Practical methodology for missing data handling in interrupted time series analysis

Juan Carlos Bazo Alvarez

Thesis submitted for the degree of Doctor of Philosophy

UNIVERSITY COLLEGE LONDON (UCL)

I, Juan Carlos Bazo Alvarez, confirm that the work presented in this thesis is my own.

Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

This PhD is partially supported by FONDECYT-CONCYTEC (grant contract number 231-2015-FONDECYT). The National Fund for Scientific, Technological Development and Technological Innovation (FONDECYT) is part of the National Council for Science, Technology and Technological Innovation (CONCYTEC), which is a Peruvian governmental agency.

Abstract

Interrupted time series (ITS) is a quasi-experimental design for evaluating the effect of an intervention or treatment by comparing the outcome trajectory over time before and after initiation of the intervention. ITS became popular for evaluating interventions at the population level (e.g. policies); thus, the development of statistical methods was mainly orientated to modelling population-level data. This thesis aims to explore the issues that emerge when population-level ITS analyses are applied to incomplete individual-level data in health research, proposing alternative analysis methods.

First, I performed a scoping review to demonstrate how the issues of missing data at the individual level have rarely been addressed in most recent ITS studies. Despite its limitations, complete case analysis is the most frequently used method for handling missing data. Individual-level data are usually transformed into population-level time-specific summaries before fitting ITS models. This method can lead to bias. Mixed effect models (MEM) can solve this, but my review demonstrates few studies have done so in the past.

I then fitted MEM to study body weight gain induced by the initiation of antipsychotics using an ITS design on electronic health records. MEM allowed fully observed covariates to inform the implicit imputation of the outcome. ITS facilitated new clinical evidence: in the long-term, typical patients do not lose the weight they gained during the first six weeks of treatment. However, the MEM alone was not ideal for handling missing covariates (i.e. dosage).

Thereafter, I used simulation studies to evaluate the performance of aggregate-level segmented regression (SR), MEM and multilevel multiple imputation (MI-JOMO) for handling data missing at random (MAR) in ITS analysis. I showed that the aggregate-level SR can over or underestimate the ITS effect. MEM is effective for handling outcomes MAR, but it should be combined with MI-JOMO when covariates are also MAR.

Finally, I applied MEM with MI-JOMO to assess how dose and age modify the antipsychotic-induced weight gain. Interaction terms in MEM helped to evaluate differences in weight trajectories over time between groups by dose or age, using MI-JOMO for handling missing dose. Clinically, older people's weight is less affected by the initiation of antipsychotic treatment than younger people's.

Impact Statement

In health research, there is an approach commonly used to evaluate how an intervention (e.g. a new health policy) or a treatment (e.g. a drug) can impact on a health outcome of interest (e.g. change in body weight). This approach is called interrupted time series (ITS) analysis and involves the use of observational outcome data before and after the initiation of the intervention. If the health outcome changes in the after-period, this change can be attributed to the intervention evaluated.

The ITS analysis requires some statistical tools for ensuring the validity of its conclusions. These tools have been gradually improved since the past century. However, these tools have not been designed to address the missing data issues that routinely collected data usually bring (e.g. digital data from primary care). In other words, to use the standard statistical methods for ITS analysis can lead to wrong conclusions when incomplete individual-level data is analysed, which is often the case in current health research.

In my thesis, I formally evaluated this problem and studied alternative statistical tools for performing valid ITS analyses when researchers face these missing data issues. For doing the evaluation, I used data from the UK primary care system to investigate the weight gain induced by second-generation antipsychotics (SGA) (i.e. an unhealthy side effect). This applied research helped me to build a practical solution in a real-life research situation. Methodologically, I proposed a two-step statistical procedure to perform ITS analyses with missing data. Clinically, I found that people prescribed SGA usually do not lose the weight they gained in the first 6 weeks, that high doses of SGA can induce more weight in the short and long-term, and that older people's weight is less affected by SGA than younger people's. These clinical conclusions were possible thanks to the new procedure for ITS analysis I proposed here.

The methodological contribution is going to impact the field of health research. The research community have now a new alternative to perform ITS analyses with routinely collected data, an alternative that is easy to use and has been tested with simulations and real-life data (ensuring its validity and usefulness). Now, more clinical research on individual-level treatments will be possible, even when missing data issues are present (a common scenario). Moreover, I am starting the conversation on why these alternative tools should be preferred over classical ITS tools when individual-level data is available. This conversation [1] will lead to new methodological researches in ITS analysis that will be beneficial even beyond the health research field; for example, in economic and social research.

The clinical contribution will impact directly on medical practice. Patients and doctors will benefit with new relevant information when decisions on SGA prescriptions are taken. Doctors will better balance the pros and cons when prescribing SGA high or low doses. Guidelines of weight control for patients treated with SGA will be adapted to the new knowledge about differences between older and younger patients.

[1] This thesis has already led to several presentations, publications and grant funded projects, which are visible in Appendix 10.5

Outline

Acknowledgements

My PhD has been a fantastic journey full of learning experiences, which could not have been possible without the generous support of many people.

Irene Petersen, my principal supervisor, has given me permanent and very dedicated support, with enough liberty to try new ideas but keeping me focused on the final goal. Thanks to Irene's thoughtful guidance, encouragement, and constructive feedback, I have completed my thesis successfully. She started as my supervisor and, with the pass of time, she became my mentor. For that, I will always be grateful.

James Carpenter, my subsidiary supervisor, has been extraordinary with his input and guidance. Each interaction with him has enriched the methodological component of my work enormously, giving me not only -brilliant- answers but also new questions to think about. I respect James much, so to work with him has been an invaluable and inspiring opportunity.

Tim Morris, my subsidiary supervisor, has been entirely outstanding with his support. Tim helped me with several issues on conducting simulations studies, as well as answering technical questions and providing fantastic feedback during all the PhD. He has been not only tremendously generous with his time but also a very positive and encouraging ally during this journey.

I like to thank Joseph Hayes and Kingshuk Pal for their great feedback on the clinical component of this thesis, Matteo Quartagno for his support on the correct use of the MI-JOMO program, and Frank Peralta for his help on the searching step during the scoping review.

The Research Department of Primary Care & Population Health (PCPH) at UCL has been a welcoming and vibrant research environment. Special thanks go to Tra My Pham, a great statistician and wonderful person, for sharing her insights and expertise, and for being so supportive during her time at PCPH.

I would also like to thank Holly Smith, Brendan Hallam, Karan Mehta, Elaine Flores, Rodrigo Carrillo, and Frank Peralta for their help with proofreading my thesis.

The support and love of my family and friends have been essential for completing this journey. I am blessed with the love of my wife, Ada Carolina, who has been my every-day pillar in a dimension I could not express with words. In the distance, my brothers and best friends Oscar and Mauro kept me full of love and positive energy to never miss home. Here in London, Elaine, Jean Paul, Rodrigo, Timesh, Josina, Cris, Paz, Nacho, Dani, Manuel, Murilo and Gauthier helped me to never forget how wonderful is to laugh and perform music with friends.

Finally, this thesis is dedicated to Yolanda and Alejandro. *En todo tiempo y lugar nunca habría tenido mayor suerte que tenerlos como padres, y por ello, doy gracias a Dios cada día. Los amo.*

Contents

List of figures

List of tables

Abbreviations

ACU	Acceptable computer usage
AMR	Acceptable mortality reporting
AP	Antipsychotic(s)
ARIMA	Autoregressive integrated moving average
ATC	Anatomical Therapeutic Chemical code
BMI	Body mass index
BNF	British National Formulary
CCA	Complete case analysis
CITS	Controlled Interrupted Time Series
CPRD	Clinical Practice Research Datalink
CVD	Cardiovascular
DGM	Data Generation Mechanism
Dx	Diagnostic
EHR	Electronic Health Records
FCS	Full Conditional Specification
GEE	Generalised estimating equations
GLM	Generalised linear model
GLMM	Generalised linear mixed models
GLS	Generalised least squares
GMS	General medical services
GP	General practice
HDL-cholesterol	High-density lipoprotein cholesterol
ICC	Intraclass Correlation Coefficient
IPS	In Practice Systems
ITS	Interrupted Time Series
LDL-cholesterol	Low-density lipoprotein cholesterol
LR test	Likelihood-ratio test
MAR	Missing at random
MCAR	Missing completely at random
MEM	Mixed effects models/modelling
MI	Multiple imputation
MICE	Multiple imputation by chained equations
MI-JOMO	Substantive model compatible multilevel joint model multiple imputation
ML	Maximum likelihood
MMI	Multilevel multiple imputation (typically MI-JOMO)
MNAR	Missing not at random
NHS	National Health Services
OLS	Ordinary Least Square
PCPH	Research Department of Primary Care & Population Health
PhD	Philosophy Doctor
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QOF	Quality and Outcomes Framework
RCT	Randomised clinical trial
REML	Restricted maximum likelihood
SBP	Systolic blood pressure
SMC JOMO	[see MI-JOMO]
SGA	Second-generation antipsychotics
SR	Segmented regression

THIN	The Health Improvement Network
UCL	University College London
UK	United Kingdom
US	United States

Notations

α	parameter(s) of the imputation model
β	parameter(s) of the analysis/substantive model
ϵ	residual error
θ	parameter(s) of the imputation model
μ	mean or mean vector
ν	a random draw from the standard normal distribution
ρ	correlation
σ	standard deviation
φ	parameter(s) of the missing data model
ω	parameter(s) of the imputation model
Σ	summation or covariance matrix
Ω	covariance matrix
B	between-imputation variance
BVN[]	bivariate normal distribution
CI	confidence intervals
e	residual error
E ()	expectation
H	hypothesis
i	indexes the follow-up time
j	indexes individuals
log	logarithm
logit()	link logit
L ()	likelihood function
m	number of imputations in MI
miss	denotes quantities missed in the observed data
n	number of individuals/observations in the dataset
N()	normal distribution
obs	denotes quantities in the observed data
Pr()	probability
P()	probability
u	a random draw from the standard normal distribution
V	variance
X	covariate(s) in the analysis model
Y	outcome variable in the analysis model
ι[ ]	indicator for the event in square brackets
ˆ	denotes an estimate of the corresponding parameter underneath
¯	denotes the mean of the corresponding quantity underneath
|	conditional probability

1

General Overview

1.1. Background and motivation

Interrupted time series (ITS) is a quasi-experimental design, recognised as one of the most robust alternatives for evaluating an intervention effect when randomised experiments are not feasible [1,2]. The idea behind ITS is relatively simple: each individual can serve as its control. If the studied outcome was recorded many times before and after the initiation (baseline) of the intervention to be evaluated; then, the pre-baseline outcome trajectory over time can work as a control of the post-baseline trajectory. Difference between the two trajectories represents the effect of the intervention and is adjusted for time-independent confounders by design. More sophisticated versions of ITS can help to control for observed time-dependent confounders [3] but, with no randomisation, any unobserved confounding remains uncontrolled.

