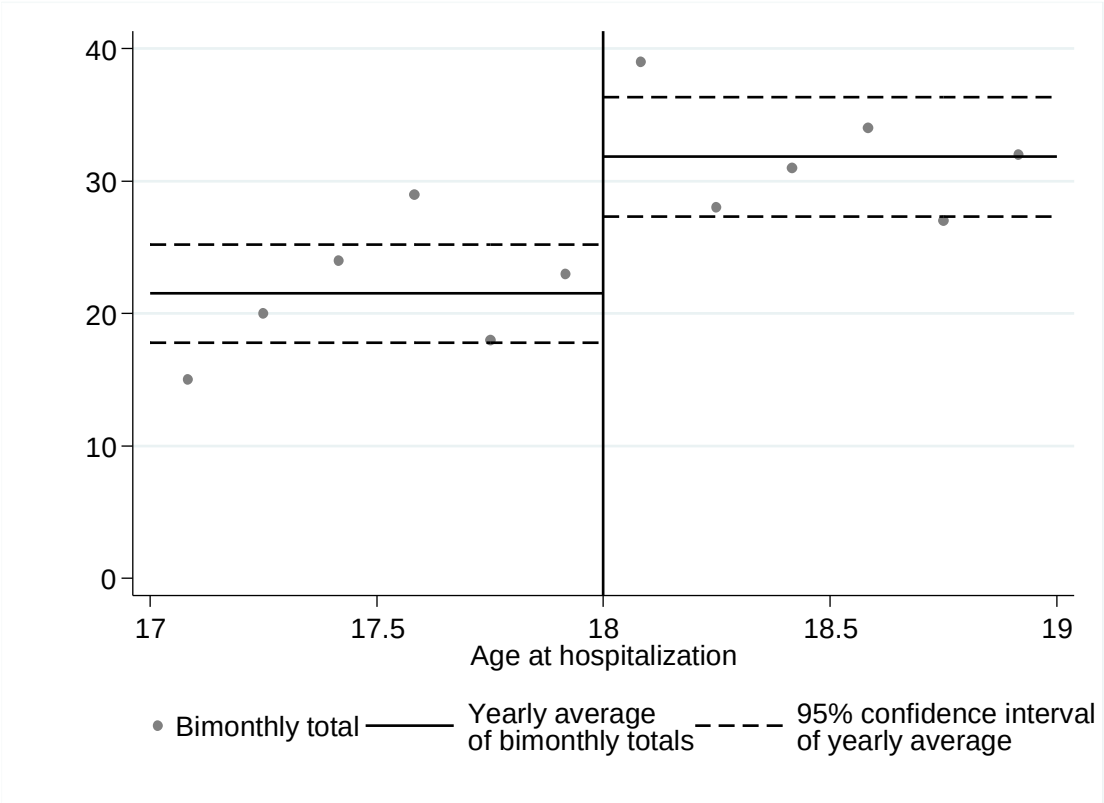


Figure 12 Number of hospitalizations to psychiatric wards by age at admission

3.3.2 Cost-consequence analysis

Figure 28 reflects the extent to which there was movement between CAMHS and adult wards within the same admission and the type of CAMHS or adult ward patients were hospitalized in. The left pair of bar charts in Figure 28 shows what percentage of bed days were spent in one of four ward types depending on the type of ward patients are admitted to. They show that the % of bed days spent in specialist wards was higher for those initially admitted to adult than to CAMHS wards. The right pair of bar charts in Figure 28 shows what percentage of bed days were spent in one of four ward types for an average patient. If the left pair of bar charts were equivalent to the right pair, then this would imply that the share of admission spent across the four ward types would be the same for patients admitted to a given ward type. However, here it appears that a small number of patients account for long admission on specialist wards, both among those initially admitted to CAMHS and to adult wards. Both parts of Figure 28 show that movement between CAMHS and adult wards

within the same admission was very limited both in terms of bed days and the number of patients. The average (Standard deviation (SD)) length of the index hospitalization was 59 (57) days following initial admission to CAMHS wards and 78 (196) days following admission to adult wards (see Figure 31 for the full distribution and Figure 32 for average rates by age of admission). After adjusting for measured confounders, length of the index admission was estimated to be 15 (95% Confidence Interval (CI) -9 to 40) higher in those admitted to adult wards but this estimate was very uncertain. The results of the sensitivity analyses were similar (see Figure 33). As shown in Figure 34, at 11% (CAMHS) and 9% (adult ward), the fraction of patients who were detained under the Mental Health Act at some point during their index hospitalization was similar across admissions to the two ward types. Figure 35 shows the results of the base case analysis at the top which corresponds to what is shown in more detail in Figure 34. The other rows in Figure 35 show the results of the sensitivity analyse. Other figures follow a similar format. Figure 35 shows that, with respect to the estimated differences in detention rates, there do not appear to be meaningful differences across the sensitivity analyses. Similarly, there was no clear evidence to suggest that difference in the length of sectioning differed across the two ward types, but the results were unstable due to skewed distributions as shown in Figure 36 and Figure 37. Within a year of discharge from both CAMHS and adult wards, about 40% of patients were estimated to have been readmitted (Figure 38). Readmission rates were estimated to be 1 (95% CI -10 to 13) percentage point higher in admissions to adult wards in the base case analysis and the results of the sensitivity analyses were very similar (Figure 39). The average (SD) number of community contacts over the study follow-up was 31 (21) following initially admission to an adult ward and 36 (26) following initial admission to a CAMHS ward (see Figure 40 for the change in the number of contacts over time). After adjusting for measured confounders and loss to follow-up, there was no evidence to suggest that there were substantial differences in community contacts between the ward types (Figure 41). There was also no evidence to indicate that there were substantial differences in the number of bed days over the study follow-up (see Figure 42, Figure 43 and Figure 13). In the base case analysis, for example, the number of bed days was estimated to be 4 (95% CI -18 to 27) days higher following admission to adult wards. There was evidence suggesting that, due to higher cost per bed day for CAMHS wards compared to adult wards, secondary mental health care costs were substantially higher following admission to a CAMHS ward (see Figure 44 and Figure 45). In

the base case analysis, I estimated costs to be -£16,774 (95% -34,648 to 1,099) lower following adult ward admission than admission to CAMHS. This estimate was, however, not robust to changes to the local randomization assumption and lower in magnitude when changing the assumption with respect to missing data.

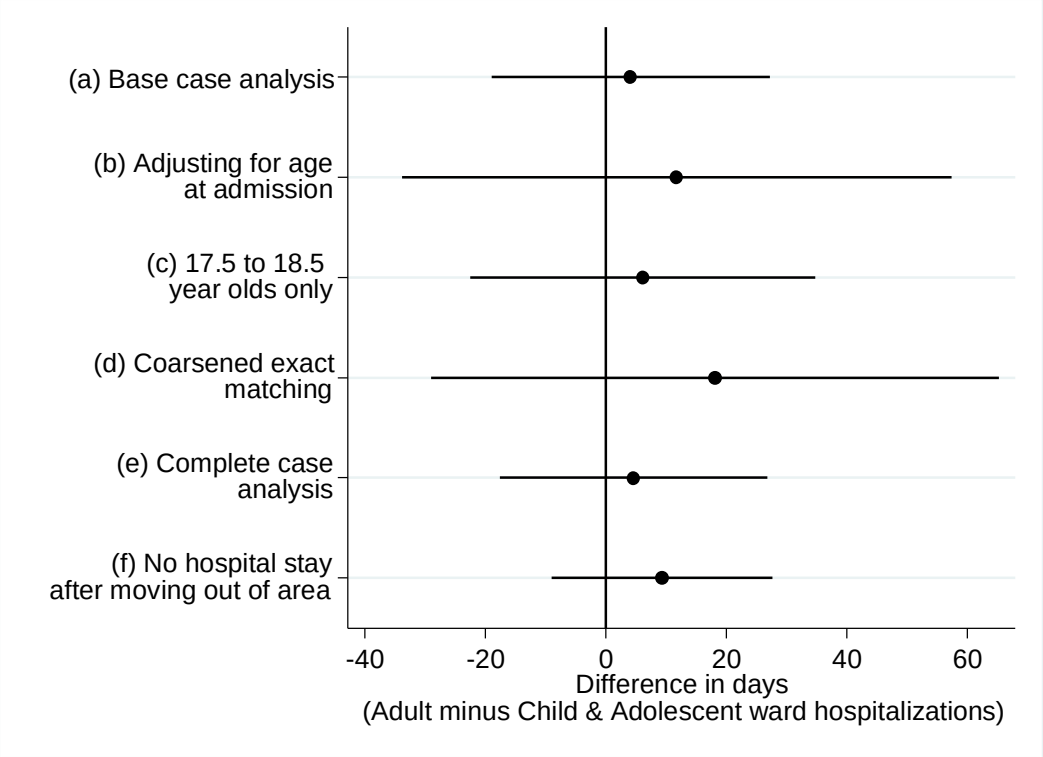


Figure 13 Length of stay on psychiatric ward within one year of the start of the index admission

3.4 Discussion

3.4.1 Key results

In this study, I used routinely collected data to conduct a retrospective cohort analysis with the aim of assessing the economic value of inpatient care modalities for young people. I found that, on average, the amount of service use, clinical outcomes and proxies thereof such as rehospitalisation rates were very similar between those admitted to a CAMHS ward and those admitted to an adult ward. On the other hand, there was evidence that cost of

service use was higher in those admitted to CAMHS wards because of the higher cost per bed day in these ward types. However, evidence suggested that the admission threshold is substantially lower for 18-year old patients. This implies that the RDD assumptions are likely to be violated, with the consequence that the comparison is likely to be biased in favour of admission to adult wards. Moreover, estimates were not estimated very precisely due to the low sample size. Therefore, I do not believe that these results can inform decision making.

3.4.2 Strengths and limitations

The key limitations of this study is that there is considerable risk of bias because there appear to be substantial differences in the admission thresholds for 17- and 18-year olds. In other words, the assumptions underlying the regression discontinuity design appear to be violated. However, a strength of the regression discontinuity design was that it allowed us to produce the quantitative evidence on which to base this judgement rather than being limited to a qualitative consideration. As discussed in Appendix B, we would argue that measurement of costs and mental health service use was of high quality. However, we only capture the benefits of inpatient care for young people through limited proxy measures. Beyond standardised clinical outcomes, a more comprehensive assessment may include the impact of inpatient care modalities on educational attainment and the patient's family or carers. As discussed in Chapter 2.4, the fact that I manually coded diagnosis based on clinical notes is likely to have been more reliable than basing the analysis on data from structured fields or NLP applications. Yet, it is likely that I misclassified some diagnoses due to lack clinical training.

Within the context of economic evaluations of child and adolescent mental health interventions, the size of the study sample is large (Beecham, 2014). However, compared to the magnitude of treatment effect that one might expect, we judge the estimates in this study to be too imprecise to draw firm conclusions even if we ignore the risk of bias. To our knowledge, it is more common for underage than adult patients with psychosis to remain on non-psychiatric wards for their care if they have previously been admitted to such wards rather than being transferred to a psychiatric ward. However, we believe that such cases are

rare so believe that our study sample was likely to be broadly representative and reflective of routine practice. This study is based on data from a single mental health trust which may differ from others in the NHS both in terms of its patient population and service provision. It should also be noted that the estimated parameter blurs the impact of care provided on a CAMHS ward compared to that provided on an adult ward with potential differences in community care provision following discharge between those aged 17 and 18 as well as the impact of transitioning between CAMHS and adult care which. Finally, we did not identify the cause for the difference in unit costs between CAMHS and adult wards which, as discussed above, would allow us to interpret the estimated difference in costs given our interest in non-specific effects of CAMHS inpatient care.

3.4.3 Comparison with the existing literature

As explained in Section 1.5.2, the existing evidence on the impact of different forms of inpatient care for people with psychosis is extremely limited and my non-systematic review only identified one study by Maxwell et al. (2019). In contrast to our analysis, Maxwell et al. (2019) find that the introduction of an integrated care services for young people leads to an increase in care contacts. It is, however, unclear whether this increase can be attributed to the service reconfiguration. It is also unclear to what extent our analysis is comparable with Maxwell et al. (2019) because, rather than expanding the CAMHS-style service model to young adults as assumed by our analysis, the findings by Maxwell et al. (2019) suggest that the service integration led to an equalisation of service provision across young people. In other words, the interventions under comparison differ between our analysis and Maxwell et al. (2019) and the comparison by Maxwell et al. (2019) may be the one that more closely mirrors the options that can realistically implemented in practice in other parts of the UK.

3.4.4 Implications for policy and research

Due to the lack of precision of the estimates and the significant risk of bias, we would argue that the findings of this study should not form the basis for decision making. Assuming that

admission patterns are similar in the rest of the country, our findings suggest that that replicating this analysis using a larger database such as Mental Health Minimum Services Data Set (MHSDS) is unlikely to be worthwhile because, while this would reduce the variance of the estimates, but the risk of bias would remain or even increase. In principle, one possibility to avoid the bias due to differential admission thresholds could be to reframe the analysis by assessing the effect of the impact of the care package offered to 17 year olds who suffer a psychotic episode compared to the care package offered to 18 year olds. However, in CRIS, I would expect that it would be difficult and labour intensive to identify all patients who are at risk of hospitalization. Instead, I would argue that a more fruitful approach would be to use a dataset that is either larger or dataset from different mental health trusts. A larger dataset may enable the analyst to compare people aged 17 who are admitted to adult wards with those admitted to CAMHS wards directly which was not a viable option in this sample given the small number of underage patients admitted to adult wards (n=14). Trusts which have undergone the reconfiguration of CAMHS services to encompass young people up to the age of 25 may yield more direct estimate of the impact of a change in the threshold between CAMHS and adult care from 18 to 25, using before-and-after, difference-in-difference or related approaches (Daw and Hatfield, 2018; O'Neill et al., 2016). Since our study suggests that underage patients are typically hospitalized to adult ward for reasons unrelated to their clinical characteristics, with a larger dataset it may also be possible to obtain a direct estimate of the effect of being admitted to adult wards rather than CAMHS wards underage patients that of acceptably precision. A potential practical barrier such an analysis could be that the type of ward that underage patients are admitted to (i.e. adult or CAMHS wards) may not be sufficiently reliably recorded in MHSDS.

Chapter 4 Economic evaluation of inpatient care for people with enduring psychosis (Analysis 2)

4.1 Introduction

In this chapter, I conduct an economic evaluation of inpatient rehabilitation compared to care as usual. In Chapter 1, I described the broader context for this analysis, its relevance to health care decision making and the existing evidence on this topic. To summarise, inpatient rehabilitation is a specialist, long-term form of inpatient care for people with complex forms of psychosis (Killaspy, 2014). In 2009, all mental health trusts in the UK had at least one inpatient rehabilitation ward or its community equivalent and, in 2017, the estimated annual expenditure for inpatient rehabilitation in England was £535 million (CQC, 2018; Killaspy et al., 2017). NICE guidance on the care for people with complex forms of psychosis including guidance on inpatient rehabilitation is currently under development (NICE, 2018). All six existing observational economic evaluations of inpatient rehabilitation argue that their findings favour the provision of this form of specialist care. However, the credibility of these findings is doubtful because of weaknesses in the study designs. For example, all except one of the previous studies use a before-and-after design. This design relies on the assumption that outcomes would have remained unchanged had patients not received care in an inpatient rehabilitation ward. Given the variable course of psychosis, it is unclear whether this assumption holds (Tandon et al., 2009). In addition, the relevance of the findings of these studies to current decision making in the UK is limited. For example, only one existing study included the considerable costs of the inpatient rehabilitation itself in its evaluation. By contrast, the two existing RCTs suggest that inpatient rehabilitation is not good value for money but these two studies also suffer from methodological limitations or are not directly generalizable to the UK setting. Like in Chapter 3, I conducted a cost-consequence analysis, that is I present costs and other outcomes alongside each other rather than combining them (Mauskopf et al., 1998). My objective was to assess the average

effect of being referred to inpatient rehabilitation on those who received such a referral compared to receiving care as usual.

4.2 Methods

4.2.1 Study design and comparison of interest

I undertook a cohort study based on routinely collected data. The estimand of interest in this study is the average effect of being referred to in-area inpatient rehabilitation ward at the start of follow-up, among those who are referred for inpatient rehabilitation at any point during a psychiatric hospitalisation, compare to continuous receipt of 'usual care'. 'Usual care' comprised any type of inpatient or community care other than inpatient rehabilitation. We censored, i.e. stopped follow up, in both groups when one of events occurred: being admitted to an out-of-area inpatient rehabilitation ward, moving out of the catchment area of the local mental health trust or reaching the end of follow-up (31st March 2019). Appendix E.3 provides a more formal description of the target quantity of interest.

As discussed in more detail below, in the base case analysis, I compared patients who were referred to in-area inpatient rehabilitation at baseline to matched patients who were not referred. This approach may be biased because of unmeasured factors (e.g. social functioning) that differ between the groups. I therefore, undertook several sensitivity analyses to assess the robustness of these findings. One of these was a before-and-after analysis which assumes that there are not time-varying factors that influence outcomes. Further, I implemented two variants of the front-door adjustment, an approach I introduced in Section 2.3.3. The front-door adjustments approaches relied on comparing subgroups of patients among those who were referred to inpatient rehabilitation, namely (1) those referred and accepted to rehabilitation with those not accepted and (2) those accepted and transferred to inpatient rehabilitation with those accepted and not transferred. The intuition behind this approach is that we do not expect that merely being referred (or being referred and accepted) to inpatient rehabilitation to change outcomes substantially.

However, patients whose referral was declined or who were not transferred to rehabilitation following acceptance to it, could reasonably be expected to be more similar in terms of unmeasured factors (e.g. social functioning) to those who did receive inpatient rehabilitative care, i.e. represent a better comparison group, than those who were never referred. This is because of the very fact that they were referred (or referred and accepted) to inpatient rehabilitation, i.e. they are known to have been under active clinically consideration to receive inpatient rehabilitation at some point. By contrast, this is not known to be the case for the matched controls.

4.2.2 Data source and setting

Our data source was the South London and Maudsley (SLaM) Biomedical Research Centre (BRC) Clinical Records Interactive Search (CRIS) database which I described more at length in Chapter 2.2. Within the time-window of this study, SLaM provided inpatient rehabilitative care in four wards. Using the typology by Killaspy (2009), we would describe all four of these wards as 'high dependency rehabilitation units' but one of them was a low-secure facility. In 2015, the care quality commission (CQC) rated the SLaM trust as a whole as providing 'good' quality care overall and gave the rating 'good' on all dimension except for safety (CQC, 2016a). The same overall and dimension-level ratings were given to SLaM rehabilitation wards for working age adults (CQC, 2016b). This suggests that this study assesses the value for money of reasonably well functioning wards

4.2.3 Study population

To be included in the analysis, patients needed to (1) be hospitalized on a SLaM ward at the start of follow-up; (2) have been originally hospitalized on a SLaM ward between 1 April 2010 and 31 March 2018; (3) be resident within the SLaM catchment area at the start of the psychiatric admission; (4) have a current primary diagnosis of schizophrenia, schizoaffective disorder or bipolar disorder at the start of follow-up, that is a diagnosis covered by the ICD-10 codes F20, F25 or F31 (WHO, 1992). From this group, we excluded observations if (1) it

was known that patients had been admitted to psychiatric inpatient rehabilitation previously; (2) patients were either below the age of 18 or above the age of 62 or if the patient's age was unknown; (3) patients were detained under a Mental Health Act Section other than Section 2 or Section 3 at the start of follow-up; (4) were not staying on psychiatric intensive care unit (PICU) or a general adult ward at the start of follow-up; (5) patients had been admitted to a specialist ward for learning disability, addictions, mental health of older adults or child and adolescent mental health care at some point prior to the start of follow-up within the same admission.

4.2.4 Potential confounders

Patient characteristics that influence referral to inpatient rehabilitation are also likely to influence clinical outcomes and costs. This can cause a distortion, known as confounding, in the comparison between patients referred to inpatient rehabilitation and those who are not. To reduce the potential bias of our results due to confounding, based on clinical experience and previous literature, we identified the following measured potential confounders in CRIS and accounted for them in the analyses (Tulloch et al., 2016a; Williams et al., 2014): (i) demographic characteristics of the patients, that is age, gender and ethnicity; (ii) history of service and medication use prior to the index admission, that is length and number of psychiatric hospitalisations in the year prior to the index admission, whether the patient was known to have had prior contact with SLaM community rehabilitation services and whether patients had a known history of clozapine use; (iii) variables characterising the length and course of hospitalisation between the start of the index psychiatric admission and start of follow-up, that is the number of days hospitalized on a general adult ward (as opposed to specialist wards) at the start of follow-up, the number of days detained under the Mental Health Act between the start of the index admission and the start of follow-up, sectioning status at the start of follow-up and the type of ward that there were cared in at the start of follow-up (i.e. general adult ward or PICU) and how many risk event the patients were known to have been involved in the 30 days prior to the start of follow-up; (iv) assessment of patient symptoms by means of (a) the closest Health of the Nation Outcome Scale (HoNOS) rating within the same psychiatric

admission (Pirkis et al., 2005). HoNOS is a clinician-rated outcome measure that has become a standard instrument for measuring patient outcomes in UK secondary mental health care in adults (Jacobs, 2016, 2009). It consists of 12 items scored on a scale from 0 (no problem) to 4 (severe problem) which assess patients' psychiatric symptoms, physical health, functioning, relationships and housing. When used in research projects, HoNOS is thought to have adequate psychometric properties (Pirkis et al., 2005). Specifically, we selected HoNOS domain 1 (aggression), 6 (hallucinations), 9 (relationships), 10 (activities of daily living), 11 (living conditions) and 12 (occupation and activities), and dichotomised the ratings into no/minor/mild problems or moderate/severe problems because we believed that this approach would make the scores more analytically tractable while being clinically meaningful; (b) description contained in clinical notes. Specifically, we focused on whether the patients was described as having poor motivation, symptoms of social withdrawal and whether the patient was self-neglecting or at risk thereof at any point in the 30 days prior to start of follow-up or since the date of hospital admission whichever period was longer.

4.2.5 Outcome measures

We measured service use and costs thereof over a three-year follow-up adopting a mental health care service perspective. As described in Appendix B, to cost service use, I used unit costs calculated using financial data obtained by SLaM. I measured costs in pound sterling (£) at 2017/2018 levels, used the hospital and community health services index to inflate prices to this financial year and discounted costs arising after the first-year of follow-up using a discount rate of 3.5% (Curtis and Burns, 2018). In terms of measuring the effectiveness of the inpatient rehabilitation, I took a patient perspective. Specifically, I assessed whether there were differences in the selected HoNOS dimensions at discharge from the index admission. In addition, I used rehospitalisation rates in the year after discharge from the index admission as a proxy indicator for potential benefits of inpatient rehabilitation (Burns, 2007). Finally, I estimated differences in length of survival over the three-year follow-up.

4.2.6 Measurement error

To make weaken the assumptions with respect to measurement error in the data while reducing the amount of manual coding required, I chose to manually code data selectively based on the following process:

(1) *Validating dates for the treatment group:* We manually verified dates related of the clinical pathway to inpatient rehabilitation (e.g. the date of referral). As explained in Appendix A.3, on inspection, these dates appeared to be incomplete in structured fields and, where available, subject to measurement error. Given that many aspects of the analysis (e.g. measurement of the outcome, appropriate choice of control group) hinged on identifying the correct inpatient rehabilitation referral/acceptance/transfer dates this was a key step in reducing the impact of measurement error. To begin, we manually verified these dates by reading the clinical notes around dates that were contained in structured fields. In this process, we identified key phrases that were used in clinical notes when discussing the pathway to inpatient rehabilitation (e.g. “Referral form for [name of rehabilitation ward]”). I then searched the entire CRIS database for records containing any of these key phrases to identify referrals to inpatient rehabilitation that were not recorded in structured fields and manually coded these additional dates. I adopted this approach because it would have been prohibitively labour intensive to read through all medical records in order to identify patients who were referred to inpatient rehabilitation but for whom no data was recorded in relation to these referrals in structured fields. Typically, referrals that were not recorded in structured fields were those that did not result in a transfer to inpatient rehabilitation

(2) *Validating confounders and diagnosis for the treatment group*: There were some confounders (and the diagnosis variable) for which we judged that, if we relied on the structured fields or outputs of natural language processing (NLP) applications, there was a considerable risk of bias due to measurement error. The treatment group was relatively small in size. Therefore, I manually coded the characteristics of the patients at the day of their referral to inpatient rehabilitation for the entire treatment. In this process, I did, however, make use of the proxy data from

structured fields or NLP outputs, to limit the amount of manual coding where this appeared possible without making noticeably stronger assumptions. For example, when verifying whether patients had a diagnosis of psychosis on the date of referral, I began with those who had a lowest number of instances of psychosis diagnoses based on structured fields or NLP outputs and continued in ascending order of diagnostic instances of psychosis until it appeared almost certain that patients truly had a diagnosis of psychosis at referral. Similarly, I used the number of mentions of clozapine in the free text as a proxy variable and only manually verified diagnoses up to a point at which I judged it to be almost certain that people had a history of clozapine use.

(3) *Reducing the number of controls*: However, manually coding the confounders for all potential controls and verifying that all potential controls had a diagnosis of psychosis would have been prohibitively labour intensive. Therefore, as described in Section 2.4.3.4, I used matching as a way to reduce the size of the control group so that these variables could feasibly be manually coded. Although the control group could have also been reduced by simple random sampling among the potential controls, as noted in Section 2.4.3.4, matching to has several advantages over simple random sample in this context: it reduces the variance of the parameter of interest compared to random sampling because the proxy variables data from structured fields and NLP applications used in the matching process go at least some way in reducing imbalance between the treatment and control group (Stuart and Ialongo, 2010), it makes the base case analysis more robust towards model misspecification and reduced the stringency of the positivity assumptions in the analysis (Kreif et al., 2013a; Petersen et al., 2012).

I used a 1:1 nearest neighbour matching approach with replacement, that is, I matched one control observation to each inpatient rehabilitation referral. I chose rank-based Mahalanobis distances as the matching metric (Rosenbaum, 2010). Rather than matching on the proxy variables themselves (e.g. the number of clozapine instances in structured fields or NLP outputs) I used the predicted values based in the data coded in step (2) in the matching process. For example, I estimate the predicted probability of having had a history of clozapine use for all potential

control observations based on the number of clozapine instances in structured fields or NLP outputs. Since patients could be referred at any point during an admission, I stratified the data sequentially, creating one row for each day that a patient was an inpatient, determining the value of the confounder for each of these days. I then matched inpatient days on which a referral took place with potential control inpatient days that had the same temporal distance to the date of admission. If, for instance, a patient was referred to inpatient rehabilitation 16 days after admission to a psychiatric ward, only control observations whose length of inpatient stay on a psychiatric wards was 16 days or longer and who fulfilled the study inclusion criteria on day 16 after admission were potential matches. These observations were then matched based on the value of the matching variables on day 16 and follow-up for both the treated patient and their matched control would begin on day 16 after admission. This approach has been described as risk-set matching (Li et al., 2014, 2001; Rosenbaum, 2010).

(4) *Manually validating confounders and diagnosis for the control group*: I manually verified whether matched observations had a diagnosis of psychosis as described above in the context of the treatment group. If a patient did not have a diagnosis of psychosis, I excluded him/her from the analysis and replaced the observation by the next closes matched and in turn checked the diagnosis until all matched controls were known or very likely to have a diagnosis of psychosis. To decrease the amount of manual coding, I excluded all patients who had never had a primary diagnosis of psychosis prior to baseline according to structured fields or NLP applications from the pool of potential matches. It appeared unlikely that this approach would lead to a significant selection bias because, among the referrals to inpatient rehabilitation there were only two patients who did not have any record of having a diagnosis of psychosis in structured fields or NLP applications but did have a diagnosis of psychosis based on inspection of clinical notes. Finally, I manually coded the potential confounders in the control group using the approach described for the treatment group.

4.2.7 Statistical analysis

To adjust for some of the remaining imbalance between those referred to inpatient rehabilitation and the matched controls, I reweighted the sequentially matched observation and using inverse probability of treatment weights (IPTW) based on a model in which I pooled the matched data (Li et al., 2018). Similarly, I used an inverse probability of censoring weighting (IPCW) with a missingness indicator for missing baseline variables to handle missing data in the analyses (Seaman and White, 2013). Given the large number of covariates relative to the sample size, particularly in year 3 of follow-up and in the sensitivity analyses described below, and the resulting risk of overfitting and/or model misspecification, I used a boosted regression approach to estimate the IPTW and IPCWs (Schonlau, 2005; Stone and Tang, 2013). Specifically, I used the default 80%/20% split between training and testing data, a two-way interaction based on discussion in Hastie et al. (2009) and a shrinkage parameter of 0.0005 as suggested by McCaffrey et al. (2004). I chose this approach instead of the more common approach of combining matching with regression adjustment or using a multiple imputation to address missing data because of the risk of over fitting given the large number of confounders relative to the sample size. To account for correlation between multiple observations of the same patient, I used cluster robust standard errors. Depending on the distribution of the outcome variables, I used a censored, logit or linear regression model. In Appendix E.4 I describe the assumptions on which the base case analysis is based in more detail.

4.2.8 Sensitivity analyses

In our base case analysis we assume that there is no unmeasured confounders. A key concern with this assumption is that we would expect that the matched controls would have had better outcomes than those referred to inpatient rehabilitation had they been referred too. This is because, typically, only some of the most unwell patients on psychiatric wards are referred to inpatient rehabilitation but some of the patient characteristics motivating referral (e.g. poor social functioning) are only poorly approximated by the variables that we measure. As a result, we would expect the base case analysis to be biased against inpatient

rehabilitation. Therefore, I undertook three sensitivity analyses to assess the robustness of the findings to alternative assumptions with regards to unmeasured confounding.

First, I undertook a before-and-after analysis in which I compared outcomes in the two years prior to referral to outcomes three years after referral to inpatient rehabilitation (Bunyan et al., 2016). This approach assumes that, if patients had not been referred to inpatient rehabilitation, on average, outcomes would have remained the same as in the two years prior to referral. Given the variable course of psychosis, it is difficult to judge to what degree or in what direction this assumption might lead to bias (Tandon et al., 2009). Intuitively, we would, however, argue that it is unlikely that patients' outcomes would have further deteriorated compared to the two years prior to referral had they not been referred to inpatient rehabilitation because, typically, they are already in a very poor clinical state in this period. Thus, in this context, a before-and-after analysis may be regarded as the analytical approach that makes the most favourable assumptions with respect to inpatient rehabilitation because it compares the post-referral outcomes with what might be considered to be the worst possible reasonable alternative scenario. This implies that we would expect the true impact of inpatient rehabilitation to be somewhere in between our base case analysis (potentially biased against rehabilitation) and the before-and-after analysis (potentially biased in favour of rehabilitation). As noted in Chapter 1, before-and-after analyses have been the most common approach in existing evaluations of inpatient rehabilitation based on observational data. Thus, this sensitivity analysis is also of value because it gives some indication as to whether the results of our base case analysis differ from the previous literature due to our choice of analytical approach or other factors.

Second, I compared outcomes between patients whose referral to inpatient rehabilitation was not accepted with those referral was accepted and multiplied this difference by the proportion of referred patients whose referrals was accepted. As explained in Section 2.3.3, this approach is known as the front-door adjustment and, for shorthand, I will refer to this sensitivity analysis as front-door adjustment #1. Front-door adjustment #1 relies on two assumptions: (1) the outcomes of patient whose referral to inpatient rehabilitation is declined (or who decline the option of inpatient rehabilitation themselves) would have been the same had they not been referred to inpatient rehabilitation to begin with; (2) the comparison between those whose referral is accepted and those whose referral is not

accepted is not distorted by unmeasured confounders or confounders measured after the date of referral (as opposed to the base case analysis in which the comparison between those referred and those not referred to inpatient rehabilitation is assumed to be undistorted).

With regards to assumption (1), one should note that, based on our clinical experience, referral to inpatient rehabilitation can delay discharge from inpatient care among those whose referral is declined. We would not expect this additional inpatient stay to have any sizeable clinically positive effect on the patient because patients referred to inpatient rehabilitation are typically judged to not be able to benefit from general inpatient care by the responsible clinician. Moreover, we would not expect this delay in discharge to have any large negative clinical effects on the patient (e.g. due to patients becoming institutionalized) because the delay in discharge is typically still relatively short. Finally, we do not believe that this delay in discharge would alter care provision after discharge. Nevertheless, we would expect assumption (1) to bias front-door adjustment #1 somewhat in favour of inpatient rehabilitation because delays in discharge lead to increases in health care costs. However, the resulting bias would, at most, be equal to the average cost of care arising between referral to inpatient rehabilitation and the date that the referral is accepted or declined, a quantity which we observe in our sample. In addition, we believe that, in practice, patients are typically referred to inpatient rehabilitation before being ready for discharge meaning that the quantity would be typically be appreciably lower than this maximum value.

With regards to assumption (2), we would argue that, given the same measured characteristics at the time of referral, patients whose referral was accepted are likely to be more comparable to patients whose referral was not accepted in terms of their unmeasured confounders than to patients who were not referred for two reasons: First, referral to inpatient rehabilitation is a selection process. Thus, we would, for example, expect that, whether their referral is accepted or not, most patients who are referred to inpatient rehabilitation have some deficits in social functioning because this is one of the key reasons for referral to inpatient rehabilitation. Second, in contrast to the base case analysis, the direction of bias in the comparison between accepted and declined referrals is less clear because referrals may both be declined because patients are too unwell or too healthy to benefit from inpatient rehabilitation. Moreover, referrals may be declined because of

reasons that we would not cause bias (e.g. shortages of inpatient rehabilitation beds). In fact, an advantage of this sensitivity analysis is that, whereas the reason for not referring a patient to inpatient rehabilitation is rarely explicitly recorded in clinical notes, the reason that a referral was declined typically is. This provides us with some circumstantial evidence to judge the magnitude and direction of bias due to violations of assumption (2).

In my fourth sensitivity analysis, I use another variant of the front-door adjustment approach which I will refer to as front-door adjustment #2. In front-door adjustment #2, I compared outcomes in patients whose referral was accepted and transferred to inpatient rehabilitation with outcomes inpatients whose referrals was also accepted but who were removed from the inpatient rehabilitation waiting list and then multiplied this difference by the probability that referrals led to a transfer to inpatient rehabilitation. Similar to front-door adjustment #1, this approach assumes that (1) neither referral nor acceptance of a referral to inpatient rehabilitation affects outcomes if they do not result in the patient being transferred to inpatient rehabilitation; (2) whether a patient is transferred to an inpatient rehabilitation ward does not depend on factors that also influence the outcomes other than those measured at the point of referral. One other one hand, front-door adjustment #2 is likely to be more biased in favour of inpatient rehabilitation with respect to the assumption (1) compared to front-door adjustment #1 because the time between referral and removal from the waiting list is likely to be larger than the time between referral and declination of the referral. One other hand, the assumption (2) appears more plausible in front-door adjustment #2 than #1 because both those accepted and transferred to inpatient rehabilitation and those accepted but removed from the waitlist have been deemed suitable for inpatient rehabilitation twice: First by the clinician referring them, second by a rehabilitation consultant following an assessment. In other words, both groups have undergone two selection stages. Thus, we would expect those transferred to inpatient rehabilitation to be more similar to those accepted but not transferred to inpatient rehabilitation in terms of unobserved variables (e.g. social functioning) than those referred with those not referred to inpatient rehabilitation. Again, identifying the reason for removal from the inpatient rehabilitation waitlist by reading the clinical notes provides us with some circumstantial evidence to judge the plausibility of front-door adjustment #2

For both front-door adjustments, I used a bootstrap approach for parameter estimation using only an IPTW approach without prior matching to adjust for measured confounders (Glynn and Kashin, 2017). Table 4 provides a simplified summary and graphical illustration to the different approaches with respect to unmeasured confounding. For a more technical description see Appendix E.4 and E.5. In addition to the sensitivity analyses motivated by the risk of unmeasured confounding, I assessed the robustness of the analysis to the assumptions with respect to missing data. To do so, I conducted a complete case analysis and an analysis in which, if data was censored, I assumed that outcome of interest did not occur, if the variable was binary, or that the outcome was 0 if the variable was continuous.

4.3 Results

4.3.1 Descriptive statistics

Figure 14 shows the study flow-chart. I identified 337 referrals to inpatient rehabilitation that fulfilled our inclusion criteria. Table 3 summarizes the characteristics of the patients referred to inpatient rehabilitation and their matched controls. The average (Standard Deviation (SD)) age at referral was 40 (12), 72% were men and 72% of the sample were non-white. About two thirds of patients had a SLaM admission in the year prior to the index admission, and the average (SD) number of days hospitalized in the year prior to the index admission across all referrals was 62 (78) days. Few (9%) had a known history of contact with community rehabilitation teams and 50% had a known history of clozapine use. The average patient spent approximately 80% of their admission until referral to inpatient rehabilitation on a general adult ward and approximately the same proportion of time detained under the Mental Health Act. Similarly, on the day of referral, about 80% were detained under the Mental Health Act. This was typically under Section 3 of the Mental Health Act. The average (SD) time from admission to a psychiatric ward to inpatient rehabilitation referral was 91 (83) days (see Figure 64 for the full distribution of this variable). Among the selected HoNOS dimensions, with 39% of patients reporting moderate or severe problems, hallucinations were the most commonly reported issue. The average (SD) time between ratings and baseline was 25 (46) days. Self-neglect or risk thereof was a

very commonly reported issue in medical notes (80%) whereas patients with poor motivation (39%) and social withdrawal (15%) were in the minority. Approximately 30% of patients were known to have been involved in a risk event in the 30 days prior to referral.

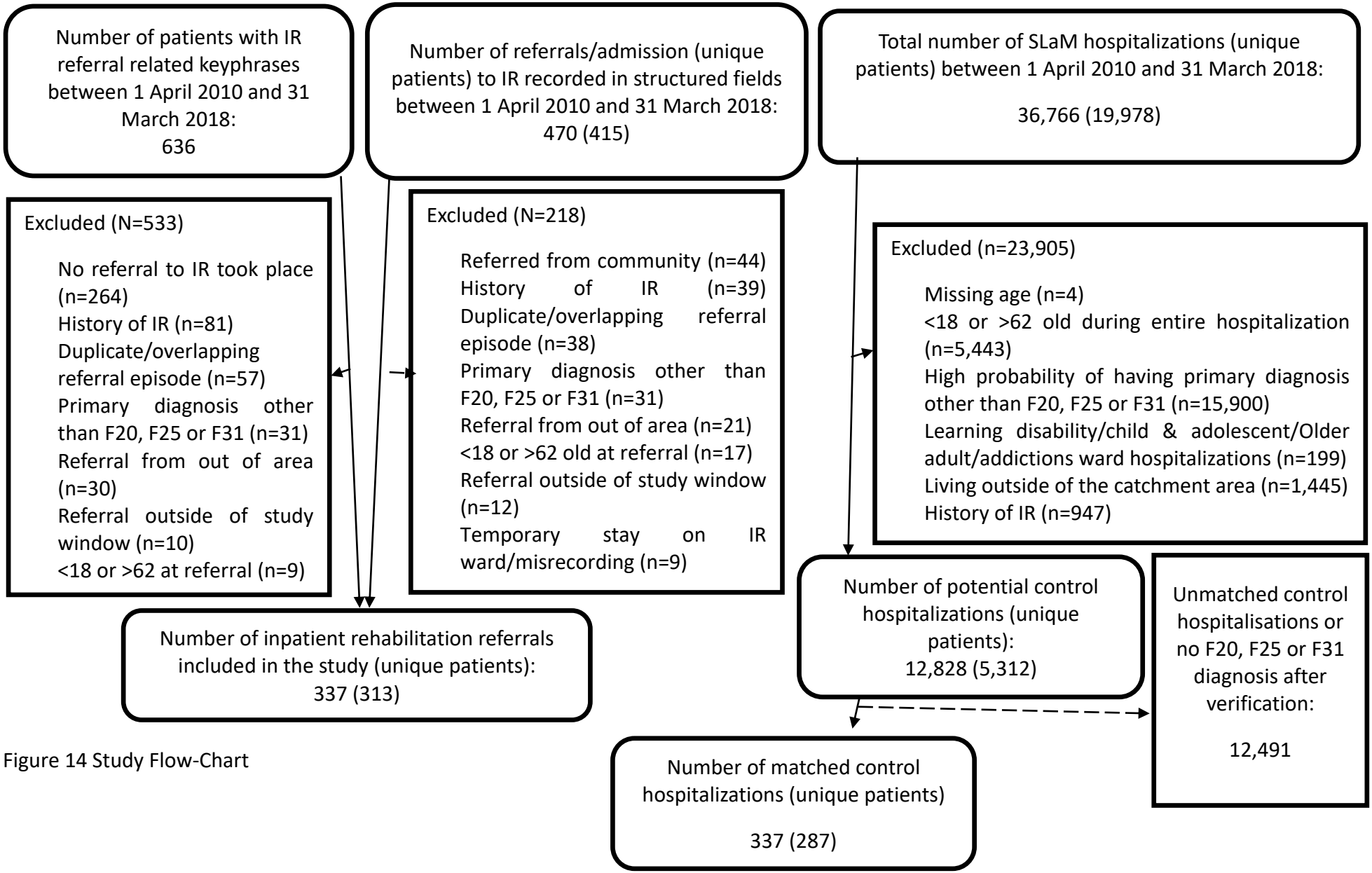


Figure 14 Study Flow-Chart

Characteristic	Referrals to inpatient rehabilitation N=337	Matched Cohort N=337	Difference in means (Ratio of variances*)
Demographics			
Mean (SD) age in years	40 (12)	40 (11)	0 (1.2)
Male (in %)	72	71	1
BME (in %)	72	77	-5
History of service/medication use			
% of observations by number of hospitalizations in the year prior to index hospitalization			
0	44	42	2
1	30	37	-7
2	14	15	-1
3 or more	12	7	6
Mean (SD) number of days hospitalised in the year prior to the index hospitalization	62 (78)	55 (71)	7 (1.2)
% with history of community rehabilitation	9	7	1
% with history of Clozapine use	50	48	1
Characteristics of index hospitalisation between initial hospitalisation and baseline			
Mean (SD) length of stay in days‡	91 (83)	91 (83)	0 (1)
Mean (SD) length of stay on General Adult Ward	75 (73)	78 (77)	-3 (1)
Mean (SD) number of days hospitalised detained under the Mental Health Act	76 (82)	76 (79)	0 (1)
Compulsory admission status at admission on day of (potential) rehabilitation ward admission (in %)	81	81	0
% of observations			

Characteristic	Referrals to inpatient rehabilitation N=337	Matched Cohort N=337	Difference in means (Ratio of variances*)
with no/minor/mind, moderate/severe problems or missing HoNOS dimensions score			
01 – Aggression	79, 18, 3	83, 15, 2	-4, 4, 0
06 – Hallucinations	58, 39, 3	62, 35, 2	-4, 4, 0
09 – Relationships	81, 14, 5	89, 7, 4	-8, 7, 1
10 – Activities of Daily Living	83, 13, 3	91, 6, 3	-7, 7, 0
11 – Living Conditions	74, 17, 9	82, 11, 7	-9, 6, 3
12 - Occupation	80, 12, 8	89, 4, 7	-9, 7, 1
Mean (SD) days between HoNOS rating and baseline date (Baseline – HoNOS rating date)	40 (45)	41 (45)	-1 (1)
% of observations with mention of symptom in clinical notes†			
Poor motivation	39	39	1
Social withdrawal	15	10	5
Self-neglecting or at risk thereof	80	78	2
Fraction of patients (in %) by number of known risk events†			
0	69	76	-6
1	21	16	5
2 or more	9	8	1

*for continuous and count data only; † between the start of the index admission and baseline or 30 days prior to baseline, whichever is larger; ‡ Identical between the groups due to sequentially stratified matching approach

SD: Standard deviation; BME: black minority ethnic group; HoNOS: Health of the Nation Outcome Scale

Table 3 Unweighted baseline Patient characteristics (base case analysis)

To complement Table 3, Figure 47 shows standardised differences between the treatment group. In addition, Figure 49, Figure 50, Figure 51, Figure 52, Figure 53 and Figure 54 show the distribution of patient characteristics for the base case analysis by treatment group graphically. All of these tables and figures suggest that, the group referred to inpatient rehabilitation and matched control were generally well balanced in terms of their measured characteristics. The most notable imbalances arose with respect to HoNOS dimensions 9, 11 and 12. As shown in Figure 50, there was also some minor lack of overlap in the upper end of the distribution of the number of days in hospital in the year prior to admission, the number of days under mental health act and the number of days between the HoNOS rating data and baseline. However, the lack of overlap in the distribution of the predicted probabilities of referral to inpatient rehabilitation was relatively limited (see Figure 55). This means that at baseline, observed differences between those referred to inpatient rehabilitation and matched controls could be balanced out. Figure 48 shows the amount of missing or censored data in the analysis. With missingness rates of around 10% or less, baseline variables were relatively complete but the proportion of missing HoNOS ratings at discharge from the index admission was high (~40-60%). In the base case analysis, the average percentage of people lost to follow-up was around 20% in the first year, and 10-15% in the second year and third year. Figure 58 shows the reasons for loss to follow-up in each year for the base case analysis and Figure 61 overlays the distribution of predicted probability of being lost to follow-up between those who were lost and those who were not in each follow-up year. There was some lack of overlap in the distribution in year 2 for those referred to inpatient rehabilitation, in year 3 for the matched controls and a substantial lack of year 3 in the group referred to inpatient rehabilitation. This means some of those lost to follow-up had no counterpart with equivalent measured characteristics in those not lost to follow-up which could have biased the analysis.

Of the referrals to inpatient rehabilitation, 23% (78/337) were declined and 24% of accepted referrals (62/259) were removed from the waiting list before transfer to a SLaM inpatient rehabilitation ward. Figure 65 shows the distribution of patients across different stages of the inpatient rehabilitation over time from the point of referral. It shows that about half of the patients are assessed for their suitability for inpatient rehabilitation within one month of referral and that the vast majority of patients who are ultimately transferred to

rehabilitation are transferred within 3 months. Similarly, Figure 66 shows the distribution of patients from the point that the referral was accepted. The average (SD) time between referral and acceptance or declination of the referral was 33 (33) days, the average time between referral and declination of the referral for those whose referral was declined was 46 (42) days, the average time between referral and transfer to a rehabilitation ward or removal from the waiting list for inpatient rehabilitation was 69 (63) days and the average time between referral and removal from the waiting list among those removed from the waiting list was 86 (66) days. Figure 67 shows the reasons that inpatient rehabilitation referrals were declined. Part (a) shows the unweighted distribution, that is the raw proportions in each category, and part (b) shows the distributions weighted by their IPTW, that is weighted by the importance of the observations in the front-door adjustments. The two distributions were very similar and the four most common reasons were patients being considered to be below the threshold for inpatient (27%), the patient himself/herself declining inpatient rehabilitation as a care option (23%), unclear reasons (19%), and patients being considered to be too unwell to benefit from inpatient rehabilitation (19%). Similarly, Figure 68 shows the weighted and unweighted distribution of reasons that referrals were removed from the waiting list for inpatient rehabilitation after being accepted. Again, the two distributions were similar and the four most common reasons were removal for unclear reasons (27%), shortages of beds on inpatient rehabilitation wards (21%), patients declining inpatient rehabilitation as a care option (18%) and patients who improved to a sufficient degree while waiting for an inpatient rehabilitation (16%).

Table 5 shows the characteristics of those whose referral to inpatient rehabilitation was accepted compared to those whose referral was declined, that is the comparison groups in front-door adjustment #1. Table 6 shows the characteristics of those who were transferred to inpatient rehabilitation compared to those whose referral was declined, that is the comparison groups in front-door adjustment #2. As indicated by Figure 56 and Figure 57, the IPTWs overlapped both in front-door analysis #1 and #2. The amount and reasons for loss to follow-up in the two front-door adjustment analysis were similar to those in the base case analysis (see Figure 59 and Figure 60). However, in the case of front-door adjustment #1, there was considerable lack of overlap in the probability of censoring (see Figure 62) whereas in the case of front-door adjustment #2 this issue was less pronounced (Figure 63).

4.3.2 Consequences

Figure 72 indicates that, at discharge, the proportion of patients rated as having moderate or severe problems on the selected HoNOS dimensions was low in both groups. After adjusting imbalances in baseline confounders and censoring, there was little evidence to suggest that referral to inpatient rehabilitation had an effect on any of the selected HoNOS dimensions (see Figure 73 and Figure 74). As shown in Figure 75, unadjusted absolute readmission rates within one year of discharge from the index admission were somewhat similar in those referred to inpatient rehabilitation and matched controls. In the base case analysis, I estimated that readmission rates were 3 percentage points (95% CI 12 to -6) lower in those referred to inpatient rehabilitation wards (Figure 15). However, the sensitivity analyses suggest that the magnitude of differences in rehospitalization rates may be lower or that the effect may in be the opposite direction. Mortality rates over the study follow-up were approximately 3% in both treatment groups (see Figure 78). There was no strong evidence to suggest that there were differences in mortality rates between the treatment groups (Figure 79).