As with other quasi-experimental designs described and studied in classic literature [2], ITS designs were not restricted to evaluate interventions applied either at the population or individual level (although tended to be used with population-level summary data). It implies that the same internal validity described above can be reached whether the intervention studied is, for example, a policy or a medical treatment. For policies, all individuals share the same starting date of intervention and similar exposure to other variables that are common at the population level (e.g. season of the year). For medical treatments, people have different starting dates of intervention and less similarity regarding those factors that vary at the population level. In both cases, policies and treatments, ITS analyses can be performed by using individual-level

data, although the conclusions derived from these analyses are always at population-level (i.e. the difference in population trajectories). For policies, it is also possible to perform ITS analyses with population-level data only.

Early access to population-level data have contributed to the evaluation of interventions at the population level (e.g. policies) with ITS designs; thus, the development of the ITS statistical methods was mainly orientated to model population-level data. Since ITS designs were born in social sciences in the middle of the 20th century, early ITS studies focused on evaluating educational or economic interventions applied at the population level [2]. In these studies, population-level data were enough to reach valid conclusions, although with significant – and sometimes downplayed – assumptions regarding the underlying individual-level data (e.g. missing data). The population was usually seen and treated as though it was made up of the same individuals over time, and the statistical analysis performed accordingly. Segmented regression with ordinary least squares was initially used to model ITS with population-level data. Other statistical methods for time-series analysis, such as autoregressive integrated moving average (ARIMA) or Fourier analysis, helped to solve key analysis issues such as autocorrelation and seasonality [2]. More sophisticated statistical tools were gradually developed to handle other methodological issues (including missing data) [3], but always for modelling population (or sub-population) level data.

Access to individual-level data of better quality is rapidly increasing, and there may be a renewed interest (e.g. routinely collected data). Nevertheless, when data are routinely collected, they often come with missing values [4] that traditional ITS methods are not designed to handle. In such a context, contemporary researchers often aggregate the individual-level data available in order to perform the ITS analysis with the population-level statistical tools they trust [4]. Without guidelines or studies describing the consequences that aggregating data with missing values can have on the validity of ITS analyses, the scientific community continued to use traditional statistical tools.

[2] These methodological issues will be explained in Chapter 2.
[3] I will mention these statistical tools in Chapter 2.
[4] Population level data frequently have missing values as well, but this is hidden in population summaries and obvious with individual data.

One of the primary purposes of this thesis is to unveil the problems that emerge when population-level ITS analyses are applied to incomplete individual-level data and to propose alternative analysis methods.

The motivation for this PhD stems from the need to address specific clinical problems, such as the impact of initiation of antipsychotic medication on short and long-term changes in body weight. While it has been demonstrated in randomised clinical trials that initiation of antipsychotic medication can substantially increase body weight over a short time [5], far less is known about the long-term effects of antipsychotic treatment initiation. I chose to examine this question by using electronic health records from UK primary care; records that include longitudinal data from clinical care. After some consideration, I decided to use the ITS design. I gradually realised that most of the ITS literature and guidelines are focused on the population-level approach [6,7], giving little attention to the evaluation of individual-level interventions with electronic health records. I also realised that electronic health records I was exploring came with different missing data issues that were challenging and should be treated with caution. With no specific literature available, I had to start methodological research to inform my approach for missing data handling.

I explored mixed effect models (MEM) as the first option to handle missing outcomes in ITS when analysing the antipsychotic-induced weight gain with electronic health records. This method helped to avoid any data aggregation, for example, across patients within the same time point or across close time points in an artificial time window. However, for fitting MEM all the cases with missing covariates had to be removed (complete case analysis), thus it was worth to explore other methods to recover information from lost cases and avoid any potential source of bias. Therefore, I decided to explore how multilevel joint modelling multiple imputation (MI-JOMO) might be used to deal with missing covariates when fitting MEM. I did this while keeping the attention on the clinical problems that motivated my PhD, ensuring that my methodological research was always connected with the improvement of the analysis of clinical questions.

1.2. Aims

The overall aim of this PhD thesis was to evaluate methods for missing data handling when the ITS approach is applied in data at the patient level and to implement the best of these methods in the analysis of health outcomes from large primary care databases.

The specific objectives were to:

1. Explore current practices in missing data handling for ITS studies, focusing on how often missing data were considered and, if so, how they were evaluated, reported and handled.

2. Apply an ITS design with MEM, as an alternative to the aggregate-level SR analysis, to model the effect of antipsychotic drug treatment initiation on weight change in general population data, in order to evaluate the potential and limitations of MEM.

3. Evaluate the performance of ITS analysis methods such as aggregate-level SR, MEM and MI-JOMO for handling missing data on health outcomes and baseline covariates for studying the effects of any treatment initiation, via simulation studies.

4. Apply MI-JOMO with MEM to model the antipsychotic-induced weight change in >40 years old population data by using an ITS design, focusing on the evaluation of how sex, age and dose may modify weight trajectories.

1.3. Outline of subsequent chapters

Chapter 2 introduces the topics of ITS designs and electronic health records (EHR). The ITS design is mainly described from the population-level approach, mentioning the standard statistical tools available and how methodological issues can be handled with these tools or design extensions. EHR are individual-level data which is routinely collected for administrative or patient management purposes. I summarise the main features of EHR, paying special attention to the missing data issues. In particular, I describe the dataset I used during my PhD, taken from The Health Improvement Network database, also known as THIN data. I end the chapter by briefly explaining some implications of performing ITS studies with EHR.

Chapter 3 gives a general overview of the missing data problem, using examples from THIN data. A general explanation of the missing data mechanisms and the common problems linked to missing data are also given. After summarising main issues related to missing data, I explain the multiple imputation (MI) approach in detail. Congeniality between the substantive and imputation models and its relevance to ITS analysis is described in detail. Finally, I explain how MI-JOMO and MEM help to handle missing values in individual-level datasets.

Chapter 4 demonstrates how the issues of missing data have rarely been addressed in most recent ITS studies performed with individual-level data. Despite its limitations, complete case analysis is the most frequently used method for handling missing data. Individual-level data are usually transformed into population-level data before fitting ITS models; for example, averaging the outcome at each time point ('averaging-step') before fitting a segmented regression ('aggregate-level' SR). I also confirmed that very few studies had applied MEM in the past. As MI-JOMO has only been developed recently, it has not been applied before in ITS studies.

Chapter 5 describes an application of MEM to the study of antipsychotic-induced weight gain by using the ITS design on electronic health records. MEM allowed fully observed covariates and partially observed outcomes to inform the implicit imputation of missing outcomes at the individual level. However, I detected that the sole use of

MEM is not ideal when covariates are missing (e.g. dose of medication). ITS facilitated new clinical evidence, especially in the long-term. Thus, I demonstrate that typical patients do not lose -in the long term - the weight they gained during the first six weeks of antipsychotic treatment.

Chapter 6 evaluates – via simulation studies – the performance of different types of methods for handling data missing at random (MAR) in ITS studies (aggregate-level SR, MI-JOMO and MEM). I showed that the averaging-step biases ITS estimates when data are MAR at the individual level. The aggregate-level SR can over- or underestimate the ITS effect, depending on how the missingness mechanism is operating. MEM are efficient and valid for handling outcomes MAR but must be combined with MI-JOMO when covariates are missing.

Chapter 7 describes an application of MEM with MI-JOMO to assess how dose and age modify the antipsychotic-induced weight gain. Interaction terms in MEM helped to evaluate differences in weight trajectories over time between groups by dose or age. MEM was combined with MI-JOMO for handling missing values of dose. Again, this provided new clinical evidence; for example, that older patients' weight is less affected by first olanzapine prescription compared to younger people.

Chapter 8 presents a discussion of the thesis. I begin by summarising the thesis findings study by study. I discuss overall the key findings and reflect on the methodological and clinical implications. I also reflect on the strengths and limitations of the thesis and describe future work (related to this thesis) and end with a conclusion.