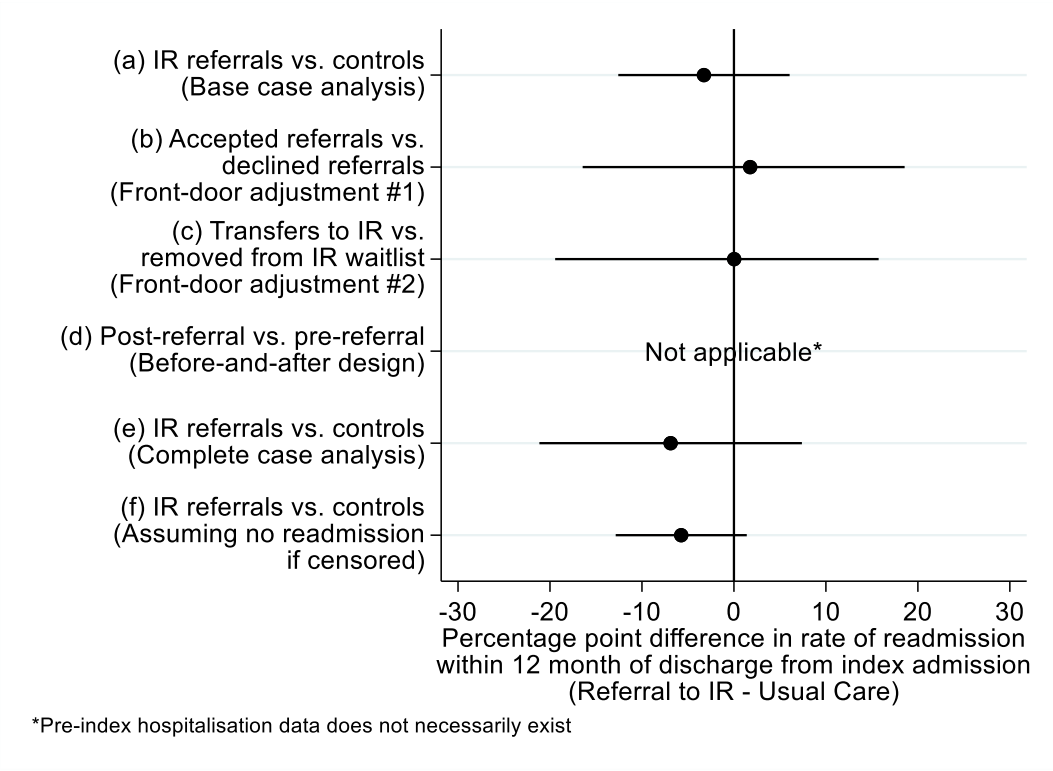


Figure 15 Difference in Readmission rates

4.3.3 Service use and costs

The mean (SD) length of stay on an inpatient rehabilitation ward among those who were transferred to inpatient rehabilitation was 353 (285) days (Figure 69). The average (SD) number of inpatient rehabilitation bed days among those referred was 140 (95% CI 125 to 160) in the first year of follow-up which decreased to 40 (95% CI 22 to 54) in the third year of follow-up (Figure 70). Figure 69 shows the distribution of the length of the inpatient rehabilitation admission. After six months, about a quarter of patients are discharged from inpatient rehabilitation, after one year about two thirds but at the end of follow-up a few patients remained hospitalised on an inpatient rehabilitation ward. At less than one per year, the number of contacts with community rehabilitation was low in absolute terms in both the treatment and the control group (Figure 76). Moreover, in most versions of the analysis, referral to inpatient rehabilitation did not appear to increase contacts with community rehabilitation or an increase compared to the pre-baseline period (Figure 77). In other words, there was no evidence that the control group substituted community

rehabilitation for inpatient rehabilitation. In the base case analysis, the average number of non-rehabilitation community contacts was estimated to be -13 (95% CI -24 to -2) lower in those referred to inpatient rehabilitation. In the sensitivity analyses, the sign and magnitude of plausible estimates was somewhat unstable but broadly similar (Figure 81). This difference appeared to be driven by reductions in non-rehabilitation community contacts in the first year after baseline (Figure 80). However, the same conclusion did not hold for use of non-rehabilitation inpatient care which was similar between the treatment groups at all follow-up time points (Figure 82). In the base-case analysis, I estimated average differences of 13 (95% CI -25 to 51) days hospitalized on a non-rehabilitation inpatient ward. As shown in Figure 83, except for the before-and-after analysis, there was no strong evidence to suggest large reductions in non-rehabilitation inpatient service use due to referral to inpatient rehabilitation. Figure 16 reflects that estimated cost of service use were higher in those referred to inpatient rehabilitation compared to matched controls in the first year of follow-up and to a lesser extent in the second and third year of follow-up. I estimated the total differences between the curve of the treatment and the curve of the control group to be £83,672 (95% CI 65,101 to 102,242). As in other cases, the top estimate in Figure 17 shows this value graphically and the estimates of the sensitivity analysis below it. The mean estimates in sensitivity analyses were consistently lower ranging between approximately £45,000 to £75,000 and there was considerable overlap in the confidence intervals (see Figure 17).

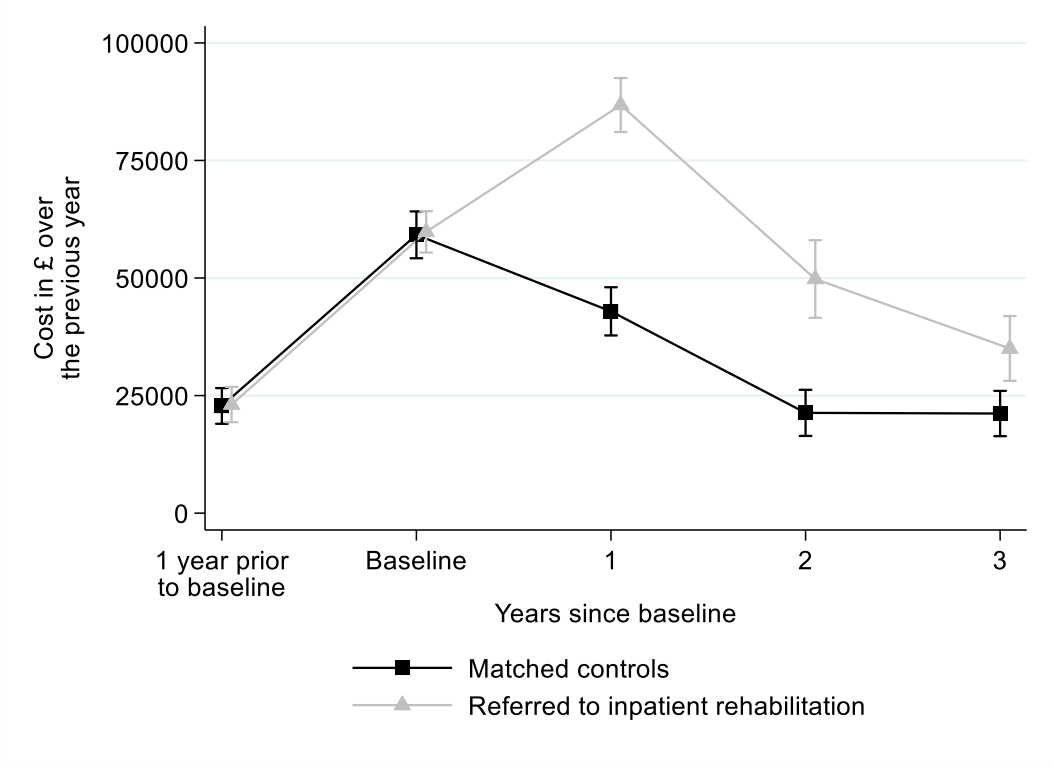


Figure 16 Secondary mental health care costs (Base case analysis)

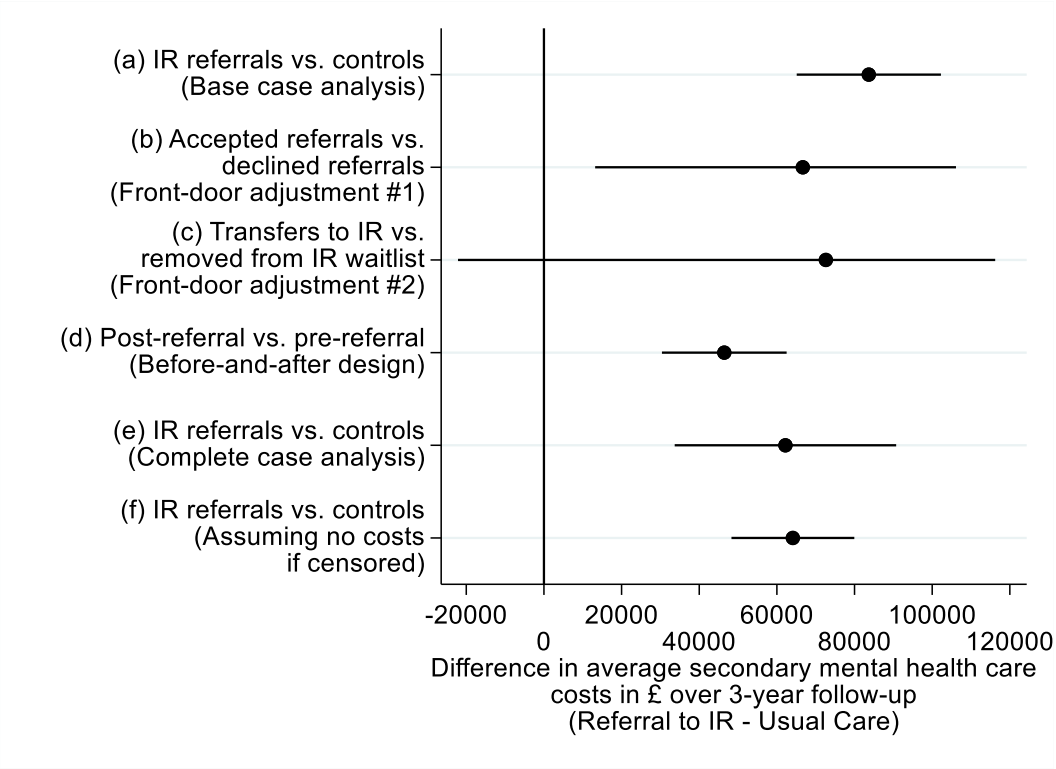


Figure 17 Estimated Differences in secondary mental health care costs

4.4 Discussion

4.4.1 Key results

In this study, I used routinely collected data to conduct a retrospective matched cohort analysis with the aim of assessing the economic value of inpatient rehabilitation for psychosis. As expected, with an average length of stay of almost one year, the length of stay on inpatient rehabilitation wards was high. However, there was no strong evidence that this investment would be offset by large reductions in other types of secondary mental health service use. In addition, there was no clear evidence that patients benefitted from inpatient rehabilitation based on rehospitalization rates, HoNOS ratings at the point of discharge from the index hospitalization or mortality rates. To assess the sensitivity of the results to unmeasured confounding, I both used an approach that is common in the literature and a novel alternative which we support with indirect evidence. In these analyses, the results were more favourable towards inpatient rehabilitation but, under any scenario, in absolute terms, the estimated effect of referral to inpatient rehabilitation on mean costs was high. This is particularly the case if one considers that only approximately 60% of patients who were referred were ultimately transferred to inpatient rehabilitation, that is only 60% would be expected to benefit from referral.

4.4.2 Strengths and limitations

Our analysis benefitted from the use of modern analytical approaches both to reduce the amount of systematic bias in our analysis and to assess the sensitivity of our results. For example, we would argue that the estimates based in the front-door adjustment approach increase our confidence in the credibility of the study findings. This is because we have reason to believe that thanks to the selection mechanisms, i.e. being referred or being referred and accepted to inpatient rehabilitation, there is reason to believe that unmeasured confounders (e.g. social functioning) are more similar in the comparisons underlying the front-door approach. The reported reasons for which patients' referrals were not accepted or the reasons that they were not transferred also do not appear to invalidate the front-door approach. Yet, some key variables, such as social functioning remain

unmeasured, the sample of those referred to inpatient rehabilitation may not be entirely representative because we indirectly identified inpatient rehabilitation referrals, and censoring may be associated with the outcomes. Indeed, we have no reason to believe that any of our approach is entirely unbiased because all rely on untestable assumptions. They fact that we were able to use several approaches only allows us to triangulate evidence. Unfortunately, conducting direct sensitivity analysis with respect to unmeasured confounding was not feasible in the context of this thesis for reasons explained in Section 2.3.1.

In terms of scope of our assessment, as noted in Appendix B, I would argue that health care service use and costs of health care can be measured in a high-quality fashion in CRIS. I did not measure costs of primary care service use but, given previous research, I expect that primary care costs are relatively low and that their omission is unlikely to have impacted the results (Jin and Mosweu, 2017). However, one of the key weaknesses of this study is that the potential benefits of inpatient rehabilitation were only measured in a limited and imperfect fashion. For example, we have two reasons to doubt that HoNOS discharge ratings are comparable between the treatment groups. First, ratings by clinicians on an inpatient rehabilitation ward are likely to have been made according to different standards compared to ratings by clinicians on an acute ward because of habituation to the levels of clinical severity they routinely encounter. Second, ratings for patients discharged from inpatient rehabilitation are likely to be based on a much more in depth knowledge of the patient. Given the goals of inpatient rehabilitation services, it would have been desirable to assess the impact of the intervention on social and occupation functioning as well as patient satisfaction which are not routinely recorded in clinical databases (Iyer et al., 2005). In addition, there is a concern that savings in health care costs in psychosis are offset by an increase in the burden to families and carers (NICE, 2014a). I was also unable to measure the cost impact beyond the health care sector. Our study has a relatively long follow-up period and there was little suggestion that by extrapolating the differences between the two groups beyond the three-year follow-up, the conclusions of the evaluation would change qualitatively. Yet, one may also argue that such a simple extrapolation of short-term effects does not adequately capture the long-term impact of inpatient rehabilitation. For example, if inpatient rehabilitation reduces the risk of self-neglect, the true benefits of this

impact may only emerge over a longer time horizon. In addition, at the end of the three-year follow-up, a small proportion of patients were still on an inpatient rehabilitation ward or had only been recently been discharged from their rehabilitation admission. The -value for money of inpatient rehabilitation may be different for these patients than for the rest of the sample.

To our knowledge, this study is the largest economic evaluation of inpatient rehabilitation to date, but its size was nonetheless insufficient to gain insight into who would benefit most from inpatient rehabilitation or what the optimal length of stay in inpatient rehabilitation might be. It should also be noted that we estimated the average effect of inpatient rehabilitation for an individual patient. Therefore, our results do not necessarily reflect the value for money of closing or opening a rehabilitation ward because this would most likely have spill over effects on the care received by all psychiatric inpatients, i.e. effects going beyond the individual patients, due to changes in the average level of patient need on general psychiatric wards.

In terms of the generalizability and relevance of our results to decision making, one of the strengths of this study that it is based on data from routine care practice and, as discussed in Chapter 1.7, the sample may be more representative than that from an equivalent randomized controlled trial investigating this question. Yet, use of data from a single trust also means that results of this study may not be generalizable to other parts of the country. One possibility is that the effect of inpatient rehabilitation is not maintained because of inadequate aftercare. We also believe in would have been more appropriate to compare referral to inpatient rehabilitation to community rehabilitation rather than receipt of any other form of care. Discharge to community rehabilitation was rare in SLaM, so this comparison was not practically feasible to make this comparison with the data at hand. Finally, we used data over a time horizon of nearly eight years over the course of which there have been changes in the service structure of SLaM (Csipke et al., 2016; Tulloch et al., 2016b). This may affect the applicability of findings for current decision making.

4.4.3 Comparison with existing literature

As noted above, contrary the conclusions drawn from our analysis, previous observational studies unanimously suggest that their evidence supports the provision of inpatient rehabilitation (Bruseker and O'Halloran, 1999; Bunyan et al., 2016; Macpherson et al., 2017; Petrie and Mountain, 2009; Tarasenko et al., 2013; Tsoutsoulis et al., 2018). There appear to be three main explanations for this: First, the scope of our assessment is broader because we include the length of stay on the inpatient rehabilitation and the cost thereof in our evaluation, and we start follow-up at the point of referral to inpatient rehabilitation. Most versions of our analysis suggest that differences in service use and costs other than inpatient rehabilitation may be low. Second, among the study designs explored in our analysis, the one used by almost all previous evaluation, that is a before-and-after approach, also results in results that are most favourable towards inpatient rehabilitation in our analysis. Like in our study, when undertaking a matched comparison, Tsoutsoulis et al. (2018) find that inpatient rehabilitation does not appear to lead to substantial reductions in service use during follow-up but they interpret these results as a return to 'normative levels' of service use. Third, variation in the patient populations and the provision of inpatient rehabilitation may explain some differences. For example, whereas in this study the average length stay on inpatient rehabilitation ward was 353 days, in Bunyan et al (2016) this figure was 701 days, in MacPherson et al. (2017) it was 380 days, in Petrie and Mountain (2009) it was 668 days and in Tsoutsoulis et al. (2018) it was 125 days. By contrast, the results of the two RCTs identified in my systematic review are very similar to the results of our study. However, the settings and time horizon of these trials are very different from this analysis.

4.4.4 Implications for policy and research

I would argue that it would be simplistic to entirely base policymaking on the findings of any single study. However, this study does challenge some of the key justifications for inpatient rehabilitation and, given the extensive annual expenditure on this form of care in the UK, calls for better evidence to justify current care provision or a circumspect exploration of alternative ways of delivering care for people with complex and enduring psychosis. It

should also be emphasised that, despite approach taken in this and previous analyses, in practice, decision makers are not faced with a binary choice between the status quo and no provision of inpatient rehabilitation. Instead, just like crisis resolution teams are not aimed to fully replace acute wards, I believe that the challenge lies in finding a sophisticated balance between institutional and community care for people with complex forms of psychosis (Cornelis et al., 2018). In fact, as indicated by a study by Tarasenko et al. (2013), even if one completely accepts the results of this study, the closure of rehabilitation wards without adequate investments into alternative forms of care for this patient group appears undesirable. My recommendation would thus be for organisations to experiment with a gradual reduction of inpatient rehabilitation in favour of increased investment in community rehabilitation on a local level without necessarily eliminating inpatient rehabilitation as a care option. Indeed, such a shift in care location could dovetail with recent calls by NICE (2019) to limit out-of-area rehabilitation. Given the relatively weak evidence base on which I make this recommendation, I would encourage decision makers to continuous evaluate the impact of such a service reconfiguration taking into account the views of a broad range of stakeholders.

I would suggest that future research should particularly focus on assessing the potential benefits of inpatient rehabilitation more extensively. Consideration should be given whether the preferred outcome measure for economic evaluations, quality-adjusted life years (QALYs), is suitable for this patient population or whether a broader measure of outcome may be warranted (Awad and Voruganti, 2016; Mavranezouli, 2010). If a prospective observational study rather than a randomized trial is to be conducted, the front-door adjustment approach appears to be a potentially advantageous design for two reasons: First, it would only require data collection in a well-defined group, that is, those referred to inpatient rehabilitation; Second, since patients already undergo an assessment for their suitability for inpatient rehabilitation in routine care practice, this assessment could be the moment in which baseline data could be collected thereby minimizing the burden of research on clinical staff. In observational studies with a larger sample size, I would also recommend considering restricting the comparison group for the front-door adjustments to patients whose referral has been declined because of bed shortages or who have been removed from the waiting list for this reason. This is because the assumption that

declination of the referral or removal from the waiting list is unconfounded may be most plausible for this subgroup. In technical terms, this may be considered an instrumental variable version of the front-door adjustment. Unfortunately, the sample size in this study was too small for such an approach. It may also be worth assessing the value of setting up a registry for patients referred to inpatient rehabilitation because this may be the only way to gather sample of sufficient size and depth to robustly identify for whom and how inpatient rehabilitation for psychosis is optimally provided (Gliklich et al., 2014).

Chapter 5 Overall discussion

5.1 Summary of findings and contributions

In Chapter 2, I identified four types of identification assumptions with respect to unmeasured confounding. Three of these, the back-door adjustment, unit-invariant fixed effects approach and instrumental variable approaches, are commonly used in the health economics literature. The fourth, the front-door adjustment approach, exploits knowledge about the treatment mechanism for causal inference and was relevant to Analysis 2. Further, I distinguished between four types of measurement error assumptions based. I then identify five different approaches to support or enable these identification assumptions, namely qualitative bias analysis, quantification of the measurement error, strategies to reduce the measurement error in proxy variables, sampling strategies and measurement error models.

In Chapter 3, I undertook a cost-consequence analysis of admission to CAMHS inpatient care compared to adult inpatient care for young people with psychosis (Analysis 1). I found that a regression discontinuity design was not a suitable approach to compare the impact of admission to CAMHS and adult wards because there was evidence suggesting that its underlying assumptions were violated.

In Chapter 4, I undertook a cost-consequence analysis of referral to inpatient rehabilitation compared to usual care for people with persistent forms of psychosis (Analysis 2). There was no strong evidence to suggest that referral to inpatient rehabilitation led to improvement on proxies of treatment effectiveness, that is rehospitalization rates, discharge HoNOS ratings and mortality rates. Apart from potential moderate reductions in community contacts, there was also no strong evidence to suggest that referral to inpatient rehabilitation led to reductions in other forms of service use. Thus, the analyses suggest that referral to inpatient rehabilitation may substantially increase cost of secondary mental health care due to the cost of inpatient rehabilitation.

To summarise, I would argue that this thesis makes two methodological contributions. It introduces an identification strategy that has not previously been used in the health

economics literature, the front-door adjustment, and illustrates its value in an empirical example. Moreover, I identify approaches to handle measurement error, a topic neglected in the existing health economics literature creating a menu of options from which future researchers can choose from. I also make two contributions to the empirical literature. I show that, despite the appeal of the idea, using a regression discontinuity design based on the age cut-off between adult and child and adolescent inpatient services is not likely to yield unbiased estimates of the differences between them. Instead I suggest that other designs in conjunction with larger datasets are needed. In addition, I produce evidence on the value for money of inpatient rehabilitation that, I argue, is more methodologically robust than all previous observational studies while being more relevant to the UK setting than existing randomised controlled trials.

5.1.1 Data source

The key strength of the CRIS database in relation to the economic evaluations in this thesis was that it allowed access to the free text clinical notes. I made relatively little use of NLP applications but, by reading medical notes, I was, for example, able to (a) establish that the age-related discontinuity in hospitalization rates in the evaluation of inpatient care modalities for young people was not an artefact of different diagnostic coding practices in CAMHS and adult services, (b) ascertain the dates of referral to inpatient rehabilitation and (c) obtain evidence to support the identification assumptions in the evaluation of inpatient rehabilitation. Currently, this is not possible with more widely used databases for health economic research such as HES, CPRD and MHSDS. The fact that database developers and administrators were collocated with CRIS researchers facilitated the linkage of new data to CRIS for the purposes of this thesis and facilitated an understanding of data structure and quality. In contrast to some other health care systems, SLaM is responsible for the provision or purchase of the vast majority of secondary mental health care consumed by the local population which allowed us to capture a greater proportion of cost data. Finally, this thesis highlights how evidence can be produce relatively inexpensively and in a relatively timely fashion thereby complementing randomized controlled trial as a vehicle for economic evaluation (Struck et al., 2014).

There were several limitations common to both economic evaluations that were inherent in the choice of the SLaM CRIS database. Despite the greater breadth of data available in CRIS, confounding remained an issue in both analyses because there were limits to the information contained in the medical records and the extent to which it was usable for analysis. Data quality was an issue in both evaluations because data in CRIS is recorded for clinical or administrative purposes, not for use in research. While I took steps to mitigate the impact of measurement errors and, to increase transparency, undertook an assessment of data quality, there were practical limits to this. Moreover, some errors are inevitable, for example, because historical records in CRIS can and are at times overwritten retrospectively or because the use of data fields and recording styles can changes over time or between individuals in routine care practice.

In addition, the nature of the database limited the extent to which broader benefits and costs of the treatments could be captured. For example, both interventions are likely to have an impact on non-healthcare costs and on the family or carers of the patients. These were not measurable in CRIS. Given the size of SLaM's catchment area, the number of people accessing specialist inpatient care was small in absolute terms, leading to a lack of precision in estimates. A limitation of CRIS that is more specific to the focus of this thesis is that in both evaluations I assessed service-level interventions, but in there was no variation in the provision of these services over time. Thus, it was not possible to capture potential spill over effects from changes in the supply of these services or short-run effects resulting from reorganization of services (Meacock, 2018). Finally, there are concerns about the generalizability of results based on CRIS data, both because of differences in the patient population and the wide range of mental health services offered by SLaM (Davis et al., 2018; Tulloch et al., 2012).

5.1.2 Collaborations

The analyses in this thesis benefitted considerably from collaboration with people who had an active role in providing the services that were evaluated. Data obtained from SLaM's finance department allowed us to cost service use much more accurately than we otherwise would have been able to. My clinical collaborators' expertise helped me in interrogate the

quality of the data, design the analysis and interpret findings to a much greater degree than might be the case in a larger and less embedded database of electronic health records such as HES. For example, a SLaM ward that was converted from being a rehabilitation unit to a general adult ward in 2010 remains classified as a rehabilitation ward in one of the tables in CRIS and therefore initially artificially doubled the number of the rehabilitation. In addition, the idea for the front-door adjustment in the evaluation of inpatient rehabilitation grew out of internal discussions around what the target parameter of interest was. Thus, this study adds to the examples of health economic research in which engagement with the subject matter gives rise to the use of non-standard identification strategies (Hammond et al., 2019). One of the weaknesses of this thesis is, however, that there was no patient and public involvement (PPI). The scope and diffusion of PPI work in relation to health economics is still relatively limited but particularly service users' input into the choice and framing of the research question would have been beneficial (Kandiyali et al., 2019; O'Shea et al., 2019).

5.1.3 Methodology

I strengthened the quantitative analyses in this thesis in three ways: First, I identified strategies to handle confounding in the analyses, including a novel approach that increased the credibility of the findings or allowed me to assess their credibility. Second, I took an informed approach to reducing the risk of bias or loss of precision due to measurement error. Third, I followed recommendations of good practice in analysis and presentation of my research (Faria et al., 2015; Gelman et al., 2002; Kreif et al., 2013b).

One of the methodological limitations of this thesis noted in Section 1.5.2, however, is that, unlike the other two systematic reviews in this thesis, the review of evidence for inpatient care in young people with psychosis was not systematic. Therefore, I may not have captured all available evidence in my discussion. In addition, weaknesses to the statistical approach in the analyses remain. There was evidence to suggest that the regression discontinuity design evaluation of inpatient care modalities for young people could be severely biased against child and adolescent wards. Alternative designs were not practically feasible because of small sample size. In the evaluation of inpatient rehabilitation, the front-door approaches

do represent an alternative to the other two more conventional approaches which makes assumption that have some plausibility. Yet, It is not certain that any of the four are close to being unbiased. As noted in Section 2.3.1, expert elicitation of bias parameters may be an approach to address this issue.

My handling of measurement error disregards errors from lack of inter-rater and intra-rater reliability and measurement error due to imprecise definition of the concepts that I intended to measure (Matt and Matthew, 2013). In Davis et al. (2018), for example, two psychiatrists who read the same CRIS records according to a structured methodology agreed in 80% of cases on a diagnosis. In the evaluation of inpatient rehabilitation, I used bootstrapping and boosted regression because, separately, these had been previously used to address statistical issues at hand but it would have been reassuring to conduct a simulation study to assess their behaviour when used in combination within the relatively small samples in some of the sensitivity analyses of Chapter 4 (Morris et al., 2019). Similarly, it would have been beneficial to explore the implications of combining a matching with an inverse probability weighting approach in more depth.

Moreover, the focus on the statistical aspects of the economic evaluations came at the expense of neglecting other methodological factors. A weakness of the economic evaluations is, for example, that I do not make reference to a theoretical framework such as a theory of change, i.e. an explanation of the process or mechanism by which inpatient rehabilitation or CAMHS inpatient care brings about causal effects on the outcomes It is increasingly recognised that both evidence for such a mechanism and of correlation between the treatment and the outcomes are necessary to establish causal claims (Jones and Schooling, 2018; Parkkinen et al., 2018). This thesis is largely limited on generating evidence of correlation. Collecting qualitative evidence may have helped in generating such evidence for the mechanism thus facilitating the interpretation of my findings (Krieger and Davey Smith, 2016). Finally, following the norm among health economic and applied quantitative research more generally, in this thesis I do not examine my personal role in the production of the research, nor do I reflect on the philosophical position of my research (Babones, 2016; Lessard, 2007; Thorpe and Holt, 2008).

5.2 Potential future research

5.2.1 Applied

In relation to the evaluation of inpatient rehabilitation, it would be of value add the costs of physical health care based on HES data. In addition, it may be possible to measure the use of supported accommodation either indirectly via patients' address data or by reading their medical notes. The use of data from CRIS system in other mental health trusts such as Camden and Islington and the Cambridge and Peterborough could address issues around the sample size and generalizability of findings to some extent (Price et al., 2017; Werbeloff et al., 2018). In addition, it could be of value to investigate the outcomes of those referred to inpatient rehabilitation in more depth. For example, one approach could be to code some salient outcomes in all those referred by in depth reading of clinical and use the front-door adjustment for parameter estimation. To reduce bias due to confounding, I limited myself to including a limited number of confounders that were chosen based on expert knowledge of my clinical collaborators. More recently, some researchers have investigate approaches that make more extensive use of data contained in electronic health records by adjusting for confounding using semi-automated high-dimensional algorithms which may strengthen the evaluation of inpatient rehabilitation (Mozer et al., 2018; Schneeweiss, 2018; Toh et al., 2011). In addition, qualitative approaches may be of value to understand contextual factors that influence when, how and for whom inpatient rehabilitation is most cost-effective, and what value dimensions are relevant for decision making in this context (Campbell et al., 2018; Raine et al., 2016).

As noted in Chapter 3, in terms of the evaluation of inpatient care for young people, HES and/or MHSDS data may be an better alternative than the CRIS systems because they are more likely to contain historical data on trusts that have change the age threshold of CAMHS care to 25 and are likely contain a larger sample of underage patients admitted to adult wards. In terms of economic evaluations of inpatient care for psychosis more broadly, another potential use of routinely collected data could be the assessment of admission to mother and baby units compared to general acute wards or community services for perinatal women with psychosis.

Finally, I would argue that, the front-door adjustment may have broader applicability as an identification strategy in the health economics literature. As discussed in Chapter 2, non-compliers can provide a credible control group, it can be used if treatment cannot be withheld from individuals. In addition, in the one-sided non-compliance scenario, the front-door adjustment approach allows evaluation of treatments which cannot be withheld from patients for practical or ethical reasons and reduces data requirements (Glynn and Kashin, 2018). For example, when evaluating the effect of clozapine, it is neither straightforward to identify which patients have treatment-resistant psychosis and would therefore be eligible to receive clozapine. However, if one is willing to assume that all those who receive clozapine have treatment-resistant psychosis, a suitable measure of compliance is measurable and one is willing to restrict the analysis to the average treatment effect on the treated (ATT), then the need to identify treatment-resistance is obviated. One of the implications of the front-door adjustment approach is that measurement of variables on the causal pathway is not only relevant if interest lies in mediation analysis, adjustment for partial compliance or sample selection but also to implement this identification strategy (Tafti and Shmueli, 2019). Similarly, there are other non-standard identification strategies such as negative controls that may merit consideration in health economic evaluations based on observational data (Glynn and Gerring, 2013; Lipsitch et al., 2010; Zhang et al., 2018).

5.2.2 Methodological

To aid applied research based on CRIS, I would argue that further steps to making the quality of data more transparent would be valuable. This could, for example, involve comprehensive, systematic and periodic internal assessment of data quality, external validation exercises and/or the production of a data dictionary (Park et al., 2012). To complement these, given the wide range of projects that CRIS is used for and the tacit knowledge clinicians working in SLaM appear to often hold about CRIS variables, it may be valuable to complement these by collecting informal information on data quality in a central repository.

Since the use of text data for medical research is increasing internationally, methodological research on ways to leverage this information more efficiently appears to be a promising avenue for future research (Chen et al., 2018). One area that I highlight in Chapter 2, is to further develop methods to reduce the impact of measurement error. For example, there appears to be little work on models for measurement error in eligibility criteria and the timing of exposures (Buonaccorsi, 2010; Carroll et al., 2006; Gustafson, 2003; Yi, 2017). Measurement error models for eligibility criteria could, for example, aid the analyses in which there is significant measurement error in diagnosis variable or when the eligibility criteria are complex such as those for treatment-resistant psychosis or treatment-resistant depression. Measurement error models for the timing of exposure, on the other hand, could, for example facilitate the evaluation of drug treatment in CRIS or analyses investigating the course of illnesses since first its first diagnosis. In addition, it appears to me that increasing the accessibility of and, where relevant, increase the scope of sampling strategies outlined in Chapter 2 would enable analyses of concepts that are currently not amenable to natural language processing and not well-characterised by measurement error models.

5.3 Conclusion

The main policy implication of this thesis is that, based on the results of the evaluation of inpatient rehabilitation, I recommend gradually and partially shifting investment on rehabilitative care for people with psychosis from an inpatient setting to the community. This fits well with the current policy of reducing out of area inpatient rehabilitative care because it would free up local inpatient beds to which patients could be returned to but, given the gaps in the existing evidence, would require cautious monitoring and involvement of all relevant stakeholders to avoid unanticipated consequences. On other hand, I would not maintain that the results from the evaluation of inpatient care modalities in young people have implications for policy making.

The main take-home messages for applied researchers are that: both a larger dataset and a different identification strategy are required for a robust evaluation of inpatient care modalities in young people; the front-door adjustment is a useful addition to the existing set

of identification strategies using in health economic evaluation based on observational data; there are a broad number of potential approaches to handling measurement error.

References

Aceituno, D., Vera, N., Prina, A.M., McCrone, P., 2019. Cost-effectiveness of early intervention in psychosis: systematic review. Br. J. Psychiatry 1–7. https://doi.org/10.1192/bjp.2018.298

Achilla, E., McCrone, P., 2013. The Cost Effectiveness of Long-Acting/Extended-Release Antipsychotics for the Treatment of Schizophrenia. Appl. Health Econ. Health Policy. 11, 95–106. https://doi.org/10.1007/s40258-013-0016-2

Arciniegas, D.B., 2015. Psychosis. Contin. Lifelong Learn. Neurol. 21, 715–736. https://doi.org/10.1212/01.CON.0000466662.89908.e7

Asaria, M., Grasic, K., Walker, S., 2016. Using Linked Electronic Health Records to Estimate Healthcare Costs: Key Challenges and Opportunities. PharmacoEconomics 34, 155–160. https://doi.org/10.1007/s40273-015-0358-8

Awad, A.G., Voruganti, L.N.P. (Eds.), 2016. Beyond Assessment of Quality of Life in Schizophrenia. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-30061-0

Babones, S., 2016. Interpretive Quantitative Methods for the Social Sciences. Sociology 50, 453–469. https://doi.org/10.1177/0038038515583637

Baltussen, R., Leidl, R., Ament, A., 1999. Real World Designs in Economic Evaluation. PharmacoEconomics 16, 449–458. https://doi.org/10.2165/00019053-199916050-00003

Barbui, C., Lintas, C., 2006. Decision models in the evaluation of psychotropic drugs. Eur. J. Health Econ. 7, 218–220. https://doi.org/10.1007/s10198-006-0348-z

Bartlett, J.W., Seaman, S.R., White, I.R., Carpenter, J.R., for the Alzheimer's Disease Neuroimaging Initiative*, 2015. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. Stat. Methods Med. Res. 24, 462–487. https://doi.org/10.1177/0962280214521348

Beecham, J., 2014. Annual Research Review: Child and adolescent mental health interventions: a review of progress in economic studies across different disorders. J. Child Psychol. Psychiatry 55, 714–732. https://doi.org/10.1111/jcpp.12216

Beesley, L.J., Salvatore, M., Fritsche, L.G., Pandit, A., Rao, A., Brummett, C., Willer, C.J., Lisabeth, L.D., Mukherjee, B., 2020. The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. Stat. Med. 39, 773–800. https://doi.org/10.1002/sim.8445

Bell, H., Wailoo, A., Hernandez, M., Grieve, R., Faria, R., Gibson, L., Grimm, S., 2016. The use of real world data for the estimation of treatment effects in NICE decision making. NICE decision support unit.

Belsley, D.A., Kuh, E., Welsch, R.E., 2005. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons.

Bentall, R., Kinderman, P., Cooke, A., 2000. Recent advances in understanding mental illness and psychotic experiences. British Psychological Society.

Bilinski, A., Hatfield, L.A., 2018. Seeking evidence of absence: Reconsidering tests of model assumptions. arXiv.org.

Blackwell, M., Honaker, J., King, G., 2017. A Unified Approach to Measurement Error and Missing Data: Details and Extensions. Sociol. Methods Res. 46, 342–369. https://doi.org/10.1177/0049124115589052

Blackwell, M., Honaker, J., King, G., 2015. A Unified Approach to Measurement Error and Missing Data: Overview and Applications. Sociol. Methods Res. https://doi.org/10.1177/0049124115585360

Boef, A.G.C., Dekkers, O.M., Vandenbroucke, J.P., le Cessie, S., 2014. Sample size importantly limits the usefulness of instrumental variable methods, depending on instrument strength and

level of confounding. J. Clin. Epidemiol. 67, 1258–1264. https://doi.org/10.1016/j.jclinepi.2014.05.019
Brakenhoff, T.B., Mitroiu, M., Keogh, R.H., Moons, K.G.M., Groenwold, R.H.H., van Smeden, M., 2018. Measurement error is often neglected in medical literature: a systematic review. J. Clin. Epidemiol. 98, 89–97. https://doi.org/10.1016/j.jclinepi.2018.02.023
Branson, Z., Mealli, F., 2018. Local Randomization and Beyond for Regression Discontinuity Designs: Revisiting a Causal Analysis of the Effects of University Grants on Dropout Rates. arXiv.org.
Briggs, A., Sculpher, M., Claxton, K., 2006. Decision Modelling for Health Economic Evaluation, Handbooks in Health Economic Evaluation. Oxford University Press, Oxford, New York.
Bromley, S., Choi, M.A., Faruqui, S., Czuchta, D., Centre for Addiction and Mental Health, 2015. First episode psychosis: an information guide.
Bruseker, R., O'Halloran, P., 1999. Preliminary results from an Australian psychosocial rehabilitation program for people with serious mental illness. Aust. N. Z. J. Ment. Health Nurs. 8, 58–64. https://doi.org/10.1046/j.1440-0979.1999.00132.x
Buchanan, M., 2014. More children on adult mental wards. BBC News.
Bunyan, M., Crowley, J., Cashen, A., Mutti, M.-F., 2017. A look at inpatients' experience of mental health rehabilitation wards. Ment. Health Pract. 20, 17–23. https://doi.org/10.7748/mhp.2017.e1163
Bunyan, M., Ganeshalingam, Y., Morgan, E., Thompson-Boy, D., Wigton, R., Holloway, F., Tracy, D.K., 2016. In-patient rehabilitation: clinical outcomes and cost implications. BJPsych Bull. 40, 24–28. https://doi.org/10.1192/pb.bp.114.049858
Buonaccorsi, J.P., 2010. Measurement error: models, methods, and applications. CRC Press.
Burns, T., 2007. Hospitalisation as an outcome measure in schizophrenia. Br. J. Psychiatry 191, s37–s41.
Burns, T., 2006. Psychiatry: A Very Short Introduction, 1 edition. ed. OUP Oxford, Oxford ; New York.
Buxton, M.J., Drummond, M.F., Hout, B.A.V., Prince, R.L., Sheldon, T.A., Szucs, T., Vray, M., 1997. Modelling in Ecomomic Evaluation: An Unavoidable Fact of Life. Health Econ. 6, 217–227. https://doi.org/10.1002/(SICI)1099-1050(199705)6:3<217::AID-HEC267>3.0.CO;2-W
Cairney, P., Oliver, K., 2017. Evidence-based policymaking is not like evidence-based medicine, so how far should you go to bridge the divide between evidence and policy? Health Res. Policy Syst. 15, 35. https://doi.org/10.1186/s12961-017-0192-x
Cameron, A.C., Trivedi, P.K., 2005. Microeconometrics: Methods and Applications. Cambridge University Press.
Campbell, D., 2016. Revealed: dozens of children still treated on adult psychiatric wards. The Guardian.
Campbell, J.A., Ezzy, D., Neil, A., Hensher, M., Venn, A., Sharman, M.J., Palmer, A.J., 2018. A qualitative investigation of the health economic impacts of bariatric surgery for obesity and implications for improved practice in health economics. Health Econ. 27, 1300–1318. https://doi.org/10.1002/hec.3776
Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M., 2006. Measurement error in nonlinear models: a modern perspective. CRC press.
Cattaneo, M.D., Idrobo, N., Titiunik, R., 2018. A Practical Introduction to Regression Discontinuity Designs: Volume II.
Chang, C.-K., Hayes, R.D., Broadbent, M., Fernandes, A.C., Lee, W., Hotopf, M., Stewart, R., 2010. All-cause mortality among people with serious mental illness (SMI), substance use disorders, and depressive disorders in southeast London: a cohort study. BMC Psychiatry 10, 77.
Chaplin, D.D., Cook, T.D., Zurovac, J., Coopersmith, J.S., Finucane, M.M., Vollmer, L.N., Morris, R.E., 2018. The Internal and External Validity of the Regression Discontinuity Design: A Meta-Analysis of 15 Within-Study Comparisons. J. Policy Anal. Manage. 37, 403–429. https://doi.org/10.1002/pam.22051

Chen, X., Xie, H., Wang, F.L., Liu, Z., Xu, J., Hao, T., 2018. A bibliometric analysis of natural language processing in medical research. BMC Med. Inform. Decis. Mak. 18. https://doi.org/10.1186/s12911-018-0594-x

Collins, B., 2016. Big Data and Health Economics: Strengths, Weaknesses, Opportunities and Threats. PharmacoEconomics 34, 101–106. https://doi.org/10.1007/s40273-015-0306-7

Cook, T.D., Shadish, W.R., Wong, V.C., 2008. Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. J. Policy Anal. Manage. 27, 724–750. https://doi.org/10.1002/pam.20375

Cornelis, J., Barakat, A., Dekker, J., Schut, T., Berk, S., Nusselder, H., Ruhl, N., Zoeteman, J., Van, R., Beekman, A., Blankers, M., 2018. Intensive home treatment for patients in acute psychiatric crisis situations: a multicentre randomized controlled trial. BMC Psychiatry 18. https://doi.org/10.1186/s12888-018-1632-z

CQC, 2018. Briefing – Mental health rehabilitation inpatient services.

CQC, 2016a. South London and Maudsley NHS Foundation Trust - Quality Report [WWW Document]. Care Qual. Comm. URL http://www.cqc.org.uk/sites/default/files/new_reports/AAAE6494.pdf (accessed 9.5.17).

CQC, 2016b. South London and Maudsley NHS Foundation Trust Long Stay Rehabilitation Wards for working age adults - Quality Report [WWW Document]. Care Qual. Comm. URL http://www.cqc.org.uk/sites/default/files/new_reports/AAAE6503.pdf (accessed 5.9.17).

Craig, T.K., 2016. Shorter hospitalizations at the expense of quality? Experiences of inpatient psychiatry in the post-institutional era. World Psychiatry 15, 91–92.

Csipke, E., Williams, P., Rose, D., Koeser, L., McCrone, P., Wykes, T., Craig, T., 2016. Following the Francis report: Investigating patient experience of mental health in-patient care. Br. J. Psychiatry 209, 35–39. https://doi.org/10.1192/bjp.bp.115.171124

Curtis, L., Burns, A., 2018. Unit Costs of Health and Social Care 2018 | PSSRU. University of Kent.

D'Amour, A., Ding, P., Feller, A., Lei, L., Sekhon, J., 2020. Overlap in observational studies with high-dimensional covariates. J. Econom. https://doi.org/10.1016/j.jeconom.2019.10.014

Davis, K.A.S., Bashford, O., Jewell, A., Shetty, H., Stewart, R.J., Sudlow, C.L.M., Hotopf, M.H., 2018. Using data linkage to electronic patient records to assess the validity of selected mental health diagnoses in English Hospital Episode Statistics (HES). PLOS ONE 13, e0195002. https://doi.org/10.1371/journal.pone.0195002

Davis, K.A.S., Sudlow, C.L.M., Hotopf, M., 2016. Can mental health diagnoses in administrative data be used for research? A systematic review of the accuracy of routinely collected diagnoses. BMC Psychiatry 16. https://doi.org/10.1186/s12888-016-0963-x

Daw, J.R., Hatfield, L.A., 2018. Matching in Difference-in-Differences: between a Rock and a Hard Place. Health Serv. Res. 53, 4111–4117. https://doi.org/10.1111/1475-6773.13017

Deb, P., Norton, E.C., Manning, W.G., StataCorp, L.L.C., 2017. Health Econometrics Using Stata. Stata Press College Station, TX.

Dhiman, P., Lee, H., Kirtley, S., Collins, G.S., 2020. A systematic review showed more consideration is needed when conducting nonrandomized studies of interventions. J. Clin. Epidemiol. 117, 99–108. https://doi.org/10.1016/j.jclinepi.2019.09.027

Diamond, A., Sekhon, J.S., 2012. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. Rev. Econ. Stat. 95, 932–945. https://doi.org/10.1162/REST_a_00318

DiazOrdaz, K., Franchini, A.J., Grieve, R., 2018. Methods for estimating complier average causal effects for cost-effectiveness analysis. J. R. Stat. Soc. Ser. A Stat. Soc. 181, 277–297. https://doi.org/10.1111/rssa.12294

Ding, P., Li, F., 2019. A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment. arXiv.org.

Dowie, J., 2004. Why cost-effectiveness should trump (clinical) effectiveness: the ethical economics of the South West quadrant. Health Econ. 13, 453–459. https://doi.org/10.1002/hec.861

Downs, J., Dean, H., Lechler, S., Sears, N., Patel, R., Shetty, H., Hotopf, M., Ford, T., Kyriakopoulos, M., Diaz-Caneja, C.M., Arango, C., MacCabe, J.H., Hayes, R.D., Pina-Camacho, L., 2018. Negative Symptoms in Early-Onset Psychosis and Their Association With Antipsychotic Treatment Failure. Schizophr. Bull. https://doi.org/10.1093/schbul/sbx197

Drummond, M.F., 2015. Methods for the Economic Evaluation of Health Care Programmes, 4 edition. ed. Oxford University Press, Oxford, United Kingdom ; New York, NY, USA.

Drummond, M.F., Sculpher, M.J., Torrance, G.W., O'Brien, B.J., Stoddart, G.L., 2005. Methods for the Economic Evaluation of Health Care Programmes. Oxford University Press.

Dutta, R., Murray, R.M., Allardyce, J., Jones, P.B., Boydell, J.E., 2012. Mortality in first-contact psychosis patients in the UK: a cohort study. Psychol. Med. 42, 1649–1661. https://doi.org/10.1017/S0033291711002807

Edwards, T., Macpherson, R., Commander, M., Meaden, A., Kalidindi, S., 2016. Services for people with complex psychosis: towards a new understanding. BJPsych Bull. 40, 156–161. https://doi.org/10.1192/pb.bp.114.050278

EUnetHTA, 2015. Methods for health economic evaluations: A guideline based on current practices in Europe Guideline.

Faria, R., Alava, M.H., Manca, A., Wailoo, A.J., 2015. NICE DSU technical support document 17: The use of observational data to inform estimates of treatment effectiveness in technology appraisal: Methods for comparative individual patient data.

Fernandes, A.C., Cloete, D., Broadbent, M.T., Hayes, R.D., Chang, C.-K., Jackson, R.G., Roberts, A., Tsang, J., Soncul, M., Liebscher, J., Stewart, R., Callard, F., 2013. Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. BMC Med. Inform. Decis. Mak. 13, 71. https://doi.org/10.1186/1472-6947-13-71

Franklin, M., Davis, S., Horspool, M., Kua, W.S., Julious, S., 2017. Economic Evaluations Alongside Efficient Study Designs Using Large Observational Datasets: the PLEASANT Trial Case Study. PharmacoEconomics. https://doi.org/10.1007/s40273-016-0484-y

Frey, S., Linder, R., Juckel, G., Stargardt, T., 2014. Cost-effectiveness of long-acting injectable risperidone versus flupentixol decanoate in the treatment of schizophrenia: a Markov model parameterized using administrative data. Eur. J. Health Econ. 15, 133–142. https://doi.org/10.1007/s10198-013-0460-9

Fusar-Poli, P., 2019. Integrated Mental Health Services for the Developmental Period (0 to 25 Years): A Critical Review of the Evidence. Front. Psychiatry 10. https://doi.org/10.3389/fpsyt.2019.00355

Gabrio, A., Mason, A.J., Baio, G., 2017. Handling Missing Data in Within-Trial Cost-Effectiveness Analysis: A Review with Future Recommendations. PharmacoEconomics - Open 1, 79–97. https://doi.org/10.1007/s41669-017-0015-6

Garrison, L.P., Neumann, P.J., Erickson, P., Marshall, D., Mullins, C.D., 2007. Using real-world data for coverage and payment decisions: the ISPOR Real-World Data Task Force report. Value Health 10, 326–335.

Gelman, A., Pasarica, C., Dodhia, R., 2002. Let's Practice What We Preach. Am. Stat. 56, 121–130. https://doi.org/10.1198/000313002317572790

Ghabri, S., Stevenson, M., Möller, J., Caro, J.J., 2018. Trusting the Results of Model-Based Economic Analyses: Is there a Pragmatic Validation Solution? PharmacoEconomics 1–6. https://doi.org/10.1007/s40273-018-0711-9

Glanville, J., Kaunelis, D., Mensinkai, S., 2009. How well do search filters perform in identifying economic evaluations in MEDLINE and EMBASE. Int. J. Technol. Assess. Health Care 25, 522–529. https://doi.org/10.1017/S0266462309990523

Gliklich, R.E., Dreyer, N.A., Leavy, M.B., 2014. Planning a Registry. Agency for Healthcare Research and Quality (US).

Glynn, A.N., Gerring, J., 2013. Strategies of Research Design with Confounding: A Graphical Description. Unpubl. Manuscr. Dep. Polit. Sci. Boston Univ.

Glynn, A.N., Kashin, K., 2018. Front-Door Versus Back-Door Adjustment With Unmeasured Confounding: Bias Formulas for Front-Door and Hybrid Adjustments With Application to a Job Training Program. J. Am. Stat. Assoc. 113, 1040–1049. https://doi.org/10.1080/01621459.2017.1398657

Glynn, A.N., Kashin, K., 2017. Front-Door Difference-in-Differences Estimators. Am. J. Polit. Sci. 61, 989–1002. https://doi.org/10.1111/ajps.12311

Golay, P., Alameda, L., Mebdouhi, N., Baumann, P., Ferrari, C., Solida, A., Progin, P., Elowe, J., Conus, P., 2017. Age at the time of onset of psychosis: A marker of specific needs rather than a determinant of outcome? Eur. Psychiatry J. Assoc. Eur. Psychiatr. 45, 20–26. https://doi.org/10.1016/j.eurpsy.2017.06.002

Griswold, K., Del Regno, P.A., Berger, R.C., 2015. Recognition and differential diagnosis of psychosis in primary care. Brain 100, 18–37.

Gronholm, P.C., Thornicroft, G., Laurens, K.R., Evans-Lacko, S., 2017. Mental health-related stigma and pathways to care for people at risk of psychotic disorders or experiencing first-episode psychosis: a systematic review. Psychol. Med. 47, 1867–1879. https://doi.org/10.1017/S0033291717000344

Grutters, J.P.C., Govers, T., Nijboer, J., Tummers, M., van der Wilt, G.J., Rovers, M.M., 2019. Problems and Promises of Health Technologies: The Role of Early Health Economic Modeling. Int. J. Health Policy Manag. 8, 575–582. https://doi.org/10.15171/ijhpm.2019.36

Guloksuz, S., Os, J. van, 2018. The slow death of the concept of schizophrenia and the painful birth of the psychosis spectrum. Psychol. Med. 48, 229–244. https://doi.org/10.1017/S0033291717001775

Gunter, T.D., Terry, N.P., 2005. The Emergence of National Electronic Health Record Architectures in the United States and Australia: Models, Costs, and Questions. J. Med. Internet Res. 7. https://doi.org/10.2196/jmir.7.1.e3

Guolo, A., 2008. Robust techniques for measurement error correction: a review. Stat. Methods Med. Res. 17, 555–580. https://doi.org/10.1177/0962280207081318

Gustafson, P., 2003. Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments. Chapman and Hall/CRC.

Hammond, J., Mason, T., Sutton, M., Hall, A., Mays, N., Coleman, A., Allen, P., Warwick-Giles, L., Checkland, K., 2019. Exploring the impacts of the 2012 Health and Social Care Act reforms to commissioning on clinical activity in the English NHS: a mixed methods study of cervical screening. BMJ Open 9, e024156. https://doi.org/10.1136/bmjopen-2018-024156

Haneuse, S., VanderWeele, T.J., Arterburn, D., 2019. Using the E-Value to Assess the Potential Effect of Unmeasured Confounding in Observational Studies. JAMA 321, 602–603. https://doi.org/10.1001/jama.2018.21554

Harrell, J.F.E., 2015. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, 2nd ed. 2015 edition. ed. Springer, Cham Heidelberg New York.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, 2nd ed. 2009, Corr. 9th printing 2017 edition. ed. Springer, New York, NY.

Hayes, D., Kyriakopoulos, M., 2018. Dilemmas in the treatment of early-onset first-episode psychosis. Ther. Adv. Psychopharmacol. 8, 231–239. https://doi.org/10.1177/2045125318765725

Hayes, R.D., Downs, J., Chang, C.-K., Jackson, R.G., Shetty, H., Broadbent, M., Hotopf, M., Stewart, R., 2015. The Effect of Clozapine on Premature Mortality: An Assessment of Clinical Monitoring and Other Potential Confounders. Schizophr. Bull. 41, 644–655. https://doi.org/10.1093/schbul/sbu120

Healthy London Partnership, 2016. Improving care for children and young people with mental health crisis in London.pdf. Healthy London Partnership's Children and Young People's Programme, London.

Hernán, M.A., Alonso, A., Logan, R., Grodstein, F., Michels, K.B., Stampfer, M.J., Willett, W.C., Manson, J.E., Robins, J.M., 2008. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiol. Camb. Mass 19, 766–779. https://doi.org/10.1097/EDE.0b013e3181875e61

Hernán, M.A., Cole, S.R., 2009. Invited Commentary: Causal Diagrams and Measurement Bias. Am. J. Epidemiol. 170, 959–962. https://doi.org/10.1093/aje/kwp293

Hernán, M.A., Robins, 2018. Causal Inference.

Hernan, M.A., Robins, J.M., 2020. Causal Inference. Boca Raton: Chapman & Hall/CRC.

Heslin, M., Khondoker, M., Shetty, H., Pritchard, M., Jones, P.B., Osborn, D., Kirkbride, J.B., Roberts, A., Stewart, R., 2018. Inpatient use and area-level socio-environmental factors in people with psychosis. Soc. Psychiatry Psychiatr. Epidemiol. https://doi.org/10.1007/s00127-018-1534-x

Higgins, J.P.T., Green, S., Cochrane Collaboration (Eds.), 2008. Cochrane handbook for systematic reviews of interventions, Cochrane book series. Wiley-Blackwell, Chichester, England ; Hoboken, NJ.

Howard, L., Flach, C., Leese, M., Byford, S., Killaspy, H., Cole, L., Lawlor, C., Betts, J., Sharac, J., Cutting, P., McNicholas, S., Johnson, S., 2010. Effectiveness and cost-effectiveness of admissions to women's crisis houses compared with traditional psychiatric wards: pilot patient-preference randomised controlled trial. Br. J. Psychiatry 197, s32–s40. https://doi.org/10.1192/bjp.bp.110.081083

Hox, J.J., Boeije, H.R., 2005. Data collection, primary versus secondary, in: Encyclopedia of Social Measurement. Elsevier, p. 593null.

Iacus, S.M., King, G., Porro, G., 2012. Causal Inference without Balance Checking: Coarsened Exact Matching. Polit. Anal. 20, 1–24. https://doi.org/10.1093/pan/mpr013

Imai, K., Kim, I.S., 2019. When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data? Am. J. Polit. Sci.

Imai, K., King, G., Stuart, E.A., 2008. Misunderstandings between experimentalists and observationalists about causal inference. J. R. Stat. Soc. Ser. A Stat. Soc. 171, 481–502. https://doi.org/10.1111/j.1467-985X.2007.00527.x

Imbens, G.W., 2010. Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). J. Econ. Lit. 48, 399–423. https://doi.org/10.1257/jel.48.2.399

Imbens, G.W., Wooldridge, J.M., 2009. Recent Developments in the Econometrics of Program Evaluation. J. Econ. Lit. 47, 5–86. https://doi.org/10.1257/jel.47.1.5

Immonen, J., Jääskeläinen, E., Korpela, H., Miettunen, J., 2017. Age at onset and the outcomes of schizophrenia: A systematic review and meta-analysis. Early Interv. Psychiatry 11, 453–460. https://doi.org/10.1111/eip.12412

Irving, J., 2019. NLP tool validation report and recommendations (Internal Report). South London and Maudsley Biomedical Research Centre, London.

Iyer, S.N., Rothmann, T.L., Vogler, J.E., Spaulding, W.D., 2005. Evaluating outcomes of rehabilitation for severe mental illness. Rehabil. Psychol. 50, 43–55. https://doi.org/10.1037/0090-5550.50.1.43

Jackson, R.G., Patel, R., Jayatilleke, N., Kolliakou, A., Ball, M., Gorrell, G., Roberts, A., Dobson, R.J., Stewart, R., 2017. Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. BMJ Open 7, e012012.

Jacobs, R., 2016. Funding of Mental Health Services: Do Available Data Support Episodic Payment.

Jacobs, R., 2009. Investigating Patient Outcome Measures in Mental Health.pdf, CHE Research Paper 48. Centre for Health Economics, University of York.

Jacobs, R., Chalkley, M., Aragón, M.J., Böhnke, J.R., Clark, M., Moran, V., 2018. Funding approaches for mental health services: is there still a role for clustering? BJPsych Adv. 1–10. https://doi.org/10.1192/bja.2018.34

Jin, H., Mosweu, I., 2017. The Societal Cost of Schizophrenia: A Systematic Review. PharmacoEconomics 35, 25–42. https://doi.org/10.1007/s40273-016-0444-6
Jin, H., Tappenden, P., Robinson, S., Achilla, E., MacCabe, J.H., Aceituno, D., Byford, S., 2020. A Systematic Review of Economic Models Across the Entire Schizophrenia Pathway. PharmacoEconomics 38, 537–555. https://doi.org/10.1007/s40273-020-00895-6
Jones, A.M., Rice, N., 2011. Econometric evaluation of health policies, in: The Oxford Handbook of Health Economics. Oxford University Press, pp. 890–923.
Jones, H.E., Schooling, C.M., 2018. Let's Require the "T-Word." Am. J. Public Health 108, 624–624. https://doi.org/10.2105/AJPH.2018.304365
Kadra, G., Stewart, R., Shetty, H., Jackson, R.G., Greenwood, M.A., Roberts, A., Chang, C.-K., MacCabe, J.H., Hayes, R.D., 2015. Extracting antipsychotic polypharmacy data from electronic health records: developing and evaluating a novel process. BMC Psychiatry 15. https://doi.org/10.1186/s12888-015-0557-z
Kandiyali, R., Hawton, A., Cabral, C., Mytton, J., Shilling, V., Morris, C., Ingram, J., 2019. Working with Patients and Members of the Public: Informing Health Economics in Child Health Research. PharmacoEconomics - Open 3, 133–141. https://doi.org/10.1007/s41669-018-0099-7
Kennedy, P., 2008. A Guide to Econometrics, 6th edition. ed. Wiley-Blackwell, Malden, MA.
Keogh, R.H., Bartlett, J.W., 2019. Measurement error as a missing data problem. ArXiv Prepr. ArXiv191006443.
Kessler, R.C., Amminger, G.P., Aguilar-Gaxiola, S., Alonso, J., Lee, S., Ustun, T.B., 2007. Age of onset of mental disorders: a review of recent literature. Curr. Opin. Psychiatry 20, 359.
Kilian, R., Losert, C., Park, A.-L., McDaid, D., Knapp, M., 2010. Cost-Effectiveness Analysis in Child and Adolescent Mental Health Problems: An Updated Review of Literature. Int. J. Ment. Health Promot. 12, 45–57. https://doi.org/10.1080/14623730.2010.9721825
Killaspy, H., 2019. Contemporary mental health rehabilitation. Epidemiol. Psychiatr. Sci. 28, 1–3. https://doi.org/10.1017/S2045796018000318
Killaspy, H., 2014. The ongoing need for local services for people with complex mental health problems. Psychiatr. Bull. 38, 257–259. https://doi.org/10.1192/pb.bp.114.048470
Killaspy, H., 2009. Enabling recovery for people with complex mental health needs.
Killaspy, H., King, M., Holloway, F., Craig, T.J., Cook, S., Mundy, T., Leavey, G., McCrone, P., Koeser, L., Omar, R., Marston, L., Arbuthnott, M., Green, N., Harrison, I., Lean, M., Gee, M., Bhanbhro, S., 2017. The Rehabilitation Effectiveness for Activities for Life (REAL) study: a national programme of research into NHS inpatient mental health rehabilitation services across England, Programme Grants for Applied Research. NIHR Journals Library, Southampton (UK).
Killaspy, H., Marston, L., Green, N., Harrison, I., Lean, M., Holloway, F., Craig, T., Leavey, G., Arbuthnott, M., Koeser, L., McCrone, P., Omar, R.Z., King, M., 2016. Clinical outcomes and costs for people with complex psychosis; a naturalistic prospective cohort study of mental health rehabilitation service users in England. BMC Psychiatry 16, 95. https://doi.org/10.1186/s12888-016-0797-6
Killaspy, H., Marston, L., Omar, R.Z., Green, N., Harrison, I., Lean, M., Holloway, F., Craig, T., Leavey, G., King, M., 2013. Service quality and clinical outcomes: an example from mental health rehabilitation services in England. Br. J. Psychiatry 202, 28–34. https://doi.org/10.1192/bjp.bp.112.114421
Kirkbride, J.B., Errazuriz, A., Croudace, T.J., Morgan, C., Jackson, D., Boydell, J., Murray, R.M., Jones, P.B., 2012. Incidence of Schizophrenia and Other Psychoses in England, 1950–2009: A Systematic Review and Meta-Analyses. PLOS ONE 7, e31660. https://doi.org/10.1371/journal.pone.0031660
Knapp, M., Ardino, V., Brimblecombe, N., Evans-Lacko, S., Iemmi, V., King, D., Snell, T., Murguia, S., Mbeah-Bankas, H., Crane, S., Harris, A., Fowler, D., Hodgekins, J., Wilson, J., 2016. Youth Mental Health: New Economic Evidence. Personal Social Services Research Unit, London School of Economics and Political Science, London.

Knapp, M., Beecham, J., McDaid, D., Matosevic, T., Smith, M., 2010. The economic consequences of deinstitutionalisation of mental health services: lessons from a systematic review of European experience: Economic consequences of deinstitutionalisation of mental health services. Health Soc. Care Community no-no. https://doi.org/10.1111/j.1365-2524.2010.00969.x

Kreif, N., Grieve, R., Radice, R., Sekhon, J.S., 2013a. Regression-adjusted matching and double-robust methods for estimating average treatment effects in health economic evaluation. Health Serv. Outcomes Res. Methodol. 13, 174–202. https://doi.org/10.1007/s10742-013-0109-2

Kreif, N., Grieve, R., Sadique, M.Z., 2013b. Statistical Methods for cost-effectiveness analyses that use observational data: A critical appraisal tool and review of current practice: Statistical methods for CEAs that use observational data. Health Econ. 22, 486–500. https://doi.org/10.1002/hec.2806

Kreif, N., Gruber, S., Radice, R., Grieve, R., Sekhon, J.S., 2016. Evaluating treatment effectiveness under model misspecification: A comparison of targeted maximum likelihood estimation with bias-corrected matching. Stat. Methods Med. Res. 25, 2315–2336. https://doi.org/10.1177/0962280214521341

Kreif, N., Sofrygin, O., Schmittdiel, J., Adams, A., Grant, R., Zhu, Z., van der Laan, M., Neugebauer, R., 2018. Evaluation of adaptive treatment strategies in an observational study where time-varying covariates are not monitored systematically. ArXiv180611153 Stat.

Krieger, N., Davey Smith, G., 2016. Response: FACEing reality: productive tensions between our epidemiological questions, methods and mission. Int. J. Epidemiol. 45, 1852–1865. https://doi.org/10.1093/ije/dyw330

Kuntz, K., Sainfort, F., Butler, M., Taylor, B., Kulasingam, S., Gregory, S., Mann, E., Anderson, J.M., Kane, R.L., 2013. Overview of Decision Models Used in Research, Decision and Simulation Modeling in Systematic Reviews [Internet]. Agency for Healthcare Research and Quality (US).

Lafave, H.G., de Souza, H.R., Gerber, G.J., 1996. Assertive community treatment of severe mental illness: a Canadian experience. Psychiatr. Serv. Wash. DC 47, 757–759. https://doi.org/10.1176/ps.47.7.757

Lamb, C., Hall, D., Kelvin, R., Van Beinum, M., 2008. Working at the CAMHS/Adult Interface: Good practice guidance for the provision of psychiatric services to adolescents/young adults. R. Coll. Psychiatr.

Lash, T.L., Fox, M.P., Fink, A.K., 2009. Applying Quantitative Bias Analysis to Epidemiologic Data, Statistics for Biology and Health. Springer-Verlag, New York.

Lash, T.L., Fox, M.P., MacLehose, R.F., Maldonado, G., McCandless, L.C., Greenland, S., 2014. Good practices for quantitative bias analysis. Int. J. Epidemiol. 43, 1969–1985. https://doi.org/10.1093/ije/dyu149

Latimer, N.R., Abrams, K.R., Lambert, P.C., Crowther, M.J., Wailoo, A.J., Morden, J.P., Akehurst, R.L., Campbell, M.J., 2014. Adjusting survival time estimates to account for treatment switching in randomized controlled trials--an economic evaluation context: methods, limitations, and recommendations. Med. Decis. Mak. Int. J. Soc. Med. Decis. Mak. 34, 387–402. https://doi.org/10.1177/0272989X13520192

Lavelle, E., Ijaz, A., Killaspy, H., Holloway, F., King, M., Keogh, F., McDonough, C., Spelman, L., Goggins, R., Daly, I., Murphy, K., McCrone, P., Drake, R., 2012. Mental health rehabilitation and recovery services in Ireland: a multicentre study of current service provision,characteristics of service users and outcomes for those with and without access to these services (Report). Mental Health Commission.

Lee, D.S., Lemieux, T., 2010. Regression Discontinuity Designs in Economics. J. Econ. Lit. 48, 281–355. https://doi.org/10.1257/jel.48.2.281

Lessard, C., 2007. Complexity and reflexivity: Two important issues for economic evaluation in health care. Soc. Sci. Med. 64, 1754–1765. https://doi.org/10.1016/j.socscimed.2006.12.006

Li, F., Morgan, K.L., Zaslavsky, A.M., 2018. Balancing Covariates via Propensity Score Weighting. J. Am. Stat. Assoc. 113, 390–400. https://doi.org/10.1080/01621459.2016.1260466
Li, Y., Schaubel, D.E., He, K., 2014. Matching Methods for Obtaining Survival Functions to Estimate the Effect of a Time-Dependent Treatment. Stat. Biosci. 6, 105–126. https://doi.org/10.1007/s12561-013-9085-x
Li, Y.P., Propert, K.J., Rosenbaum, P.R., 2001. Balanced risk set matching. J. Am. Stat. Assoc. 96, 870–882.
Lipsitch, M., Tchetgen, E.T., Cohen, T., 2010. Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. Epidemiol. Camb. Mass 21, 383–388. https://doi.org/10.1097/EDE.0b013e3181d61eeb
Liu, N.H., Choi, K.H., Reddy, F., Spaulding, W.D., 2011. Heterogeneity and the longitudinal recovery of functioning during inpatient psychiatric rehabilitation for treatment-refractory severe mental illness. Am. J. Psychiatr. Rehabil. 14, 55–75. https://doi.org/10.1080/15487768.2011.546293
Liu, W., Brookhart, M.A., Schneeweiss, S., Mi, X., Setoguchi, S., 2012. Implications of M bias in epidemiologic studies: a simulation study. Am. J. Epidemiol. 176, 938–948. https://doi.org/10.1093/aje/kws165
Lloyd, K., White, J., 2011. Democratizing clinical research. Nature 474, 277–278.
Lohr, S., 2009. Sampling: design and analysis. Nelson Education.
Macpherson, R., Calciu, C., Foy, C., Humby, K., Lozynskyj, D., Garton, C., Steer, H., Elliott, H., 2017. A service evaluation of outcomes in two in-patient recovery units. BJPsych Bull. 41, 330–336. https://doi.org/10.1192/pb.bp.116.055137
Makady, A., de Boer, A., Hillege, H., Klungel, O., Goettsch, W., (on behalf of GetReal Work Package 1), 2017. What Is Real-World Data? A Review of Definitions Based on Literature and Stakeholder Interviews. Value Health J. Int. Soc. Pharmacoeconomics Outcomes Res. 20, 858–865. https://doi.org/10.1016/j.jval.2017.03.008
Mangalore, R., Knapp, M., 2007. Cost of schizophrenia in England. J. Ment. Health Policy Econ. 10, 23–41.
Marschner, I., 2006. Measurement error bias in pharmaceutical cost-effectiveness analysis. Appl. Stoch. Models Bus. Ind. 22, 621–630. https://doi.org/10.1002/asmb.644
Martinic, M.K., Pieper, D., Glatt, A., Puljak, L., 2019. Definition of a systematic review used in overviews of systematic reviews, meta-epidemiological studies and textbooks. BMC Med. Res. Methodol. 19, 203. https://doi.org/10.1186/s12874-019-0855-0
Mathur, R., Bhaskaran, K., Chaturvedi, N., Leon, D.A., vanStaa, T., Grundy, E., Smeeth, L., 2014. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. J. Public Health 36, 684–692. https://doi.org/10.1093/pubmed/fdt116
Matt, V., Matthew, H., 2013. The retrospective chart review: important methodological considerations. J. Educ. Eval. Health Prof. 10. https://doi.org/10.3352/jeehp.2013.10.12
Mauskopf, J.A., Paul, J.E., Grant, D.M., Stergachis, A., 1998. The role of cost-consequence analysis in healthcare decision-making. PharmacoEconomics 13, 277–288. https://doi.org/10.2165/00019053-199813030-00002
Mavranezouli, I., 2010. A review and critique of studies reporting utility values for schizophrenia-related health states. Pharmacoeconomics 28, 1109–1121.
Maxwell, S., Ugochukwu, O., Clarke, T., Gee, B., Clarke, E., Westgate, H., Wilson, J., Lennox, B.R., Goodyer, I.M., 2019. The effect of a youth mental health service model on access to secondary mental healthcare for young people aged 14–25 years. BJPsych Bull. 43, 27–31. https://doi.org/10.1192/bjb.2018.70
McCaffrey, D.F., Ridgeway, G., Morral, A.R., 2004. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol. Methods 9, 403–425. https://doi.org/10.1037/1082-989X.9.4.403

McNamee, R., 2005. Optimal design and efficiency of two-phase case-control studies with error-prone and error-free exposure measures. Biostatistics 6, 590–603. https://doi.org/10.1093/biostatistics/kxi029
Meacock, R., 2018. Methods for the economic evaluation of changes to the organisation and delivery of health services: principal challenges and recommendations. Health Econ. Policy Law 1–16. https://doi.org/10.1017/S1744133118000063
Mental Welfare Commission, 2012. Good practice guide - Young people in adult mental health wards.
Mihaylova, B., Briggs, A., O'Hagan, A., Thompson, S.G., 2011. Review of Statistical Methods for Analysing Healthcare Resources and Costs. Health Econ. 20, 897–916. https://doi.org/10.1002/hec.1653
Morgan, S.L., Winship, C., 2014. Counterfactuals and Causal Inference. Cambridge University Press.
Morris, T.P., White, I.R., Crowther, M.J., 2019. Using simulation studies to evaluate statistical methods. Stat. Med. 38, 2074–2102. https://doi.org/10.1002/sim.8086
Moscoe, E., Bor, J., Bärnighausen, T., 2015. Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. J. Clin. Epidemiol. 68, 132–143. https://doi.org/10.1016/j.jclinepi.2014.06.021
Mozer, R., Miratrix, L., Kaufman, A.R., Anastasopoulos, L.J., 2018. Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality. arXiv.org.
Murcott, W., 2016. A scoping review of care received by young people aged 16-25 when admitted to adult mental health hospital wards. J. Public Ment. Health 15, 216–228. https://doi.org/10.1108/JPMH-05-2016-0025
Myers, J.A., Rassen, J.A., Gagne, J.J., Huybrechts, K.F., Schneeweiss, S., Rothman, K.J., Joffe, M.M., Glynn, R.J., 2011. Effects of Adjusting for Instrumental Variables on Bias and Precision of Effect Estimates. Am. J. Epidemiol. 174, 1213–1222. https://doi.org/10.1093/aje/kwr364
Naysmith, S., 2018. Children admitted to adult mental health wards 98 times last year in Scotland. HeraldScotland.
Neumann, P.J., Ganiats, T.G., Russell, L.B., Sanders, G.D., Siegel, J.E. (Eds.), 2016. Cost-Effectiveness in Health and Medicine, Second Edition. ed. Oxford University Press, Oxford, New York.
NHS Digital, 2018. Children and young people (CYP) in adult mental health in-patient wards [WWW Document]. NHS Digit. URL https://digital.nhs.uk/data-and-information/find-data-and-publications/supplementary-information/2018-supplementary-information-files/children-and-young-people-cyp-in-adult-mental-health-in-patient-wards (accessed 3.18.19).
NHS England, 2014. Child and Adolescent Mental Health Services (CAMHS) Tier 4 Report.
NHS improvement, 2018. NHS Reference costs 2017/18 [WWW Document]. Ref. Costs NHS Improv. URL https://improvement.nhs.uk/resources/reference-costs/ (accessed 4.4.19).
NICE, 2020. The NICE methods of health technology evaluation: the case for change.
NICE, 2018. Final Guideline scope - Rehabilitation in adults with severe and enduring mental illness. National Institute of Health and Care Excellence, London.
NICE, 2014a. Psychosis and Schizophrenia in adults - The NICE guideline on treatment and management - Updated Edition 2014, National Clinical Guideline Number 178. National Collaborating Centre for Mental Health commissioned by National Institute of Health and Care Excellence.
NICE, 2014b. Bipolar disorder: assessment and management. London.
NICE, 2013. Psychosis and schizophrenia in children and young people: recognition and management. Rcpsych Publiciations.
Nordentoft, M., Øhlenschlæger, J., Thorup, A., Petersen, L., Jeppesen, P., Bertelsen, M., 2010. Deinstitutionalization revisited: a 5-year follow-up of a randomized clinical trial of hospital-based rehabilitation versus specialized assertive intervention (OPUS) versus standard

treatment for patients with first-episode schizophrenia spectrum disorders. Psychol. Med. 40, 1619–1626. https://doi.org/10.1017/S0033291709992182
Office of National Statistics, 2017. Mortality statistics in England and Wales - Quality and Methodology Information [WWW Document]. URL https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/qmis/mortalitystatisticsinenglandandwalesqmi (accessed 6.21.18).
OHE, 1997. The Pros and Cons of Modelling in Economic Evaluation.
O'Keeffe, A.G., Baio, G., 2016. Approaches to the Estimation of the Local Average Treatment Effect in a Regression Discontinuity Design. Scand. J. Stat. Theory Appl. 43, 978–995. https://doi.org/10.1111/sjos.12224
Oldenburg, C.E., Moscoe, E., Bärnighausen, T., 2016. Regression Discontinuity for Causal Effect Estimation in Epidemiology. Curr. Epidemiol. Rep. 3, 233–241. https://doi.org/10.1007/s40471-016-0080-x
O'Neill, S., Kreif, N., Grieve, R., Sutton, M., Sekhon, J.S., 2016. Estimating causal effects: considering three alternatives to difference-in-differences estimation. Health Serv. Outcomes Res. Methodol. 16, 1–21. https://doi.org/10.1007/s10742-016-0146-8
O'Shea, E., Ogbebor, F., Queally, M., Murphy, E., 2019. Knowledge of public patient involvement among health economists in Ireland: a baseline audit. HRB Open Res. 2, 4. https://doi.org/10.12688/hrbopenres.12896.1
Ouzzani, M., Hammady, H., Fedorowicz, Z., Elmagarmid, A., 2016. Rayyan—a web and mobile app for systematic reviews. Syst. Rev. 5. https://doi.org/10.1186/s13643-016-0384-4
Owen, J., 2018. Children are still being forced to travel far for mental healthcare. BMJ k3769. https://doi.org/10.1136/bmj.k3769
Park, C., McDermott, B., Loy, J., Dean, P., 2011. Adolescent Admissions to Adult Psychiatric Units: Patterns and Implications for Service Provision. Australas. Psychiatry 19, 345–349. https://doi.org/10.3109/10398562.2011.601311
Park, H.S., Lloyd, S., Decker, R.H., Wilson, L.D., Yu, J.B., 2012. Overview of the Surveillance, Epidemiology, and End Results database: evolution, data variables, and quality assurance. Curr. Probl. Cancer 36, 183–190. https://doi.org/10.1016/j.currproblcancer.2012.03.007
Parkkinen, V.-P., Wallmann, C., Wilde, M., Clarke, B., Illari, P., Kelly, M.P., Norell, C., Russo, F., Shaw, B., Williamson, J., 2018. Evaluating Evidence of Mechanisms in Medicine: Principles and Procedures, SpringerBriefs in Philosophy. Springer International Publishing.
Patel, R., Jayatilleke, N., Broadbent, M., Chang, C.-K., Foskett, N., Gorrell, G., Hayes, R.D., Jackson, R., Johnston, C., Shetty, H., others, 2015. Negative symptoms in schizophrenia: a study in a large clinical sample of patients using a novel automated method. BMJ Open 5, e007619.
Patel, R., Oduola, S., Callard, F., Wykes, T., Broadbent, M., Stewart, R., Craig, T.K.J., McGuire, P., 2017. What proportion of patients with psychosis is willing to take part in research? A mental health electronic case register analysis. BMJ Open 7, e013113. https://doi.org/10.1136/bmjopen-2016-013113
Pearl, J., 2010. An Introduction to Causal Inference. Int. J. Biostat. 6. https://doi.org/10.2202/1557-4679.1203
Pearson, S.D., Dreitlein, W.B., Towse, A., Hampson, G., Henshall, C., 2018. A framework to guide the optimal development and use of real-world evidence for drug coverage and formulary decisions. J. Comp. Eff. Res. 7, 1145–1152. https://doi.org/10.2217/cer-2018-0059
Perera, G., Broadbent, M., Callard, F., Chang, C.-K., Downs, J., Dutta, R., Fernandes, A., Hayes, R.D., Henderson, M., Jackson, R., others, 2016. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) case register: current status and recent enhancement of an electronic mental health record-derived data resource. BMJ Open 6, e008721.

Petersen, M.L., Porter, K.E., Gruber, S., Wang, Y., van der Laan, M.J., 2012. Diagnosing and responding to violations in the positivity assumption. Stat. Methods Med. Res. 21, 31–54. https://doi.org/10.1177/0962280210386207
Petrie, R., Mountain, D., 2009. An Observational Study of the Impact of a Rehabilitation Admission on Readmission Data. Scott. Med. J. 54, 32–34. https://doi.org/10.1258/rsmsmj.54.1.32
Petrou, S., Gray, A., 2011. Economic evaluation alongside randomised controlled trials: design, conduct, analysis, and reporting. BMJ 342, d1548. https://doi.org/10.1136/bmj.d1548
Pirkis, J.E., Burgess, P.M., Kirk, P.K., Dodson, S., Coombs, T.J., Williamson, M.K., 2005. A review of the psychometric properties of the Health of the Nation Outcome Scales (HoNOS) family of measures. Health Qual. Life Outcomes 3, 76. https://doi.org/10.1186/1477-7525-3-76
Price, A., Farooq, R., Yuan, J.-M., Menon, V.B., Cardinal, R.N., O'Brien, J.T., 2017. Mortality in dementia with Lewy bodies compared with Alzheimer's dementia: a retrospective naturalistic cohort study. BMJ Open 7, e017504. https://doi.org/10.1136/bmjopen-2017-017504
Raftery, J., Young, A., Stanton, L., Milne, R., Cook, A., Turner, D., Davidson, P., 2015. Clinical trial metadata: defining and extracting metadata on the design, conduct, results and costs of 125 randomised clinical trials funded by the National Institute for Health Research Health Technology Assessment programme. Health Technol. Assess. 19, 1–138. https://doi.org/10.3310/hta19110
Raine, R., Fitzpatrick, R., Barratt, H., Bevan, G., Black, N., Boaden, R., Bower, P., Campbell, M., Denis, J.-L., Devers, K., Dixon-Woods, M., Fallowfield, L., Forder, J., Foy, R., Freemantle, N., Fulop, N.J., Gibbons, E., Gillies, C., Goulding, L., Grieve, R., Grimshaw, J., Howarth, E., Lilford, R.J., McDonald, R., Moore, G., Moore, L., Newhouse, R., O'Cathain, A., Or, Z., Papoutsi, C., Prady, S., Rycroft-Malone, J., Sekhon, J., Turner, S., Watson, S.I., Zwarenstein, M., 2016. Challenges, solutions and future directions in the evaluation of service innovations in health care and public health, Health Services and Delivery Research. NIHR Journals Library, Southampton (UK).
Reed, S.I., 2008. First-episode psychosis: A literature review. Int. J. Ment. Health Nurs. 17, 85–91. https://doi.org/10.1111/j.1447-0349.2008.00515.x
Reilly, M., 1996. Optimal sampling strategies for two-stage studies. Am. J. Epidemiol. 143, 92–100.
Richardson, G., 2010. Child and adolescent mental health services: An operational handbook. RCPsych Publications.
Rose, D., Evans, J., Laker, C., Wykes, T., 2015. Life in acute mental health settings: experiences and perceptions of service users and nurses. Epidemiol. Psychiatr. Sci. 24, 90–96. https://doi.org/10.1017/S2045796013000693
Rosenbaum, P.R., 2010. Design of Observational Studies, Springer Series in Statistics. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4419-1213-8
Rubin, D.B., 2008. For objective causal inference, design trumps analysis. https://doi.org/10.1214/08-AOAS187
Sainfort, F., Kuntz, K.M., Gregory, S., Butler, M., Taylor, B.C., Kulasingam, S., Kane, R.L., 2013. Adding decision models to systematic reviews: informing a framework for deciding when and how to do so. Value Health J. Int. Soc. Pharmacoeconomics Outcomes Res. 16, 133–139. https://doi.org/10.1016/j.jval.2012.09.009
Saunders, C.L., Abel, G.A., Turabi, A.E., Ahmed, F., Lyratzopoulos, G., 2013. Accuracy of routinely recorded ethnic group information compared with self-reported ethnicity: evidence from the English Cancer Patient Experience survey. BMJ Open 3, e002882. https://doi.org/10.1136/bmjopen-2013-002882
Schneeweiss, S., 2018. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. Clin. Epidemiol. 10, 771–788. https://doi.org/10.2147/CLEP.S166545

Schneeweiss, S., 2006. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. Pharmacoepidemiol. Drug Saf. 15, 291–303. https://doi.org/10.1002/pds.1200
Schonlau, M., 2005. Boosted regression (boosting): An introductory tutorial and a Stata plugin. Stata J. 5, 330–354.
Sculpher, M.J., Claxton, K., Drummond, M., McCabe, C., 2006. Whither trial-based economic evaluation for health care decision making? Health Econ. 15, 677–687. https://doi.org/10.1002/hec.1093
Seaman, S.R., White, I.R., 2013. Review of inverse probability weighting for dealing with missing data. Stat. Methods Med. Res. 22, 278–295. https://doi.org/10.1177/0962280210395740
Sernyak, M.J., Rosenheck, R., 2003. Risk adjustment in studies using administrative data. Schizophr. Bull. 29, 267–271.
Shaffer, D., Gould, M.S., Brasic, J., Ambrosini, P., Fisher, P., Bird, H., Aluwahlia, S., 1983. A Children's Global Assessment Scale (CGAS). Arch. Gen. Psychiatry 40, 1228–1231. https://doi.org/10.1001/archpsyc.1983.01790100074010
Simmons, P., Hawley, C.J., Gale, T.M., Sivakumaran, T., 2010. Service user, patient, client, user or survivor: describing recipients of mental health services. The Psychiatrist 34, 20–23. https://doi.org/10.1192/pb.bp.109.025247
Singh, S.P., Tuomainen, H., Girolamo, G. de, Maras, A., Santosh, P., McNicholas, F., Schulze, U., Purper-Ouakil, D., Tremmery, S., Franić, T., Madan, J., Paul, M., Verhulst, F.C., Dieleman, G.C., Warwick, J., Wolke, D., Street, C., Daffern, C., Tah, P., Griffin, J., Canaway, A., Signorini, G., Gerritsen, S., Adams, L., O'Hara, L., Aslan, S., Russet, F., Davidović, N., Tuffrey, A., Wilson, A., Gatherer, C., Walker, L., 2017. Protocol for a cohort study of adolescent mental health service users with a nested cluster randomised controlled trial to assess the clinical and cost-effectiveness of managed transition in improving transitions from child to adult mental health services (the MILESTONE study). BMJ Open 7, e016055. https://doi.org/10.1136/bmjopen-2017-016055
Singleton, N., Bumpstead, R., O'Brien, M., Lee, A., Meltzer, H., 2003. Psychiatric morbidity among adults living in private households, 2000. Int. Rev. Psychiatry Abingdon Engl. 15, 65–73. https://doi.org/10.1080/0954026021000045967
Sinha, S., Peach, G., Poloniecki, J.D., Thompson, M.M., Holt, P.J., 2013. Studies using English administrative data (Hospital Episode Statistics) to assess health-care outcomes--systematic review and recommendations for reporting. Eur. J. Public Health 23, 86–92. https://doi.org/10.1093/eurpub/cks046
Siomi, A.B., Razzouk, D., 2017. Costing Psychiatric Hospitals and Psychiatric Wards in General Hospitals, in: Mental Health Economics. Springer, Cham, pp. 225–237. https://doi.org/10.1007/978-3-319-55266-8_14
Slade, M., Byford, S., Barrett, B., Lloyd-Evans, B., Gilburt, H., Osborn, D.P.J., Skinner, R., Leese, M., Thornicroft, G., Johnson, S., 2010. Alternatives to standard acute in-patient care in England: short-term clinical outcomes and cost-effectiveness. Br. J. Psychiatry 197, s14–s19. https://doi.org/10.1192/bjp.bp.110.081059
Smith, A.F., Messenger, M., Hall, P., Hulme, C., 2018. The Role of Measurement Uncertainty in Health Technology Assessments (HTAs) of In Vitro Tests. PharmacoEconomics 1–13. https://doi.org/10.1007/s40273-018-0638-1
Soares, M.O., Sharples, L., Morton, A., Claxton, K., Bojke, L., 2018. Experiences of Structured Elicitation for Model-Based Cost-Effectiveness Analyses. Value Health 21, 715–723. https://doi.org/10.1016/j.jval.2018.01.019
Spiegelhalter, D.J., Abrams, K.R., Myles, J.P., 2004. Bayesian approaches to clinical trials and health-care evaluation. John Wiley & Sons.
Sterne, J.A., Hernán, M.A., Reeves, B.C., Savović, J., Berkman, N.D., Viswanathan, M., Henry, D., Altman, D.G., Ansari, M.T., Boutron, I., Carpenter, J.R., Chan, A.-W., Churchill, R., Deeks, J.J.,

Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y.K., Pigott, T.D., Ramsay, C.R., Regidor, D., Rothstein, H.R., Sandhu, L., Santaguida, P.L., Schünemann, H.J., Shea, B., Shrier, I., Tugwell, P., Turner, L., Valentine, J.C., Waddington, H., Waters, E., Wells, G.A., Whiting, P.F., Higgins, J.P., 2016. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ i4919. https://doi.org/10.1136/bmj.i4919

Stevenson, A., 2010. Oxford Dictionary of English. OUP Oxford.

Stewart, R., Davis, K., 2016. 'Big data' in mental health research: current status and emerging possibilities. Soc. Psychiatry Psychiatr. Epidemiol. 51, 1055–1072. https://doi.org/10.1007/s00127-016-1266-8

Stewart, R., Soremekun, M., Perera, G., Broadbent, M., Callard, F., Denis, M., Hotopf, M., Thornicroft, G., Lovestone, S., 2009. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. BMC Psychiatry 9. https://doi.org/10.1186/1471-244X-9-51

Stoll, C.R.T., Izadi, S., Fowler, S., Green, P., Suls, J., Colditz, G.A., 2019. The value of a second reviewer for study selection in systematic reviews. Res. Synth. Methods 10, 539–545. https://doi.org/10.1002/jrsm.1369

Stone, C.A., Tang, Y., 2013. Comparing Propensity Score Methods in Balancing Covariates and Recovering Impact in Small Sample Educational Program Evaluations. Pract. Assess. Res. Eval. 18.

Struck, R., Baumgarten, G., Wittmann, M., 2014. Cost-efficiency of knowledge creation: randomized controlled trials vs. observational studies. Curr. Opin. Anaesthesiol. 27, 190–194. https://doi.org/10.1097/ACO.0000000000000060

Stuart, E.A., 2010. Matching Methods for Causal Inference: A Review and a Look Forward. Stat. Sci. 25, 1–21. https://doi.org/10.1214/09-STS313

Stuart, E.A., Ialongo, N.S., 2010. Matching methods for selection of subjects for follow-up. Multivar. Behav. Res. 45, 746–765. https://doi.org/10.1080/00273171.2010.503544

Sullivan, M.E., Richardson, C.E., Spaulding, W.D., 1991. University-state hospital collaboration in an inpatient psychiatric rehabilitation program. Community Ment. Health J. 27, 441–453. https://doi.org/10.1007/bf00752667

Sutton, M., Garfield-Birkbeck, S., Martin, G., Meacock, R., Morris, S., Sculpher, M., Street, A., Watson, S.I., Lilford, R.J., 2018. Economic analysis of service and delivery interventions in health care. Health Serv. Deliv. Res. 6, 1–16. https://doi.org/10.3310/hsdr06050

Swanson, S.A., Hernán, M.A., 2018. The challenging interpretation of instrumental variable estimates under monotonicity. Int. J. Epidemiol. 47, 1289–1297. https://doi.org/10.1093/ije/dyx038

Szymczynska, P., Walsh, S., Greenberg, L., Priebe, S., 2017. Attrition in trials evaluating complex interventions for schizophrenia: Systematic review and meta-analysis. J. Psychiatr. Res. 90, 67–77.

Tafti, A.R., Shmueli, G., 2019. Beyond Overall Treatment Effects: Leveraging Covariates in Randomized Experiments Guided by Causal Structure. Available SSRN 3331772.

Tandon, R., Nasrallah, H.A., Keshavan, M.S., 2009. Schizophrenia, "just the facts" 4. Clinical features and conceptualization. Schizophr. Res. 110, 1–23. https://doi.org/10.1016/j.schres.2009.03.005

Tarasenko, M., Sullivan, M., Ritchie, A.J., Spaulding, W.D., 2013. Effects of eliminating psychiatric rehabilitation from the secure levels of a mental-health service system. Psychol. Serv. 10, 442–451. https://doi.org/10.1037/a0030260

Textor, J., van der Zander, B., Gilthorpe, M.S., Liśkiewicz, M., Ellison, G.T., 2016. Robust causal inference using directed acyclic graphs: the R package 'dagitty.' Int. J. Epidemiol. 45, 1887–1894. https://doi.org/10.1093/ije/dyw341

Thomas, K.A., Rickwood, D., 2013. Clinical and Cost-Effectiveness of Acute and Subacute Residential Mental Health Services: A Systematic Review. Psychiatr. Serv. 64, 1140–1149. https://doi.org/10.1176/appi.ps.201200427

Thorn, J.C., Coast, J., Cohen, D., Hollingworth, W., Knapp, M., Noble, S.M., Ridyard, C., Wordsworth, S., Hughes, D., 2013. Resource-Use Measurement Based on Patient Recall: Issues and Challenges for Economic Evaluation. Appl. Health Econ. Health Policy. 11, 155–161. https://doi.org/10.1007/s40258-013-0022-4

Thorpe, R., Holt, R., 2008. The SAGE Dictionary of Qualitative Management Research. SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom. https://doi.org/10.4135/9780857020109

Toh, S., García Rodríguez, L.A., Hernán, M.A., 2011. Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: an application to electronic medical records. Pharmacoepidemiol. Drug Saf. 20, 849–857. https://doi.org/10.1002/pds.2152

Torgerson, D.J., Torgerson, C.J., 2003. Avoiding Bias in Randomised Controlled Trials in Educational Research. Br. J. Educ. Stud. 51, 36–45. https://doi.org/10.1111/1467-8527.t01-2-00223

Trevillion, K., Shallcross, R., Ryan, E., Heslin, M., Pickles, A., Byford, S., Jones, I., Johnson, S., Pawlby, S., Stanley, N., Rose, D., Seneviratne, G., Wieck, A., Jennings, S., Potts, L., Abel, K.M., Howard, L.M., 2019. Protocol for a quasi-experimental study of the effectiveness and cost-effectiveness of mother and baby units compared with general psychiatric inpatient wards and crisis resolution team services (The ESMI study) in the provision of care for women in the postpartum period. BMJ Open 9, e025906. https://doi.org/10.1136/bmjopen-2018-025906

Tsoutsoulis, K., Maxwell, A., Padinjareveettil, A.M.T., Zivkovic, F., Rogers, J.M., 2018. Impact of inpatient mental health rehabilitation on psychiatric readmissions: a propensity score matched case control study. J Ment Health 0, 1–9. https://doi.org/10.1080/09638237.2018.1466049

Tulloch, A., 2010. Length of stay on acute psychiatric wards. University of London, London.

Tulloch, A.D., Khondoker, M.R., Fearon, P., David, A.S., 2012. Associations of homelessness and residential mobility with length of stay after acute psychiatric admission. BMC Psychiatry 12, 121.