2

Interrupted Time Series & Electronic Health Records

2.1. Interrupted Time Series (ITS)

2.1.1. Introduction to Interrupted Time Series

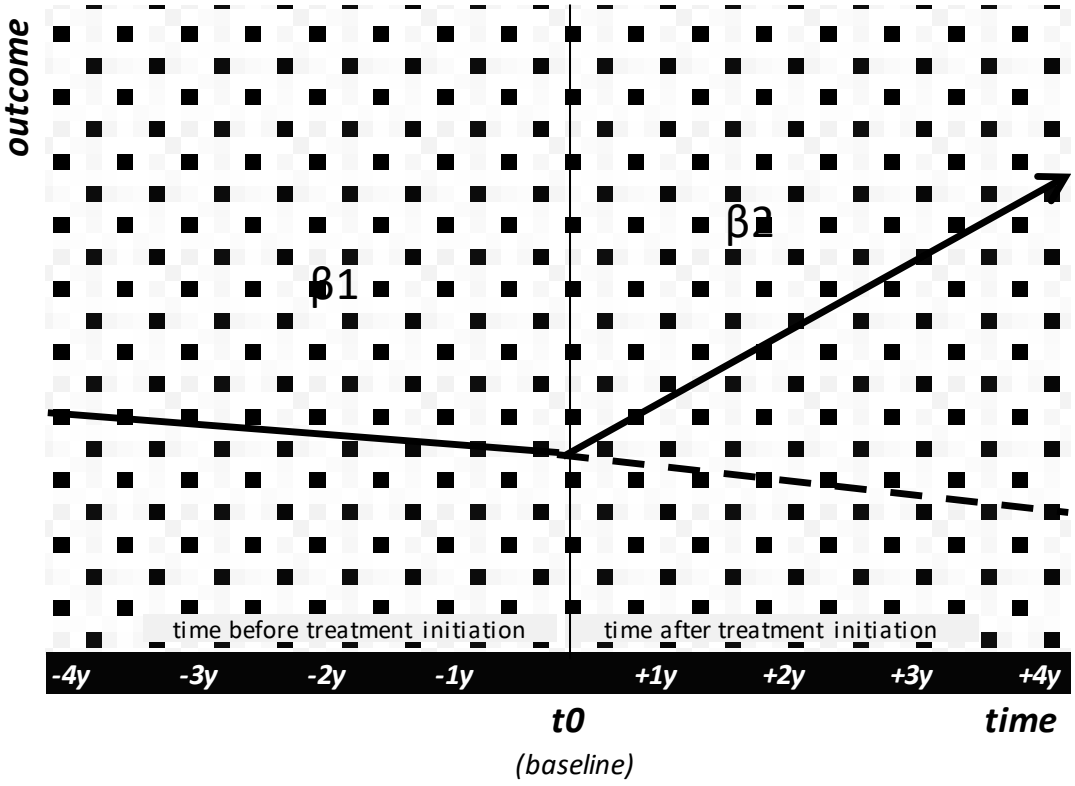
Interrupted Time Series (ITS) is a quasi-experimental approach for evaluating how an intervention can affect an outcome of interested [2]. ITS requires various outcome measurements before and after the intervention initiation for drawing their trajectory change overtime. Pre-intervention trajectory serves as a control of the post-intervention trajectory, and the difference between trajectories is the estimated intervention effect [6].

For modelling a time series in a population, we need the outcome to be sequentially measured over time and – typically – with equal intervals between measurements [6]. Conceptually, ITS estimates the underlying trend behind the observed outcomes, detecting an 'interruption' of this trend at a predefined time-point (e.g. when the intervention starts) (Figure 2.1.). Hypothetically, if the intervention did not occur, the trend observed before intervention should continue similarly after the intervention initiation, serving as 'counterfactual' for the actual trend observed since the

intervention starts. The central assumption for attributing any trend 'interruption' to the sole intervention is that any potential confounder must be stable over-time (time-invariant covariates) [3].

Figure 2.1 represents an example where lines describe outcome trajectory: β1 line represents the pre-intervention trend, β2 line represents the post-intervention trend, and dash lines -as an extension of the β1 line- represents the theoretical 'counterfactual' for the post-intervention trend.

Figure 2.1 Visual representation of a single ITS design



For any ITS study, we need the pre-intervention and intervention periods to be clearly differentiated. At the population level, this is not always an easy task since some interventions can have different phases to be 'completed' or uncontrolled time variability of the valid starting point for each individual of the population. Nevertheless, such issues can be handled by appropriately defining 'impact models' [8] or modelling trajectories at the individual level (see chapters 5, 6 and 7), respectively. An impact model describes how the outcome could be affected by an intervention; for example, if

there is an abrupt change at the time of intervention (i.e. intercept change) followed by a trajectory over time that is different from the pre-treatment trajectory (i.e. slope change). An impact model is appropriately defined if it incorporates available evidence on the studied effect (e.g. from clinical trials) and how the intervention has been applied (e.g. duration). In practice, the impact model helps to draft a basic ITS shape (e.g. defining inflexion points) before performing the ITS analysis itself. Different feasible impact models can be contrasted during the ITS analysis, but all of them should be proposed before the analysis and based on a rationale [8].

ITS outcomes can be continuous, counts or binary variables [7]. For example, it is common to see outcome prevalence (proportions) to be calculated at each time point and then used to draw the ITS trends over time. Although there is not a minimal fixed number of data points recommended for ITS [6], it is essential to take into consideration both the shape of the trend and the statistical power. Estimating linear trends requires fewer observations than estimating non-linear trends. Power increases with the number of measurement times as well as if these measurement times are equally distributed before and after the initiation of an intervention [9].

2.1.2. Statistical analysis for ITS

Standard ITS analyses (i.e. population summary level analysis) use segmented regression models [6,7]. Having an outcome Y_i observed at each time point of the period '$time_i$ $(-k \leq i \leq k)$' for which an intervention has started at t_0, the trajectory of Y_i can be modelled by:

[Equation 2.1]

$$Y_i = \beta_0 + \beta_1 time_i \times 1[time_i < 0] + \beta_2 time_i \times [time_i \geq 0] + \varepsilon_i$$

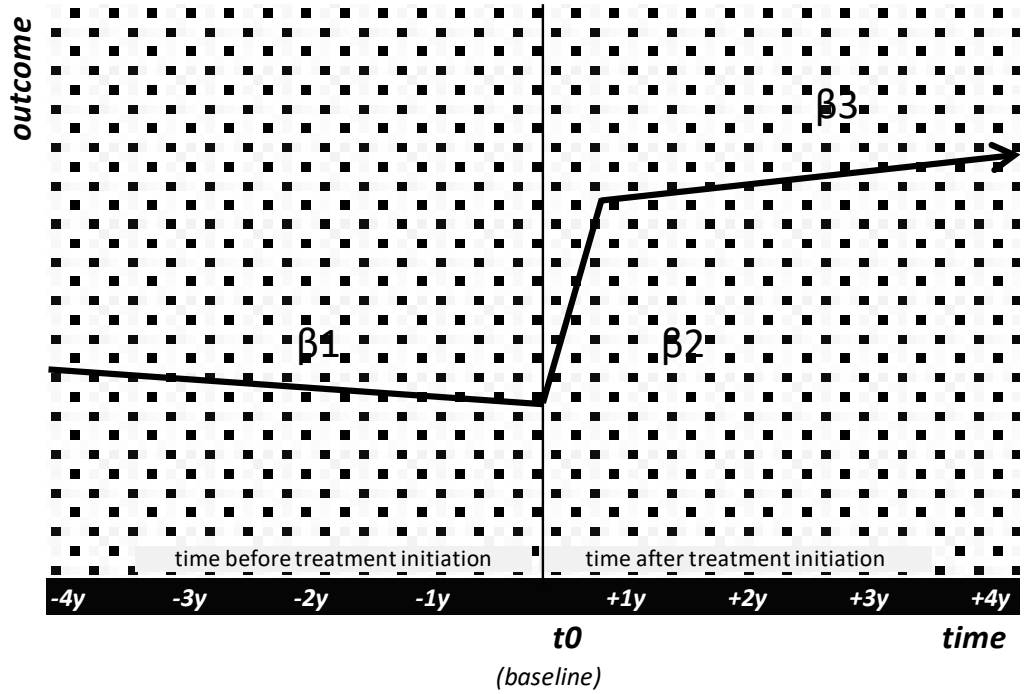
$$\varepsilon_i \sim N(0, \sigma^2),$$

where i denotes the follow-up time, and 1[ ] is an indicator for the event in square brackets. This model can be represented by Figure 2.1, where β1 represents the pre-intervention trajectory, β2 the trajectory after the initiation of intervention, and β0 the

average Y_i when the intervention started (t_0). Parameters are usually estimated by maximum likelihood estimators, like for any generalised linear model (GLM) [6], although ITS is not inherently linked to maximum likelihood estimation. The use of GLM facilitates modelling different type of outcomes (continuous, counts, binary) with a similar segmented and – commonly – linear prediction component on the right of the equation. These GLM type models only include population-level fixed effects; thus, no random intercept and slopes at the individual level are considered in standard guidelines for ITS [6]. The implications of this restriction are shown and discussed in different sections of the thesis, but mainly in Chapter 6.