Tulloch, David, A.S., Thornicroft, G., 2016a. Exploring the predictors of early readmission to psychiatric hospital. Epidemiol. Psychiatr. Sci. 25, 181–193. https://doi.org/10.1017/S2045796015000128

Tulloch, Soper, B., Görzig, A., Pettit, S., Koeser, L., Polling, C., Watson, A., Khondoker, M., Rose, D., McCrone, P., Tylee, A., Thornicroft, G., 2016b. Management by geographical area or management specialised by disorder? A mixed-methods evaluation of the effects of an organisational intervention on secondary mental health care for common mental disorder. Health Serv. Deliv. Res. 4, 1–114. https://doi.org/10.3310/hsdr04090

Turner, R.M., Spiegelhalter, D.J., Smith, G.C.S., Thompson, S.G., 2009. Bias modelling in evidence synthesis. J. R. Stat. Soc. Ser. A Stat. Soc. 172, 21–47. https://doi.org/10.1111/j.1467-985X.2008.00547.x

Twomey, C., Prina, A.M., Baldwin, D.S., Das-Munshi, J., Kingdon, D., Koeser, L., Prince, M.J., Stewart, R., Tulloch, A.D., Cieza, A., 2016. Utility of the Health of the Nation Outcome Scales (HoNOS) in Predicting Mental Health Service Costs for Patients with Common Mental Health Problems: Historical Cohort Study. PLOS ONE 11, e0167103. https://doi.org/10.1371/journal.pone.0167103

Tyrer, P., Sharfstein, S., O'Reilly, R., Allison, S., Bastiampillai, T., 2017. Psychiatric hospital beds: an Orwellian crisis. The Lancet 389, 363. https://doi.org/10.1016/S0140-6736(17)30149-6

Uddin, Md.J., Groenwold, R.H.H., Ali, M.S., de Boer, A., Roes, K.C.B., Chowdhury, M.A.B., Klungel, O.H., 2016. Methods to control for unmeasured confounding in pharmacoepidemiology: an overview. Int. J. Clin. Pharm. 38, 714–723. https://doi.org/10.1007/s11096-016-0299-0

Ungar, W.J., Santos, M.T., 2003. The Pediatric Economic Database Evaluation (PEDE) Project: establishing a database to study trends in pediatric economic evaluation. Med. Care 41, 1142–1152. https://doi.org/10.1097/01.MLR.0000088451.56688.65

van Eck, N., Waltman, L., 2009. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 84, 523–538. https://doi.org/10.1007/s11192-009-0146-3
van Os, J., Rutten, B.P., Poulton, R., 2008. Gene-Environment Interactions in Schizophrenia: Review of Epidemiological Findings and Future Directions. Schizophr. Bull. 34, 1066–1082. https://doi.org/10.1093/schbul/sbn117
VanderWeele, T.J., Arah, O.A., 2011. Bias Formulas for Sensitivity Analysis of Unmeasured Confounding for General Outcomes, Treatments, and Confounders: Epidemiology 22, 42–52. https://doi.org/10.1097/EDE.0b013e3181f74493
VanderWeele, T.J., Hernan, M.A., 2013. Causal inference under multiple versions of treatment. J. Causal Inference 1, 1–20. https://doi.org/10.1515/jci-2012-0002
VanderWeele, T.J., Hernán, M.A., 2012. Results on Differential and Dependent Measurement Error of the Exposure and the Outcome Using Signed Directed Acyclic Graphs. Am. J. Epidemiol. 175, 1303–1310. https://doi.org/10.1093/aje/kwr458
VanderWeele, T.J., Tchetgen Tchetgen, E.J., 2017. Mediation analysis with time varying exposures and mediators. J. R. Stat. Soc. Ser. B Stat. Methodol. 79, 917–938. https://doi.org/10.1111/rssb.12194
Velentgas, P., Dreyer, N.A., Nourjah, P., Smith, S.R., Torchia, M.M., 2013. Developing a protocol for observational comparative effectiveness research: a user's guide. Government Printing Office.
Velupillai, S., Suominen, H., Liakata, M., Roberts, A., Shah, A.D., Morley, K., Osborn, D., Hayes, J., Stewart, R., Downs, J., Chapman, W., Dutta, R., 2018. Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances. J. Biomed. Inform. 88, 11–19. https://doi.org/10.1016/j.jbi.2018.10.005
von Schéele, B., Mauskopf, J., Brodtkorb, T.-H., Ainsworth, C., Berardo, C.G., Patel, A., 2014. Relationship between modeling technique and reported outcomes: case studies in models for the treatment of schizophrenia. Expert Rev. Pharmacoecon. Outcomes Res. 14, 235–257. https://doi.org/10.1586/14737167.2014.891443
Waffenschmidt, S., Knelangen, M., Sieben, W., Bühn, S., Pieper, D., 2019. Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. BMC Med. Res. Methodol. 19, 132. https://doi.org/10.1186/s12874-019-0782-0
Weidmann, B., Miratrix, L., 2020. Lurking Inferential Monsters? Quantifying Selection Bias in Evaluations of School Programs. J. Policy Anal. Manage. n/a. https://doi.org/10.1002/pam.22236
Weiskopf, N.G., Weng, C., 2013. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J. Am. Med. Inform. Assoc. 20, 144–151. https://doi.org/10.1136/amiajnl-2011-000681
Werbeloff, N., Osborn, D.P.J., Patel, R., Taylor, M., Stewart, R., Broadbent, M., Hayes, J.F., 2018. The Camden & Islington Research Database: Using electronic mental health records for research. PLOS ONE 13, e0190703. https://doi.org/10.1371/journal.pone.0190703
WHO, 1992. The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. World Health Organization.
Willan, A.R., Briggs, A.H., Hoch, J.S., 2004. Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. Health Econ. 13, 461–475. https://doi.org/10.1002/hec.843
Williams, P., Csipke, E., Rose, D., Koeser, L., McCrone, P., Tulloch, A.D., Salaminios, G., Wykes, T., Craig, T., 2014. Efficacy of a triage system to reduce length of hospital stay. Br. J. Psychiatry 204, 480–485. https://doi.org/10.1192/bjp.bp.113.141051
Wilson, S.T., Stanley, B., 2006. Ethical Concerns in Schizophrenia Research: Looking Back and Moving Forward. Schizophr. Bull. 32, 30–36. https://doi.org/10.1093/schbul/sbj023

Wing, J.K., Beevor, A.S., Curtis, R.H., Park, S.G.B., Hadden, J., Burns, A., 1998. Health of the Nation Outcome Scales (HoNOS): Research and development. Br. J. Psychiatry 172, 11–18. https://doi.org/10.1192/bjp.172.1.11

Wong, V.C., Steiner, P.M., 2018. Designs of Empirical Evaluations of Nonexperimental Methods in Field Settings. Eval. Rev. 42, 176–213. https://doi.org/10.1177/0193841X18778918

Wong, V.C., Steiner, P.M., Anglin, K.L., 2018. What Can Be Learned From Empirical Evaluations of Nonexperimental Methods? Eval. Rev. 42, 147–175. https://doi.org/10.1177/0193841X18776870

World Bank, 2019. World Bank Country and Lending Groups – World Bank Data Help Desk [WWW Document]. URL https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups (accessed 10.29.19).

Yeomans, D., 2014. Clustering in mental health payment by results: a critical summary for the clinician. Adv. Psychiatr. Treat. 20, 227–234. https://doi.org/10.1192/apt.bp.113.011320

Yi, G.Y., 2017. Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application. Springer.

Zhang, X., Faries, D.E., Li, H., Stamey, J.D., Imbens, G.W., 2018. Addressing unmeasured confounding in comparative observational research. Pharmacoepidemiol. Drug Saf. 27, 373–382. https://doi.org/10.1002/pds.4394

Zhou, J., Millier, A., Toumi, M., 2018. Systematic review of pharmacoeconomic models for schizophrenia. J. Mark. Access Health Policy 6, 1508272. https://doi.org/10.1080/20016689.2018.1508272

Appendices

Appendix A Data quality assessment

A.1 Approach to data quality assessment

To structure my data quality assessment, I used as widely cited framework created for this purpose by Weiskopf and Weng (2013). This framework is based on a review and synthesis of approaches to data quality assessment in the context of electronic health records. To provide some context, I will first explain the way(s) that each variable can be derived from the CRIS database before describing what is known about their quality in terms of following five dimensions of data quality that Weiskopf and Weng identified:

- Completeness: whether the truth about a patient is contained in the electronic health record
- Correctness: whether the element presented in the electronic health record is true
- Concordance: whether there is agreement between elements in the electronic health record or between the electronic health record and another data source
- Plausibility: whether the element in the electronic health record makes sense in light of other knowledge about what the element is measuring
- Currency: whether the element in the electronic health record is a relevant presentation of the patient at a given point in time

I refer to both existing evidence on data quality and selectively expanded the assessment of data quality. Specifically, I used one or more of following approaches: (i) I compare manually coded data, which I considered to be the 'gold standard', with data from structured fields, NLP outputs and/or keyword searches or referenced previous CRIS research that has done so. More specifically, I reference both published studies and a recent internal report by Irving (2019); (ii) I assessed whether the same variable agrees across CRIS and HES and assessing the stability of standardised ratings over time and between raters; (iii) I compared whether related elements agreed with each other; (iv) I assessed whether the distribution or summary statistics of aggregate data corresponded to the expect distribution of the

concept of interest; (v) When more robust methods are not practically feasible or not available, I make note of anecdotal evidence and subjective judgements.

A.2 Inclusion criteria (G) and the running variable (W)

Patient locality	
Data source description	CRIS allows access to the first half of the postcode of patients' residence over time which, in principle, makes it possible to determine whether a patient was living inside of the SLaM catchment area. I mad use of a version of CRIS address data that was 'cleaned' by Tulloch et al. (2012) and also indicates periods during which the patient was homeless. (Tulloch et al. (2012) suggest that approximately 15% of patients admitted to SLaM inpatient care were either homeless just prior to hospitalisation or discharged to homelessness.) In addition, I developed an alternative approach to identifying patient locality. Specifically, I explored whether the fact that for every health care contact in HES, the primary care trust (PCT) that the patient's general practitioner was registered at ('gpprpct') or the PCT that was responsible for that patient's care ('pctcode02' and 'pctcode06') could be used to infer periods in which patients were living in the SLaM catchment area care more reliably. My rationale was that, unlike the address data in CRIS, there are financial incentives to reliably record these variables better because the data that is used to create HES is also used for reimbursement. (With the 2013 NHS restructuring, Clinical Commissioning Groups (CCGs) now commission health care instead of PCTs but only historically 'frozen' PCT data was available in the version of HES linked to CRIS and there was a one to one mapping from CCGs to PCTs in the relevant areas). More specifically, for this alternative approach to determining a patient's locality I (i) created 'treatment spells', i.e. periods of time in which dates of health care contacts or inpatient admissions overlapped or took place on consecutive days; (ii) I

	determined whether 'gpprpct', 'pctcode02' or 'pctcode06' indicated that any of the contacts or admissions within these spells were associated with PCTs within the SLaM catchment area; (iii) based on this, I identified periods in which patients continuously access health services while living in the SLaM catchment area, ignoring one-off treatment spells in areas outside of the SLaM catchment area to avoid falsely inferring that someone moved away if they received specialist care elsewhere without having moved or care while temporarily located in other parts of the country (e.g. emergency care). Conditional on these exceptions, I assumed that a patient lived in the SLaM catchment area from the day of their first in-area contact recorded in HES until the day prior to a health care contact associated with an out-of-area PCT. If the last contacts recorded in HES was associated with one of the PCTs in the SLaM catchment area, I assumed that the patient remained under SLaM care until the end of data collection.
Completeness	Unclear. It was not straightforwardly possible to determine whether address data is missing or whether a patient lives out of area.
Correctness	As noted by Heslin et al. (2018) and my clinical collaborators, there are doubts about the validity of the address data in CRIS. This is because in routine practice it is only of importance to be able to contact the patient when they are currently under care but there is little incentive to record historical residence, keep address records up to date when the patient is discharged from SLaM care (e.g. to primary care), reconcile overlapping address spells or record the patient address when they are an inpatient. In addition, one would expect data on homelessness to be less reliable and it is not clear whether the patient is homeless and (predominantly) under the responsibility of SLaM or itinerant. Figure 19 and Figure 18 the differences between HES alternative approach and a version of CRIS address data that was 'cleaned' by Tulloch et al. (2012) for the evaluation of inpatient rehabilitation. Figure 18 shows that in about 85% of hospitalization the two sources of

address data agree on whether a patient resides in the SLaM catchment area or not but, on average, patients are more likely to be classified as residing in the SLaM catchment area according to CRIS address data. Figure 19 reflects that in most cases, the length of follow-up indicated by the two sources agrees in approximately 90% of cases but on average time to move out the catchment area is longer based on HES derived address data.

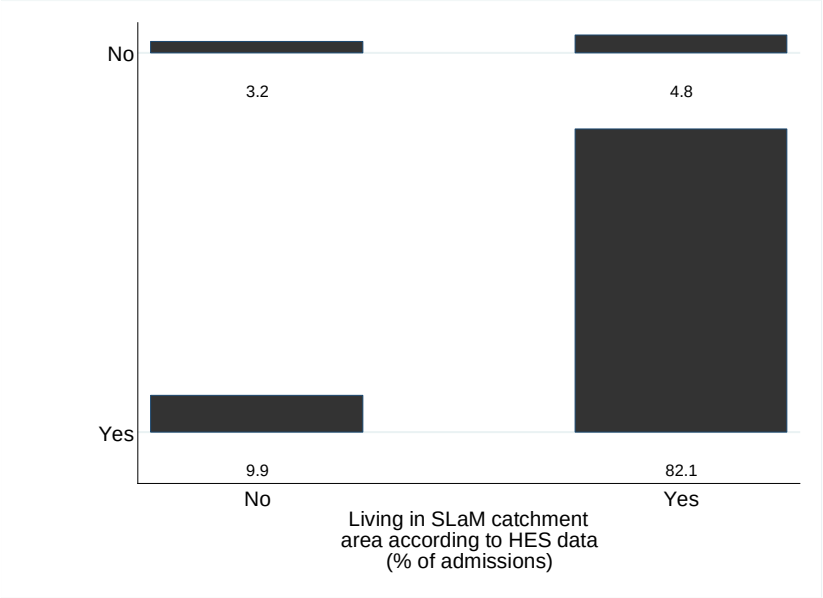


Figure 18 Agreement between CRIS and HES-derived address data in terms of residency at baseline (Analysis 2)

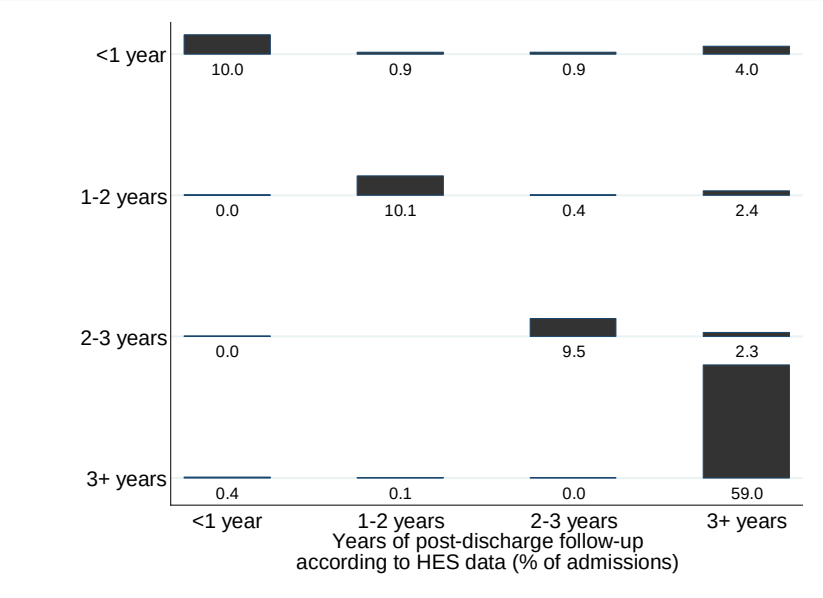


Figure 19 Agreement between CRIS and HES-derived address data in

	terms of length of follow-up (Analysis 2)
Concordance	Not assessed
Plausibility	Unclear
Currency	See above, significant concerns
Medical diagnoses	
Data source description	There are four potential sources of diagnostic data: (i) diagnoses recorded in 'structured fields' in HES, (ii) diagnoses recorded in 'structured' fields in CRIS, (iii) diagnoses extracted from medical notes using NLP approaches, (iv) reading through the medical notes (Perera et al., 2016). Unlike (i) and (ii), (iii) does not allow us to differentiate between primary or secondary diagnoses
Completeness	Unclear. It was not straightforwardly possible to distinguish between missing diagnoses and whether a patient did not have diagnosis of psychosis
Correctness	Davies et al. (2018) suggests that the positive predictive value and sensitivity of HES data with respect to lifetime diagnosis of psychotic spectrum disorders can be up to 90%. For any given diagnosis, there is also evidence that the positive predictive value is high in administrative databases but that agreement with a reference standard is relatively low in administrative databases (Davis et al., 2016). In other words, when a diagnosis of psychosis is made, it is likely that, at that point in time, the diagnosis is correct, but not that all people with such diagnosis are identified as having one. Based on a sample of 51 documents sampled from the whole of CRIS, Irving (2019) estimated that positive predictive value of the NLP application for psychosis was 95% and the sensitivity was 43% at the document level. Downs et al. (2018) found that combining approach (ii) and (iii) and requiring a mention of an antipsychotic, yields a positive predictive value of 0.98 for a lifetime diagnosis of a psychotic spectrum disorder at the patient level in people between 10 to 17 years old. They

	do not report the sensitivity of this approach.
Concordance	Not assessed
Plausibility	Unclear
Currency	See above, significant concerns
Patient age	
Data source description	The month and year of birth are available in CRIS. In addition, age of the patient rounded to years is recorded in HES, but it is a deterministic matching variable for the HES-CRIS linkage so, by construction, it is identical across the databases.
Completeness	Available for more than 99% of patients
Correctness	When searching the clinical notes for the phrase “year old” or “yr old” preceeded by a number between 15 and 99, I found that in 93% of all instances the age stated in the clinical notes agreed with the age recorded in stuctured fields within ± 2 years and in only 3 out of 20 random instances of the remaining 7% that I inspected did the statement refer to the patient rather than a third person.
Concordance	Not assessed
Plausibility	The age distribution appears broadly plausible given my collaborators clinical knowledge of the patients accessing SLaM
Currency	Not applicable
Compulsory admission status	
Data source description	Compulsory admission, that is sectioning status is routinely recorded in CRIS. As with address data, I made use of a version of the sectioning data that was ‘cleaned’ by Tulloch et al. (2016b).
Completeness	Unclear. It was not straightforwardly possible to determine whether sectioning status data is missing or whether someone was not detained under the Mental Health Act.
Correctness	Given legal requirements under the Mental Health Act, there is reason to believe that this information is relatively well, but I did not undertake any manual validation
Concordance	Not assessed

Plausibility	Based on my clinical collaborators' experience, the data had some face validity
Currency	Unclear

A.3 Treatment variables (A)

Psychiatric inpatient admission data for young people	
Data source description	As discussed further in the next section, non-SLaM inpatient service use is not comprehensively recorded in CRIS and there were some errors in the inpatient service use data.
Completeness	It is difficult to distinguish between missing data and no admission to inpatient care.
Correctness	However, among those who were admitted to inpatient care, only 50% had contact with such services in the week prior to admission suggesting that this approach would underestimate this group; (ii) If they were referred for inpatient care but this referral was rejected. However, given the reality of clinical practice and the referral data for inpatient rehabilitation, there is reason to believe that such cases are under-recorded in CRIS.
Concordance	Not assessed
Plausibility	Not assessed
Currency	Not assessed
Pathway to and history of inpatient rehabilitation	
Data source description	There are structured fields that contain data on referral dates, whether and when a referral was accepted, and whether and when a patient was transferred to inpatient rehabilitation. This data can be supplemented by examination of clinical notes. HES does not contain any referral data with respect transfers between different types of inpatient care.
Completeness	Some SLaM rehabilitation wards appear to routinely record instances in which a referral is rejected whereas others only appeared to create a record when patients are ultimately transferred to inpatient rehabilitation and dates are frequently missing.
Correctness	The majority of entries in the structured field for the date of referral as well as the date that the referral was accepted or rejected appeared unreliable upon inspection. A structured field called 'planned start date', for example, appeared to sometimes report the referral date, sometimes

	the referral acceptance/rejection date and have no discernible meaning in other cases. The same applied to the date on which the record for the patient episode was created in PJS. Another complication is that at times patients are referred to multiple rehabilitation inpatient wards around the same time or a referral is rejected from one rehabilitation ward only to be redirected to another rehabilitation ward. There was some data on non-SLaM inpatient rehabilitation placements in CRIS. Based on a subset of the out of area placement data that we obtained from SLaM and linked to CRIS, it appeared that this data was reasonably reliable and complete from 2011 onwards. Only the date of admission not the date of referral is recorded in structured fields for non-SLaM inpatient stays
Concordance	Not assessed
Plausibility	The lengths of admission to inpatient rehabilitation were broadly in line with national averages
Currency	See above

A.4 Outcome variables (Y)

Mental health care service use and costs	
Data source description	Clinical activity is routinely recorded in CRIS as part of service delivery. For the analyses in this thesis, I again used a version of the service use data that was to some extent 'cleaned' for previous studies (Tulloch et al., 2016b; Tulloch, 2010). Due to the complexity of this 'cleaning' process, I am not familiar with all its details. I am aware that another working group at the BRC has developed a different approach to 'cleaning' the inpatient data but I do not know which of the two approaches performs better. It appears that the service use (and cost) data is of better quality from 2010 onwards because SLaM services underwent a process of restructuring and consolidation in this year. In addition, we became aware that data for use of some, usually more peripheral, SLaM services was not comprehensively accessible in CRIS because parallel recording systems exist that have not been fully incorporate into CRIS yet. These include psychotherapy, substance use, and some forensic services but we were particularly concerned with private sector bed use and out-of-area placement data given their cost and because the latter included out-of-area inpatient rehabilitation services. Data on private, so-called 'overspill', inpatient service use was only systematically recorded in CRIS between 2013 and 2016 but my collaborators obtained a dataset from SLaM, with which we were able add data for 2011 and 2012 and verify that the already available data that was accurate. Likewise, a dataset on out of area placements that my collaborators obtained from SLaM indicated that the quality of placement data in CRIS was likely to be comprehensive and, where available, of high quality but we were able to add some missing episode end dates and increase the number of placements by 4%. HES also contains some SLaM service use data but, among those who have been linked to HES, the number of inpatient days is 17% lower in HES. No SLaM outpatient data is available in HES, other mental health trusts

	stopped submitting such data in 2010 and the concept itself does not map well to mental health. Thus, data on mental health service use is only complete when the patient is living in the SLaM catchment area. It is possible to improve the quality of the service use data by examining clinical records using data triangulated from different sources in CRIS to guide this process.
Completeness	It is not straightforward to distinguish whether service use data is missing or a service has not been used.
Correctness	Given my experience of doing economic evaluations in mental health, overall, my sense is that the quality of this service use data is probably significantly higher than what could be achieved through self-report in a prospective study. However, as would be expected in a data source of this kind, it also contains errors. Team and ward names change over time, sometimes data is changed retrospectively, the date the activity as recorded does not always correspond to the date that the activity took place, inpatient spells can overlap, service use is sometimes attributed to the wrong patient and, more generally, both over and under-recording of service use can occur. I
Concordance	Not assessed
Plausibility	Not assessed
Currency	Not assessed
HoNOS	
Data source description	HoNOS ratings both are recorded in structured fields in CRIS. This includes different versions of HoNOS such as HoNOS for adults or the child and adolescent version of HoNOS, known as HoNOS-CA.
Completeness	HoNOS ratings should be completed at admission to inpatient care, at discharge, if a mental health cluster expires, every 28 days if the patient held as an inpatient under section 2 of the mental health act and when prompted for the so-called care program approach, i.e. every 3 months. There is also a child and adolescent version of HoNOS, known as HoNOS-CA. Completion is audited by SLaM and it is known to be one of the

	mental health trusts in the UK in which recording of HoNOS scores alongside clinical assessment is most widespread. Nonetheless, large amounts of HoNOS data in CRIS are missing and, in practice, the timing of ratings is much more variable (see, for example, Figure 50 for a distribution of the time between referral to inpatient rehabilitation and the closest preceding HoNOS rating). CRIS contains very few HoNOS-CA ratings.
Correctness	There is some anecdotal evidence that HoNOS ratings may not be used as intended in routine clinical practice. For example, item 3 is defined as reflecting actual drug use around the time of rating. This should rarely happen on a psychiatric ward, so it is often scored to reflect latent problems with drug use. Similarly, item 11 is scored to reflect problems with the patient's accommodation outside the ward. In addition, I found some evidence that HoNOS scores in relevant dimensions are significantly poorer just prior to admission to inpatient care than just after admission (see Figure 20). This may partly reflect habituation of staff in different clinical settings to a certain level of symptomatology or community teams may deliberately inflate ratings to bring about an inpatient admission. There was not enough data investigate whether there may be similar differences in ratings between consultants on acute psychiatric wards and rehabilitation wards. Figure 21 shows the stability of HoNOS ratings over time.

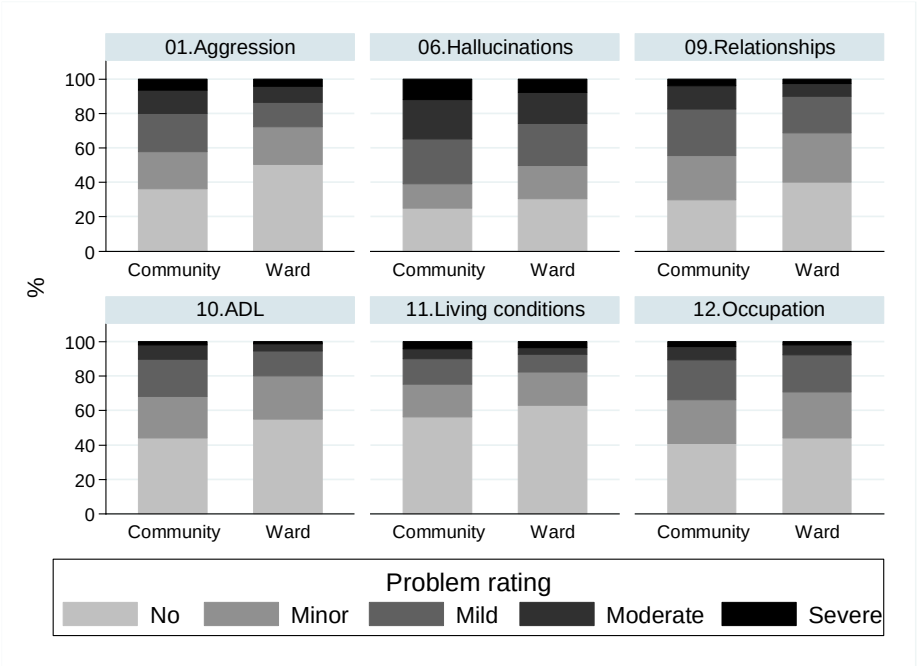


Figure 20 Distribution of HoNOS scores when ratings have been taken for the same patient prior to hospitalisation and after hospitalisation and within less than 4 days.

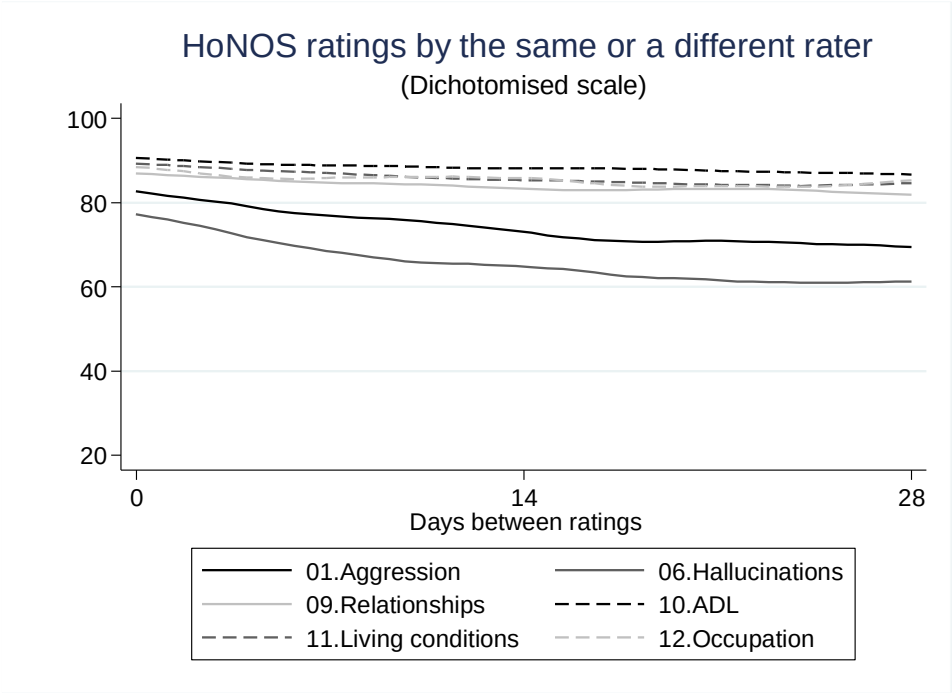
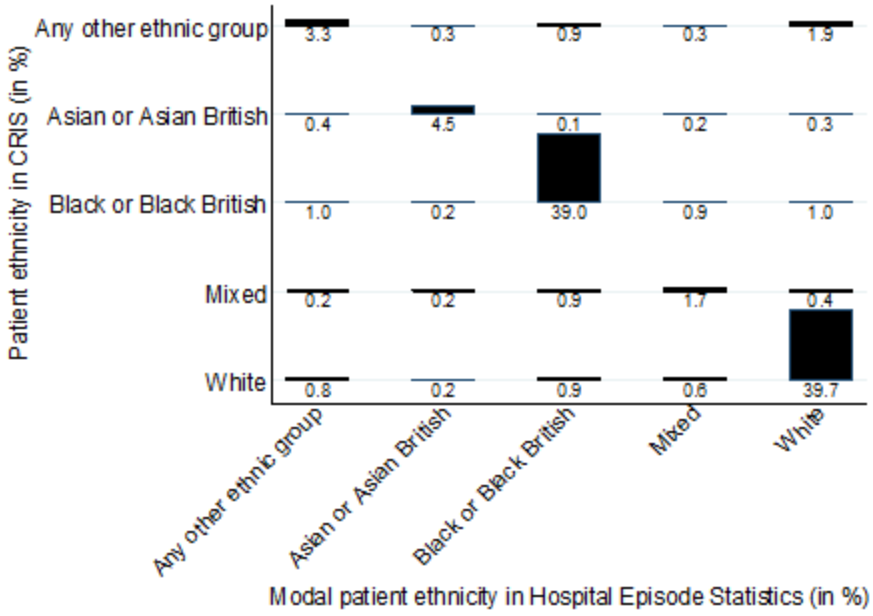


Figure 21 Locally smoothed curves of HoNOS intra- or inter-rater agreement by dimension over time with ratings dichotomised between no/minor/mild problems and moderate/sever problems (Inpatient ratings only)

Concordance	Not assessed
Plausibility	Appeared broadly plausible based on my clinical collaborators' views
Currency	Not assessed
Mortality data	
Data source description	Mortality data from the Office of National Statistics (ONS) are available in CRIS. See Chang et al. (2010) for further details regarding mortality data in CRIS.
Completeness	It is not straightforwardly possible to distinguish between whether mortality data is missing or a patient who has not died yet.
Correctness	This data has some known limitations. Overall, however, the quality of ONS mortality data appears to be relatively high and it is routinely used in medical research (Office of National Statistics, 2017; Stewart et al., 2009).
Concordance	Unclear
Plausibility	Broadly speaking
Currency	There can be a lag between a person's death and the recording of their date of death in official records, and in a recent update of the linkage between ONS and CRIS some people who were previously thought to have died were in fact still alive

A.5 Variables only used as confounders (X)

Gender	
Data source description	Reported in a structured field in CRIS
Completeness	This variable was available for all patients
Correctness	The relative proportion of CRIS documents that contained a female pronoun to those containing a male pronoun, agreed with what one would expect given this structured field suggesting that this variable was well-recorded
Concordance	Not assessed
Plausibility	Appeared broadly plausible based on my clinical collaborators' views
Currency	Not applicable
Patient ethnicity	
Data source description	Ethnicity data could either be extracted from structured fields in CRIS or in HES
Completeness	There was no missing data
Correctness	Not assessed
Concordance	As shown in Figure *22*, HES ethnicity agreed with CRIS ethnicity about 90% of patients. As one might expect, the proportion of patients who were classified differently was higher if one of the data sources suggested that the ethnic background of a patient was non-white. This degree of measurement error in ethnicity data is similar to that in other sources of routinely collected data (Mathur et al., 2014; Saunders et al., 2013).

	 Figure 22 Agreement between CRIS and HES in terms of ethnicity
Plausibility	The distribution of ethnicity appeared to broadly correspond with my clinical collaborators' expectations.
Currency	Not applicable
Symptom indicators	
Data source description	As described above, there are again three ways to extract symptom data: (i) simple keyword searches (e.g. for "self-neglect"), (ii) natural language processing approaches which reduce the proportion of false positives compared to (i). Most relevantly, Jackson et al. (2017) have developed a machine-learning based natural language processing approach to identify some symptoms of severe thus reducing the measurement error to low levels.
Completeness	It is not straightforward to distinguish between the absence of symptoms or underrecording of symptoms
Correctness	In all cases what is reported in clinical notes can differ from what the clinicians observe (and act on) in clinical practice because symptoms are unlikely to be recorded regularly or according to a consistent standard and are more likely to be recorded the more contacts a patient has with

	mental health services. Unlike HoNOS, one cannot necessarily determine whether a symptom was present or absent over a specified period or its level of severity. Based on a sample of 30 randomly annotated instances, Irving (2019) finds that the positive predictive value of the NLP application for social withdrawal is 90% and 50% for the poor motivation application. The sensitivity of the application is unclear.
Concordance	Not assessed
Plausibility	Not assessed
Currency	See above, of concern even in the
Clozapine use	
Data source description	There are five potential ways to obtain data on clozapine use: (i) pharmacy data in CRIS, (ii) keyword search, (iIi) NLP approaches which extract information on medication use from clinical notes (see Hayes et al. (2015) and Kadra et al. (2015) for a more detailed discussion of this approach), (iv) data on clozapine use in ZTAS, (v) reading through medical notes and manually coding data on medication use. Typically, these approaches allow the analyst to obtain data on instances of antipsychotic use, discussion or prescription rather than spells of medication use. Therefore, this data needs to be further processed for use in evaluation (e.g. to obtain, the first date a drug was administered). No data on prescribing in primary care is available.
Completeness	It is not straightforwardly possible to differentiate between missingness and measurement error in antipsychotic use, however, medication use prior to 2007 is fragmentary.
Correctness	Based on a sample of 200 random documents, Irving (2019) suggests that at the patient level, the sensitivity of all sources of data combined in identifying people who have ever had a history of clozapine use is 92% and the positive predictive value is 100%. This evidence is, however, not directly related to the variable of interest, that is whether at the patient has had a history of clozapine use.
Concordance	Not assessed

Plausibility	Not assessed
Currency	See above
Risk events	
Data source description	Risk events can be recorded in up to three different locations in SLaM clinical practice.
Completeness	It is not straightforward to distinguish between missing data and mismeasured data. However, anecdotal evidence suggests that among the three potential recording locations, the most commonly used is not linked with CRIS
Correctness	Not assessed
Concordance	Not assessed
Plausibility	Not assessed
Currency	Unclear

Appendix B Costing of secondary mental health care use

B.1 Rationale for costing approach

There are two sources which are commonly used by health economists in the UK to cost health care service use: NHS reference costs which are used for activity-based payments in England and Wales, so-called Payment by Results (PbR) or the compendium of costs published by the Personal Services Research Unit (PSSRU) (Curtis and Burns, 2018; NHS improvement, 2018). However, about half of the SLaM budget are not covered by the proposed mental health PbR system, the PbR system has been regarded as simplistic and evidence suggests that it does not perform well (Jacobs et al., 2018; Twomey et al., 2016; Yeomans, 2014). Similarly, the PSSRU compendium contains very limited data on the cost of mental health professionals and services. However, we were able to obtain financial data from SLaM's internal accounting system which Dr Alexander Tulloch and I used to calculate a service and year-specific unit cost by service using a top-down approach. Some authors have recommended a bottom-up costing approach over a top-down approach because it provides unit costs depending on the patients' profiles and we are aware that there are plans of introducing a so-called patient-level costing system (PLICS) in SLaM (Siomi and Razzouk, 2017). However, for such an approach to be operationalised, more time and resources are likely necessary. A top-down approach also has the advantage that, that, to some extent, it allows for allowing for biases in recording.

B.2 Implementation of costing approach

We calculated unit costs per ward or team or group of services for the financial year 2008/09 to 2016/17, using a currency of cost per single contact for community locations and cost per day for inpatient locations. This required creating an iterative algorithm than combined all budgets and activity data of services that were financed by multiple budgets or budgets that financed multiple services. In addition, we undertook several rounds of error checking excluding financial and activity data for services for which data quality was not adequate. For example, we excluded budgets if the activity provided by these services were unlikely to be adequately recorded in CRIS (e.g. peer-led groups) or because of activity

recorded in CRIS was funded by sources outside of the SLaM budget (e.g. some forensic services) because they were partially funded by non-NHS resources and contacts with these were more commonly recorded on a separate system. Figure 23 for a schematic overview of the process). Figure 24, Figure 25 and Figure 26 provide a breakdown of the percentage of the SLaM budget captured in the analyses and an overview of the service use and budget data excluded because of reasons of data quality. I inflated costs to 2018/19 levels using the hospital and community health services (HCHS) index (Curtis and Burns, 2018). Due to commercial confidentiality I cannot report any unit costs or any cost data that could be used to infer unit costs.

.

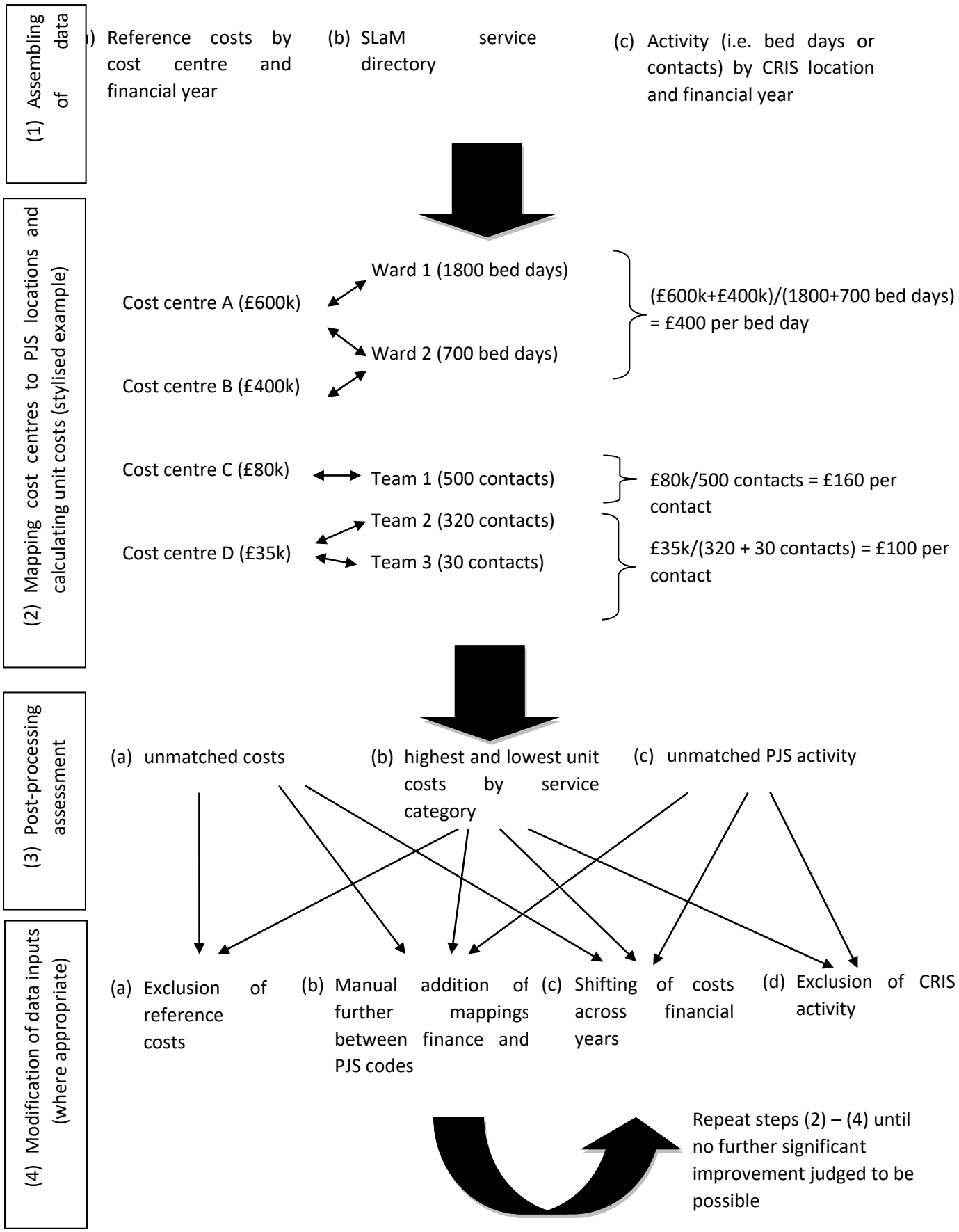


Figure 23 Schematic overview of top-down approach to construct unit costs of secondary mental health care service use

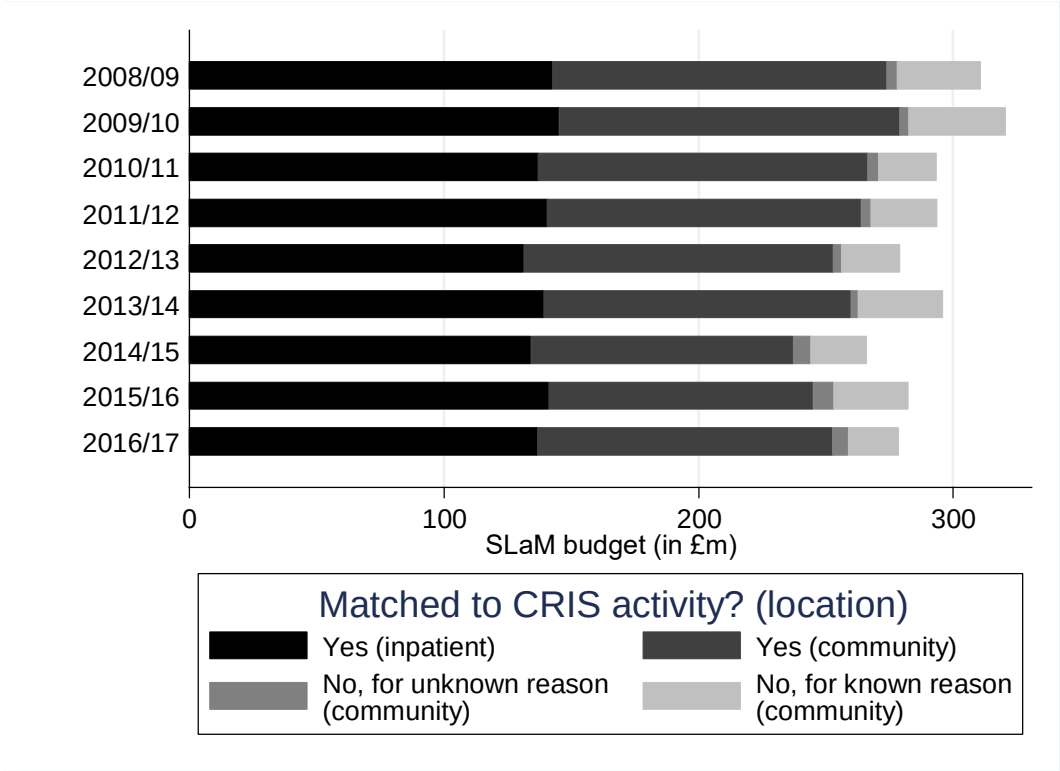


Figure 24 Secondary mental health care budget costing gaps in terms of budget

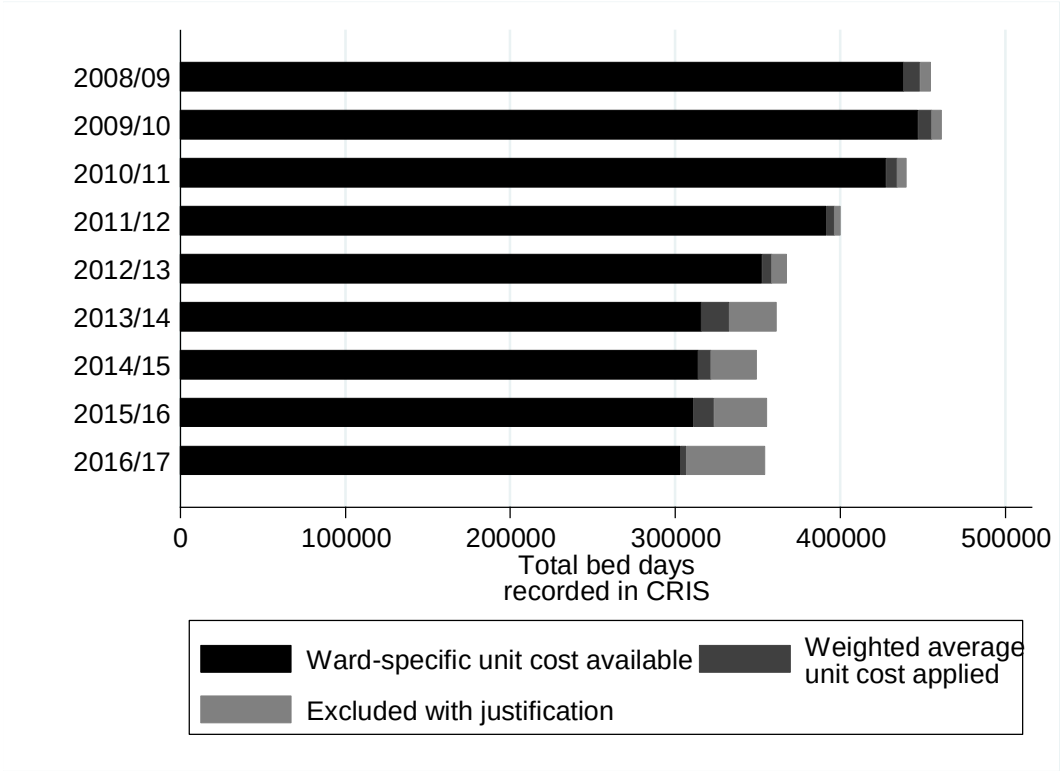


Figure 25 Secondary mental health care budget costing gaps in terms of inpatient bed days

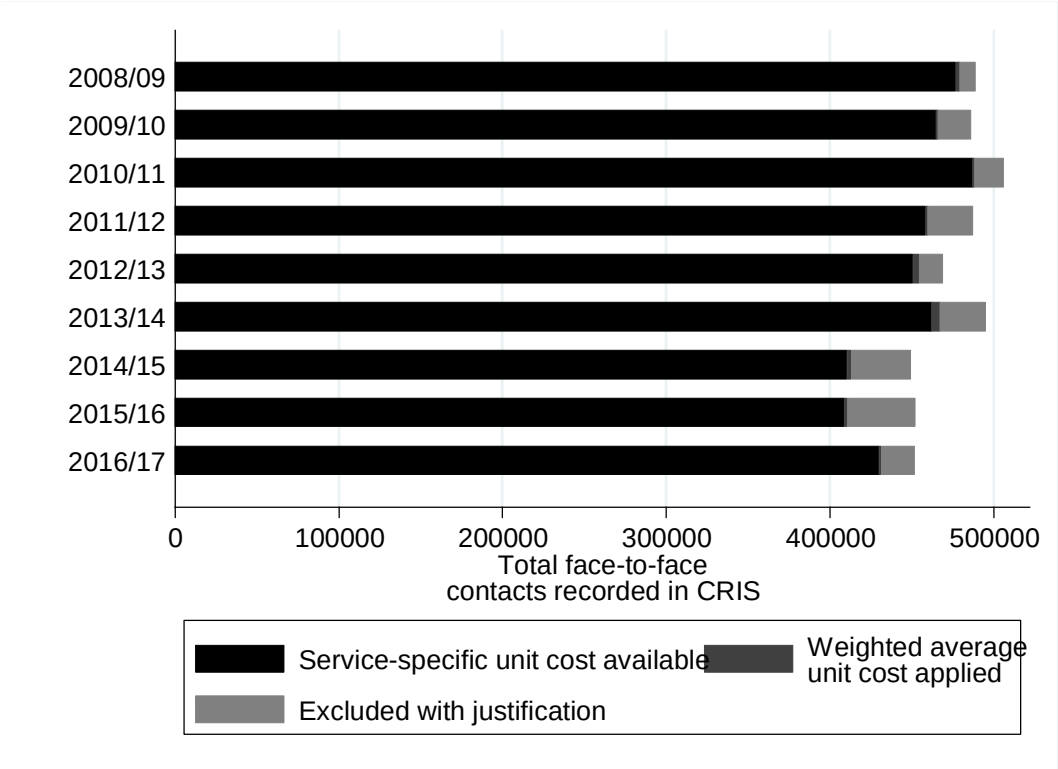


Figure 26 Secondary mental health care budget costing gaps in terms of community contacts

Appendix C Methodological details to Analysis 1

C.1 Notation

During an acute episode of psychosis, broadly speaking, in England young people can receive one of five treatments: (i) He/she may be admitted to a child and adolescent mental health (CAMHS) ward or continue to receive care on a paediatric ward, which I define as $A = 0$; (ii) be admitted to an adult psychiatric ward ($A = 1$); (iii) a clinical judgement may be made that the situation can be managed with the patient staying at home through intense care by a CAMHS community team ($A = 2$); (iv) the patient may receive care at home from an adult community team ($A = 3$); (v) the patient may not receive any care if he/she does not come in contact with health care services or that the psychotic crisis is not recognised as such ($A = 4$). I denote age at admission as W, which is also referred to as the running variable in the RDD literature, and the threshold variable as $Z = \mathbb{I}(W \geq 18)$ where $\mathbb{I}(\,)$ is an indicator function taking the value 1 if the condition is true and 0 otherwise. I let the set of (partially) measured covariates confounders other than age be represented by L, unmeasured confounders by U and the outcomes by Y. For convenience, I let L be a vector $L = (G, X)$ where G is a vector of study eligibility criteria, including, for example, an indicator for whether the patient has a primary working diagnosis of psychosis, and X is a vector of other confounders such as number of previous psychiatric inpatient admission, which would be expected to vary between patients. I denote the potential outcome under treatment strategy D as Y^D and C is a censoring indicator equal to 1 if the patient. As in Chapter 2, I assume that there are three versions of each variable: the true value, indicated by the absence of superscripts (e.g. Y), the version of the variable that can be easily extracted from structured fields, through natural language processing applications or keyword search indicated by the superscript asterisk (e.g. Y^*), and the version of the variable that can be obtained by reading the clinical notes (e.g. Y^+).

C.2 Causal model and observed data

Figure 27 shows the diagram assumed to underly the analysis. As suggested, a patient's age, W, known as the running variable, is believed to strongly determine the type of care the

patient received. Those who are 18 years or older (and younger than 65), should be referred to adult inpatient care whereas CAMHS wards are designated for those younger than 18. In other words, this effect is largely mediated via the instrumental variable Z. I assume that L, W and Z are defined as being measured before A which, in turn, is measured before C which is measured before Y.

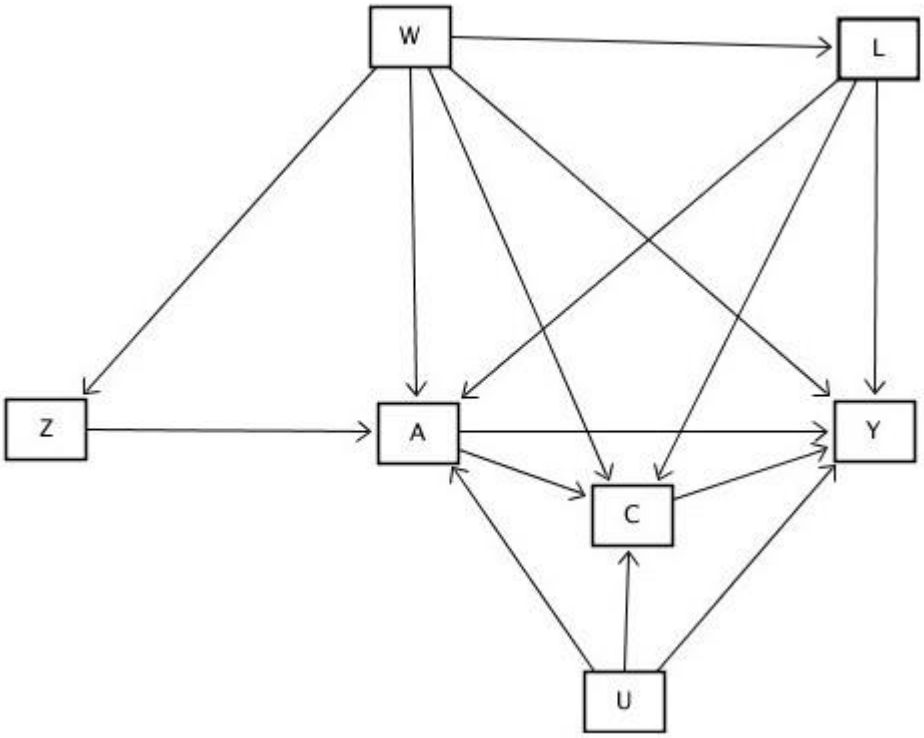


Figure 27 Causal diagram for Analysis 1

C.3 Target causal parameter

Given the notation defined in Section C.1, our target parameters can be formally defined as,

$$\psi_{ATU} = \mathbb{E}(Y^{A=1,D=0} - Y^{A=0,D=0} | A < 2{,}17 \leq W < 18)$$

and

$$\psi_{ATE} = \mathbb{E}(Y^{A=1,D=0} - Y^{A=0,D=0} | A < 2{,}18 \leq W < 25)$$

C.4 Identifying assumptions

Confounding: The key assumption in this analysis was that outcomes were independent of whether patients were admitted to a child and adolescent ward conditional on the fact that

they were 17 or 18 at admission, the decision was made that they were admitted to inpatient care and conditional on other measured confounders. More formally,

$$(Y^1, Y^0) \perp A | (L, 17 \leq W < 19, A < 2)$$

This assumption is stronger than the most common RDD assumption in two respects: First, it is based on what is known as the local randomization approach (Cattaneo et al., 2018). This approach assumes that interventions are randomized within a window around the cut-off threshold, here for people aged 17 or 18 at admission, rather than that the regression function is continuous around the cut-off, known as the continuity-based approach to RDD. I invoked the local randomization rather than the continuity-assumption for three reasons: (1) based on subject-matter knowledge, we had reason to believe that developmentally there would not be substantial difference between those hospitalized when they were 17 compared to those hospitalized when they were 18, that is there was some reason to support this assumption; (2) the local randomization assumption allows the use of statistical approaches that yield more precise parameter estimate which was important given the relatively small sample size of this analysis; (3) although the running variable, that is age, is measured on a continuous scale in CRIS, one way to scale up this analysis to the whole of the UK, would be to use HES data which, typically, only contains information on patients' age of admission in years. In other words, the running variable would be continuous such that a local randomization assumption would be the only possible option.

Homogeneity/bounds of treatment effects: In order to be able to generalize the estimates obtained using a population of patients who were 17 or 18 at admission to the target populations of interest, for each of the target parameters defined in the previous section (Appendix C.3), assumptions are needed. With respect to ψ_{ATT}, I assumed that

$$\mathbb{E}(Y^1 - Y^0 | 17 \leq W < 19, A < 2) = \mathbb{E}(Y^1 - Y^0 | 17 \leq W < 18, A < 2)$$

meaning that the treatment effect is homogenous across 17- and 18-year olds. A reasonable alternative could be to assume that whatever benefit patients who are 17 and are admitted to an adult ward derive from being admitted to a CAMHS ward instead, is at least as high as the benefit that 18-year olds derive. This assumption would yield bounds in the target causal parameter.

Similarly, with respect to ψ_{ATE}, I assumed that

$$\mathbb{E}(Y^1 - Y^0 | 17 \leq W < 19, A < 2) = \mathbb{E}(Y^1 - Y^0 | 18 \leq W < 25, A < 2)$$

meaning that the treatment effect among 17- to 18- year olds is the same as among 18- to 25-year olds. Again, if one believes that, the older patients are at admission, the less they are likely to benefit

from CAMHS inpatient care, then the parameter estimates represent upper (or lower) bounds, depending on the type of outcome measure.

Positivity: While, by construction, there is not necessarily any overlap in terms of the running variable, W, in an RDD design, I did assume that there was overlap in the distribution of other measured confounders, that is

$$0 < \mathbb{P}(A = 1|L, A < 2) < 1$$

Sampling: I assumed that the sample was obtained through simple random sampling, that is

$$\mathbb{P}(Y = y, A = a, L = l, W = w) = \frac{1}{N}$$

where N is the sample size and lowercase letters are the realization of the uppercase random variables.

Incomplete variables: I assumed that movement out of the catchment area was independent of the outcome conditional on measured confounders and the treatment

$$Y \perp C|(L, 17 \leq W < 19, A < 2)$$

I assess the sensitivity of the results with respect to this assumption by assuming that if follow-up was censored that the outcome did not occur after censoring (e.g. there was no more service use after censoring).

Model misspecification: I assumed that the estimation models were correctly specified. I used a coarsened exact matching approach in a sensitivity analysis to assess the robustness of the findings with respect to model specification.

Interference: I assumed that there was no interference between hospitalizations. In other words, whether one admission was to a child and adolescent ward or to an adult ward did not affect outcomes in other hospitalizations.

Measurement error: I assumed that there was no measurement error in the measured confounders as extracted from the structured field, that is $L = L^*$ and $W = W^*$. By implication, I assumed that the instrumental variable was measured without error, that is $Z = Z^*$. As discussed in Chapter 2, I considered the assumption that readily available diagnosis variable were measurement without error to be too strong, so I coded all

diagnosis assuming that $G = G^{+}$. With respect to outcomes I assumed that $\mathbb{E}(Y) = \mathbb{E}(Y^{*})$ for continuous outcome measures and $Y = Y^{*}$ for binary outcome measures.

Appendix D Supplementary figures to Analysis 1

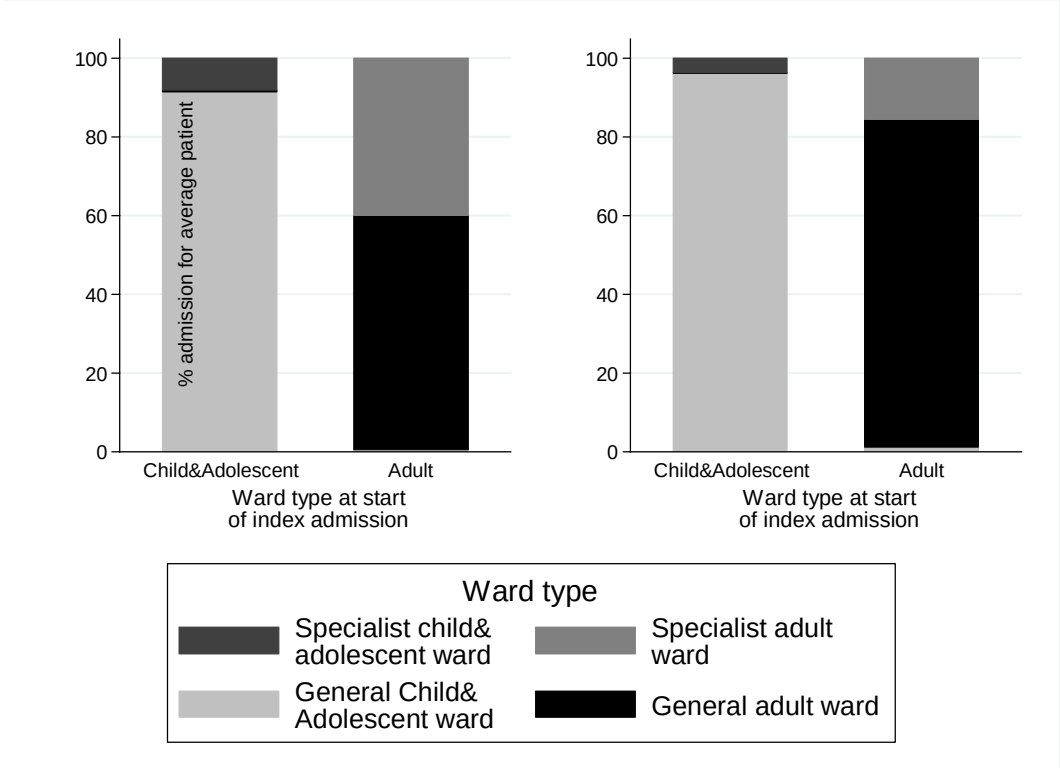


Figure 28 Fraction of inpatient stays across ward subtypes by ward type at the start of the index admission

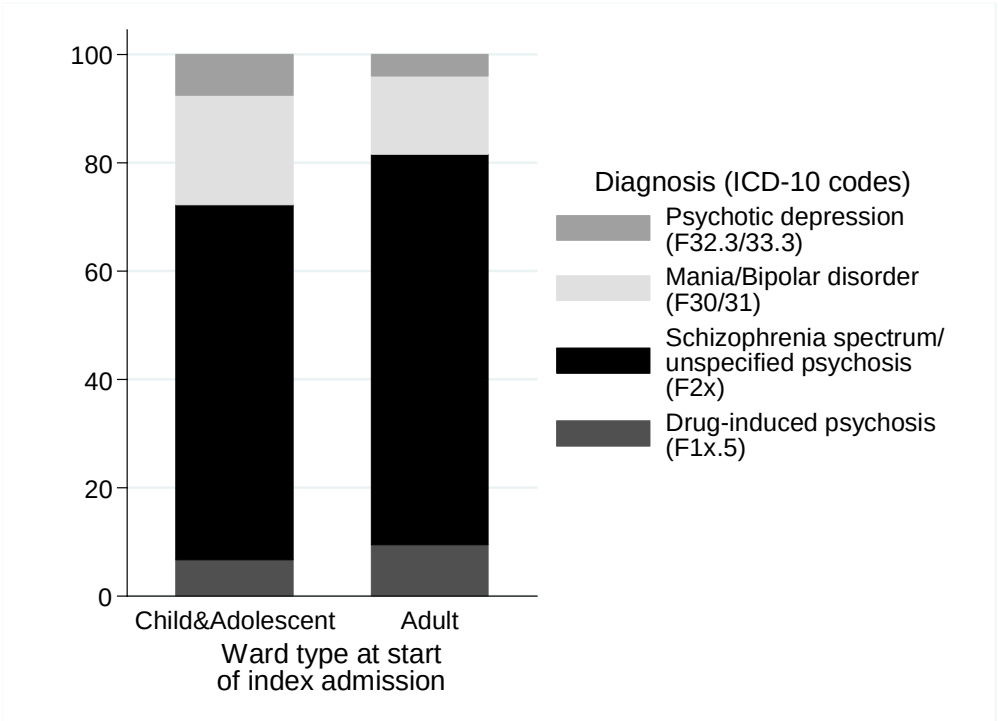


Figure 29 Distribution of working diagnoses at admission by ward type at the start of the index admission

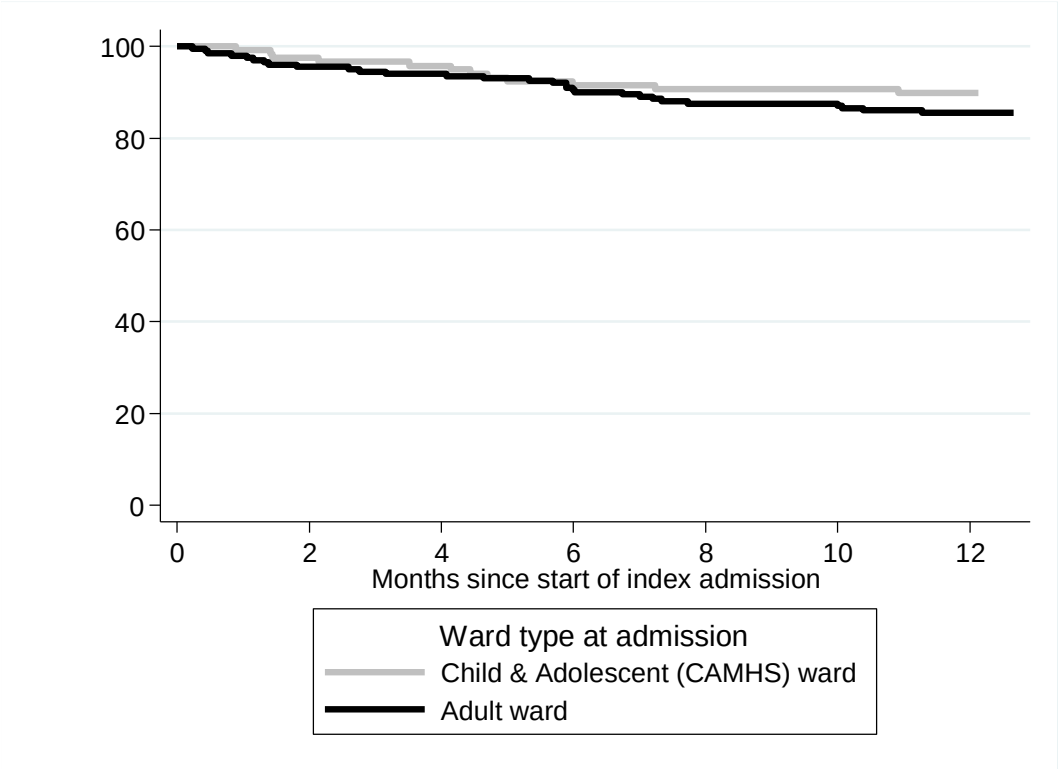


Figure 30 Distribution of time to censoring by ward type at the start of the index admission

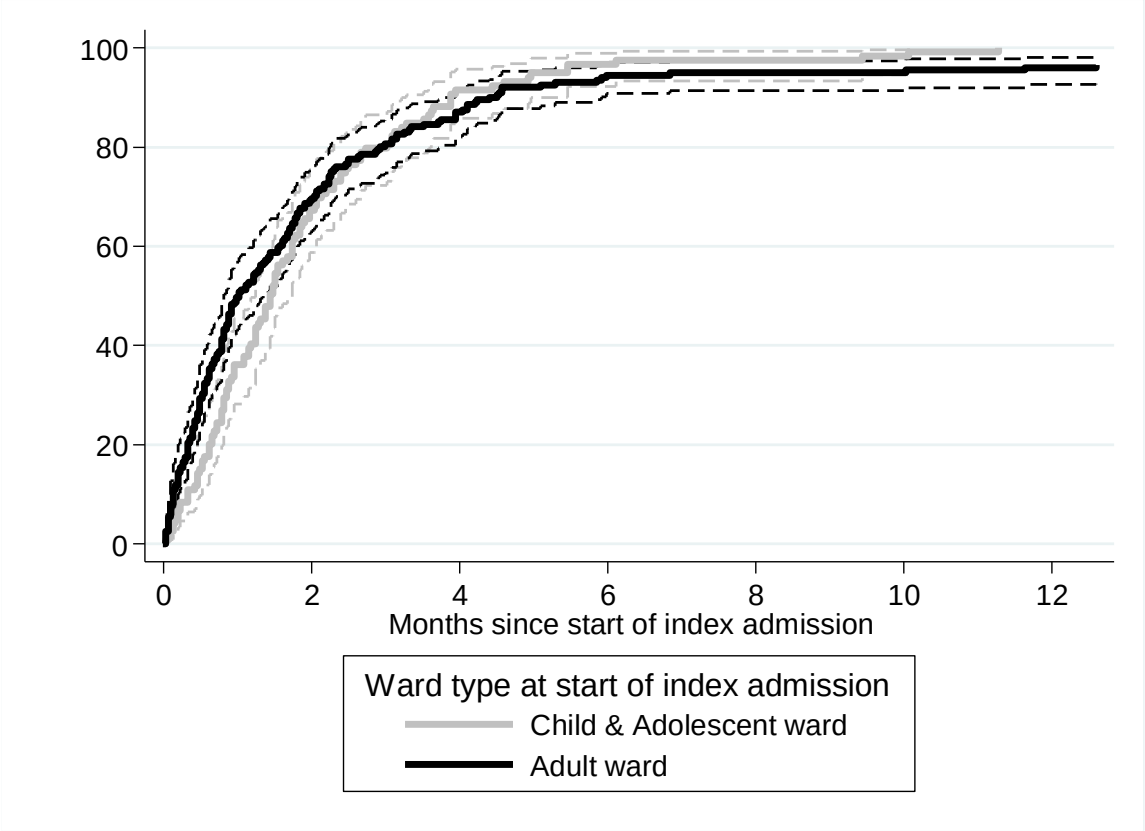


Figure 31 Unadjusted rates of discharge from the index admission

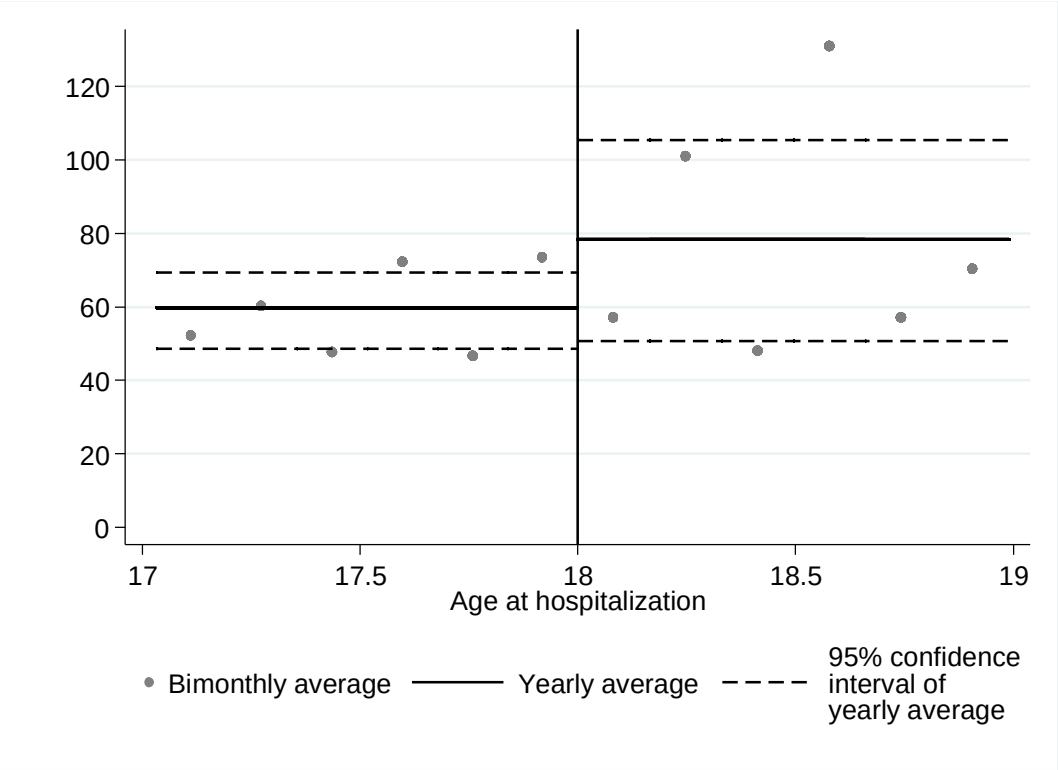


Figure 32 Length of the index hospitalization by age at the start of the index hospitalization

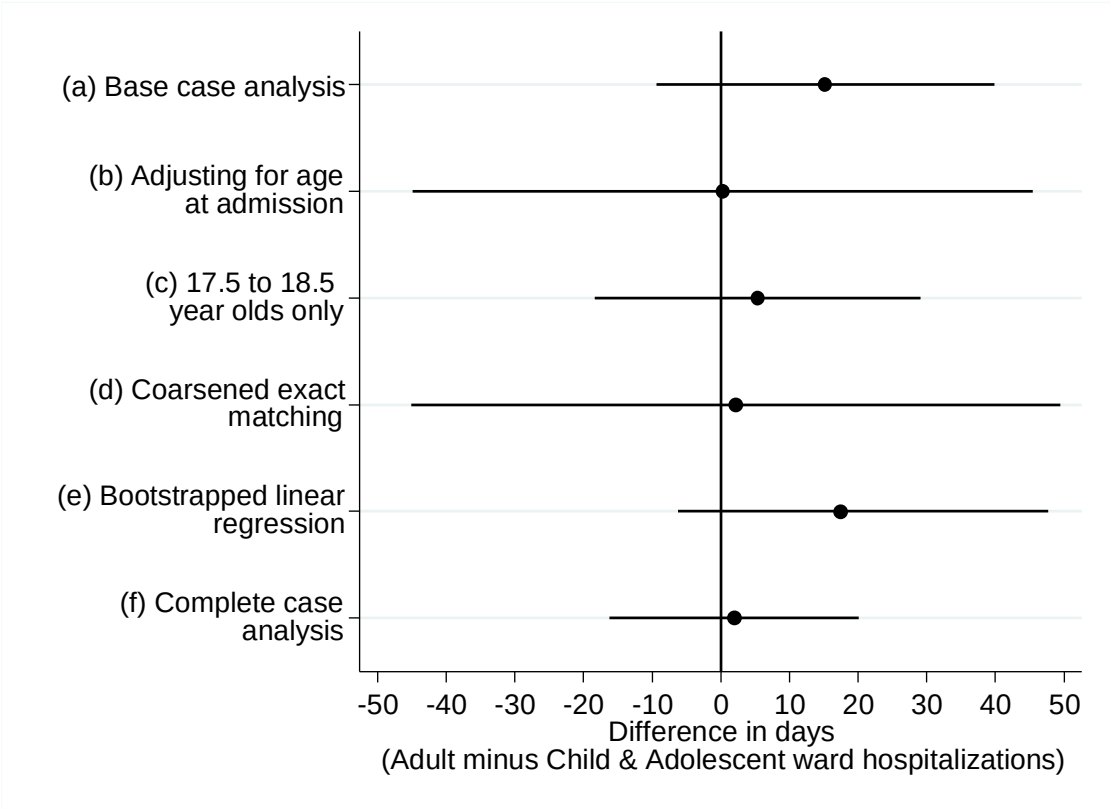


Figure 33 Difference in the length of the index hospitalization

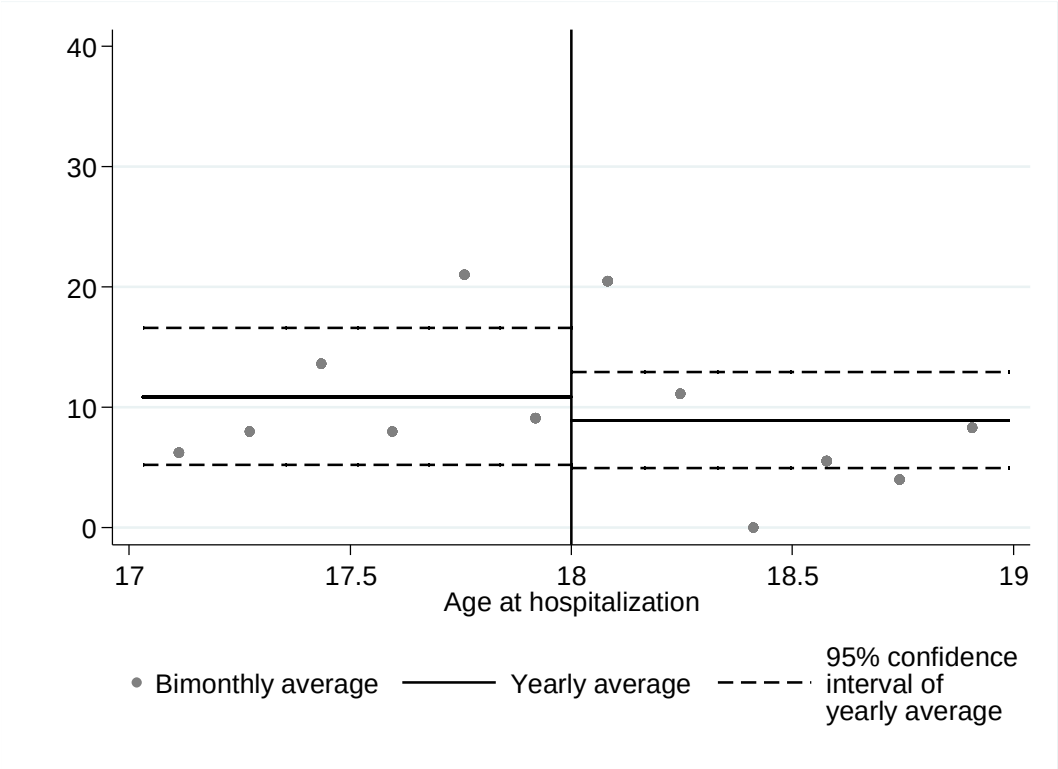


Figure 34 Probability of being detained under the Mental Health Act during the index hospitalization by age at the start of hospitalization

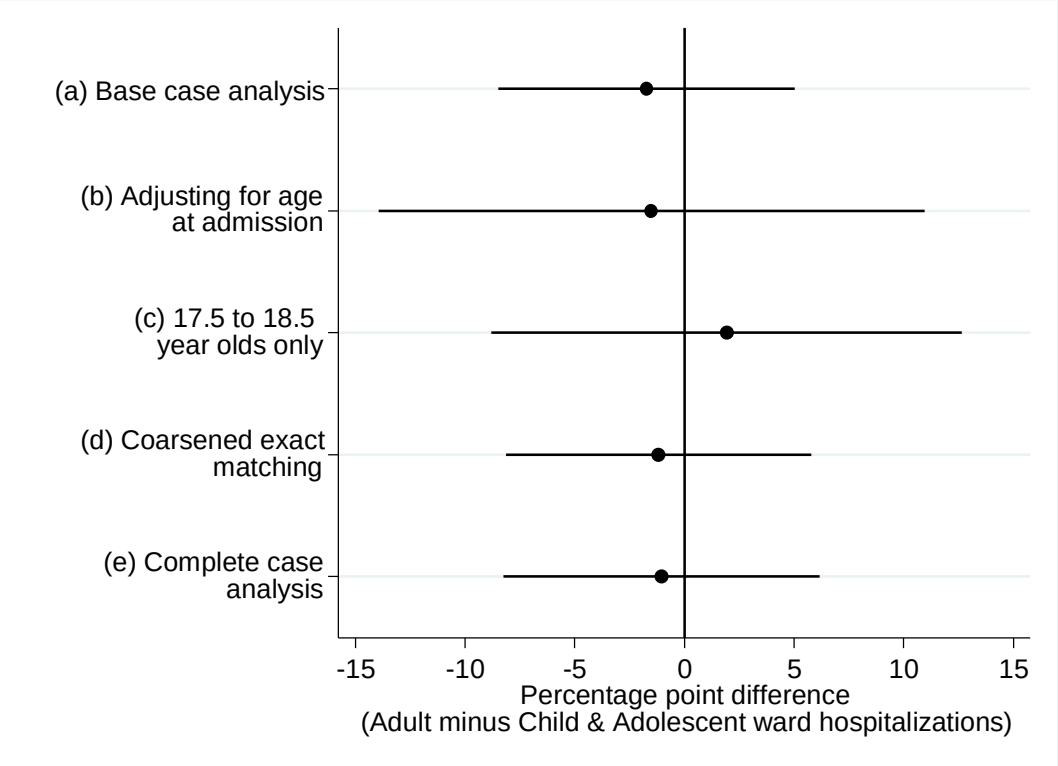


Figure 35 Difference in the probability of being sectioning during the index hospitalization

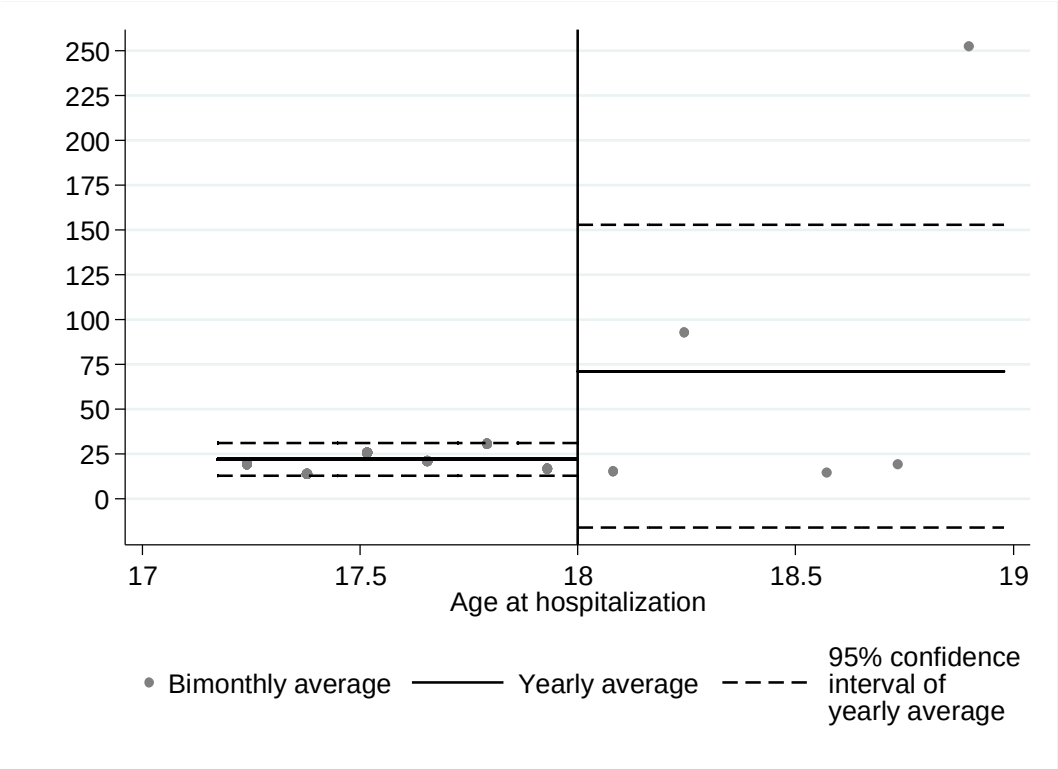


Figure 36 Number of days under section among those detained under the Mental Health Act during the index hospitalization by age at the start of the index hospitalization

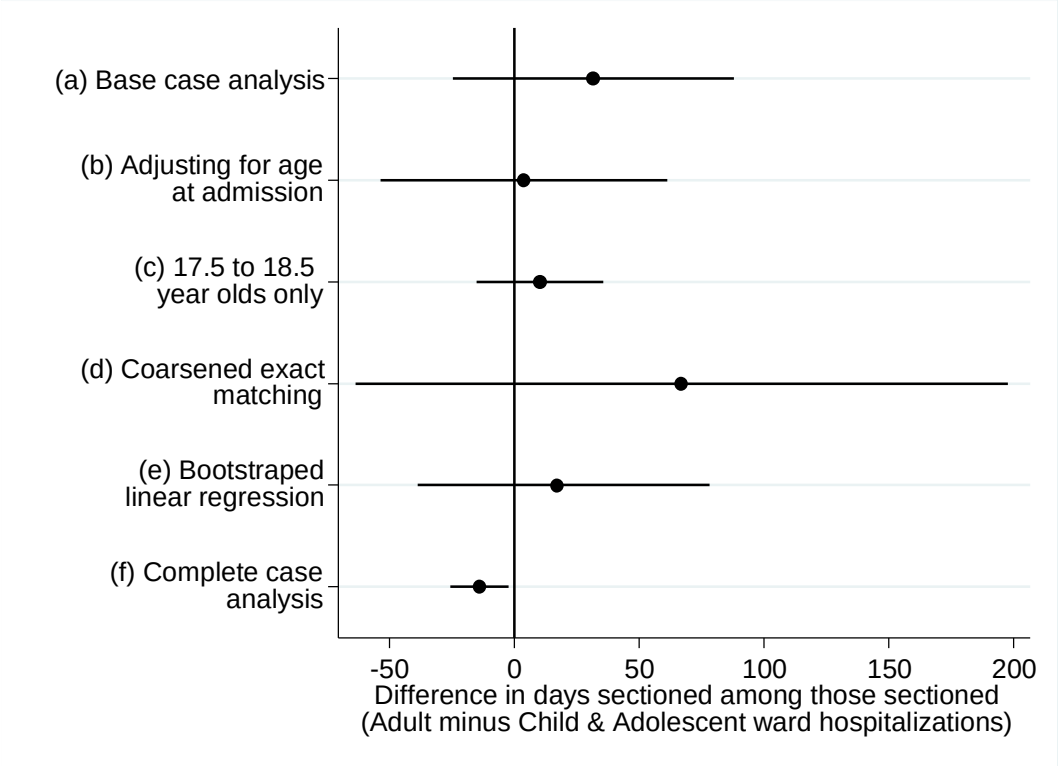


Figure 37 Difference in the length of detention under the Mental Health Act among those detained

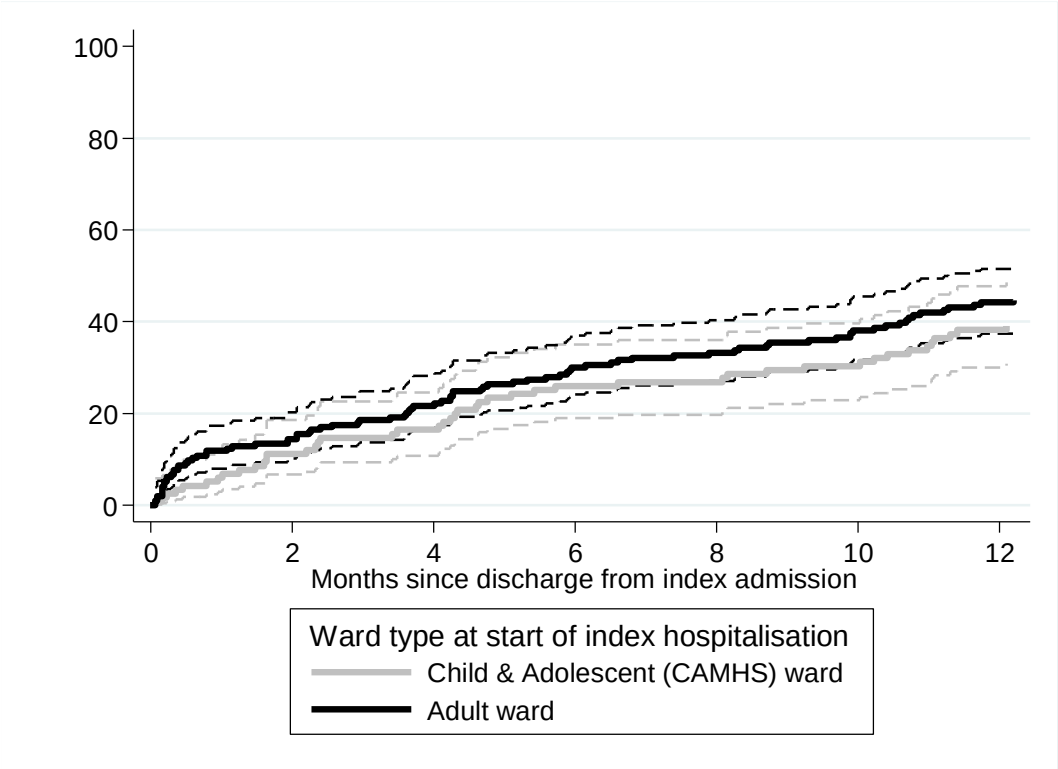


Figure 38 Rehospitalization rates over the course of the follow-up (base case analysis)

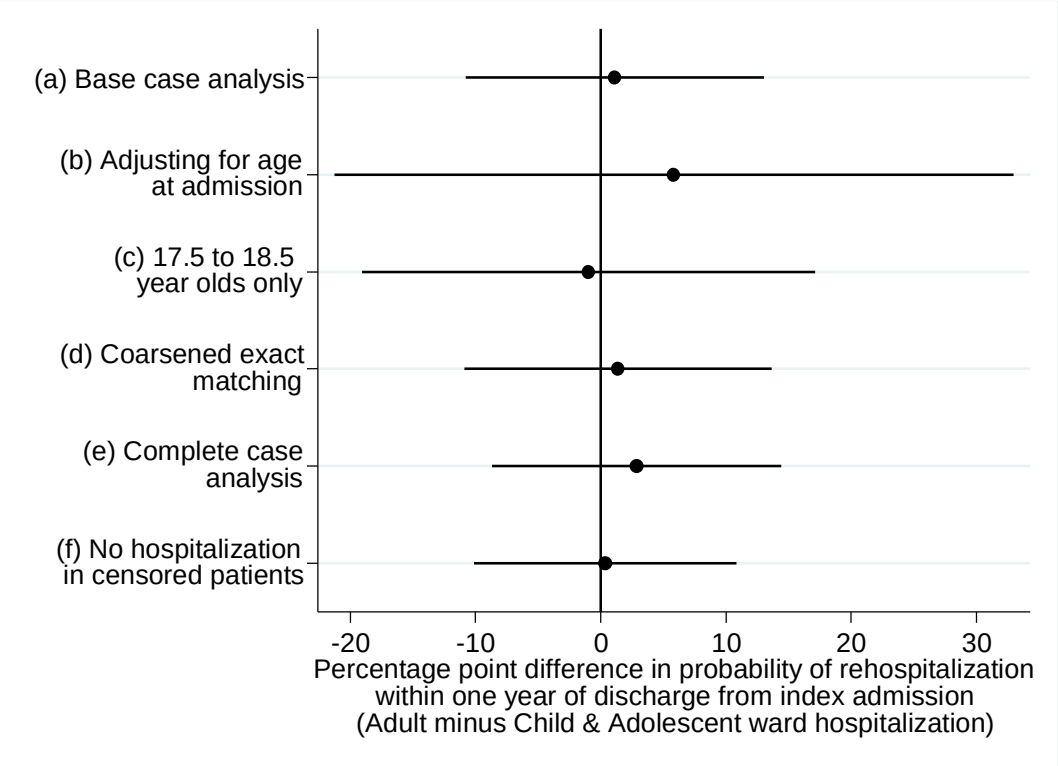


Figure 39 Difference in probability of rehospitalization within one year of discharge from index admission

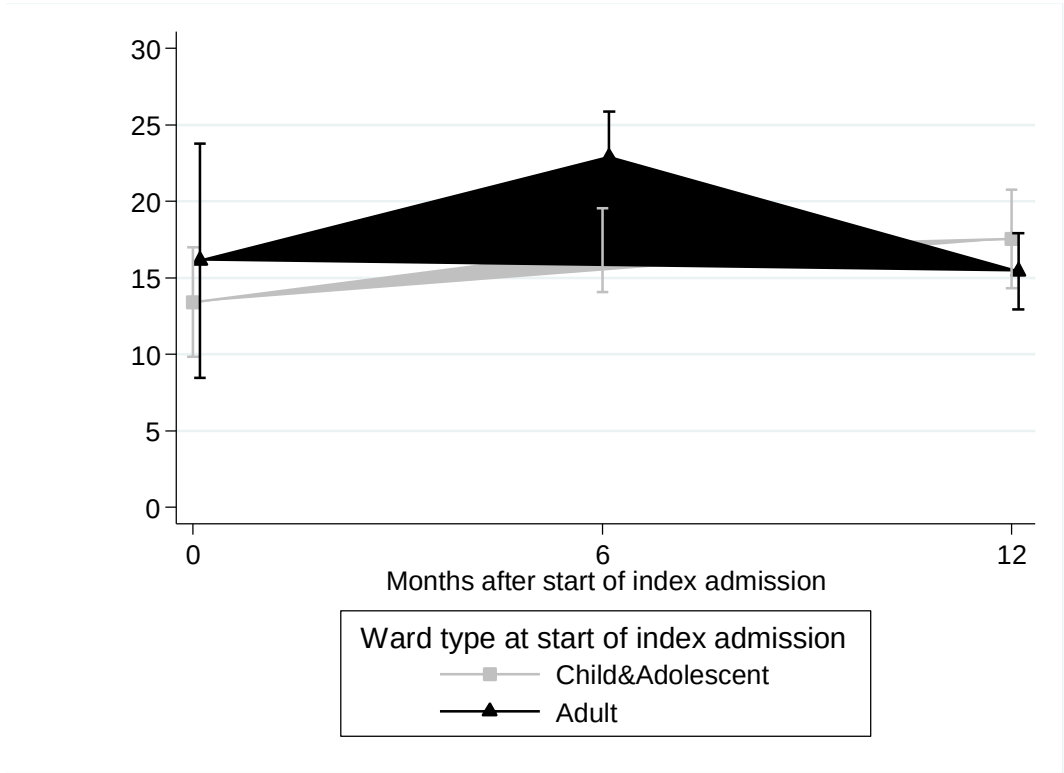


Figure 40 Number of face-to-face community contacts over time (Base case analysis)

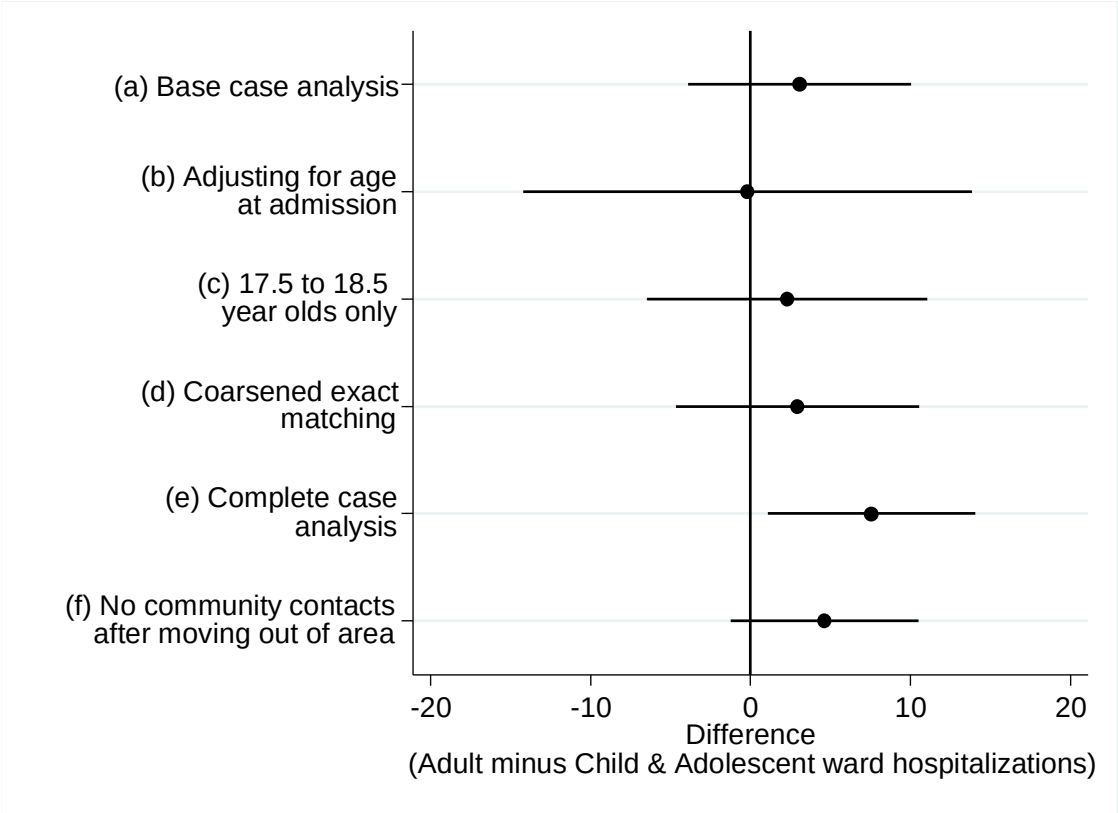


Figure 41 Difference in the number of face-to-face community contacts over one-year follow-up

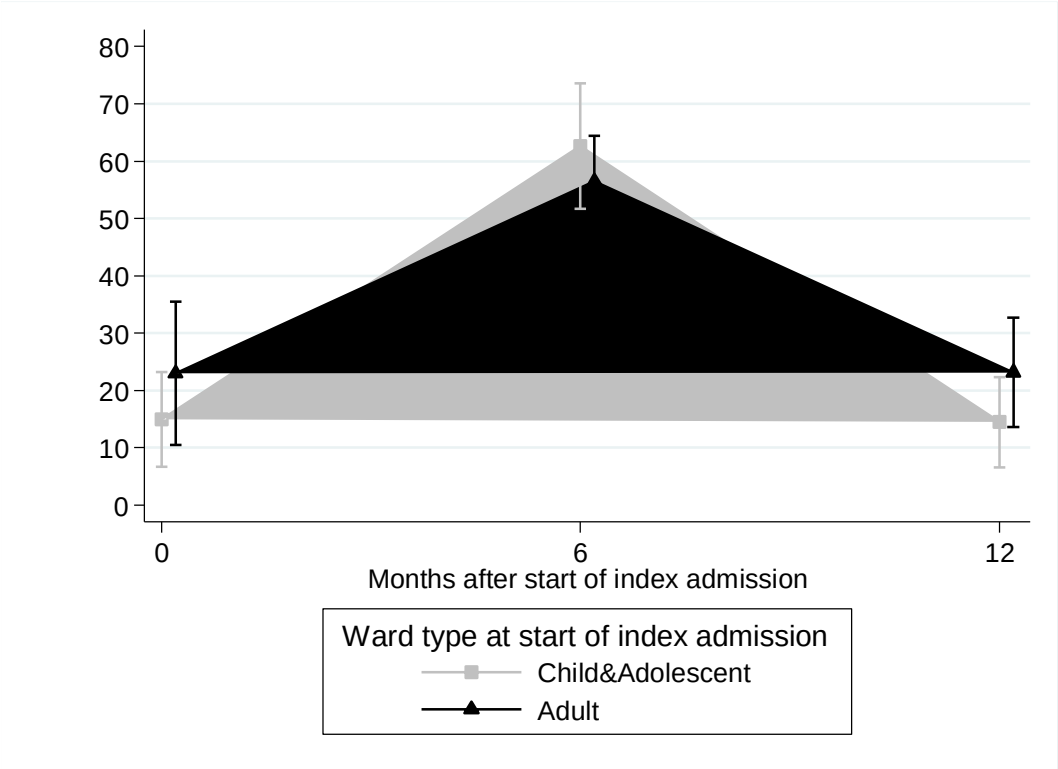


Figure 42 Number of psychiatric bed days over time (Base case analysis)

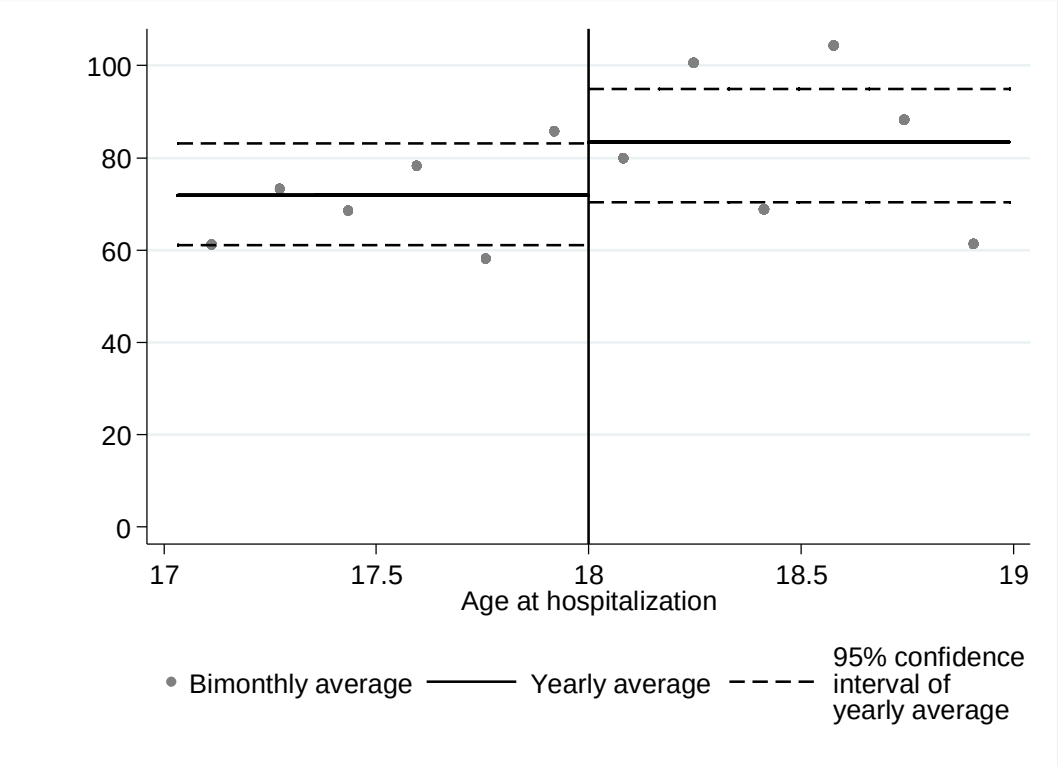


Figure 43 Length of stay on psychiatric wards by age at the start of the index hospitalization