The selection of impact model is a crucial step in ITS statistical analysis. As mentioned above, impact models define the type of effect we are expecting on the outcome due to the intervention [8]. [Equation 2.1] is the simplest impact model in ITS, which reflects only an immediate change in the outcome trajectory. As a slightly more complex example, the impact model can specify that the intervention will have a huge and quick impact on the outcome in a relatively short time, and later the outcome will change gradually in the long-term (Figure 2.2) (e.g. as in Chapter 5). In such an impact model, we would want to define not two but three slopes of change over time (β_1 as pre-intervention, β_2 as short-term impact, and β_3 as long-term impact). The selection of an impact model is theory-based; thus, it should be based on substantive knowledge of the research topic. In practice, this implies that the overall shape of the outcome change over time (i.e. number of slopes and their time-point of origin in linear trajectories) should be defined before any statistical analysis. If there is no robust background to support the models, alternatives are to use external data for defining the intervention [10], to perform some sensitivity analyses [8] or just initial exploratory studies (e.g. identify change points [11]). In particular, sensitivity analysis can be useful when different theories explain the same studied effect; thus, different feasible models or assumptions can be tested (e.g. diverse ranges of outcomes or lag periods for the intervention effect) [8]. If non-linear trajectories could best explain smoother effects, non-linear models can be contrasted against linear models with visual tools or goodness of fit tests. However, enough and reliable data points should be used for avoiding overfitting and non-replicable results.

Figure 2.2 Visual representation of a single ITS design with three slopes



2.1.3. Methodological issues in ITS

As any other observational design, ITS approach relies on assumptions that are hard to verify in practice or that raise issues requiring special statistical treatment. There are at least three well-studied sources of bias in the literature that need to be considered in ITS studies: seasonality, autocorrelation and time-varying confounding.

Seasonality describes a pattern where outcomes over time follow wave trajectories, with consecutive crests and troughs separated by similar intervals. This wave pattern can overlap an underlying outcome trend which is linear or non-linear. In population-level data, where time variable is usually calendar time (e.g. weeks, months or years), wave patterns can be associated with events that are regular at specific weeks or months (e.g. winter or summer time). Seasonality is a problem in short-term ITS when a higher proportion of some specific weeks/months are overrepresented (e.g. summer months), leading to biased estimates. Consecutive months within a wave section are more correlated than non-consecutive months from other wave sections, which brings issues such as autocorrelation or overdispersion. Seasonality can be assessed visually or tested with autocorrelation functions or partial autocorrelation functions [12]. Adjustment for

seasonality can be achieved by using a covariate, seasonal ARIMA (autoregressive integrated moving average) [13], splines or Fourier terms (pairs of sine and cosine functions) [14].

Autocorrelation (or serial correlation) is common in longitudinal data and means that consecutive outcome measurements are more correlated than non-consecutive measurements. This occurs because ITS is intended to take measurements from the same units over time. Segmented regression with GLM or OLS assumes that all these consecutive measurements are independent, which is violated in ITS data (both at the population or individual level). When individual data is available, mixed effect models can address the within-individual correlation over time, controlling the autocorrelation problem. In the second study (Chapter 5), I show an application of these models. At the population level, autocorrelation can be assessed by plotting the residuals, using a partial correlation function or running formal tests (e.g. Breusch-Godfrey [15]). For controlling autocorrelation in time series, a factor that explains the autocorrelation (e.g. seasonality) can be included in the model as a predictor, or other methods such ARIMA can also be applied [16].

Time-varying confounders are not controlled by ITS design, which is particularly problematic in long-term studies. If there are only time-invariant confounders, the 'counterfactual' of the intervention period – which is an extension of the pre-intervention trend – is by definition unconfounded because the confounders will be identical for the pre-and post-intervention periods, making a comparison between intervention and its 'counterfactual' valid for conclusions about the intervention effect. However, in many observational scenarios, it is not unusual to have confounders that change over time, especially in long periods of observation (e.g. years). If the time-varying confounder has been identified and recorded, it can be included as a covariate in the segmented regression model for control purposes. There are other options. For example, given a population that has received the intervention, we can match a similar population and observe it during the same period and use it as a control group. This design is known as Controlled Interrupted Time Series (CITS) [3,6]. In terms of analysis, we can either model each group separately or fit just one single model. One single model can include indicator variables for de control or the intervention series as interaction terms which allows for formal group comparisons. If we decided to model

ITS separately, the interpretation is mainly descriptive. For ITS results in the control group that mirrors results from the intervention group, it is more likely that outcome changes are attributable to the intervention [3]. In the final thesis study (Chapter 7), I will go into further details on interaction models for ITS studies with electronic health records, which is relevant to the control of time-varying confounders.

In this section, I have summarised the key characteristics of the ITS design. ITS is a robust quasi-experimental alternative when randomisation is not an option, which can control for any time-invariant confounder and observed time-variant confounders. ITS design preserves its internal validity, whether the intervention evaluated is applied at the population or individual level. Most statistical tools have been designed to model ITS with population-level data. These traditional tools and more complex ITS designs help to account for methodological issues such as autocorrelation, seasonality and time-variant confounders. However, there are no standard practices on how to handle missing data for individual-level data in ITS studies.

In health research, access to individual-level data for research purposes is becoming widespread due to the increased access to routinely collected data. Electronic Health Records (EHR) are an excellent example of data that are useful for health research in general, and ITS studies in particular. In the next section, I make an introduction to EHR, describing some issues about data recording. Later on, I explain some advantages and disadvantages of using EHR for ITS studies.

2.2. Electronic health records (EHR)

2.2.1. Introduction to EHR

Electronic health records (EHR) are data routinely obtained from patients or the general population and stored in a digital format. EHR are largely used as a source of data in health research in Asia [17], Europe and the US [18]; while in Latin America [19] and Africa [20], EHR are an emerging alternative due to the early development of their health informatics systems. EHR are usually longitudinal at the individual level and can be collected from different care settings (and sometimes linked between them). For example, hospitals, insurance companies, health surveys, health devices and applications, and others. In this thesis, I focus my attention on EHR from the UK primary care.

Primary care in the UK is free for all and is provided in a national network of general practices (GPs). Each person must register in the GP closest to their home in order to receive medical attention and regular health monitoring (e.g. receive opportune vaccination or regular prescription of contraceptives). In the UK, more than 95% of the population are registered in a GP [21].

EHR in primary care are recorded when patients receive treatment by general practitioners or other primary care staff, as well as phone and online consultations [22]. In the UK, there are three main sources of primary care EHR: QRESEARCH [23], the Clinical Practice Research Datalink (CPRD) [24] and The Health Improvement Network (THIN) [25–27], which are all primary care databases. Typically, these EHR includes demographical information (e.g. Townsend score [5], age and sex), diagnostics and symptoms (e.g. diabetes, schizophrenia or hypertension), treatment and drug prescriptions (e.g. contraceptives), laboratory results (e.g. HbA1c or cholesterol), lifestyle information (e.g. smoking status) and other health indicators which are relevant for health research (e.g. blood pressure or weight).

[5] Townsend score is a measure of material deprivation based on four indicators: unemployment, non-car ownership, non-home ownership and household overcrowding. It can be measured for a population living in a specific area, then easily connected to patients' ZIP codes. In THIN data, Townsend scores are given in quintiles (Q1-Q5), from the least to the most deprived.

For most of EHR, available data allows retrospective cohorts at the individual level. For example, it is possible to identify patients' demographic variables together with health indicators that change over time, such as weight, drinking or smoking status. Specific clinical symptoms or diagnostics can be identified by using the Read Code System, a hierarchical classification of symptoms and diagnoses [28]. Drug treatment prescribed to patients is also easy to identify by using the BNF or ATC coding systems, a comprehensive and organised list of pharmaceutical drugs [29]. Prescription codes usually come with extra information about date and dosage, facilitating patient follow up and the longitudinal analysis. Some databases also provide the opportunity to link those patient records with additional relevant data. For example, CPRD can be linked to hospital data, allowing to detect relevant episodes in the history of primary care patients, such as major surgeries or highly specialised medical attention and treatment.

2.2.2. Issues about data recording in primary care

In the UK, the evolution of EHR quality included progressive initiatives for improving data collection. In 1995, the NHS established a contract between the government and GP's centred on financial compensation for registering new patients from target groups (e.g. aged over 75 years), as well as for providing them with standard services. In practical terms, the agreement implied that "the patient records" were vital for giving that standard service and having a proof of the provided attention. That strategy proved to be useful for recruiting patients and collecting relevant information but was not cost-effective [30]. In 2004, the General Medical Services (GMS) created the Quality and Outcomes Framework (QOF) –which is defined as "an annual reward and incentive program" - in order to improve the quality of the delivery of primary medical services [31]. The QOF helps to focus the work of NHS employees on a set of measurable achievements, which implies rigorous monitoring of selected health indicators with public health significance. In practice, QOF monitors people with certain chronic conditions, and they are more likely to have health indicators recorded on a regular basis, whereas people without these conditions are less likely to have health indicators recorded on a regular basis [32]. It makes the handling of the missing data complex, so statistical methods such as multiple imputation should be considered [32].

Missing data is a common problem with EHR, mainly because the original data are not collected for scientific research. In the UK, the National Health Service (NHS) got the mission of being able to provide to patients their health profile (with information about diagnostics, prescriptions, and others) and health care interactions, all in real-time, by 2020 [33]. All their efforts are focused on the goal of making the NHS "paper-free at the point of care" by the same year. In those circumstances, missing data can be generated when: i) a value was not measured by health care staff, ii) was measured, but not recorded, or iii) was measured, but recorded wrongly (so it would have to be omitted in its later analysis) [6]. In these scenarios, missing data may have a higher impact on the research, which is based on this data rather than the clinical and administrative use.

2.2.3. The Health Improvement Network (THIN)

The Health Improvement Network is one of the largest sources of longitudinal electronic health records in the UK. It has good data quality and is roughly representative of the UK population [27]. The data are collected via the In Practice Systems (IPS) called Vision GP. Data from Vision GP system run regularly in the GP´s systems and is sent to Cegedim, the owners of THIN, who then supply this data to IQVIA under license. After anonymising data, IQVIA provides and supports access to this data for health research. The data that I have used from THIN for this PhD has data from 2003 and to January 2017, including data from 15.6 million of patients, of which 3 million are active patients from 711 practices and can be prospectively followed [34]. There are different quality markers for THIN data. The acceptable computer usage (ACU) which is defined as "the year in which a general practice was continuously entering on average at least two therapy records, one medical record and one additional health data record per patient per year" [35]. Likewise, there is a quality marker for acceptable mortality reporting (AMR) which indicates when GP's mortality records are consistent with the official national statistics [36]. Moreover, it has been shown that THIN is roughly representative at the national level in terms of demographics, deprivation and chronic diseases [27].

[6] In most missing data definitions [53], a value that never was intended to be recorded is not considered missing data.

2.3. ITS and EHR

Since its origins, the ITS design has mainly been applied to evaluate interventions with population-level data. For example, to assess the discontinuity of national or regional prevalence trajectories over time before and after the evaluated intervention. The prevalence at each time point is usually aggregated at convenient time units (e.g. monthly or annually). In the past, the individual-level data that generated those prevalence numbers were not accessible, challenging to obtain or – in many cases – not considered relevant for the standard ITS analysis. In the ITS approach, population-level interventions can be studied with population-level data directly [2,6]. However, in recent years, individual-level data has become available in some settings [37]. In such context, ITS studies are increasingly being applied in health research based on data routinely collected at the individual level [4], including EHR. EHR can provide data for evaluating both individual- and population-level interventions by using the ITS design. As I explained above, EHR come with missing data issues that need to be considered before performing any ITS analysis. In this thesis, I explore methods for handling missing data in EHR when ITS design is applied to the study of individual-level interventions.

The use of EHR in ITS studies brings advantages and disadvantages to be considered. On the positive side, EHR improve the amount of longitudinal data available. Access to EHR helps to improve the statistical power by increasing the number of observations at each time point [38], which, in typical ITS population-level data, can be as small as one (e.g. when data brings many observation time points). However, the advantage of using EHR comes with a cost: missing data issues. ITS usually request the outcome regularly recorded at each time point, but this is hard to achieve in most observational data such as EHR. Often, researchers tend to handle this problem by aggregating all the available outcome data in some sort of time windows (e.g. months or years), but usually ignoring the potential consequences of this operation when data are missing. Missing data can affect ITS estimates in many ways, as I will show throughout the thesis. Lopez-Bernal et al. [6] recommend to assess any change in data collection recording; for example, missing data proportion changing over time. Particularly, when the data collection patterns change at the time the intervention initiates, ITS analysis could lead to biased results [2].

Since these missing data issues associated with data routinely collected are widespread and relevant for ITS studies, I will give an introduction in the next Chapter 3 and examine this in further detail in Chapter 6.

3

Missing Data Handling

3.1. Missing data

3.1.1. Missing data in THIN

As any other longitudinal electronic health record which has mainly an administrative and management purpose, THIN presents issues with missing data. More general demographics data, such as age, sex or Townsend (deprivation index), usually do not represent a significant problem [39,40]. However, other variables such as ethnicity are poorly recorded, and its practical use in statistical analyses requires particular caution [32]. Anthropometric and behavioural measures like weight, height, smoking status [41] and alcohol consumption [42] also present different patterns and volumes of missing data. However, the cumulative recording can partially solve the missing data problem for some of them (i.e. the height is relatively stable over time and can be passively imputed after the age of 18-20 years). Statistical analysis with health outcomes like blood pressure [43], diabetes [44], cholesterol [45], and body mass index [46] is also challenging due to missing data [32].

Table 3.1. Missing data in relevant variables for the cohorts of Olanzapine, Quetiapine and Risperidone, stratified by sex.

Variable in THIN data	OLANZAPINE (N=9499)		QUETIAPINE (N=19965)		RISPERIDONE (N=9401)	
	Women N=5004 (52.7%)	Men N=4495 (47.3%)	Women N=12149 (60.9%)	Men N=7816 (39.1%)	Women N=5153 (54.8%)	Men N=4248 (45.2%)
	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)
Smoking Status	591 (11.8)	686 (15.3)	1386 (11.4)	973 (12.4)	472 (9.2)	490 (11.5)
Drinking Status	598 (12.0)	876 (19.5)	1511 (12.4)	1136 (14.5)	398 (7.7)	518 (12.2)
Height	1786 (35.7)	1481 (32.9)	4665 (38.4)	2832 (36.2)	2021 (39.2)	1439 (33.9)
SBP	1659 (33.2)	2099 (46.7)	3601 (29.6)	2631 (33.7)	1280 (24.8)	1484 (34.9)
LDL-Cholesterol	1901 (38.0)	1757 (39.1)	4783 (39.4)	2797 (35.8)	1792 (34.8)	1544 (36.3)
HDL-Cholesterol	1459 (29.2)	1385 (30.8)	3596 (29.6)	1995 (25.5)	1307 (25.4)	1089 (25.6)
First Dose	1369 (27.4)	1138 (25.3)	4865 (40.0)	3164 (40.5)	1735 (33.7)	1396 (32.9)
Body Weight *	2156 (43.1)	1858 (41.3)	5540 (45.6)	3438 (44.0)	2389 (46.4)	1771 (41.7)

(*) Body weight has been calculated as the average of the wieght records available up to 12 months before treatment initiation. As sex, age, deprivation (Townsend) and Diabetes Diagnostic were fully observed in the dataset. SBP = systolic blood pressure; THIN = The Health Improvement Network.

Table 3.1 presents a summary of missing data percentages in THIN data, for crucial variables that will be included in most analyses thorough the thesis (especially Chapters 5, 6 and 7). With a few exceptions, all these variables have >11% of missing values. In these cohorts, smoking and drinking status are less affected by missing data (<20%), while other health indicators such as systolic blood pressure or cholesterol have in general >30% of missing data. Particular attention should be paid to the first dose, which is never <25% and can be up to 40% missing. This variable is central in the study of antipsychotic-induced weight gain (the clinical topic of this thesis); thus, in Chapter 7 I will apply the approach validated in Chapter 6 to handle missing data in the first dose.

Bodyweight represents a particular challenge. Table 3.1 shows >40% of missing data of weight at baseline, making hard any ITS analysis stratified by weight at baseline. In Chapter 5, I show how to address this problem by fitting mixed-effects models in ITS designs. Nevertheless, weight is also the outcome of interest, adding new challenges to missing data handling. Since THIN data is routinely collected, weight records are irregularly recorded in each patient history. It means that, for any time-window defined for the ITS analysis (e.g. week or month), the same patients are not going to be visible in all the time-windows. Table 3.2 describes that, for 8 years of observation (416 weeks), the median number of weight records is around 9 (interquartile range around 10). However, the total amount of available outcome data is enough to try alternative ITS models (see Chapters 5, 6 and 7).). Figure 3.1 shows the missing data distribution of the outcome before and after initiation of antipsychotic treatment. It demonstrates that, close to treatment initiation, there are more weight records, but these records become

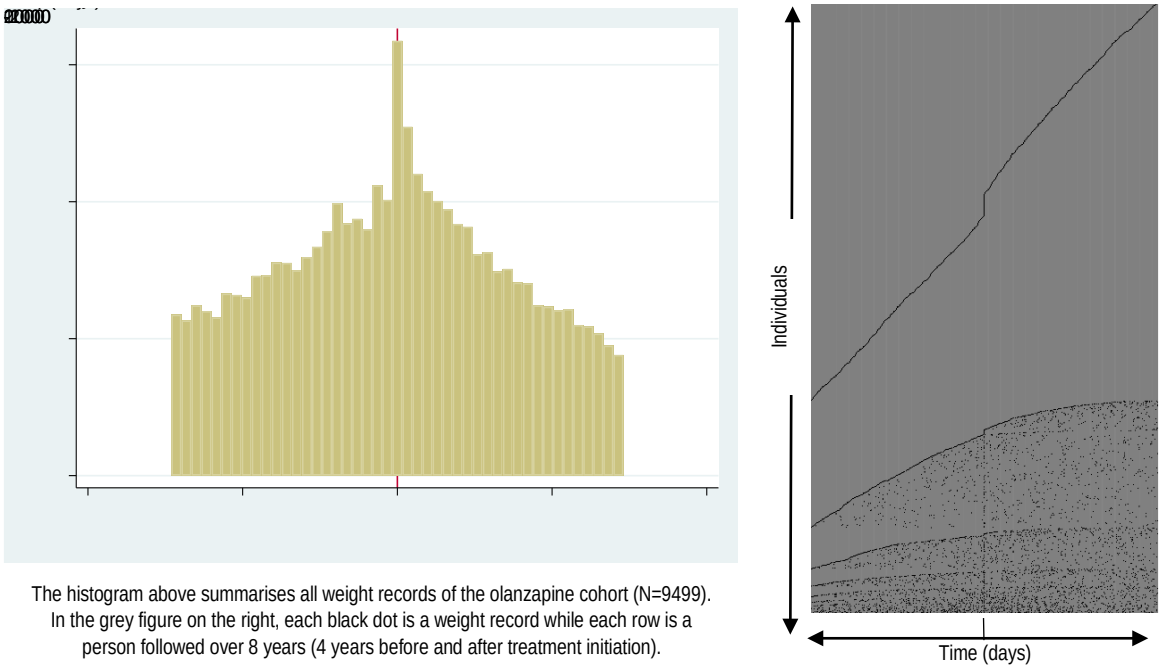
progressively less frequent after the initiation date. This distribution and the pattern of missing data visible in Figure 3.1 were carefully considered in the simulation study of this thesis (Chapter 6), as well as the percentage of missing data described above (Tables 3.1 and 3.2) and the associations found in Chapter 5 (see also relevant equations for the data-generation mechanism in Chapter 6).