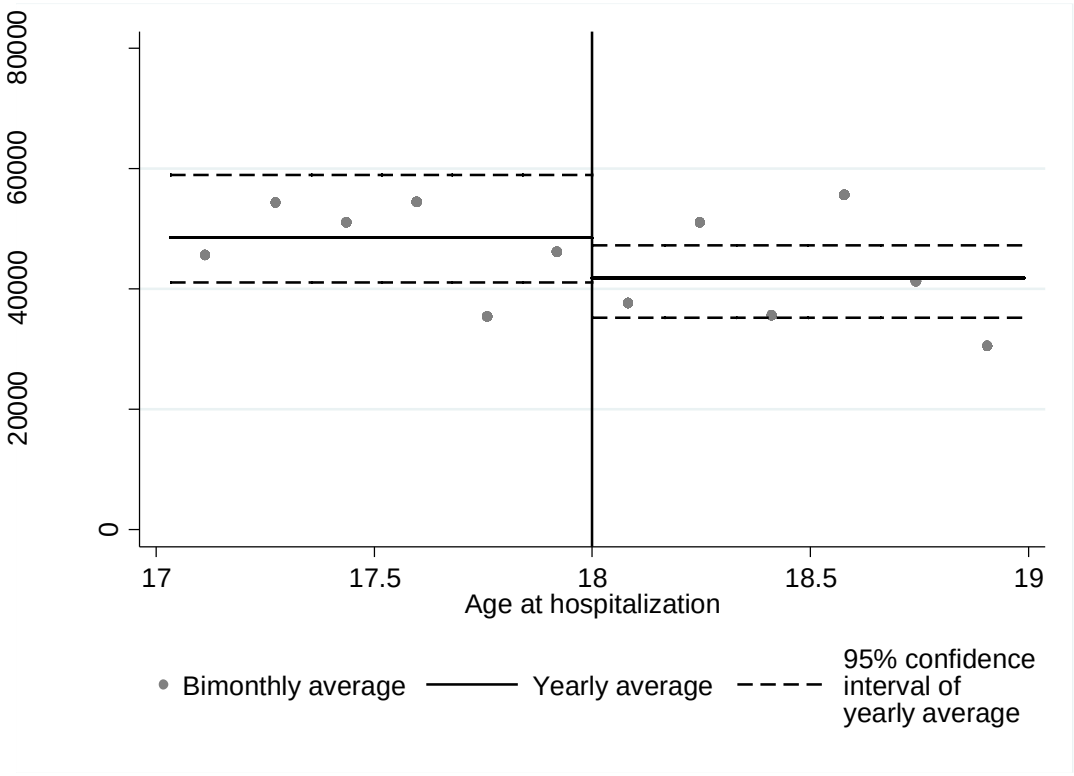


Figure 44 Cost of secondary psychiatric care within one year of the start of the index admission by age at the start of the index admission

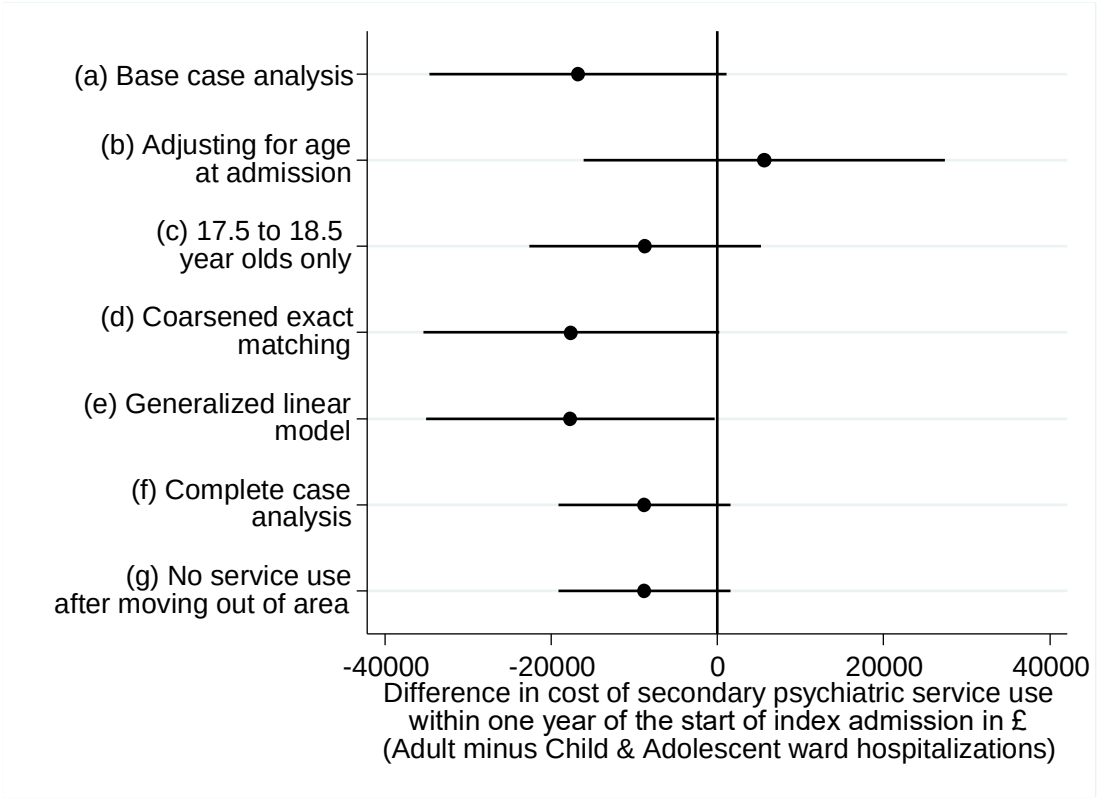


Figure 45 Cost of secondary psychiatric service use within one year of the start of the index admission

Appendix E Methodological details to Analysis 2

E.1 Notation

In principle, referrals to inpatient rehabilitation can be made at any time during a psychiatric admission. Moreover, as discussed in Chapter 1, the pathway to inpatient rehabilitation is a longitudinal process. Decisions in this process are responsive to changes in measured and unmeasured confounders over time. Therefore, in contrast to Analysis 1, the causal structure for the evaluation of inpatient rehabilitation is better characterised by using a temporal index t to reflect when variables are measured. I let t be measured at daily intervals and $t = 0$ be the start of respective psychiatric hospitalisations and T the length of the index psychiatric hospitalisation. $A(t)$ is a binary treatment indicator which is equal to 1 if a patient has been referred to in-area inpatient rehabilitation and waiting for the referral to be accepted or declined following an assessment and 0 otherwise. $M_1(t)$ is a binary mediator variable equal to 1 if the referral has been accepted and the patients is waiting to be transferred to an inpatient rehabilitation ward and equal to 0 otherwise. $M_2(t)$ is a second mediator variable equal to 1 when a patient is staying on an in-area inpatient rehabilitation ward and equal to 0 otherwise. Thus, if, for example, $A = (0,0,1,1,0,0,0, \dots)$, $M_1 = (0,0,0,1,1,1,0, \dots)$ and $M_2 = (0,0,0,0,0,0,0, \dots)$, then this patient is referred to in-area inpatient rehabilitation on day three of the psychiatric admission, the admission is accepted on day four but he/she is removed from the waiting list on day 6, i.e. not transferred to inpatient rehabilitation in this referral episode. In addition, I define $C(t)$ to be a binary indicator equal to 1 if a patient experiences a censoring event and 0 otherwise. Given these indicator variables, for notational convenience, I define D_0 as the number of days between $t = 0$ and the first censoring event that a patient experiences, D_1 as the number of days between psychiatric admission and referral, D_2 as the number of days between admission and the point at which a referral is either accepted or declined, and D_3 as the number of days between admission and the point at which a patient is either transferred to an inpatient rehabilitation ward or removed from the waiting list. If no censoring event occurs, then $D_0 = \infty$, if a patient is never referred to inpatient rehabilitation, then $D_1 = \infty$ and if a patient is referred but not accepted, then I let $D_2 = D_3$. By definition, referral to inpatient rehabilitation precedes acceptances/declination to inpatient rehabilitation which in turn

precedes transfers/removals from wait listing, i.e. $D_1 \leq D_2 \leq D_3$. In the above example, $D_1 = 3$, $D_2 = 4$ and $D_3 = 6$. Given our inclusion criteria, in our sample referrals needs to precede discharge, that is if $D_1 \neq \infty$ then $D_1 \leq T$, but in some cases patients are discharged while waiting for inpatient rehabilitation assessment or while waiting for an inpatient rehabilitation bed to become available. I let $Y(t)$ represent the outcome, whereas $L(t) = \big(G(t), X(t)\big)$ is a vector of the confounders $X(t)$ and binary indicator variable $G(t)$ equal to 1 if a patient fulfils the study eligibility criteria and 0 otherwise. I collectively refer to exogenous unmeasured random variables with respect to each observed variable as $U(t) = \big(U_{Y(t)}, U_{L(t)}, U_{A(t)}, U_{C(t)}, U_{M_1(t)}, U_{M_2(t)}\big)$. I denote the history of the variables using overbars such that, for example, the treatment history admission to psychiatric hospital to time t is represented by $\bar{A}(t)$ and $\overline{\boldsymbol{a}}$ is a vector denoting the treatment regime from the baseline to the end of follow-up. The length of follow-up over which outcomes are compared is denote as t_0. I let $\tilde{Y}\{t,h\}$ be outcome Y evaluated between time t and h, that is $\tilde{Y}\{t,h\} = f_{\tilde{Y}}\big(Y(t), Y(t+1), \dots, Y(h)\big)$. For example, if $\tilde{Y}\{t,h\}$ is the total cost of service use costs between t and h, then $\tilde{Y}\{t,h\} = \sum_{p=t}^{h} Y(p)$. Further, I denote the potential outcome under treatment strategy D as $\tilde{Y}^{D}\{t,h\}$. I let $R_L(t)$ and $R_Y(t)$ be indicators for missing L and Y respectively. As in Chapter 2, I assume that there are three versions of each variable V: the true value V indicated by the absence of superscripts, V^* the version of the variable that can be easily extracted from structured fields, through natural language processing applications or keyword search, and V^+ the version of the variable that can be obtained by reading the clinical notes. Further, $S_v(t)$ is an indicator variable denoting whether V^+ was sampled. Finally, $\mathbb{E}(X)$ denotes the expected value of a random variable X, whereas $\mathbb{P}(X = x)$ denotes the probability that the random variable X takes a certain value x.

E.2 Causal Model and observed data

For clarity, Figure 46 only shows a one-period version of the assumed causal diagram although it is assumed to hold for $t = 0, \dots, T + t_0$ and each variable is potentially a function of the entire history variables prior to t. The only exceptions from this are that, by definition, referrals can only be accepted if a referral has taken place previously, patients can only be transferred if their referral has been accepted previously. In other words, if $A(t) = 0$ and $M_1(t-1) = 0$ then $M_1(t) = 0$, and if $M_1(t) = 0$ and $M_2(t-1) = 0$ then

$M_2(t) = 0$. As discussed in Chapter 2, we do not observe the version of the variables without measurement error, $V(t)$, but two types of proxy variables: $V^*(t)$ the version of the variable that can be from structured fields or from NLP applications and $V^+(t)$ the manually coded version of the variable. $V^*(t)$ is observed if $R_V(t) = 0$, $R_V(t)$ being a missingness indicator, and $V^+(t)$ is observed if $S_v(t) = 1$, $S_v(t)$ being an indicator for whether the observation was manually coded.

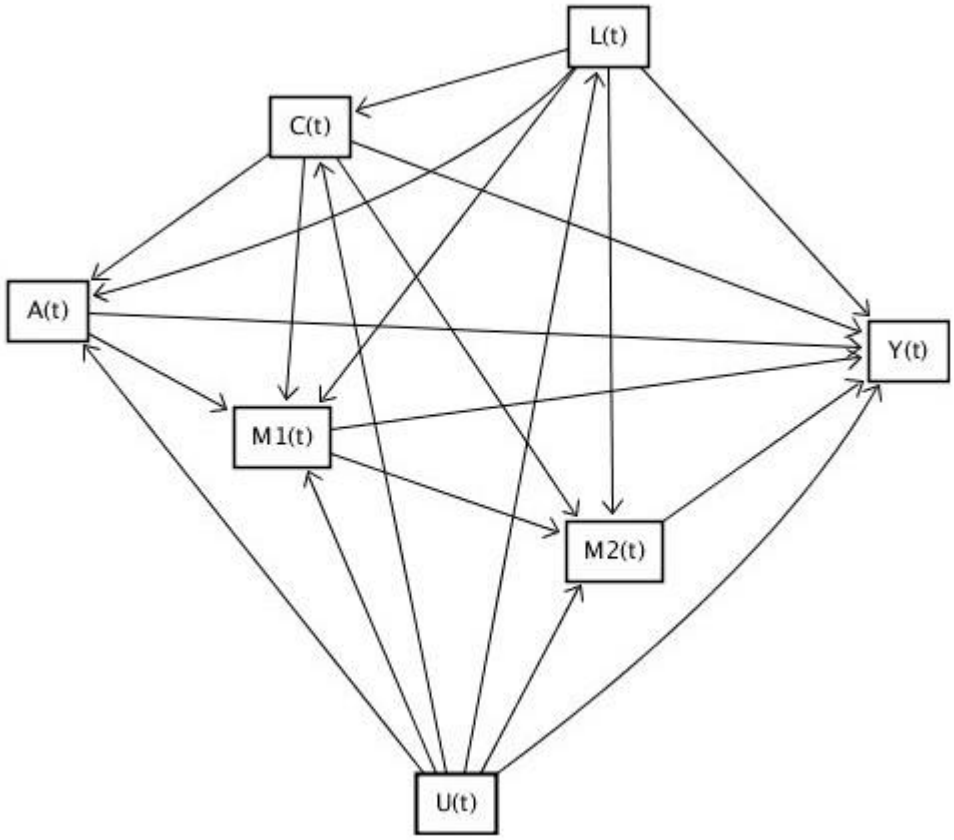


Figure 46 One-period version of the assumed causal model

E.3 Target causal parameter

Our comparison of interest was between the longitudinal treatment strategies $\bar{\boldsymbol{a}} = (a(D_1), \dots, a(D_1 + t_0))$ and $\bar{\boldsymbol{c}} = (0, \dots, 0)$ compared to $\bar{\boldsymbol{a}} = (0, \dots, 0)$ and $\bar{\boldsymbol{c}} = (0, \dots, 0)$. Formally, our target parameter can be defined as

$$\psi_{ATT} = \mathbb{E}\left(\begin{matrix} \tilde{Y}^{\,D_1 = d_1, D_0 > d_1 + t_0}\{D_1, D_1 + t_0\} \\ -\tilde{Y}^{\,D_1 > d_1 + t_0, D_0 > d_1 + t_0}\{D_1, D_1 + t_0\} \end{matrix} \middle| G(D_1) = 1, D_1 < T \right)$$

In line with the literature, I refer to this parameter as the average treatment effect on the treated (ATT), but since some patients who are referred to inpatient are not transferred to an inpatient rehabilitation ward, a more exact description may be the intention-to-treat effect on the intended (Glynn and Kashin, 2018).

E.4 Identifying assumptions (base case analysis)

Confounding: The main identifying assumptions was that the sequential randomisation assumption holds when referrals are made, that is

$$\tilde{Y}^{D}\{D_1, D_1 + t_0\} \perp A(D_1) | L(D_1), \tilde{Y}\{t, h\}$$

where t to h is some meaningful period prior to referral, i.e. $t \leq h \leq D_1$. In the case of cost data for example, we chose $t = -365$ and $h = 0$, i.e. the year prior to hospitalisation. In other words, based on our clinical knowledge, we made the simplifying assumption that conditional on the current value of observed confounders and the lagged value of the outcome over the period t and h, the potential outcome is independent of the decision to refer to inpatient rehabilitation. While a subset of $L(D_1)$ contains the cumulative history of some variables (e.g. number of days on section from admission to referral) and some variables are time-invariant within a hospitalisation (e.g. number of psychiatric inpatient days in the year prior to the index admission), this assumes that clinical decisions are primarily based on patients' health status at referral rather than the disease trajectory between admission and referral. We also make the simplifying but, based on our clinical knowledge, plausible assumption that prior referrals within an admission that did not result in transfer to inpatient rehabilitation have no long-term effect on the potential outcome, after conditioning on the history of service use. In other words, conditioning on $\bar{A}(D_1 - 1)$ and $\overline{M_1}(D_1 - 1)$ is assumed to be unnecessary. We investigate the sensitivity of results to departures from this assumption in the sensitivity analyses.

Positivity: I assume that probability of treatment assignment is non-negative at all time points and covariate combinations value of the covariates or that there is overlap in the prognostic score

$$\mathbb{P}\big(A(D_1) = 0 \big| \mathbb{P}\big(A(D_1) = 1 \big| L(D_1), \tilde{Y}\{t, h\}\big) > 0, L(D_1), \tilde{Y}\{t, h\}\big) > 0$$

This assumption is weaker than the standard positivity assumption because the target parameter is the ATT not the average treatment effect. In the base case analysis, I use a matching approach to make this assumption more realistic.

If an outcome of interest is measured between time h' and t', then I also assumed that

$$\mathbb{P}\left(D_0 > t' \middle| \begin{matrix} \mathbb{P}\left(D_0 < t' \middle| L(D_1), \tilde{Y}\{t,h\}, \tilde{Y}\{D_1, D_0\}\right) > 0, \\ L(D_1), \tilde{Y}\{t,h\}, \tilde{Y}\{D_1, h'\} \end{matrix}\right) > 0$$

meaning that for any combination of baseline confounders and outcomes until h' for which the outcome of some patients was censored there was a positive probability that someone with the same characteristics and outcomes until that point was uncensored at that point.

Sampling: I assumed that the sample was obtained through simple random sampling, that is

$$\mathbb{P}\big(Y(t) = y(t), A(t) = a(t), L(t) = l(t)\big) = \frac{1}{N}$$

where N is the sample size. Since, as noted in Chapter 2, it appeared that referrals that were declined were systematically underrecorded in structured field, we attempted to make this assumption more realistic by searching for keywords in medical notes to identify patients whose inpatient rehabilitation referral was not recorded in structured records. Moreover, this assumptions violated to some degree because the distribution of time to referral to inpatient rehabilitation within a given admission, D_1, is to some extent unrepresentative because D_1 is censored by D_0 (Li et al., 2014). In practice, I believe that this violation of the assumption is unlikely to have a noticeable bias on the results because (a) the length of follow-up is at least one year for all patients and as shown below the vast majority of referrals to inpatient rehabilitation are referred within the first 6 months of admission; (b) by definition, patients cannot move outside of the catchment area when hospitalized and (c) it is rare for patients are referred to out of area inpatient rehabilitation without initial referral to in-area rehabilitation.

Incomplete variables: I assumed that

$$Y(t) \perp R_Y(t) | L(D_1) \cdot (1 - R_L(D_1)), A(D_1), \tilde{Y}\{t,h\}, R_L(t)$$

meaning that outcomes were missing at random, conditional on the treatment variable, the missingness indicator and the baseline confounders when they are not missing. This is

stronger than the standard missing at random assumption but, as shown in Figure *48*, in practice, there were only missing baseline values in the HoNOS scores and in most dimensions missingness rates were low.

Model misspecification: I assumed that the models to estimate treatment weights and the censoring/missingness weights were correctly specified. I used the matching and boosted regression approach to make this assumption more realistic.

Measurement error: With respect to the outcome, I assumed that

$$\mathbb{E}\big(Y^{+}(t) \cdot S_Y(t) + Y^{*}(t) \cdot \big(1 - S_Y(t)\big) | A(D_1), L(D_1), \tilde{Y}\,\{t, h\}\big) = \mathbb{E}\big(Y(t) | A(D_1), L(D_1), \tilde{Y}\,\{t, h\}\big)$$

meaning that, in expectation, the mixture of the outcome variable that were manually verified and variables that were not, is not systematically different from the expected value of the true outcomes $Y(t)$.

With respect to age, gender, ethnicity, number of days on a general adult ward, sectioning status, number of days detained under the Mental Health Act, history of community rehabilitation, the history of service use, the number of risk events and HoNOS ratings, I assumed that there was no measurement error in the structured fields, that is $V(t) = V^{*}(t)$.

With respect to whether a patient had a diagnosis of psychosis or a history of clozapine use, I assumed that $V(t) = (1 - S_V(t)) \cdot V^{*} + S_V(t) \cdot V^{+}(t)$, that is I assumed that

With respect to whether a patient was described as having poor motivation, was socially withdrawn or at risk of self-neglect I assumed $V(t) = V^{+}(t)$.

With respect to the history of clozapine use and the symptoms manually coded by reading the clinical notes, I also assumed that, if $S_X(D_1) = 1$, the difference between the manually coded version of the confounder and the readily available version of X, that is $X^{*}(D_1) - X^{+}(D_1)$, was neither an instrumental variable nor a collider caused by a instrumental variable and a variable that only predicts outcomes. In other words, I assumed that conditioning on $X^{*}(D_1) - X^{+}(D_1)$ does not lead to Z- or M-bias (Liu et al., 2012; Myers et al., 2011).

E.5 Identifying assumptions (sensitivity analyses)

Before-and-after analysis: I replaced the sequential randomisation assumption by

$$\tilde{Y}^{D}\{D_1, D_1 + t_0\} - \frac{t_0}{D_1 - h'}\tilde{Y}\{h', D_1\} \perp A(D_1)|A(D_1) = 1$$

where h' is again some meaningfully chosen period. In the case of costs, following the example of Bunyan et al. (2016), we chose $h' = D_1 - 730$, i.e. the two years prior to referral The term $\frac{t_0}{D_1 - h'}$ rescales the specified pre-referral period so that it corresponds with the length of the post-referral period of interest. The positivity assumption with respect to treatment assignment was not necessary in this analysis given that, by construction, all observations included in this analysis were referred:

Front-door adjustment #1: Two assumptions are needed to in place of the sequential randomisation assumption (Glynn and Kashin, 2018). First, that referral to inpatient rehabilitation does not affect the outcome directly, an assumption known as the exclusion restriction, which can be expressed as follows

$$\begin{aligned} &E\left(\tilde{Y}\{D_1, D_1 + t_0\}\middle|\bar{A}(D_1) = 1, M_1(D_2) = 0, \bar{L}(D_1), \bar{U}(D_1)\right) \\ &\quad = E\left(\tilde{Y}^{D_1 > d_1}\{D_1, D_1 + t_0\}\middle|\bar{A}(D_1) = 1, M_1(D_2) = 0, \bar{L}(D_1), \bar{U}(D_1)\right) \\ &\quad = E\left(\tilde{Y}^{D_1 = d_1}\{D_1, D_1 + t_0\}\middle|\bar{A}(D_1) = 1, M_1(D_2) = 0, \bar{L}(D_1), \bar{U}(D_1)\right) \end{aligned}$$

In words, the observed outcome of patients who were referred but not accepted to inpatient rehabilitation is assumed to be equivalent to their outcome had they been referred at this point. Second, whether a referral to inpatient rehabilitation is accepted or not needs to be conditionally randomised at the point at which the referral is accepted or rejected, i.e.

$$\tilde{Y}^{D}\{D_2, D_1 + t_0\} \perp M_1(D_2)|A(D_1) = 1, L(D_1), \tilde{Y}\{t, h\}, D_2 - D_1$$

As above, this is stronger than minimum identification assumption for the intervention effect of time-varying mediators because I assume that conditioning on treatment and mediator history, i.e. $\bar{A}(D_2)$ and $\overline{M_1}(D_2 - 1)$, is not necessary after conditioning on the history of service use (VanderWeele and Tchetgen Tchetgen, 2017). Moreover, for simplicity, I only condition on $L(D_1)$ rather than $L(D_2)$

Appendix 4 In addition to these two assumptions, modified versions of the positivity assumptions are required. With respect to the treatment, this analysis assumes

$$\mathbb{P}\left(M_1(D_2) = 0 \middle| \begin{matrix} \mathbb{P}\big(M_1(D_2) = 1 \big| L(D_1), \tilde{Y}\{t,h\}, A(D_1) = 1\big) > 0, \\ L(D_1), \tilde{Y}\{t,h\}, A(D_1) = 1 \end{matrix}\right) > 0$$

whereas with respect to censoring, this analysis assumes that

$$\mathbb{P}\left(D_0 > t' \middle| \begin{matrix} \mathbb{P}\big(D_0 < t' \big| L(D_1), \tilde{Y}\{t,h\}, \tilde{Y}\{D_1, D_0\}, A(D_1) = 1\big) > 0, \\ L(D_1), \tilde{Y}\{t,h\}, \tilde{Y}\{D_1, h'\}, A(D_1) = 1 \end{matrix}\right) > 0$$

Front-door adjustment #2: In the second version of the front-door adjustment approach, the exclusion restriction is identical except that D_3 rather than D_2 and M_2 rather than M_1 are the relevant variables/time points in this context

$$\begin{aligned} &E\left(\tilde{Y}\{D_1, D_1 + t_0\} \middle| \bar{A}(D_1) = 1, M_2(D_3) = 0, \bar{L}(D_1), \bar{U}(D_1)\right) \\ &\quad = E\left(\tilde{Y}^{D_1 > d_1}\{D_1, D_1 + t_0\} \middle| \bar{A}(D_1) = 1, M_2(D_3) = 0, \bar{L}(D_1), \bar{U}(D_1)\right) \\ &\quad = E\left(\tilde{Y}^{D_1 = d_1}\{D_1, D_1 + t_0\} \middle| \bar{A}(D_1) = 1, M_2(D_3) = 0, \bar{L}(D_1), \bar{U}(D_1)\right) \end{aligned}$$

Put differently, this assumes that, whether referred patients are accepted to inpatient rehabilitation or not, as long as they are not transferred to inpatient rehabilitation, in expectation, their outcomes are identical to what they would have been had they never been referred. Analogously, in this sensitivity analysis, whether a patient is transferred to inpatient rehabilitation or removed from the waiting list is assumed to be conditionally randomised, that is

$$\tilde{Y}^{D}\{D_3, D_1 + t_0\} \perp M_2(D_3) | A(D_1) = 1, L(D_1), \tilde{Y}\{t,h\}, D_3 - D_1$$

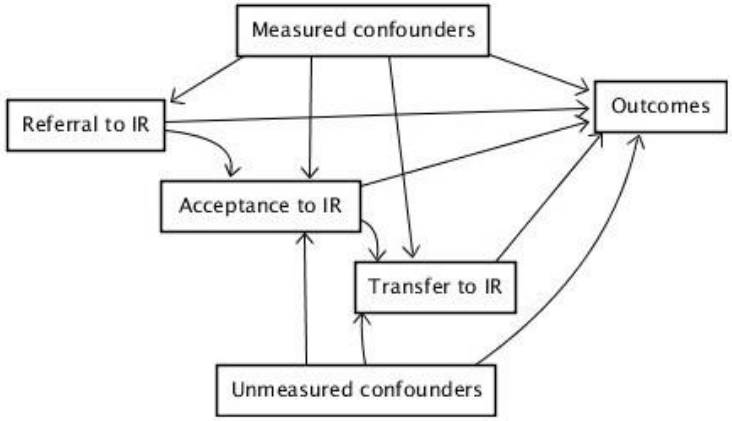
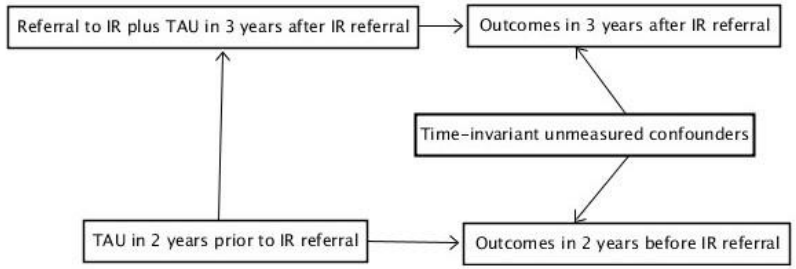
Again, I do not condition on $\bar{A}(D_3)$, $\overline{M_1}(D_3)$ or $L(D_3)$ and alternative positivity assumptions are required are required with respect to the treatment,

$$\mathbb{P}\left(M_2(D_3)=0 \middle| \begin{matrix}\mathbb{P}\big(M_2(D_3)=1 \big| L(D_1), \tilde{Y}\{t,h\}, A(D_1)=1, M_1(D_2)=1\big)>0, \\ L(D_1), \tilde{Y}\{t,h\}, A(D_1)=1, M_1(D_2)=1\end{matrix}\right)>0$$

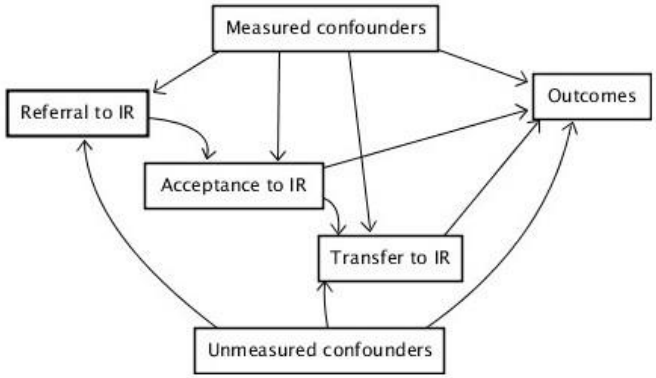
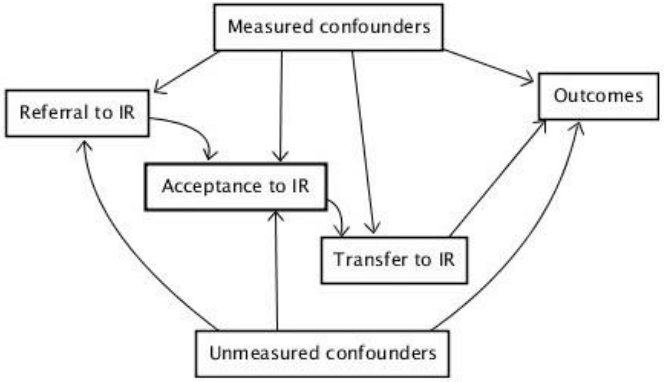
and with respect to censoring

$$\mathbb{P}\left(D_0>t' \middle| \begin{matrix}\mathbb{P}\big(D_0<t' \big| L(D_1), \tilde{Y}\{t,h\}, \tilde{Y}\{D_1,D_0\}, A(D_1)=1, M_1(D_2)=1\big)>0, \\ L(D_1), \tilde{Y}\{t,h\}, \tilde{Y}\{D_1,h'\}, A(D_1)=1, M_1(D_2)=1\end{matrix}\right)>0$$

Appendix F Supplementary figures and tables to Analysis 2

Analysis	Base case analysis (Back-door adjustment)	Sensitivity Analysis 1 (Before-and-after analysis)
Estimation	Outcomes in those referred to IR – outcomes in matched controls after adjusting for observed confounders	Outcomes (post referral) – outcomes prior to referral in those referred to IR
Key assumptions	Whether patient is referred to IR or not does not depend on factors that also influence the outcomes after adjusting for observed confounders	If patients had not been referred to IR, the outcomes would have stayed unchanged over the study follow-up compared to the two-year period prior to referral to IR
Graphical illustration of key assumptions†		

Analysis	Sensitivity Analysis 2 (Front-door adjustment #1)	Sensitivity Analysis 3 (Front-door adjustment #2)
Estimation	(Outcomes in patients who are referred and accepted – outcomes in patients who are referred but not accepted to IR after adjusting for observed confounders) x Proportion of patients who are accepted among those referred	(Outcomes in those who are referred, accepted and transferred to IR – outcomes in patients who are referred, accepted but not transferred to IR after adjusting for observed confounders) x Proportion transferred to IR among those referred
Key assumptions	Whether a patient is accepted to IR or not given that they have been referred not does not depend on factors influencing the outcome after adjusting for observed	Whether a patient is transferred to IR or not given that they have been referred and accepted not does not depend on factors influencing the outcome after adjusting for observed

	confounders Had patients who were referred but not accepted to IR not been referred, they would have had the same outcomes	confounders Had patients who were referred but not transferred to IR not been referred, they would have had the same outcomes
Graphical illustration of assumptions†		

† Arrows between two variables indicates that causal effects may exist between variables without distorting the comparisons (e.g. $X \rightarrow Y$, indicate that X may cause Y). Missing arrows between variables indicate that causal effects need to be absent

Table 4 Simplified summary of approaches to handling unmeasured confounding (Analysis 2)

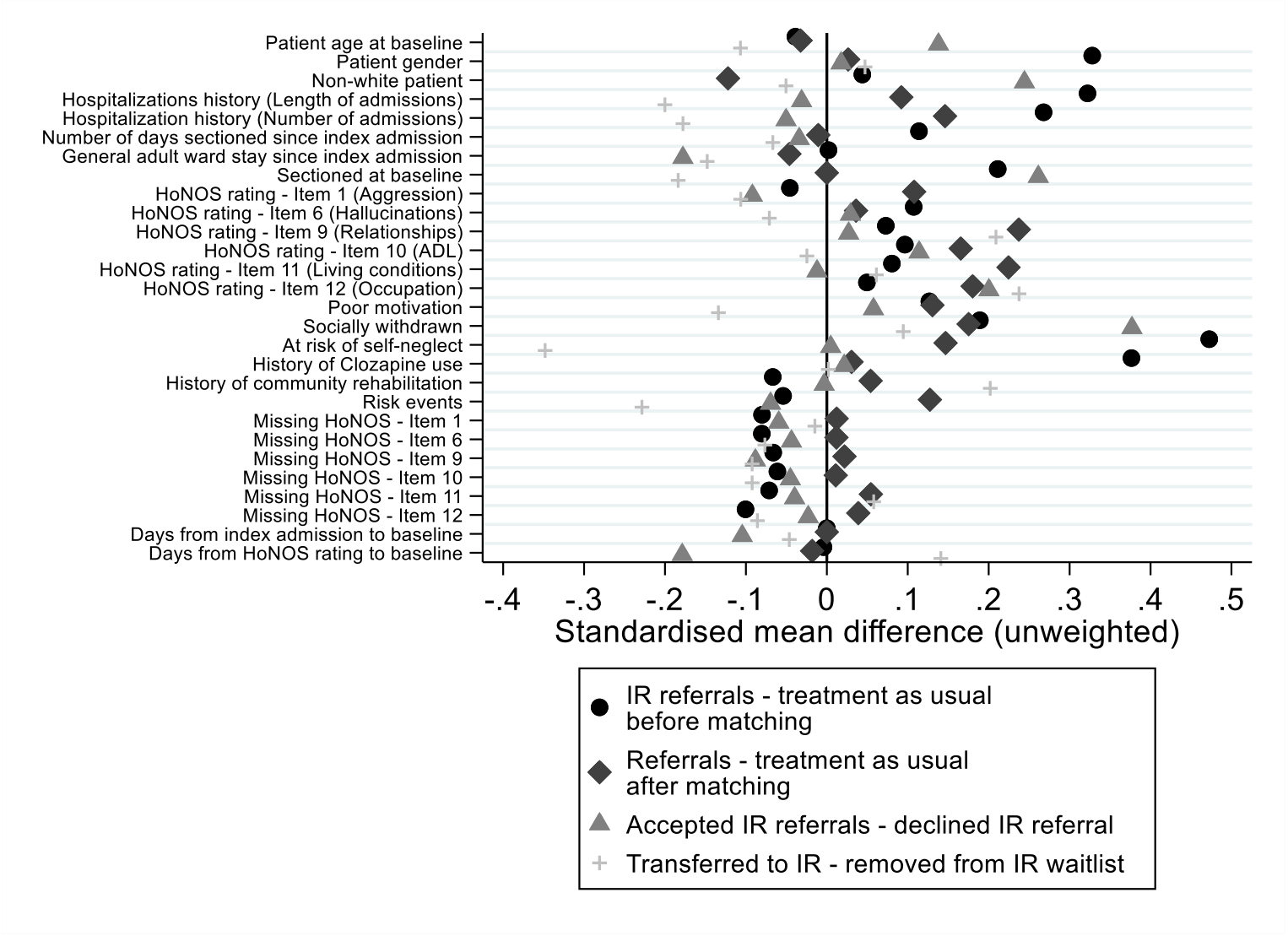


Figure 47 Standardized differences between inpatient rehabilitation (IR) treatment groups at baseline

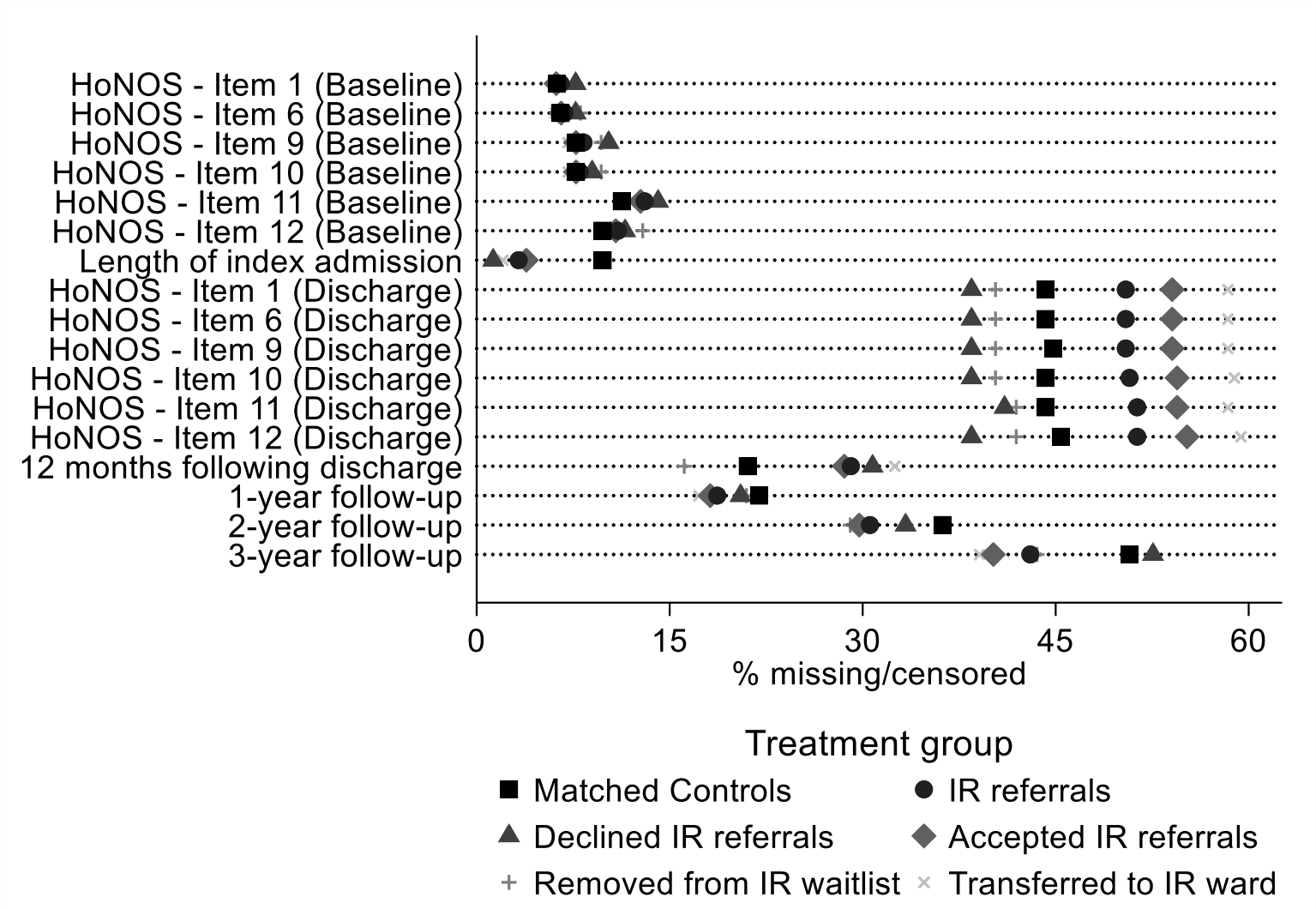


Figure 48 Percentage of missing data among incomplete variables by inpatient rehabilitation (IR) treatment group

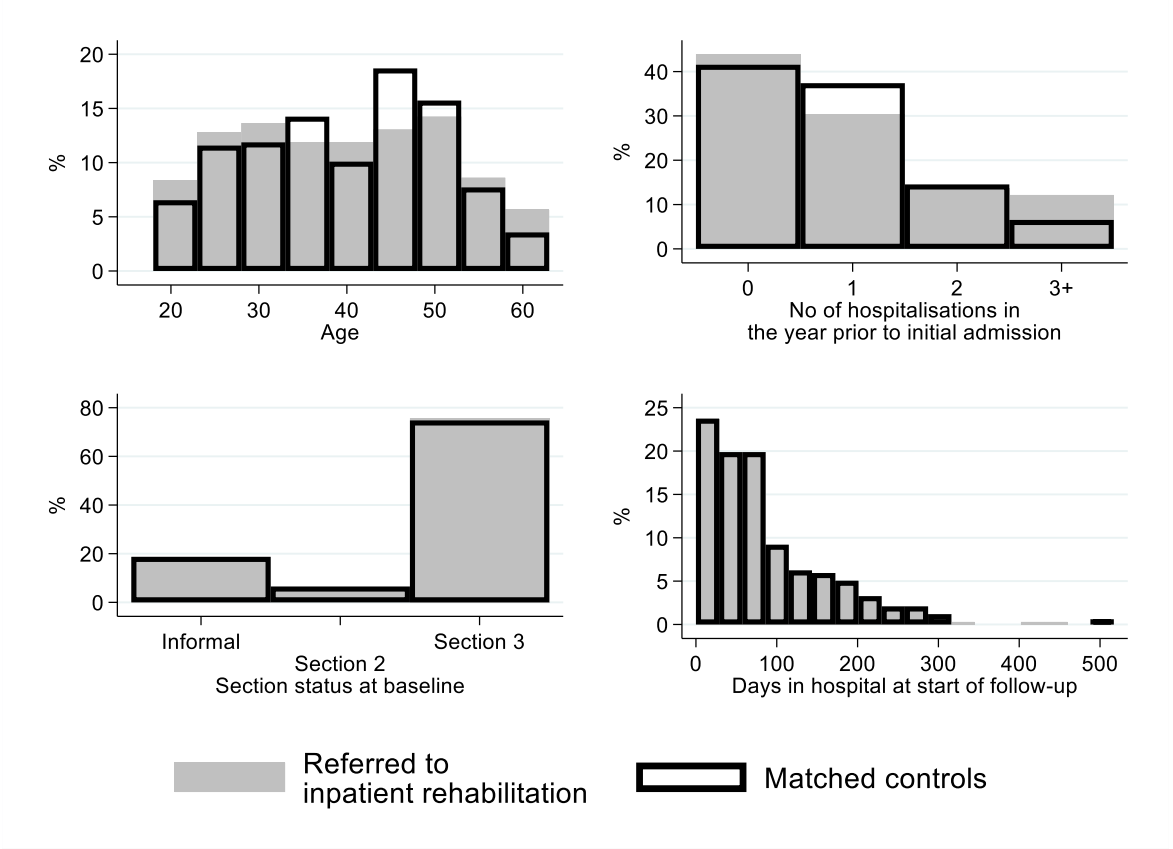


Figure 49 Distribution of continuous/categorical confounders at baseline by treatment group – part 1 (base case analysis)

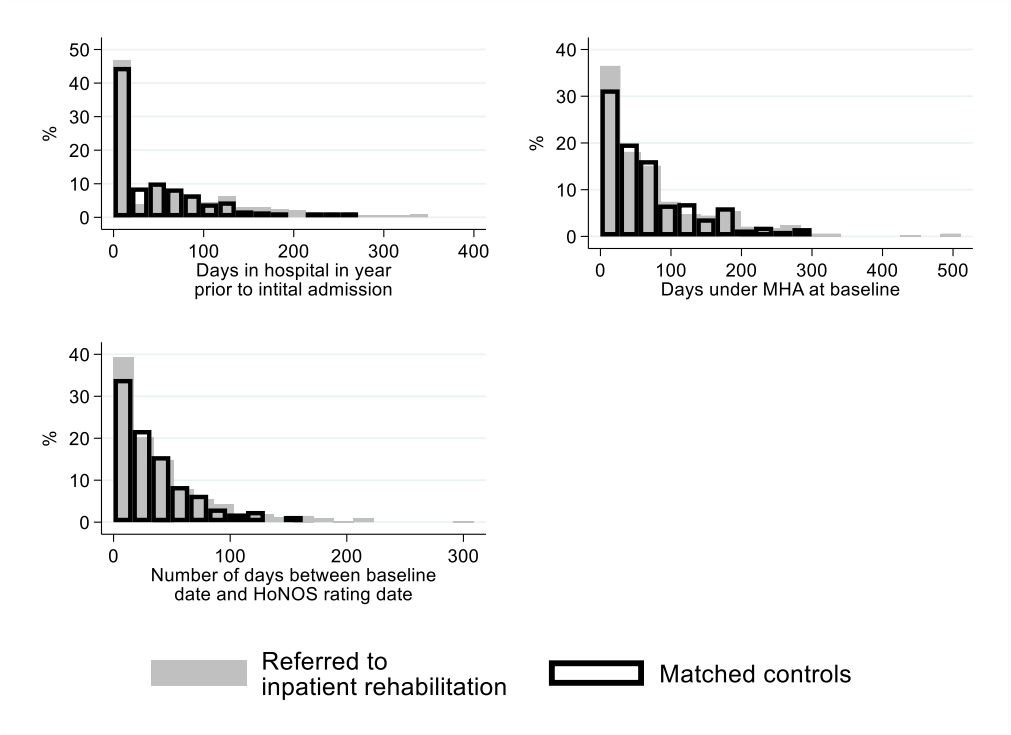


Figure 50 Distribution of continuous/categorical confounders at baseline by treatment group – part 1 (base case analysis)

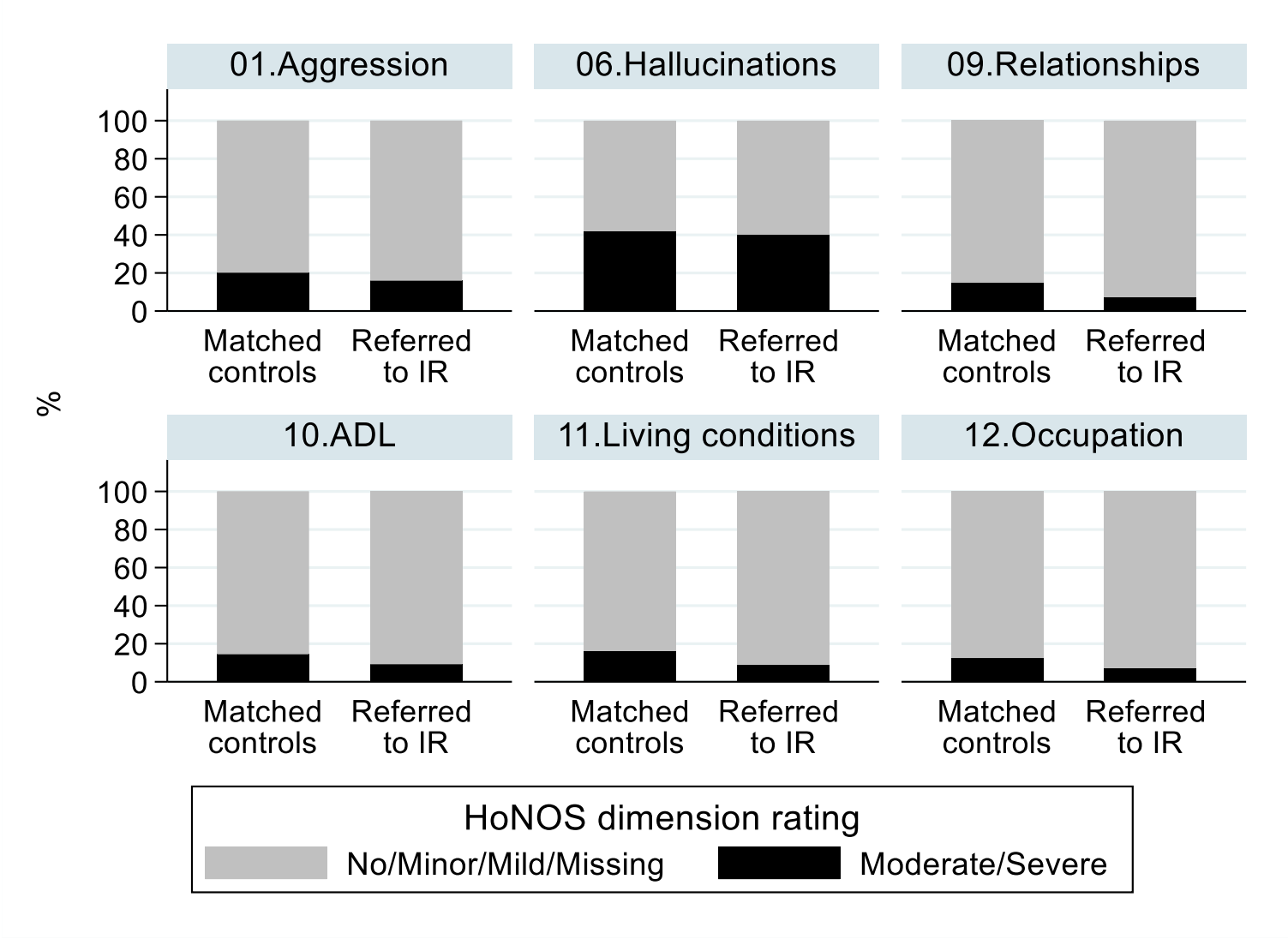


Figure 51 Distribution of HoNOS ratings at baseline by treatment group (base case analysis)