Table 3.2. Number of weight records for the cohorts of Olanzapine, Quetiapine and Risperidone, stratified by sex.

Drug Cohort	Sex	Number of weight records (n) before and after treatment initiation								
		4 years before			4 years after			8 years (total)		
		n	median	iqr	n	median	iqr	n	median	iqr
Olanzapine	Male	6360	4	5	6950	4	5	13387	8	7
	Female	11269	6	8	10786	5	7	22159	10	11
Quetiapine	Male	14100	5	7	11923	5	5	26100	9	10
	Female	27174	6	8	23880	5	7	51205	10	11
Risperidone	Male	7333	5	6	6328	4	5	13766	8	9
	Female	11138	6	7	9353	5	6	20597	10	10

iqr = interquartile range

Figure 3.1 Distribution and patterns of weight records in THIN data



The histogram above summarises all weight records of the olanzapine cohort (N=9499). In the grey figure on the right, each black dot is a weight record while each row is a person followed over 8 years (4 years before and after treatment initiation).

It is worth mentioning that, when a complete case analysis is considered, the combined effect of missing data in outcome and covariates reduces the sample size to extreme levels. That is one of the key reasons why developing effective and efficient alternatives to handle missing data is so essential in these situations.

3.1.2. Missing data mechanisms

Missing data have a cause, a generation mechanism that can be more stochastic or more systematic. Initially described by Rubin [47], in general, three generic mechanisms can cause missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

To clarify what these mechanisms are, I will use an explanation slightly adapted from van Buuren [48]. Let us say that we have a matrix *X* with the shape *n x p*, where *n* is the number of individuals and *p* the number of variables. Then denote X_{obs} as those with observed values in the data matrix and X_{miss} otherwise. The matrix *R* will reproduce the same structure than *X*, but with r_{ij}=0 when the value is missing and r_{ij}=1 otherwise. Graphically, an example of both matrixes could be:

$$X = \begin{bmatrix} 15 & 22 & \\ & 4 & 3 \\ 10 & 51 & \\ & 34 & 49 \end{bmatrix} \quad R = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

Finally, let φ denotes the parameters of the missing data model, which is $\Pr(R = 0|X_{obs}, X_{miss}, \varphi)$. Then, the data in *X* is MCAR if:

$$\Pr(R = 0|X_{obs}, X_{miss}, \varphi) = \Pr(R = 0|\, \varphi) \qquad \text{[Equation 3.1]}$$

Here, the probability of being missing does not depend on the observed or unobserved data; it just depends on φ. A different mechanism is MAR, which occurs if:

$$\Pr(R = 0|X_{obs}, X_{miss}, \varphi) = \Pr(R = 0|\, X_{obs}, \varphi) \qquad \text{[Equation 3.2]}$$

Here, the distribution of the missing values is conditioned on observed data; in other words, missingness does not depend on unobserved variables given observed variables. Finally, a definition of MNAR is:

$$\Pr(R = 0|X_{obs}, X_{miss}, \varphi) = \Pr(R = 0|X_{obs}, X_{miss}, \varphi) \qquad \text{[Equation 3.3]}$$

Here, the probability of being missing is also related to missing values, making complex any empirical verification (which is also the case for MAR). We could test the hypotheses Ho: MAR and Ho: MNAR only if we knew the values of X_{miss} which, by definition, we do not know. That is why MAR and MNAR are just assumed based on

both the identified associations – recognised on the dataset – and previous knowledge[7]. Missing data mechanisms have direct implications on statistical analyses and inferences [48]. In the next section, I will examine part of those implications.

3.1.3. Issues linked to missing data

Some specific issues arise from missing data handling related to bias, inefficiency and power. Both the estimates and standard errors can be biased when the missing data is not completely at random (not MCAR) [49]. For example, let us say that we want to analyse a dataset from primary care that contains data of body weight and sex of registered persons. We are interested in calculating the overall average weight in a specific month (e.g. January), but we have missing data issues on weight. In general, there are at least three possible consequences related to the mechanisms described above:

- First, if the weight records are missing completely at random (MCAR), there will be no bias on the calculated average, so any deviation from the true value (i.e. parameter) will be non-systematic.
- Second, if the weight records are missing in a higher proportion for men, then we can say that weight records are missing at random (MAR) on sex. Since we know that men weigh -on average – more than women, we expect a biased overall average estimate from the observed records (i.e. a systematic deviation from the true value).
- Third, if the weight records from patients with the lowest weight are missing, then we are facing a missing not at random mechanism (MNAR) because the missingness depends on unobserved weight values. In this circumstance, we expect the overall average estimate to be biased (again, a systematic deviation from the true value).

In general, methods for handling missing data can provide unbiased estimates (i.e. valid inferences) in the second scenario (MAR) but doing so in the third scenario (MNAR) is harder to achieve. This example with cross-sectional data can be extended to longitudinal data analysis, with similar issues and solutions linked to missing data.

[7] Sometimes we can go one step further and compare the actual values of the observed data to external data sources which may give an indication on whether the data are MAR or MNAR.

The inefficiency is related to a common practice called listwise deletion or complete case analysis (CCA), where all cases (subjects) with a gap in any variable -relevant to the analysis - are dropped. That procedure can severely reduce the sample size, especially in longitudinal studies [50], and much useful information is simply lost from the dataset. Listwise deletion produces unbiased estimates when missing values are completely at random [48], but always implies a sacrifice of statistical precision and power [51].

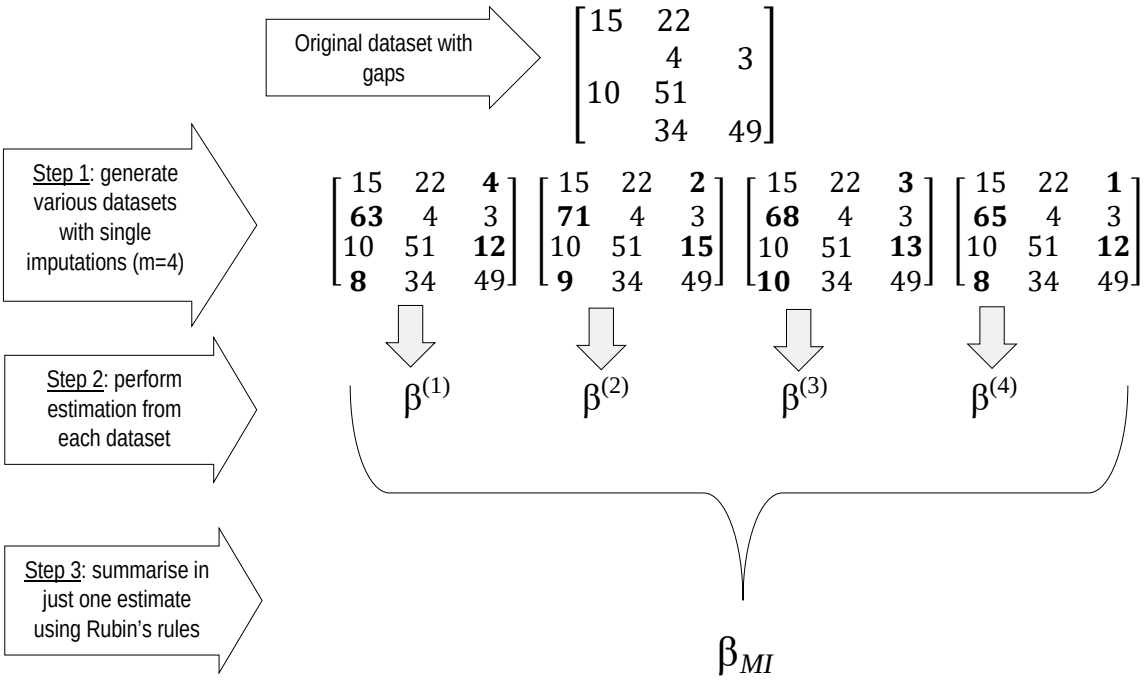
3.2. Multiple Imputation (MI)

3.2.1. Introduction to MI

Conceived initially by Rubin [52], MI was the natural development of single imputation strategies. From the very basic mean imputation to the more sophisticated stochastic regression imputation, simple imputations had an intrinsic limitation: a fixed imputed value always omits a critical source of variability, the uncertainty (we do not know the actual value we missed). Rubin solved this issue after proposing to use more than one fixed value for filling each missing data gap. That strategy helps us to preserve the uncertainty of missing data in the inferences made.

MI method has three steps: 1) to generate *m* datasets with imputations; 2) to perform the estimation of the parameter from each *m* dataset generated; 3) to summarise each of the *m* estimates in just one estimate, following Rubin's rules [53] (see below). Figure 3.1 illustrates the principles of MI:

Figure 3.2 Visual representation of the multiple imputation procedure



3.2.2. Rubin's rules for MI inference

The combined estimate shown in Figure 3.1, as well as its standard error, can be obtained after using two formulas named as Rubin's rules [50]. If *k* is the index for imputed datasets (k=1, 2, …, K):

i) The combined estimate averages the *m* separate estimates:

$$\hat{\beta}^{MI} = \frac{1}{m}\sum_1^m \hat{\beta}^{(k)} \qquad \text{[Equation 3.4]}$$

ii) Combined variance estimate:

$$\hat{V}^{MI} = \overline{V} + \left(1 + \frac{1}{m}\right) B \qquad \text{[Equation 3.5]}$$

Where:

$$\widehat{\overline{V}} = \sum_1^m V^{(k)}/m$$

$$\hat{B} = \sum_1^m \left(\hat{\beta}^{(k)} - \hat{\beta}^{(MI)}\right)^2/(m-1)$$

B is called the between-imputation variance. Confidence intervals and test statistics are performed under the assumption that $(\hat{\beta}^{MI} - \beta)/\sqrt{V^{MI}}$ has either t or standard normal distribution. These rules can be extended to the estimation of multiple parameters, as explained elsewhere [53].