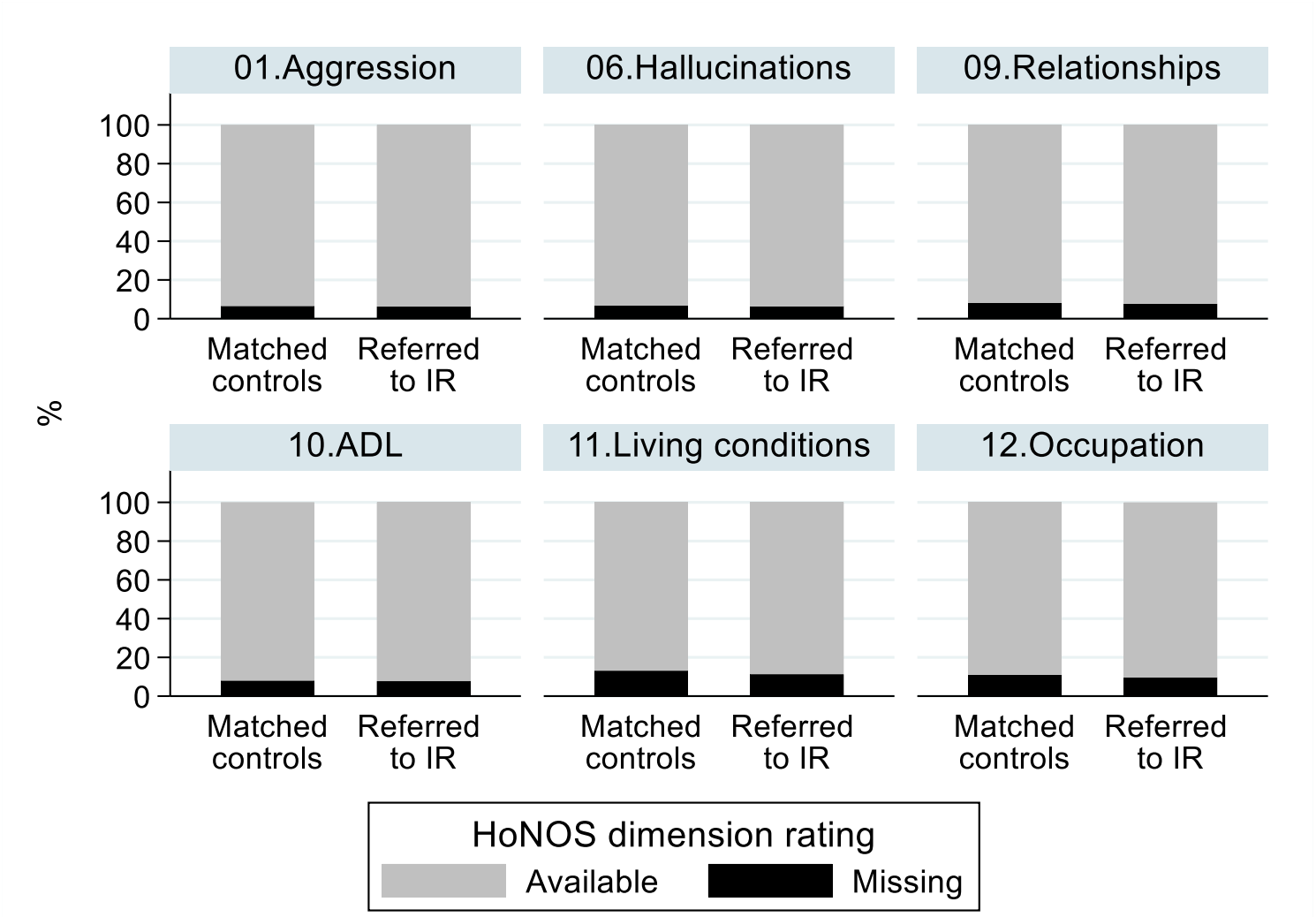


Figure 52 Distribution of missing HoNOS scores at baseline by treatment group (base case analysis)

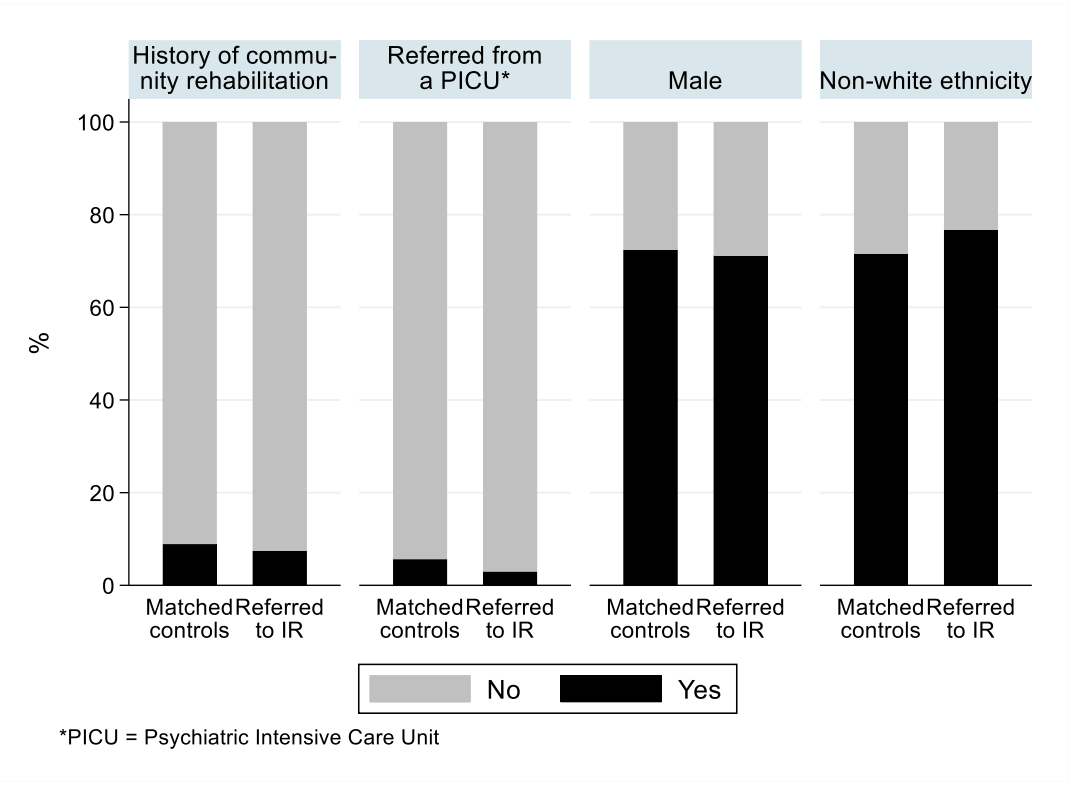


Figure 53 Distribution of other binary confounders at baseline by treatment group – part 1 (base case analysis)

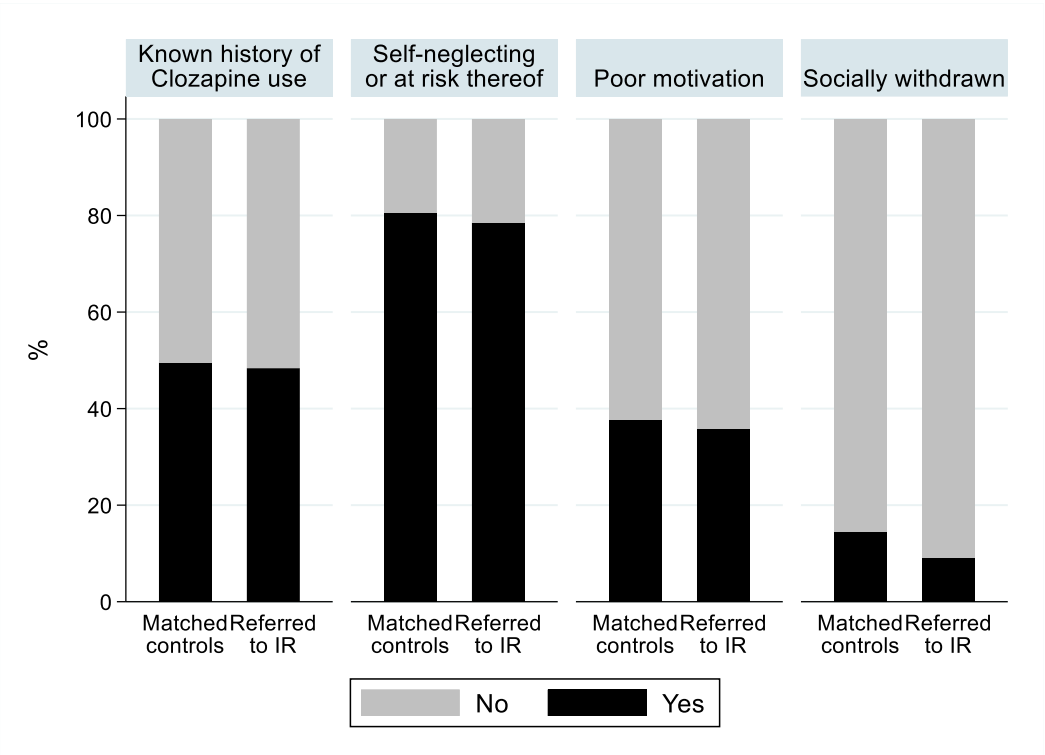


Figure 54 Distribution of other binary confounders at baseline by treatment group – part 2 (base case analysis)

Characteristic	Accepted referrals N=259	Declined referrals N=78	Difference in means (Ratio of variances*)
Demographics			
Mean (SD) age in years	40 (12)	38 (11)	2 (1.1)
% male	73	72	1
% BME	74	63	11
History of service/medication use			
% by number of hospitalizations in the year prior to index hospitalization			
0	44	45	-1
1	30	32	-2
2	15	9	6
3 or more	12	14	-3
Mean (SD) number of days hospitalised in the year prior to the index hospitalization	61 (76)	64 (85)	-3 (0.8)
% with history of community rehabilitation	9	9	0
History of Clozapine use	50	49	1
Characteristics of index hospitalisation between initial hospitalisation and baseline			
Mean (SD) length of stay in days	89 (80)	98 (91)	-9 (0.8)
Mean (SD) length of stay on General Adult Ward	72 (68)	86 (89)	-14 (0.6)
Mean (SD) number of days hospitalised detained under a Mental Health Act	75 (80)	78 (91)	-3 (0.8)
% detained at baseline	84	73	11
% with no/minor/mind, moderate/severe problems or missing HoNOS dimensions score			
01 – Aggression	81, 17, 2	73, 23, 4	8, -6, -2
06 – Hallucinations	56, 41, 2	63, 33, 4	-6, 8, -2

09 – Relationships	83, 13, 4	76, 18, 6	7, -5, -2
10 – Activities of Daily Living	83, 14, 3	83, 13, 4	0, 1, -1
11 – Living Conditions	75, 16, 9	69, 19, 12	6, -3, -3
12 - Occupation	80, 13, 8	83, 9, 8	-4, 4, 0
Mean (SD) days between HoNOS rating and baseline date (Baseline – HoNOS rating date)	26 (46)	21 (43)	5 (1.2)
% with mention of symptom in clinical notes†			
Poor motivation	39	39	0
Social withdrawal	17	6	11
Self-neglecting or at risk thereof	80	80	1
% by number of known risk events†			
0	71	65	5
1	20	24	-4
2 or more	9	10	1

*for continuous and count data only; † between the start of the index admission and baseline or 30 days prior to baseline, whichever is larger

SD: Standard deviation; BME: black minority ethnic group; HoNOS: Health of the Nation Outcome Scale

Table 5 Unweighted baseline Patient characteristics (Front-Door adjustment #1)

Characteristic	Transferred to inpatient rehabilitation N=197	Removed from inpatient rehabilitation waiting list N=62	Difference in means (Ratio of variances*)
Demographics			
Mean (SD) age in years	40 (12)	41 (10)	-1 (1.3)
% male	73	71	2
% BME	74	71	2
History of service/medication use			
% by number of hospitalizations in the year prior to index hospitalization			
0	46	37	9
1	29	32	-3
2	15	16	-1
3 or more	11	15	-4
Mean (SD) number of days hospitalised in the year prior to the index hospitalization	57 (75)	73 (80)	-16 (0.9)
% with history of community rehabilitation	10	5	5
History of Clozapine use	50	50	0
Characteristics of index hospitalisation between initial hospitalisation and baseline			
Mean (SD) length of stay in days	88 (80)	92 (82)	-4 (1)
Mean (SD) length of stay on General Adult Ward	69 (70)	79 (63)	-10 (1.2)
Mean (SD) number of days hospitalised detained under a Mental Health Act	74 (80)	80 (81)	-5 (1)
% detained at baseline	82	89	-6

% with no/minor/mind, moderate/severe problems or missing HoNOS dimensions score			
01 – Aggression	81, 16, 3	79, 19, 2	2, -3, 1
06 – Hallucinations	58, 40, 3	52, 47, 2	6, -7, 1
09 – Relationships	80, 16, 4	90, 5, 5	-10, 11, -1
10 – Activities of Daily Living	84, 13, 3	81, 16, 3	4, -3, 0
11 – Living Conditions	74, 17, 9	79, 13, 8	-5, 4, 1
12 - Occupation	80, 13, 7	79, 11, 10	1, 2, -3
Mean (SD) days between HoNOS rating and baseline date (Baseline – HoNOS rating date)	27 (50)	23 (39)	4 (1.6)
% with mention of symptom in clinical notes†			
Poor motivation	39	41	-2
Social withdrawal	18	15	3
Self-neglecting or at risk thereof	79	84	-5
% by number of known risk events†			
0	73	63	10
1	18	26	-8
2 or more	9	11	-3

*for continuous and count data only; † between the start of the index admission and baseline or 30 days prior to baseline, whichever is larger

SD: Standard deviation; BME: black minority ethnic group; HoNOS: Health of the Nation Outcome Scale

Table 6 Unweighted baseline Patient characteristics (Front-Door adjustment #2)

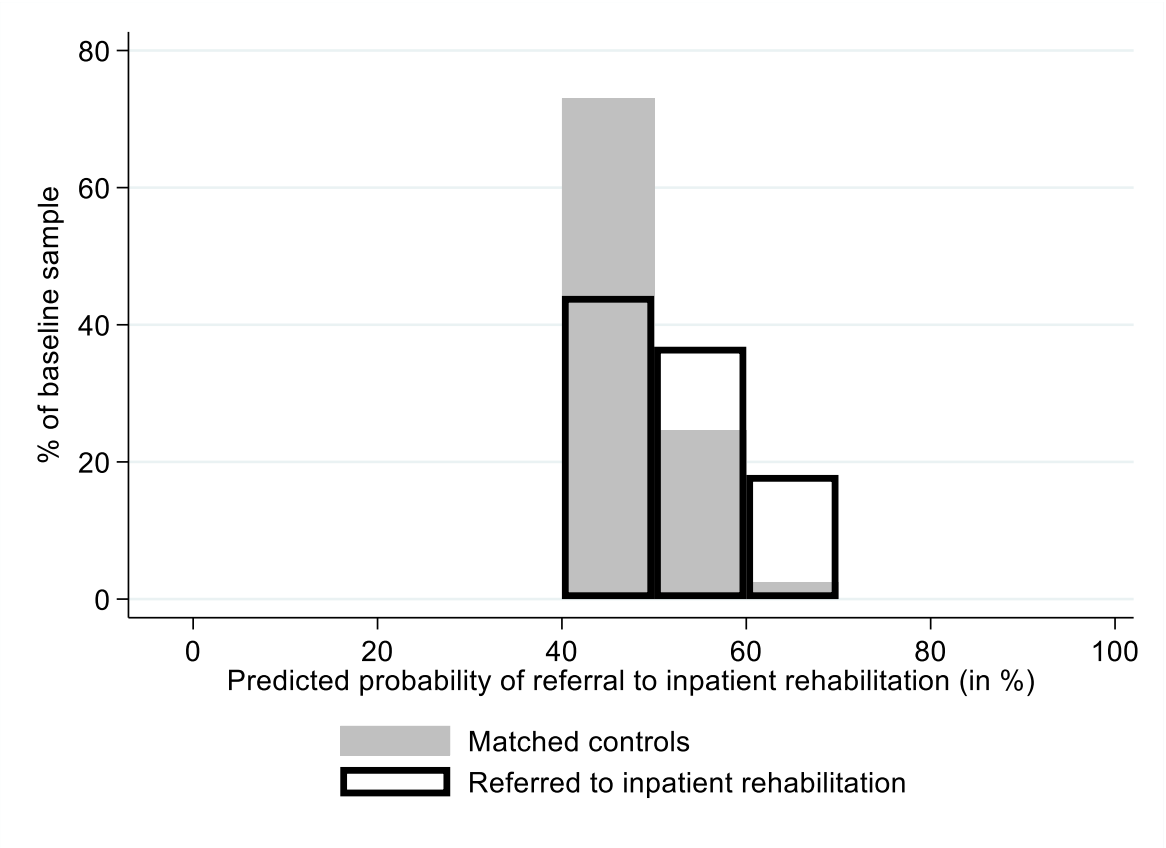


Figure 55 Distribution of predicted probability of referral to inpatient rehabilitation (IR) by treatment group (base case analysis)

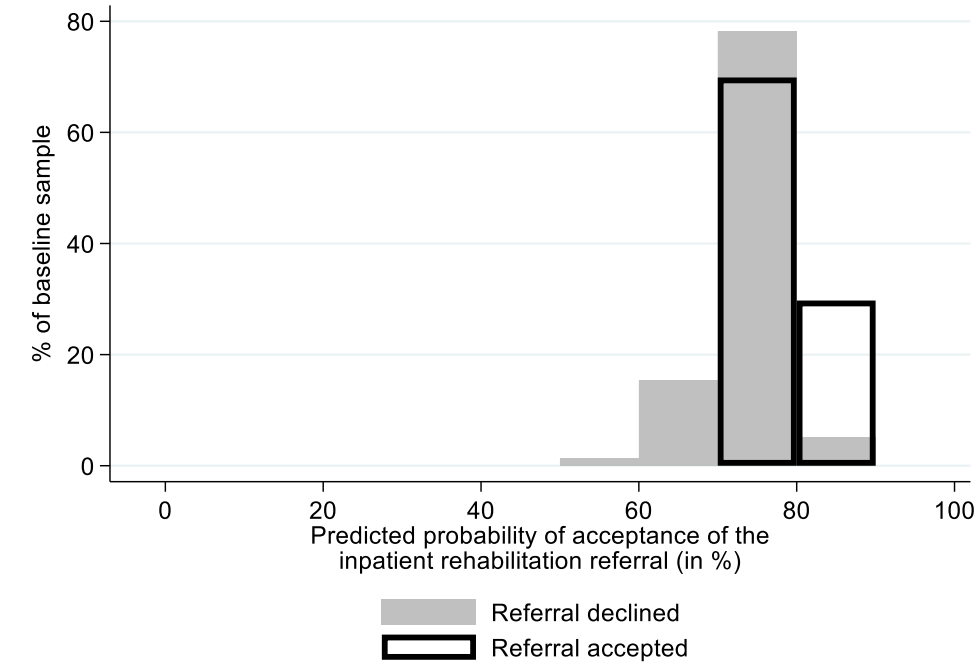


Figure 56 Distribution of predicted probability of acceptance to inpatient rehabilitation (IR) by treatment group (front-door adjustment #1)

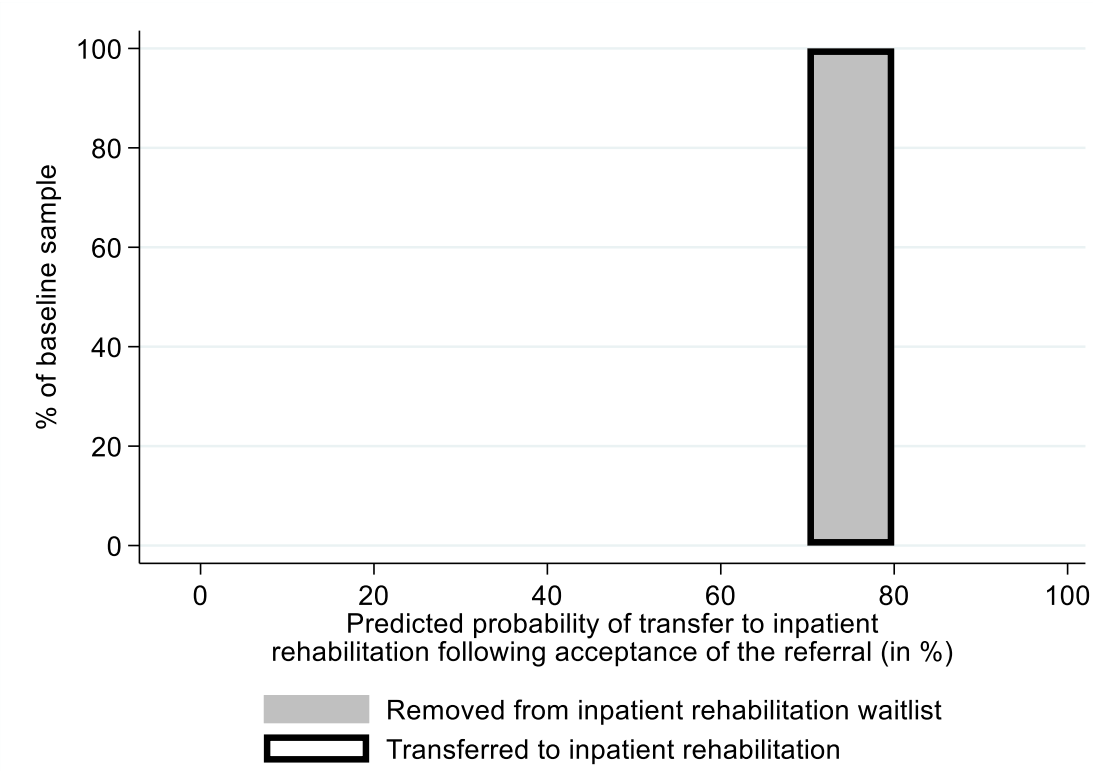


Figure 57 Distribution of predicted probability of referral to inpatient rehabilitation (IR) by treatment group (front-door adjustment #2)

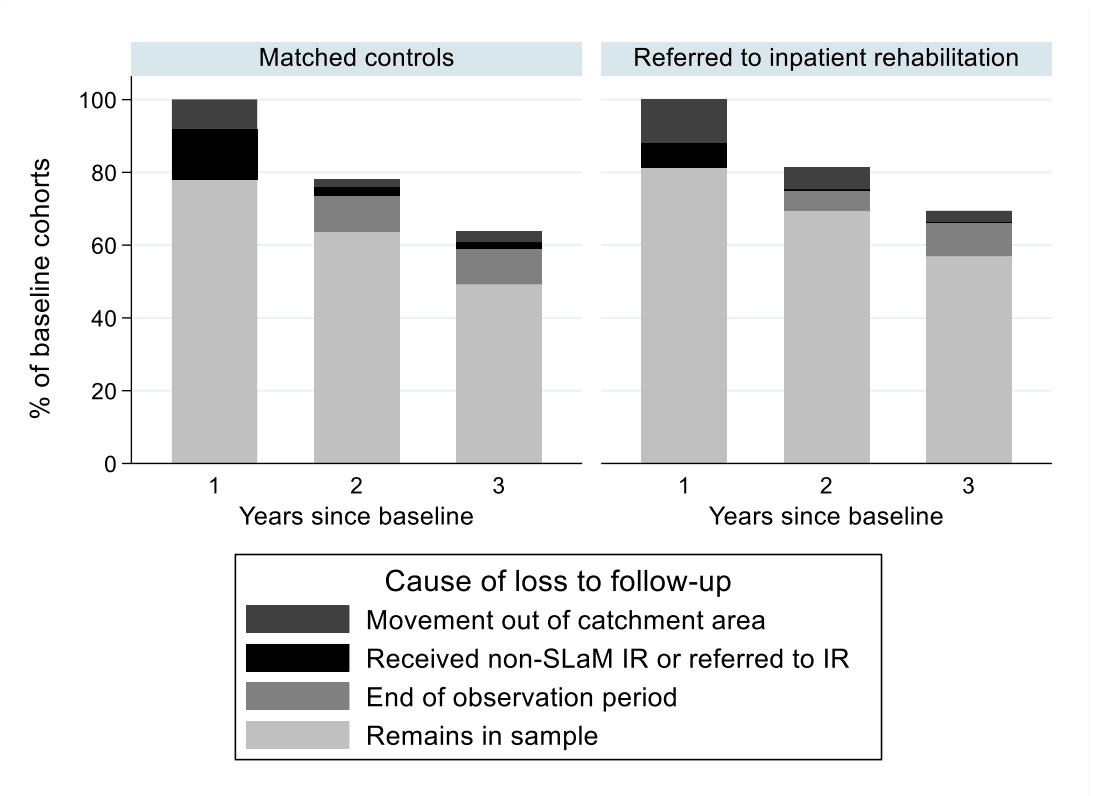


Figure 58 Reasons for loss to follow-up by time point (base case analysis)

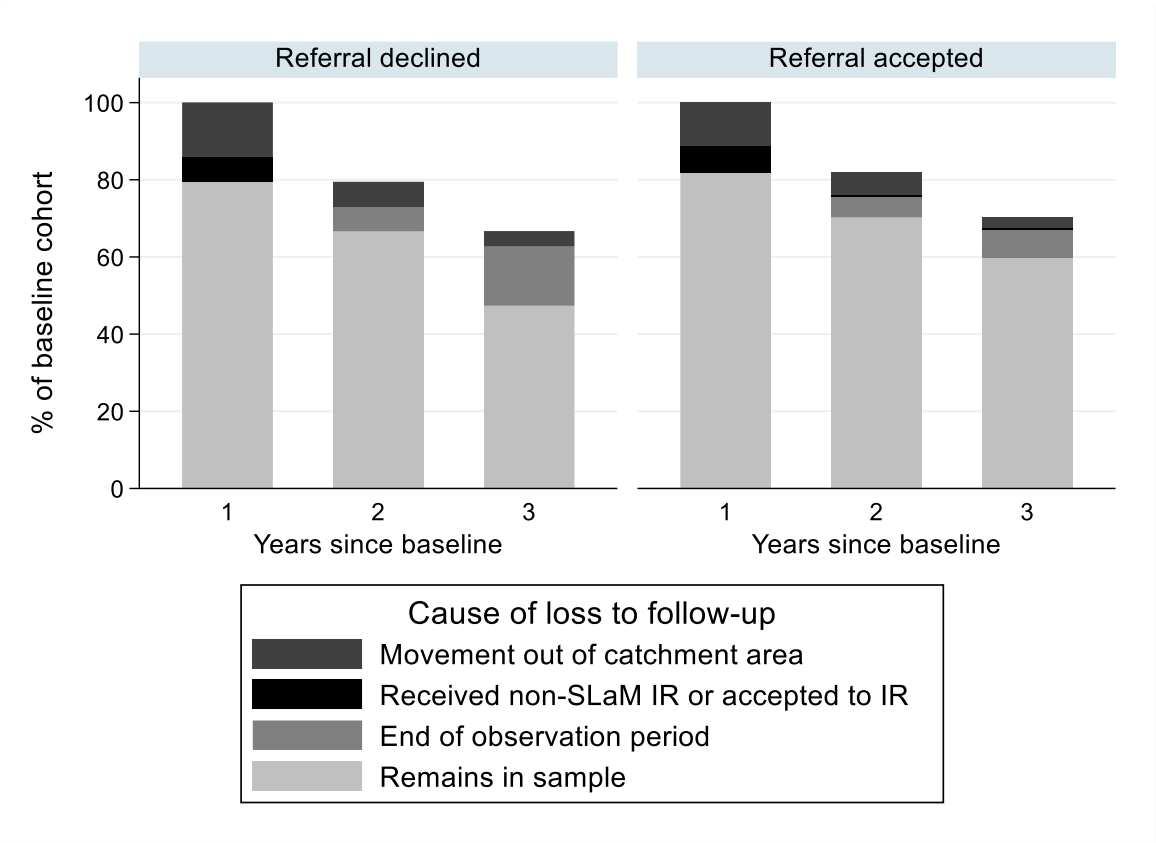


Figure 59 Reasons for loss to follow-up by time point (front-door adjustment #1)

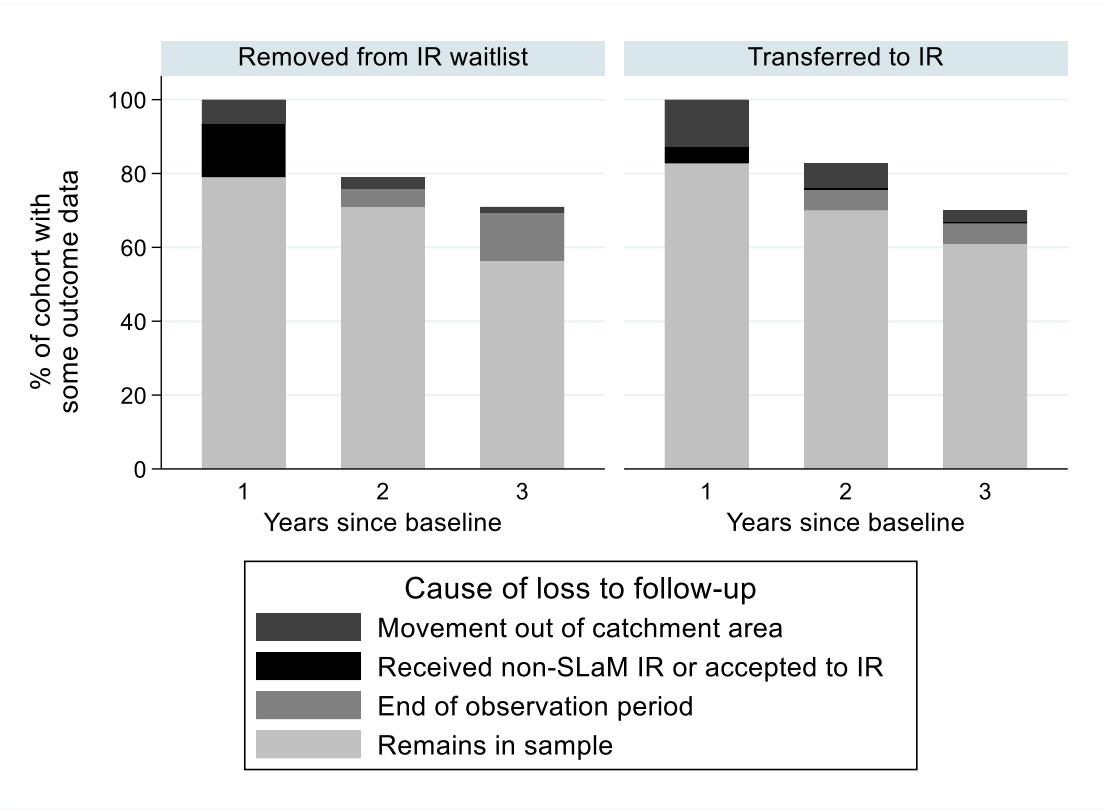


Figure 60 Reasons for loss to follow-up by time point (front-door adjustment #2)

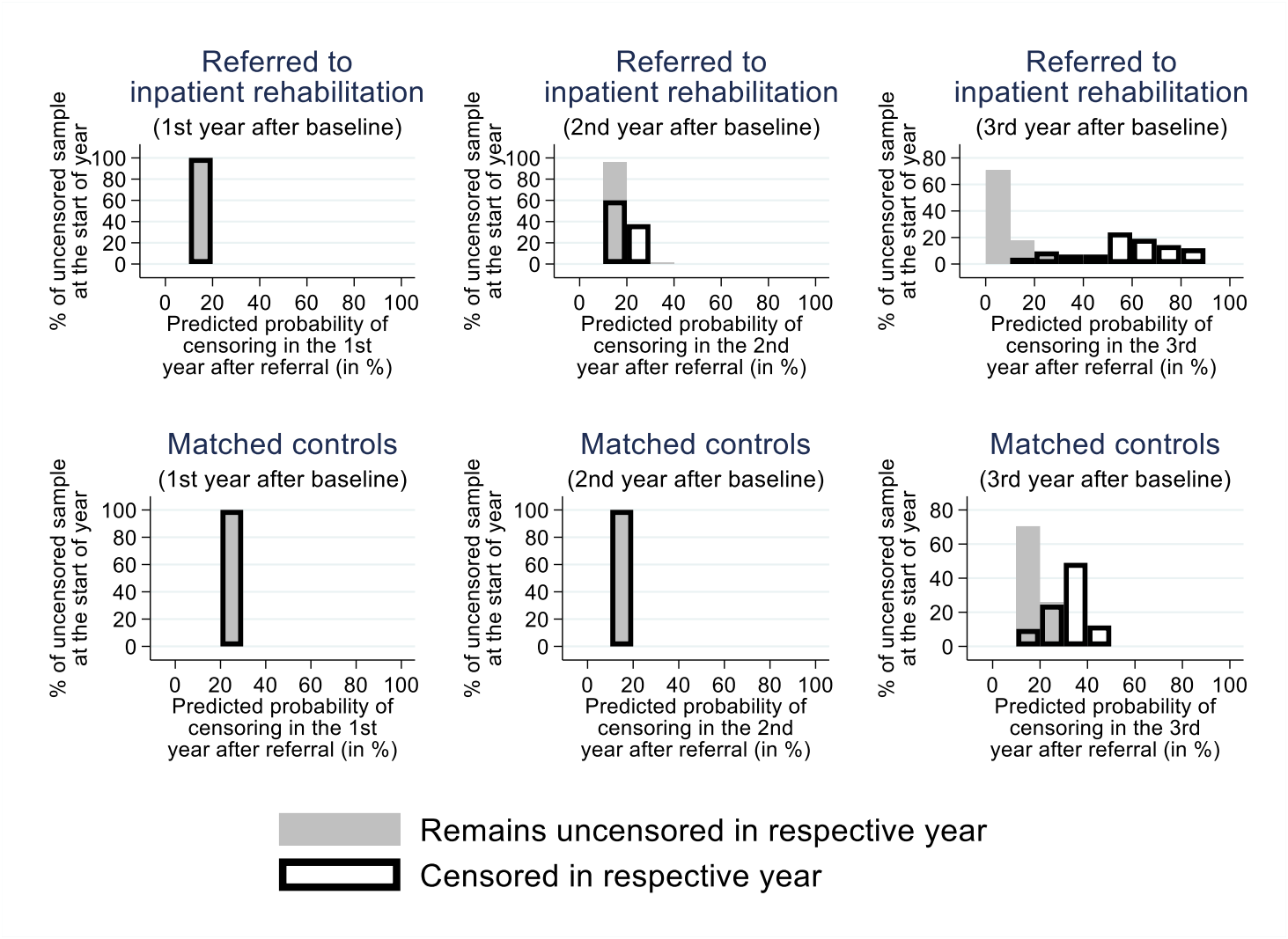


Figure 61 Distribution of predicted probability of censoring by follow-up time point and treatment group (base case analysis)

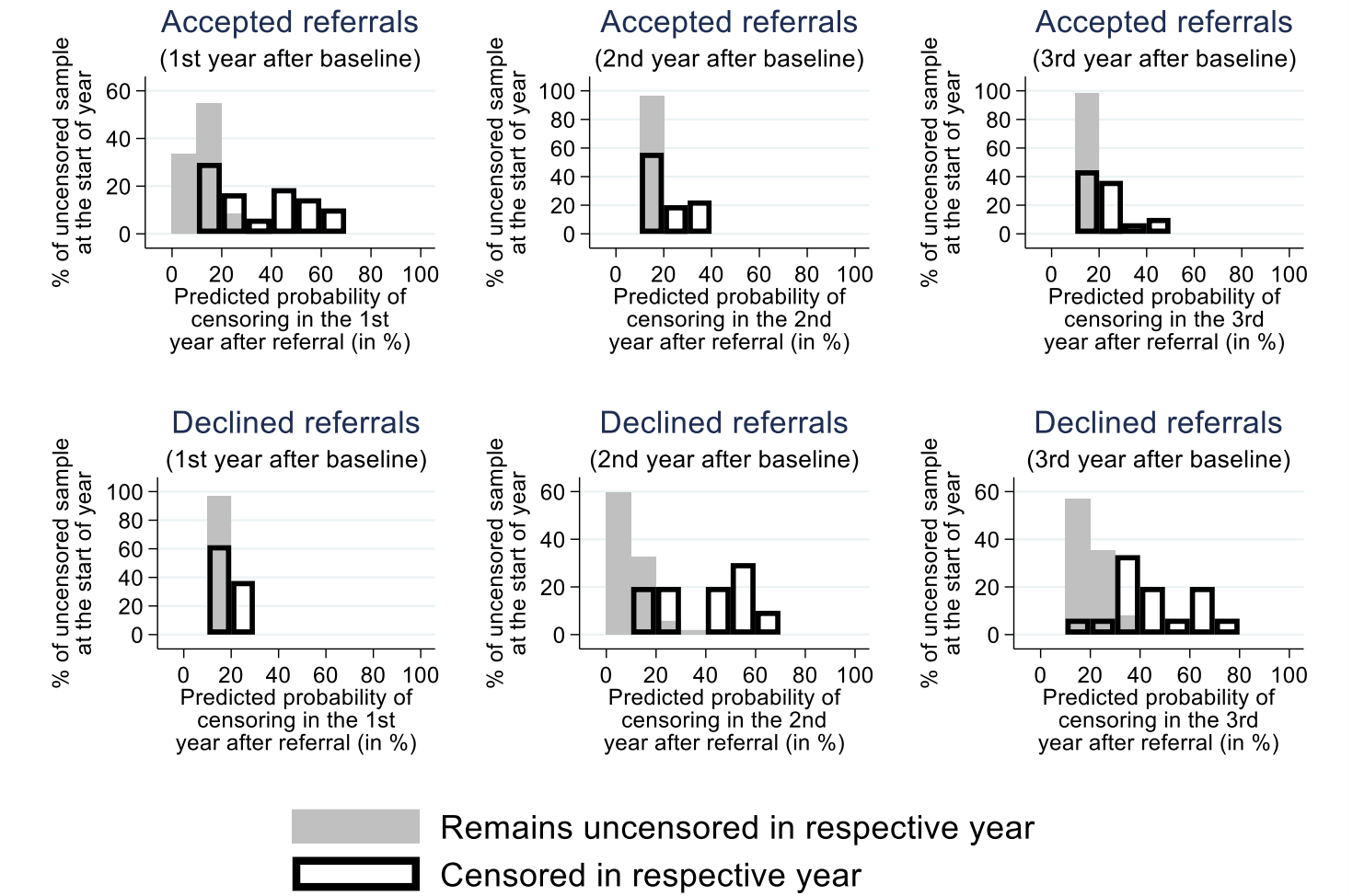


Figure 62 Distribution of predicted probability of censoring by follow-up time point and treatment group (front-door adjustment #1)

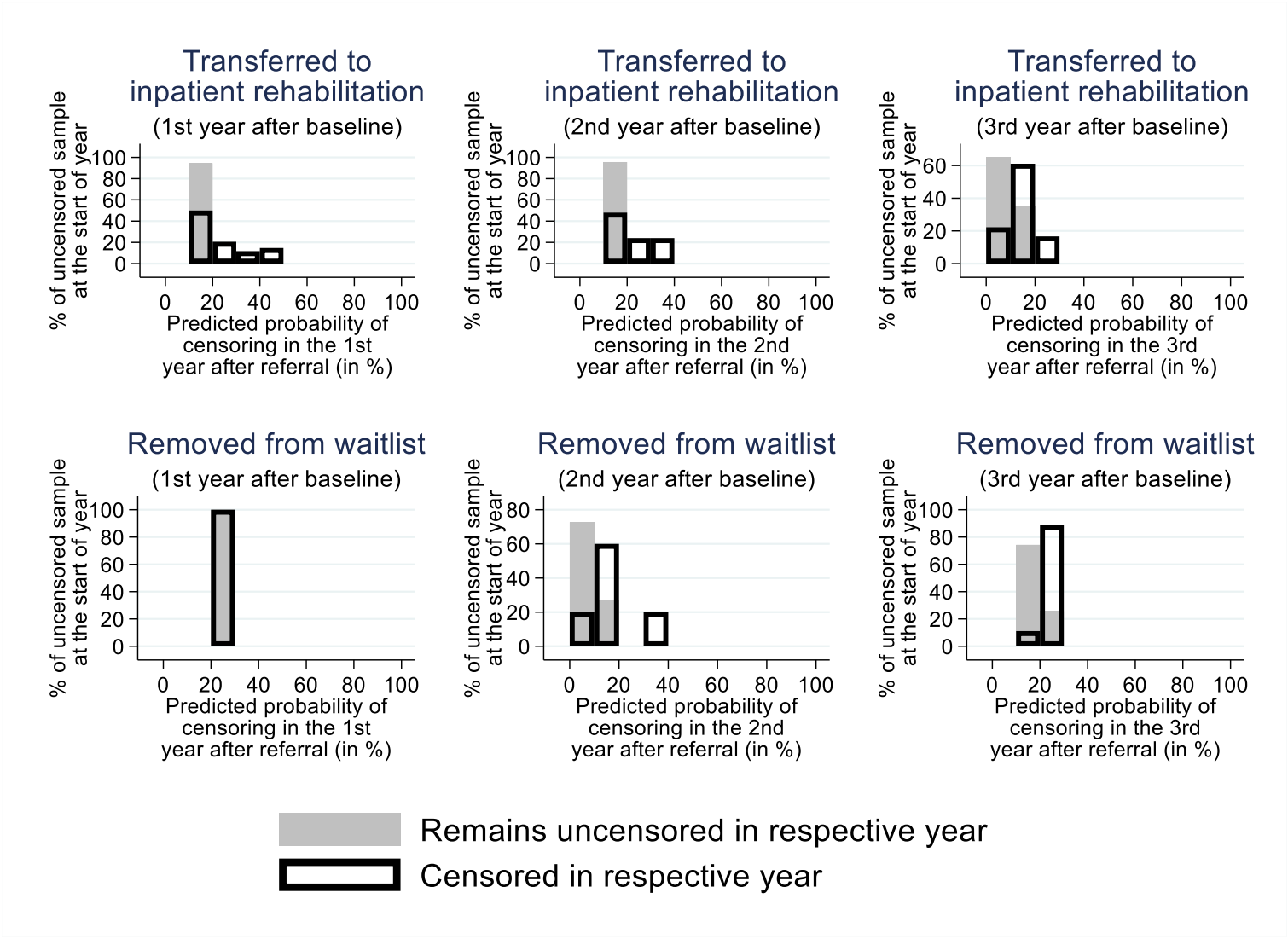


Figure 63 Distribution of predicted probability of censoring by follow-up time point and treatment group (front-door adjustment #2)

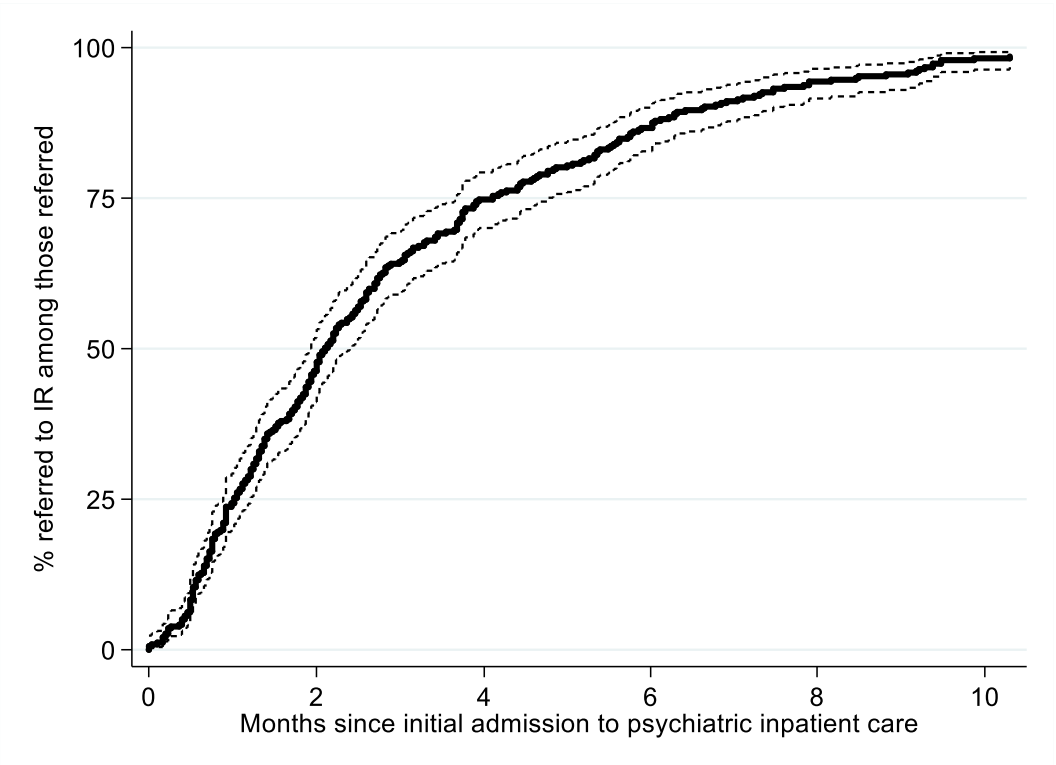


Figure 64 Time from admission to referral

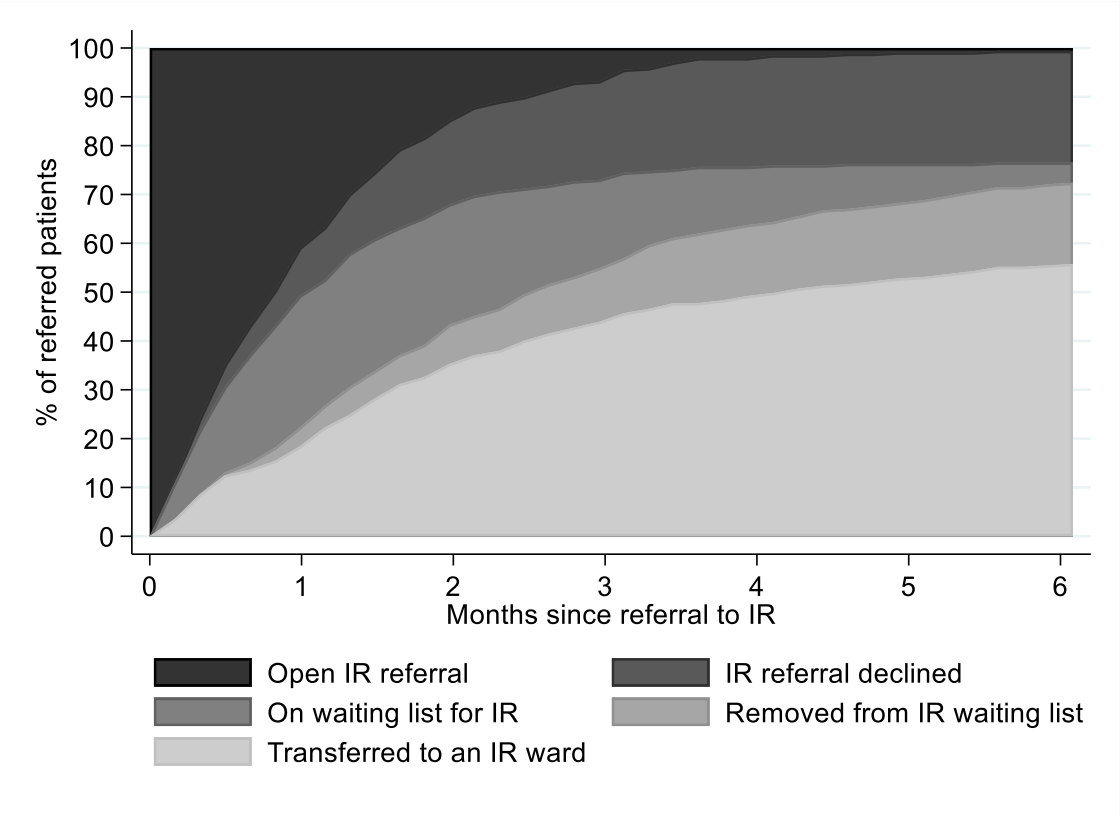


Figure 65 Distribution of patients by stage of inpatient rehabilitation referral pathway over time