3.2.3. Compatible substantive and imputation models

In MI, there are 'imputation' and 'substantive' models. The substantive model "is the analysis procedure if there had been no missing data" [54]. The imputation model may be used to impute missing values on any studied variable. Ideally, the imputation model should be compatible with the substantive model for Rubin's rules to provide valid estimates. The two models are termed 'compatible' if there is a joint model which yields the imputation and substantive models as conditionals. The simplest example [8] can help to illustrate what compatibility means. An outcome Y (fully observed) and a covariate X (incomplete) follow a bivariate normal distribution:

$$\binom{Y}{X} \sim BVN\left[\binom{\mu_Y}{\mu_X}, \begin{pmatrix} \sigma_Y^2 & \rho\sigma_Y\sigma_X \\ \rho\sigma_Y\sigma_X & \sigma_X^2 \end{pmatrix}\right]$$

The substantive model (linear regression of Y on X) can be derived as a conditional distribution from the joint distribution:

$$Y|X \sim N\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X), \sigma_Y^2(1 - \rho^2)\right)$$

This is, E(Y|X) is a linear function of X with slope $\rho\frac{\sigma_y}{\sigma_x}$. Similarly, the imputation model for X (linear regression of X on Y) can be derived as a conditional distribution:

$$X|Y \sim N\left(\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(Y - \mu_Y), \sigma_X^2(1 - \rho^2)\right)$$

Then, the imputation and substantive models are compatible. Compatibility ensures that, with large m, inference on MI data approximates a bivariate normal model for X,Y estimated by maximum likelihood [55].

[8] Example given by Tra My Pham in the UCL course 'Missing data and multiple imputation for cross-sectional and longitudinal data'.

In practice, perfect compatibility is hard to achieve, and some flexibility is usually needed. There has been a long debate about how incompatible models can lead to subtle deficiencies of Rubin's variance estimator [53]. Nevertheless, there are two broad forms of incompatibility to consider [55]: 1) if the imputation model wrongly omits a variable, then the analysis is typically biased; but 2) if the imputation model rightly includes one or more extra-variables, then the analysis can be benefited by bias correction. Carpenter and Kenward [53] make an extensive discussion on the topic and conclude convenience of including right extra-variables. These extra variables, called auxiliary variables, are those not included in the substantive model but associated with X and explicative of its missingness. Thus, an imputation model should include all variables (including interactions, non-linearities, and random effects) from the substantive model plus other optional auxiliary variables that could be relevant for the imputations of X. Under MAR assumptions, auxiliary variables can help to improve precision and reduce bias.

3.2.4. Joint modelling imputation for ITS

In ITS performed with individual-level data, to make the imputation and substantive models compatible is not straightforward. For example [9], we can expand [Equation 2.1] by adding two fixed effects, X1 and X2 (time-invariant covariates at treatment initiation), as well as a random intercept (u_{0j}) and two random slopes ($u_{1j}\, u_{2j}$), as follows:

[Equation 3.6]

$$Y_{\text{ij}} = \beta_0 + u_{0j} + (\beta_1 + u_{1j})time_{ij} \times 1[time_{ij} < 0] + (\beta_2 + u_{2j})time_{ij} \times [time_{ij} \geq 0] + \text{X}_{1j} + \text{X}_{2j} + \epsilon_{ij}$$

$$\begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{pmatrix} \sim N \begin{pmatrix} 0 \\ 0, \Sigma_u \\ 0 \end{pmatrix} \qquad \epsilon_{ij} \sim N(0, \sigma_e^2),$$

This is the substantive model of scientific interest, which is a mixed-effects model that suits an ITS design. Y, X1 and X2 are continuous variables each containing some missing

[9] I am adapting the very good explanation from Quartagno and Carpenter [58] to the specific case of ITS designs. This example gives some technical details for the application of MI-JOMO which will be useful in Chapters 6 and 7.

values (an issue typical in EHR). To solve the missing data issue on Y, X1 and X2, we could assume a trivariate normal joint model for the three variables:

[Equations 3.7]

$$\begin{cases} Y_{i,j} \\ \qquad = \alpha_0 + (\theta_0) time_{ij} \times 1\left[time_{ij} < 0\right] + (\omega_0) time_{ij} \\ \qquad \times \left[time_{ij} \geq 0\right] + v_{0,j} + e_{0,i,j} \\ X_{1,j} = \alpha_1 + v_{1,j} + e_{1,j} \\ X_{2,j} = \alpha_2 + v_{2,j} + e_{2,j} \end{cases}$$

$$\begin{pmatrix} v_{0,j} \\ v_{1,j} \\ v_{2,j} \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{\Omega_u}) \qquad \begin{pmatrix} e_{0,i,j} \\ e_{1,j} \\ e_{2,j} \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{\Omega_e})$$

Note that because X1 and X2 are time fixed, then time is not included in as predictor of the imputation model of X1 and X2. This model [Equation 3.7] can be fitted with a standard Gibbs sampler for generating *m* complete datasets. Thus, [Equation 3.6] can be fitted in each of the *m* datasets, and the estimates of interest (usually β_1 and β_2) can be summarised by applying Rubin's rules. However, the imputation model [Equation 3.7] is not compatible with the substantive model [Equations 3.6] since it does not include the random effects (i.e. the conditional distribution of Y given X1 and X2 derived from [Equation 3.7] is not [Equation 3.6].

3.2.5. Substantive Model Compatible Joint Modelling MI for ITS

The substantive model compatible joint modelling multiple imputation (MI-JOMO) is a novel solution for imputing missing values [56–58], that is well suited for ITS studies performed with individual-level data. MI-JOMO is an imputation method that assumes a joint multilevel multivariate normal model for the partially observed data. Broadly, MI-JOMO applies a Bayesian approach to factorise the joint distribution in two terms: a

joint model for the covariates of the analysis model and a conditional model for the outcome given the covariates. MI-JOMO defines an imputation model compatible with [Equation 3.6] by factorising the joint distribution of the three variables ($Y_{i,j}$ $X_{1,j}$ $X_{2,j}$) in two terms: 1) a joint model for the covariates and 2) a conditional model for the outcome given the covariates:

[Equation 3.8]

[Subcomponent 3.8.1]

$$\begin{cases} X_{1,j} = \alpha_1 + v_{1,j} + e_{1,j} \\ X_{2,j} = \alpha_2 + v_{2,j} + e_{2,j} \end{cases}$$

$$\begin{pmatrix} v_{1,j} \\ v_{2,j} \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{\Omega_u}) \qquad \begin{pmatrix} e_{1,j} \\ e_{2,j} \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{\Omega_e})$$

[Subcomponent 3.8.2]

$$Y_{ij} = \beta_0 + u_{0j} + (\beta_1 + u_{1j}) time_{ij} \times 1[time_{ij} < 0] + (\beta_2 + u_{2j}) time_{ij} \times [time_{ij} \geq 0] + X_{1j} + X_{2j} + \epsilon_{ij}$$

$$\begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{pmatrix} \sim N \begin{pmatrix} 0 \\ 0, \Sigma_u \\ 0 \end{pmatrix} \qquad \epsilon_{ij} \sim N(0, \sigma_e^2),$$

As before, a standard Gibbs sampler is needed to create *m* different imputed datasets. To perform the imputations of the covariates ($X_{1,j}$ $X_{2,j}$), an additional Metropolis-Hastings[10] step within the Gibbs sampler is needed. These imputations are drawn from a proposal distribution, and the Metropolis ratio (which takes into account the substantive regression model for Y) defines acceptance probabilities. With a symmetrical proposal distribution, the Metropolis ratio is equal to the likelihood of the model with the new proposed imputed value divided by the likelihood of the model with the previous imputed value. This iterative procedure forces the conditional model in [Equations 3.8] to be congenial with the substantive model [Equation 3.6]. The method can be readily extended to include interactions and non-linearities in the

[10] The Gibbs sampler is a special case of the Metropolis-Hastings (MH) sampler, set up so that each proposal is accepted. The software uses a Gibbs sampler with additional MH steps to handle categorical data through the latent normal model.

conditional model of the outcome given the covariates in [Equations 3.8, Subcomponent 3.8.2]. This is an important feature that allows ITS studies to include other characteristics; for example, adding a control group (for which interactions are useful) or making the change over time smoother (non-linearity).

3.3. Restricted Maximum Likelihood for ITS

In practice, MI-JOMO can be very useful for handling missing covariates (e.g. $X_{1,j}$ $X_{2,j}$) in ITS studies performed with individual-level data. On the other hand, it can be computationally inefficient for imputing outcome values (Y_{ij}) that are missing due to the irregular recording over time that is expected in EHR. However, since MI-JOMO is compatible with the mixed-effects model of interest [Equation 3.6], it can be used for imputing missing covariates only. On the *m* datasets imputed by MI-JOMO, now containing covariates fully observed, the [Equation 3.6] can be fitted using a Restricted Maximum Likelihood (REML) estimator. REML implicitly imputes outcome values (Y_{ij}), simplifying the overall procedure (i.e. no explicit multiple imputation of the outcome is needed). Then, Rubin's formulas will allow summarising the final ITS estimates of interest.