Figure 66 Distribution of patients by stage of inpatient rehabilitation referral pathway over time from acceptance

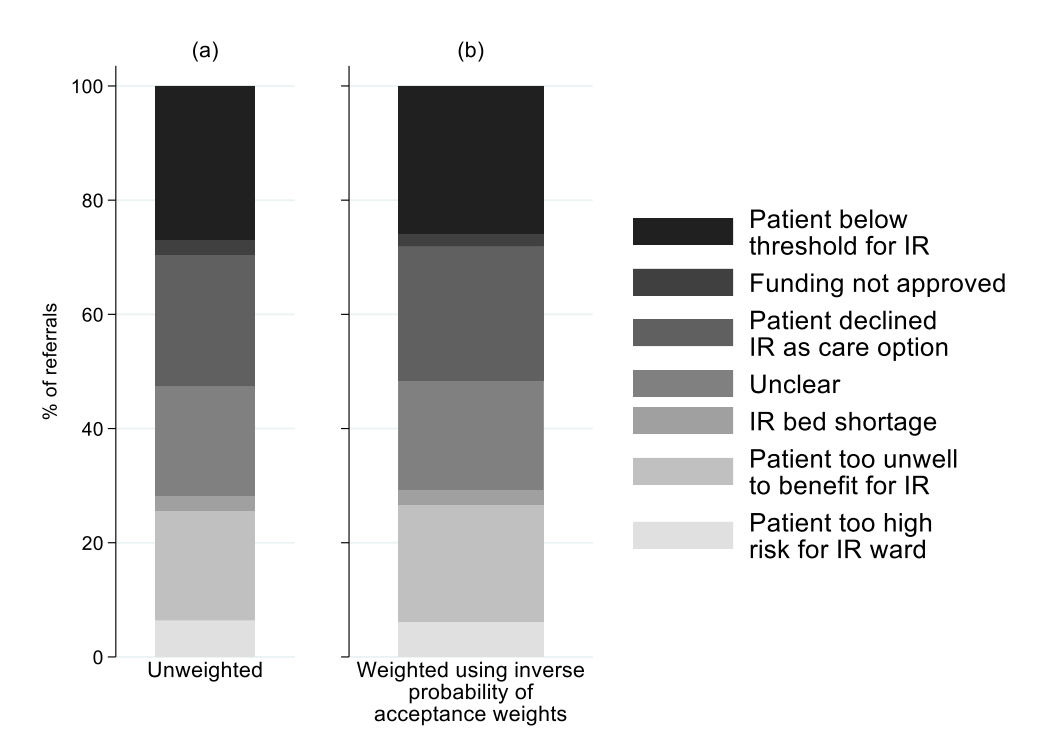


Figure 67 Reasons for declining inpatient rehabilitation referrals

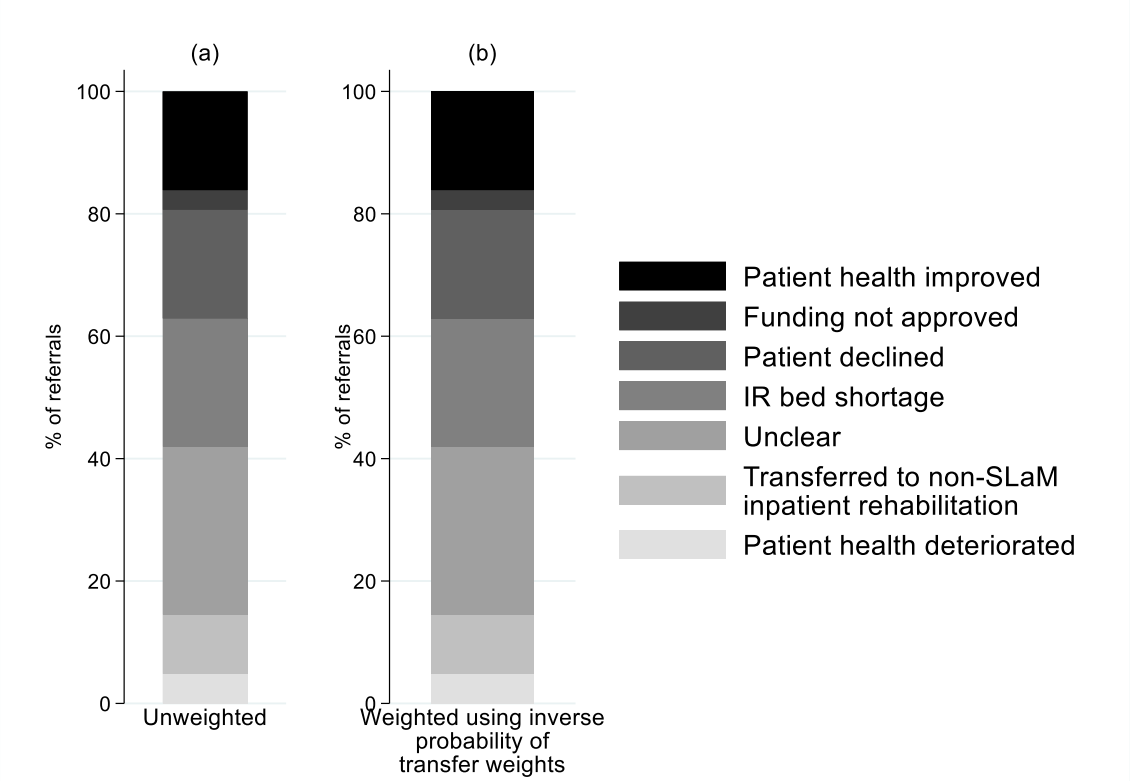


Figure 68 Reasons for removing patients from inpatient rehabilitation waiting list after acceptance

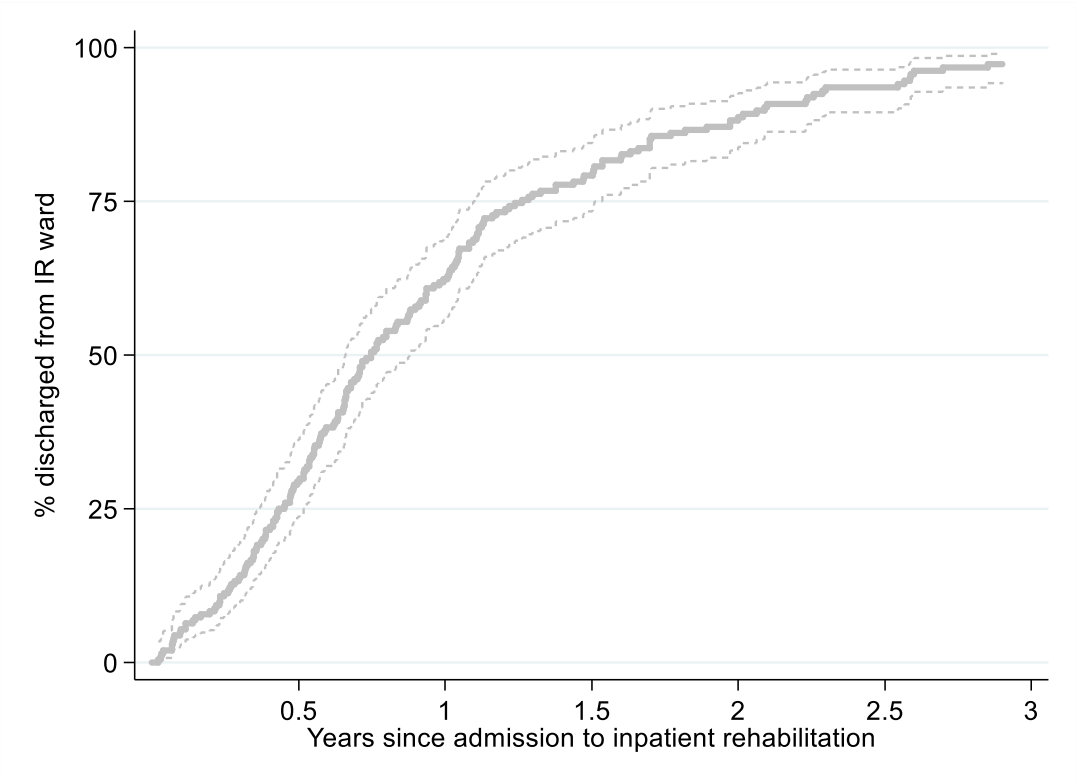


Figure 69 Time from start to end of first inpatient rehabilitation admission after referral

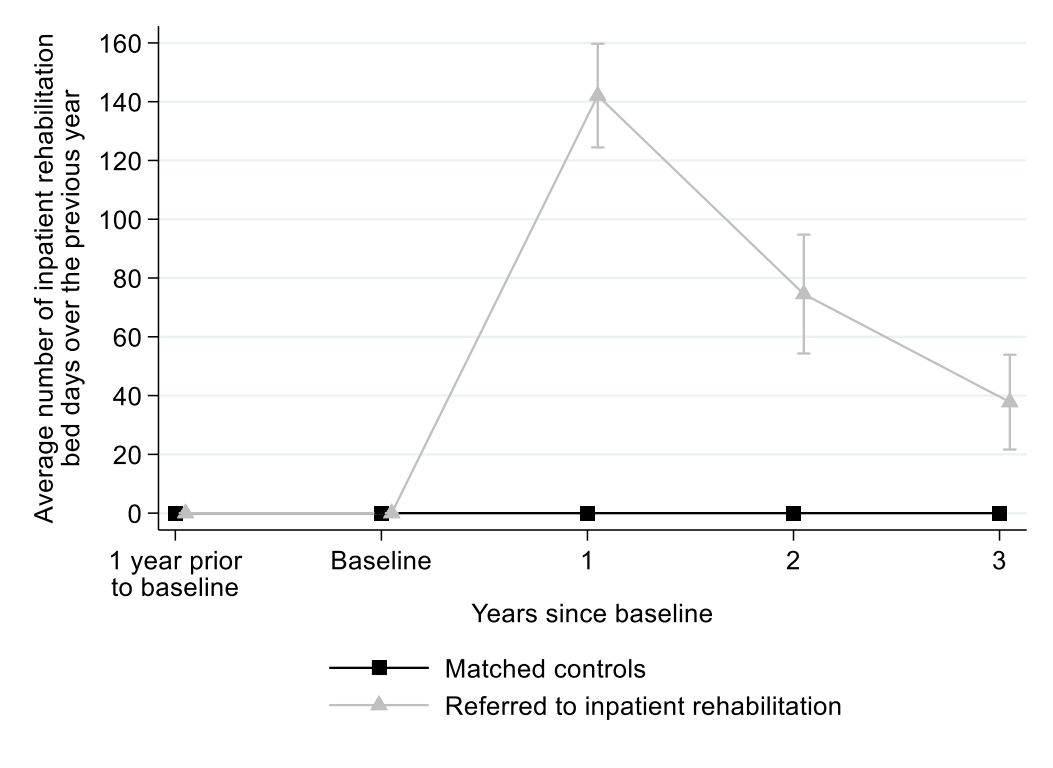


Figure 70 Inpatient rehabilitation bed days by treatment group over time (unadjusted)

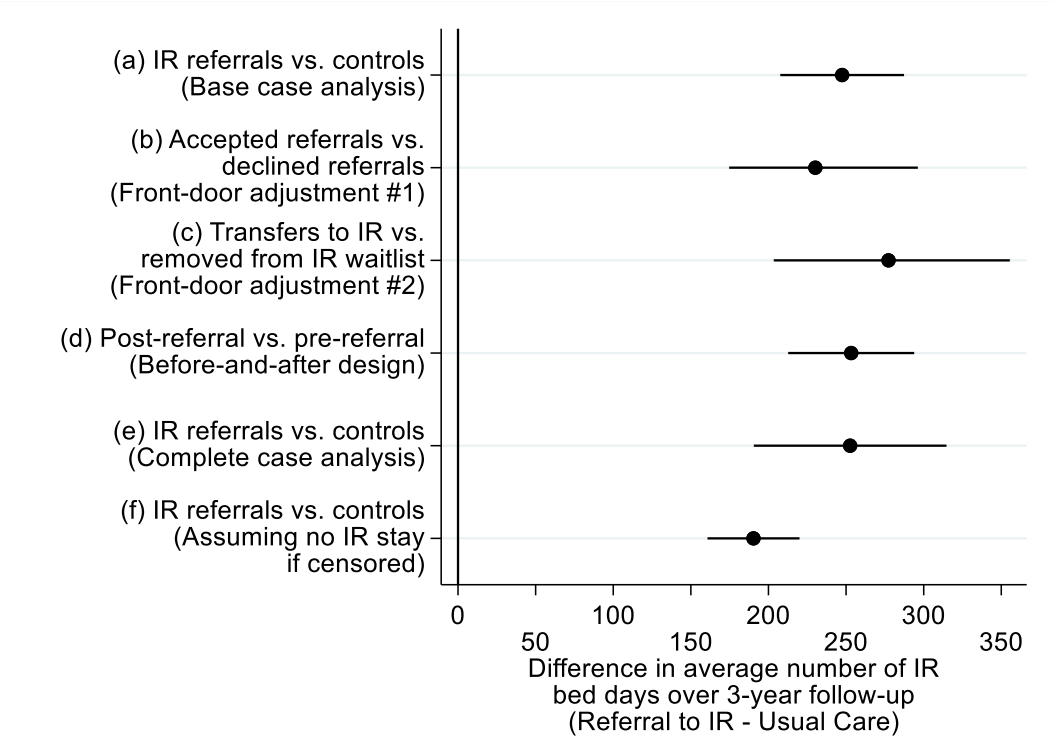


Figure 71 Difference in inpatient rehabilitation bed days

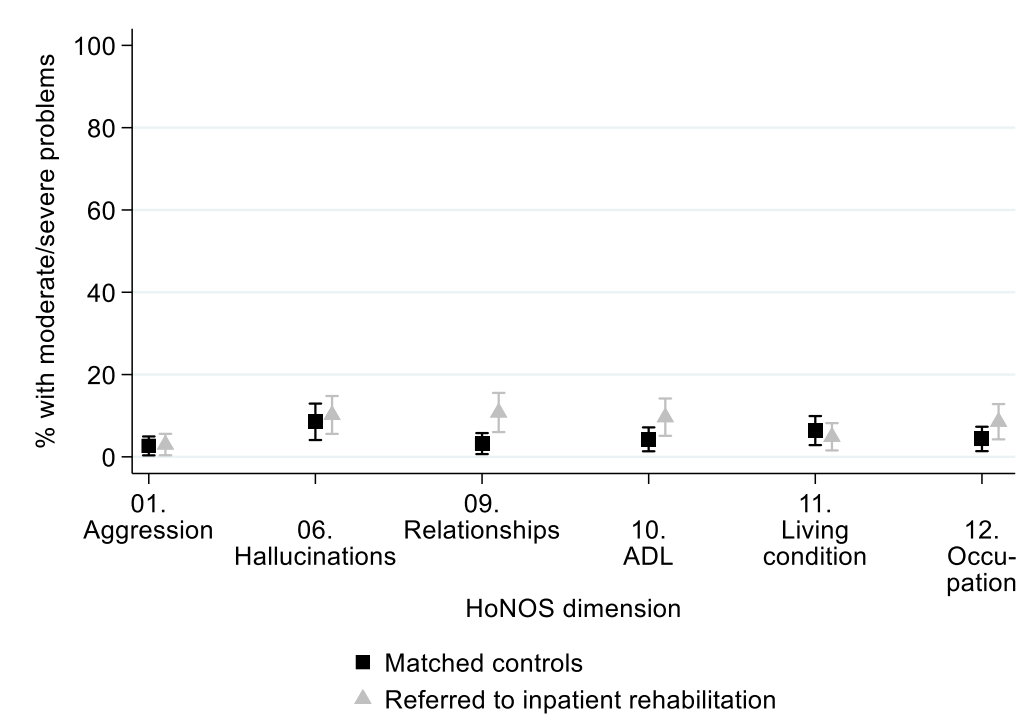


Figure 72 Unadjusted distribution of HoNOS dimension scores at discharge from the index admission by treatment group (unadjusted)

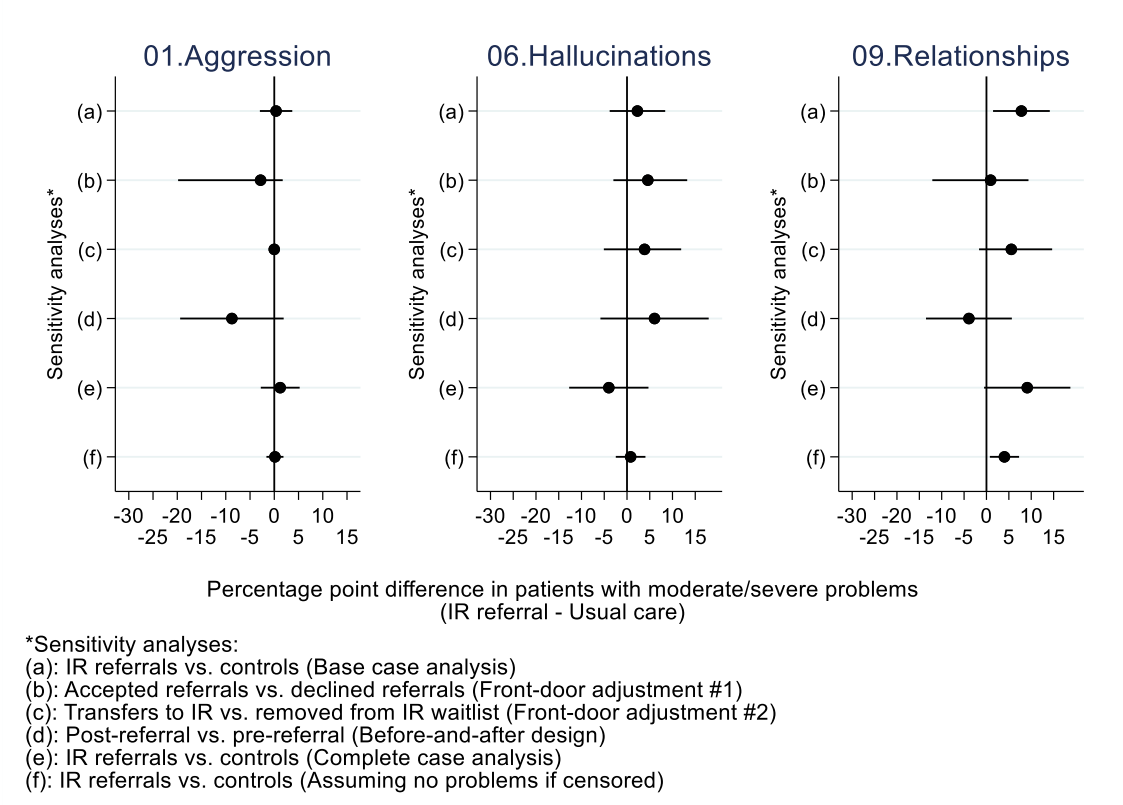


Figure 73 Difference in HoNOS dimension scores at discharge from the index admission (dimension 1, 6 and 9)

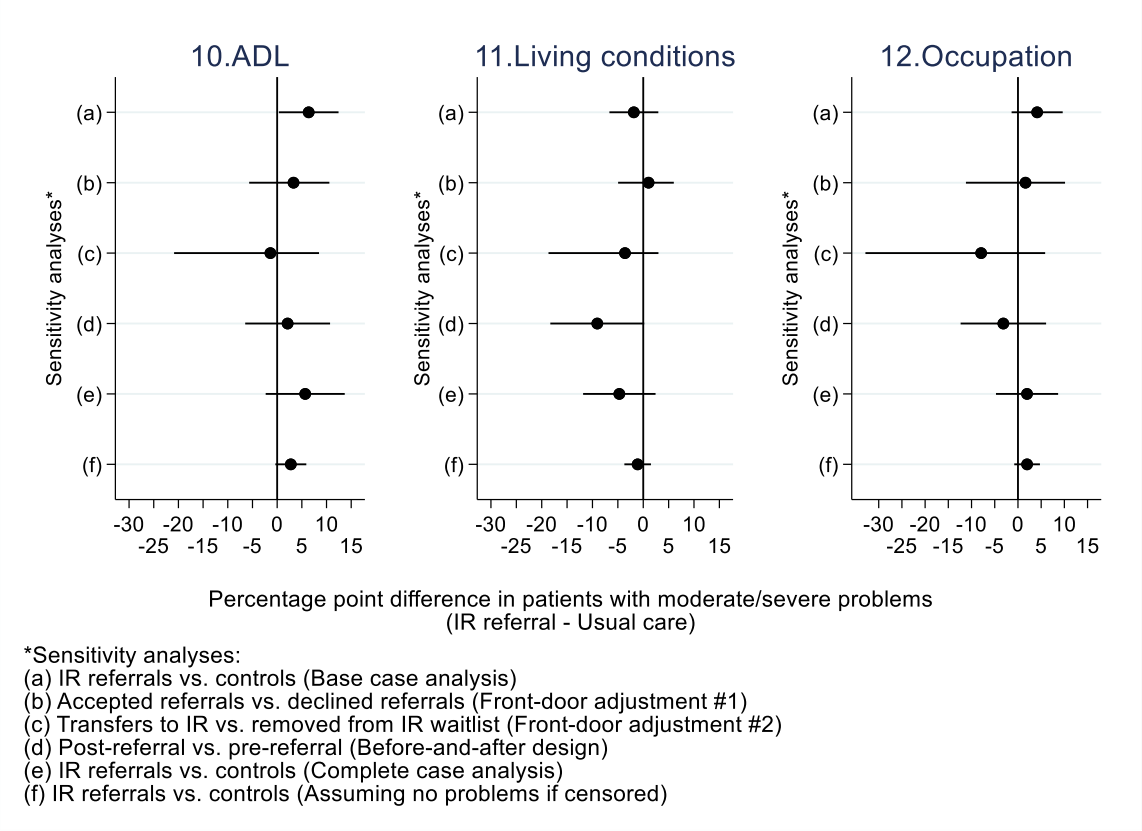


Figure 74 Difference in HoNOS dimension scores at discharge from the index admission (dimension 10, 11 and 12)

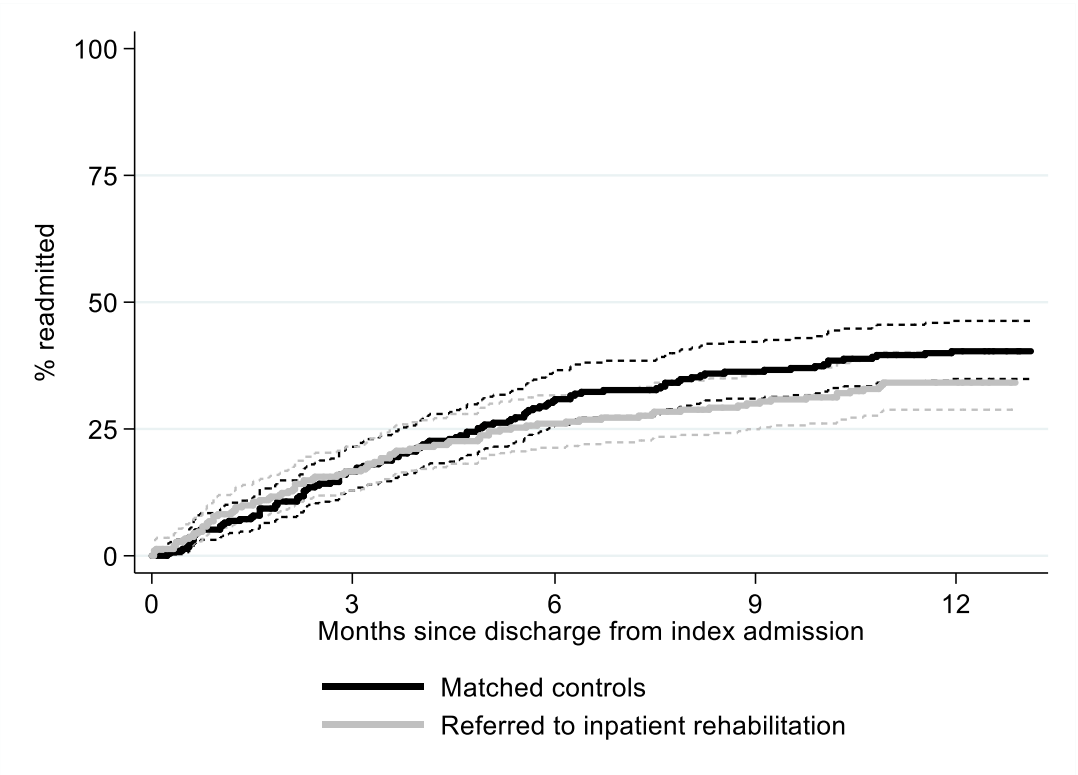


Figure 75 Unadjusted absolute rates of readmission by treatment group (unadjusted)

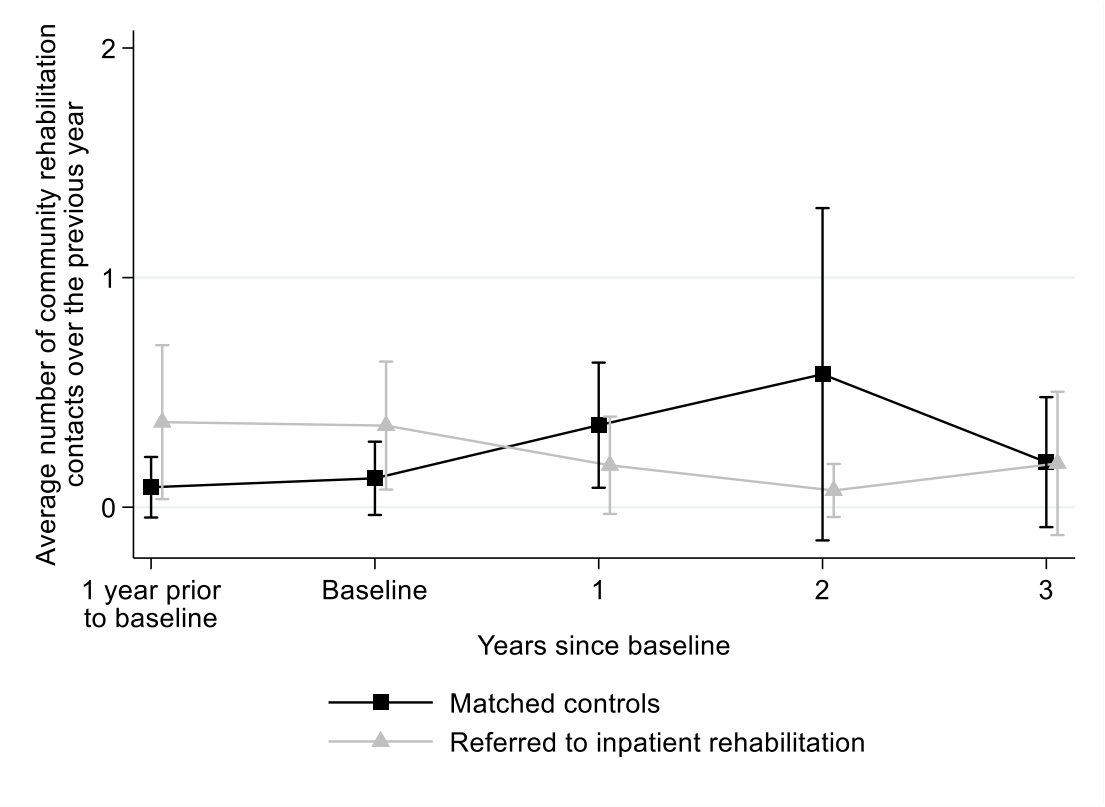


Figure 76 Community rehabilitation contacts by treatment group over time (unadjusted)

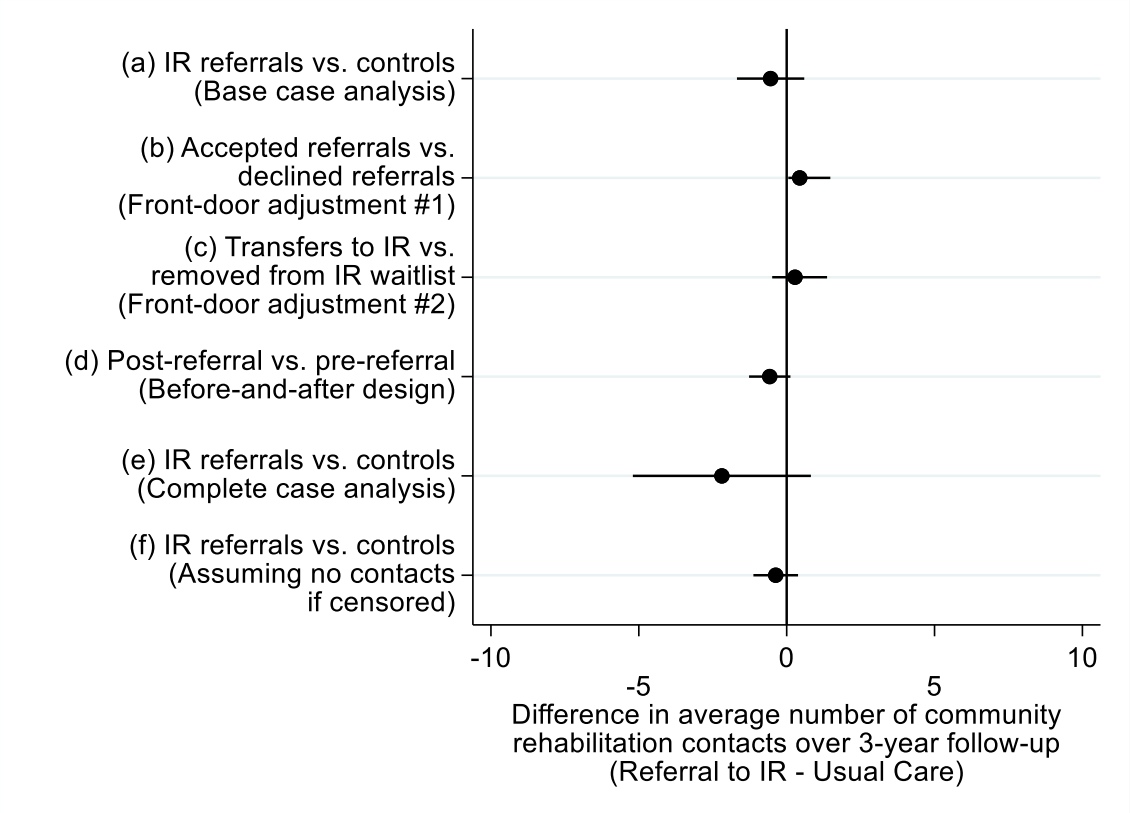


Figure 77 Difference in the number of community rehabilitation contacts

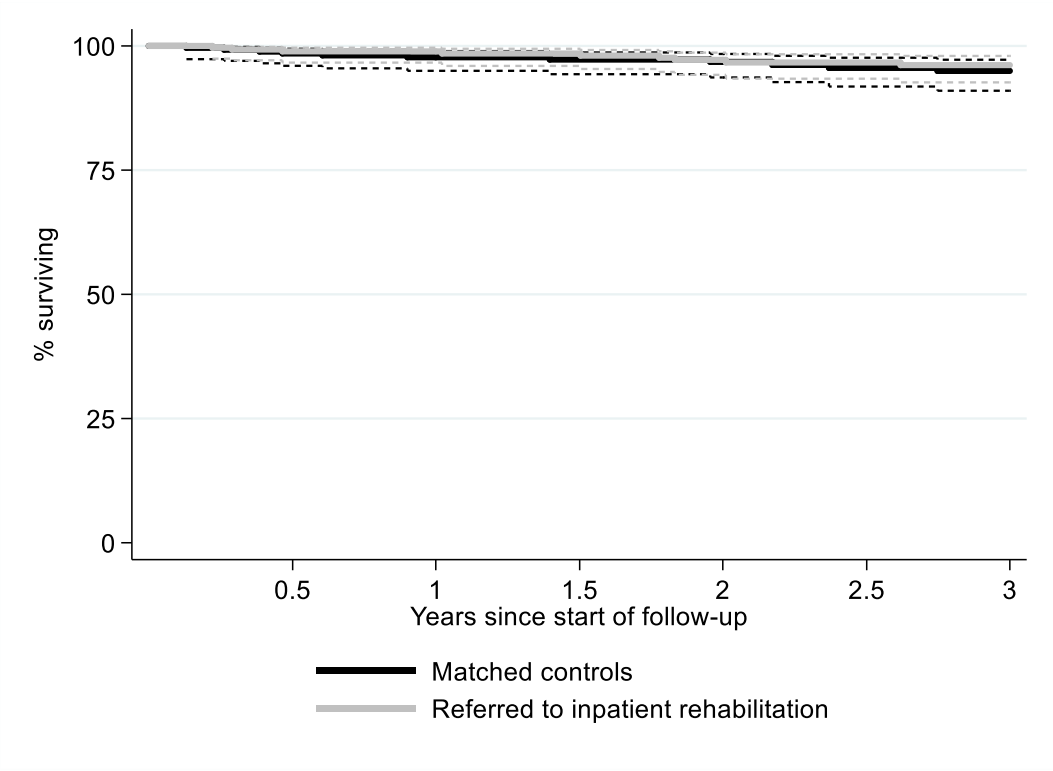


Figure 78 Unadjusted probability of survival over time (unadjusted)

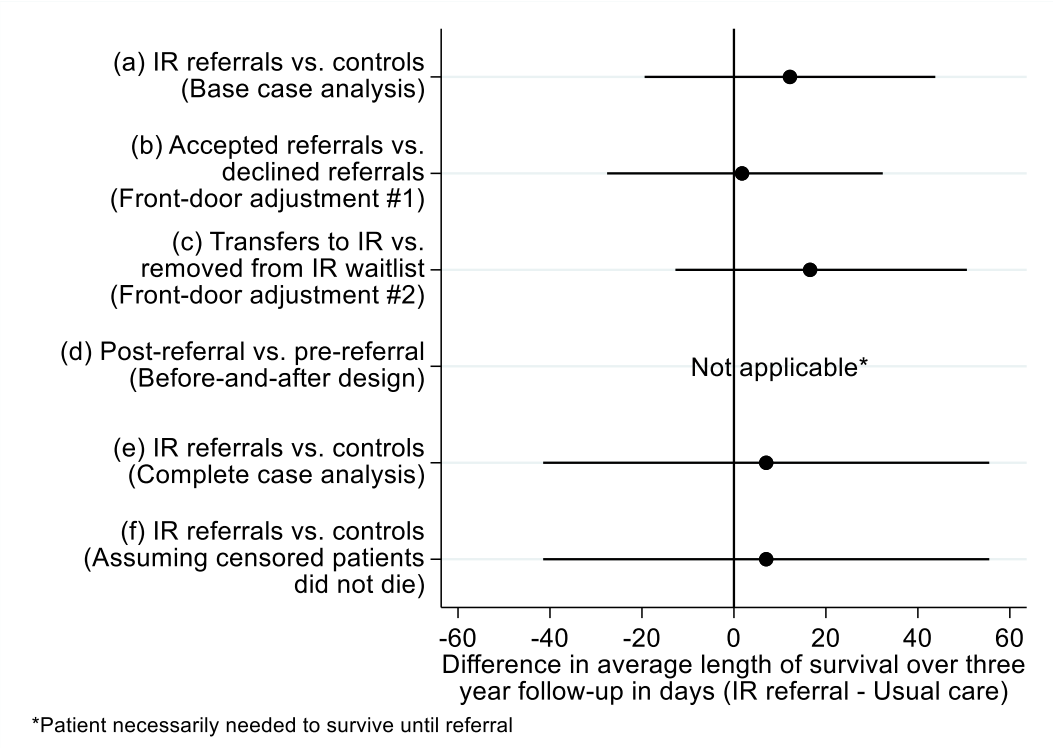


Figure 79 Difference in length of survival

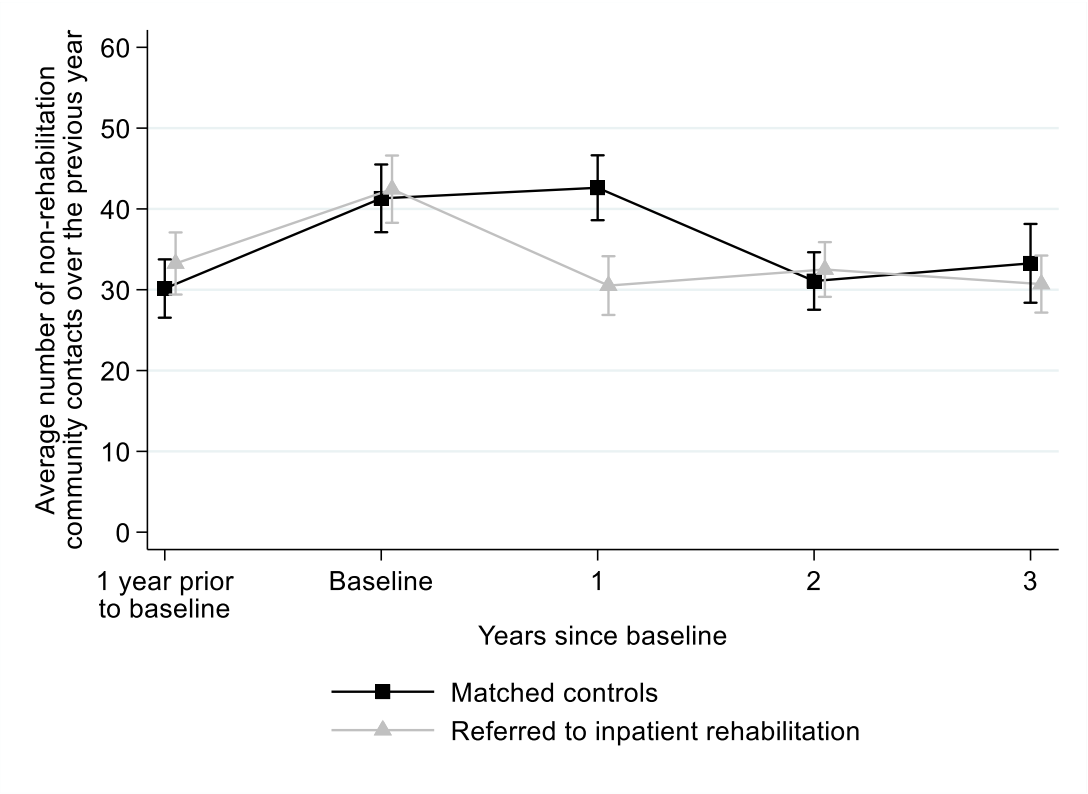


Figure 80 Non-rehabilitation community contacts over time (unadjusted)

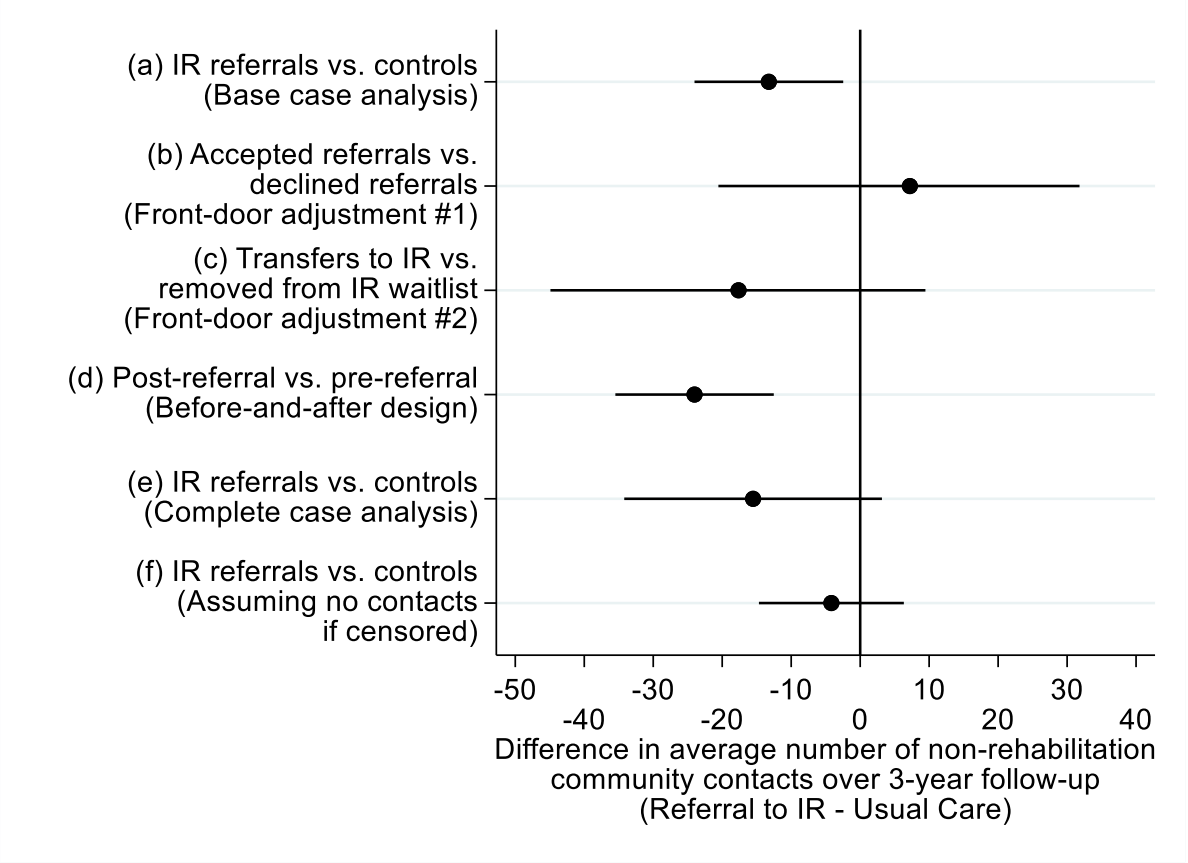


Figure 81 Difference in non-rehabilitation community contacts

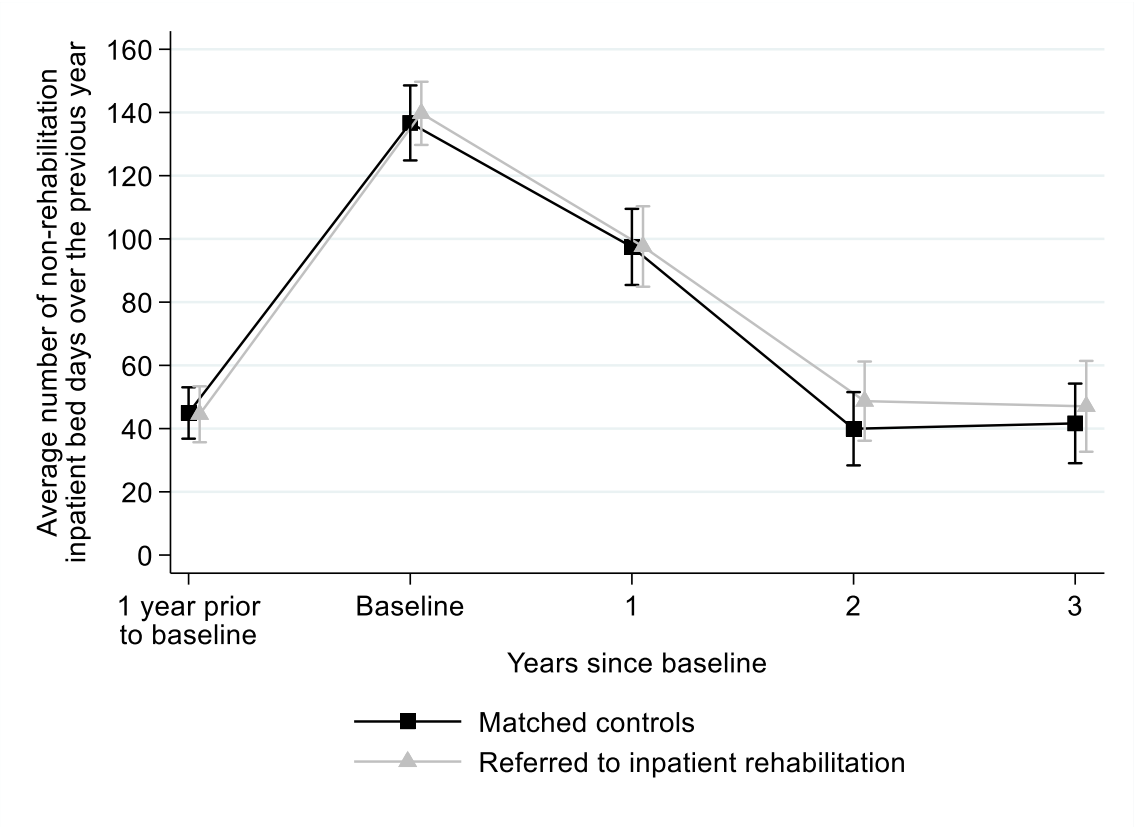


Figure 82 Non-rehabilitation psychiatric inpatient care use over time (unadjusted)

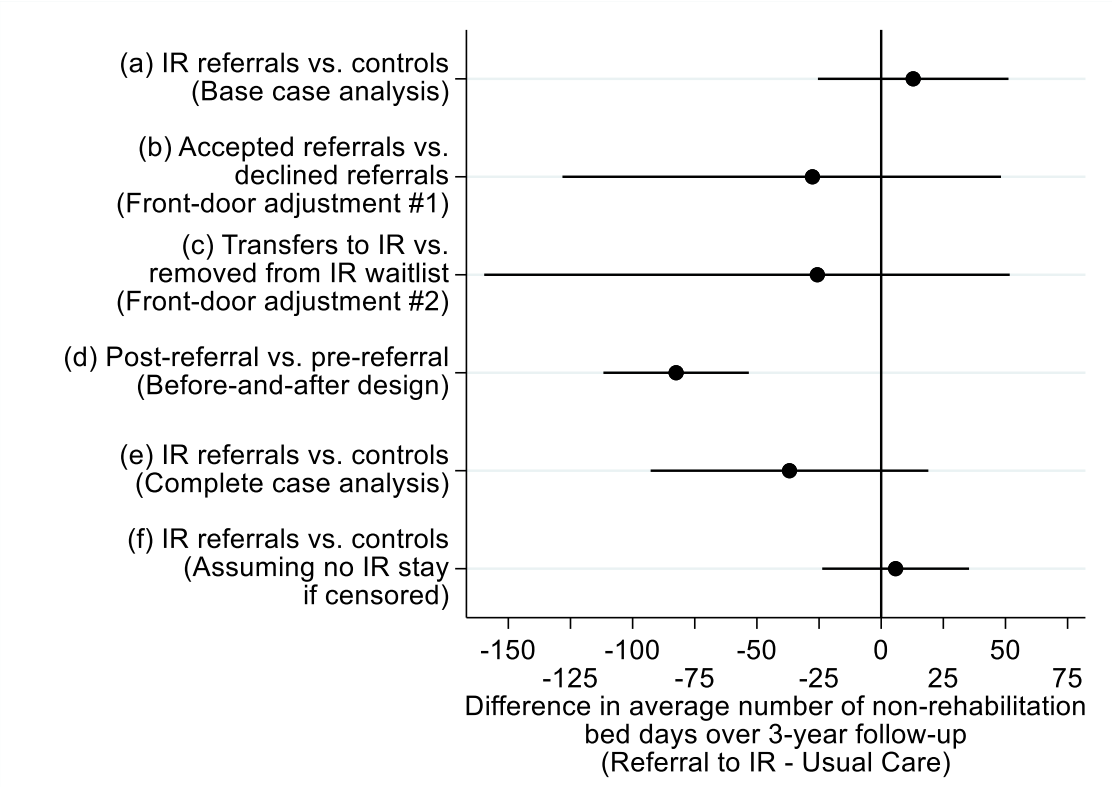


Figure 83 Difference in non-rehabilitation inpatient care use