Since REML is closely related to the maximum likelihood (ML) estimator, I will first explain how ML makes the implicit imputations and then the difference between REML and ML.

In a dataset with an outcome Y_{ij} irregularly recorded over time, ML estimator will use all the available data to identify the population parameters with the highest probability of producing the sample data [59]. It means that the estimator does not discard individuals with incomplete Y_{ij}. For doing this, ML assumes a multivariate normal distribution and computes the log-likelihood of each individual *j* as follows:

[Equation 3.9]

$$logL_j = -\frac{k_j}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(Y_j - \mu)^T\Sigma^{-1}(Y_j - \mu)$$

where k_j is the number of observed scores for individual j, Y_j is the vector of observed values over time, and μ and Σ are estimates of the population mean vector and covariance matrix (at a particular computational cycle), respectively. We can easily calculate the marginal likelihood of each individual's observed data, from the fully likelihood [Equation 3.9], and use this to include their observed data in the analysis. Based on [Equation 3.9], the estimation process is mainly driven by:

[Equation 3.10]

$$(Y_j - \mu)^T\Sigma^{-1}(Y_j - \mu)$$

[Equation 3.10] quantifies the standardised distance between an individual's observed outcome value and the parameter estimates. A small distance (i.e. a high log-likelihood value) results when an individual's outcome value is close to the outcome mean. As ordinary least squares estimation does, ML estimator finds the parameter estimates that minimise the sum of all the individual standardised distances to the data. This final likelihood is calculated by aggregating all individual log-likelihood values across the complete sample, as follows:

[Equation 3.11]

$$log\ L = \sum \log L_j$$

Commonly, ML uses an iterative optimisation algorithm that repeatedly tests different parameter values (i.e. a modified Newton-Raphson algorithm in Stata), until it locates the estimates that maximise the log-likelihood [Equation 3.11].

The final (overall) log-likelihood come from [Equation 3.11], but the key of missing data handling relies on the way data from each individual is analysed by [Equation 3.9]. In this equation, the data and parameter vectors belong to the same individual. This

individual contribution to the log-likelihood is calculated with all the available outcome data at the individual level, meaning the number of outcome values and the pattern of its missingness can vary across individuals.

This estimation process based on individual likelihoods does not explicitly impute the missing values but borrows information from the observed outcome values [59]. For example, given the multivariate normal distribution assumption, an individual with high outcome values at previous time points will be more likely to have high outcome values in latter time points. If these latter points are missing data, the observed values and the normality assumption will be enough to compensate the information lost. There is no explicit imputation in this process, but to reach consistent results, the multivariate normality assumption is essential [59].

The REML is a particular form of ML that uses a likelihood function calculated on a transformed dataset. First, the algorithm generates contrasts of the outcome (linear combination of variables whose coefficients add up to zero) that do not depend on the fixed effects but instead depend on the variance components to be estimated. Then, the likelihood is calculated from the probability distribution of the contrasts. The procedure of computing individual likelihood, and then adding all of them into an overall likelihood, is similar to ML. The justification for REML is that having used the dataset of contrasts, it can produce unbiased estimates of variance and covariance parameters, for many models, which ML cannot do in general.

ML estimates of variance components tend to be biased downward because they do not incorporate the degrees of freedom used by REML to estimate the fixed effects. On the other hand, likelihood-ratio (LR) tests based on ML are correct, whereas, with REML, the LR-test can only compare models with similar fixed-effects specifications [60].

3.4. Missing data handling in ITS with EHR

In this chapter, I reviewed how missing data is a real problem in electronic health records (e.g. THIN data) and three generic mechanisms that can explain missingness (MCAR, MAR and MNAR). In the analysis, some issues on validity and efficiency can emerge with data MAR or MNAR, so they need special handling. Complete case analysis can provide unbiased estimates in some cases, but with the risk of losing a large amount of data (e.g. in longitudinal studies). Conversely, multiple imputation (MI) can provide valid inferences -without sacrificing data - when the imputation model is compatible with the substantive model. For ITS performed with individual-level data, the compatibility is not easy to achieve due to the necessity of including random intercepts and slopes. If the data is MAR in the outcome only, the implicit imputation performed by REML provides valid inferences in an efficient manner. If the data is MAR on covariates, the substantive model compatible joint modelling multiple imputation (MI-JOMO) is a novel a flexible method that can provide valid inferences.

In Chapter 5, I will show an application of ITS analysis with mixed-effects models (MEM with REML estimator). In Chapter 6, I will test an efficient way to combine MI-JOMO with MEM in ITS analysis. In Chapter 7, I will show an application of the two-step procedure (MI-JOMO plus MEM) to handle missing data in ITS with individual-level data. Before doing all these, I will review how this missing data problem is addressed in recent ITS studies on health topics, in the next Chapter 4.

4

Current practices in missing data handling for Interrupted Time Series studies: A scoping review

4.1. Introduction

Interrupted time series (ITS) is a widely used quasi-experimental approach that evaluates the potential impact of an intervention over time, using longitudinal data [6]. ITS is becoming more widespread in health research in the last decade [7,61]. The use of observational patient-level data is also more frequent in ITS [7,62–64], but routinely collected health data usually bring missing data issues [32]. Hudson et al. [7] detected that only 5% of the ITS studies in healthcare reported how missing data were handled. Current recommendations in the ITS literature [6] focus the attention on autocorrelation, seasonality and sample size as potential sources of bias, whereas little advise is given on reporting and handling of missing data.

Missing data management and statistical analysis can be crucial for any ITS study. In a preliminary search [7,62–64], I have identified two practices among researchers that could affect the validity of ITS estimates. First, before any statistical analysis, researchers opt to aggregate individual-level data into population-level data. For example, they average all the available outcome values at each predefined time point

(e.g. month) and use these averages as population-level outcome values in the subsequent time-series analyses. I call this the 'averaging-step' and, as I will explain later, this can lead to bias in the ′aggregate-level′ data analysis. Second, researchers are using statistical tools/approaches for modelling these aggregate-level data (e.g. ARIMA) that have not been designed to account for missing data at the individual level.

ITS guidelines recommend controlling for other potential confounders, such as autocorrelation or seasonality, using tools designed for population-level analyses and ignoring the missing data problem at the individual level [6]. Autocorrelation, whereby two consecutive data points can be more correlated with each other, could appear smaller than it is if there are different people at different time-points (i.e. missing outcome records). Then, missing data will impact on the trajectories estimates. Seasonality, which is defined by cyclic patterns on the outcome over time at population-level, can also be distorted by unseasonal missing data patterns at individual-level. Thus, traditional approaches to control for seasonality could be insufficient as well.

Four systematic reviews exploring methodological characteristics of health research with ITS designs have been performed in the past [4,7,61,65]. They have contributed to detecting gaps in reporting and the use of standard ITS analyses. However, these studies did not focus on the particular problem of applying these standards (mainly designed for population-level data [6]) to the analysis of individual-level records with missing values. In particular, the missing data issues related to the averaging-step and the selected statistical approach have been ignored by previous methodological or review studies of ITS.

I decided to look further into the practices in missing data handling and analysis that are prevalent in the ITS literature; particularly, how researchers are addressing the problem of having missing data at the individual level. Thus, I conducted a scoping review.

4.2. Objectives

The general aim was to determine the current practices in missing data handling for interrupted time series studies performed for health research. Particularly, I was interested in those studies that had access to individual-level data. For achieving this aim, I reviewed ITS investigations on health topics to:

1. Determine the data management strategies and statistical analysis performed in these studies
2. Determine how often missing data were considered and, if so, how they were evaluated, reported and handled in the analysis

4.3. Methods

I performed this study following the steps previously specified in a scoping review protocol (see Appendix 4A), and paying attention to standard recommendations from the *PRISMA Extension for Scoping Reviews* [66]. 'Scoping reviews' are alternatives to systematic reviews, especially suitable for investigating more general questions such as common practices in research [67]. For that reason, scoping reviews usually omit any critical appraisal within the reviewed articles, do not assess the risk of bias across the studies, and should not end with meta-analyses [68].

4.3.1. Inclusion and exclusion criteria

I included studies of ITS that assessed any intervention relevant to health care (e.g. policies or programs), with no restrictions on participants, the language of publication or the type of outcome. I excluded grey literature (e.g. government reports), systematic reviews, meta-analyses, randomised controlled trials, protocols, editorials, letters to editors, retraction papers, methodological studies, studies with no access to individual-level data, and studies that used Google Trends data only. Studies whose full text was not available -after trying several avenues - were also excluded. The access to individual-level data was verified in the methods section of each article, usually in the subsection describing the settings, sample or population studied (e.g. if they reported

data routinely collected from patients). If the authors did not describe their data as individual-level data, then the article was excluded.

4.3.2. Search strategy

I used the MEDLINE(R) and Epub Ahead of Print, In-Process & Other Non-Indexed Citations and Daily database (Ovid version) on 8 February 2020, to identify ITS studies published from 01 January 2019 to 31 December 2019. Search strategies utilised a combination of free-text terms and subject headings, and co-authors were consulted for the revision of search terms. The search strategy was reviewed by an information specialist from the UCL Library. Details of the search strategy are in Appendix 4B.

I performed the automatic search with the Ovid tool, removed duplications using EndNote, and screened titles and abstracts manually by the search for inclusion. My colleague Frank Peralta (FP) double assessed 10% of same titles and abstracts and, if we were in agreement, I would proceed to screen using a full-text version of the publications. FP also assessed 10% of the full-text and, in the absence of disagreement; I would proceed to randomly select 60 publications for final data extraction, using a random number generator with Stata. A third colleague was available to help in any disagreement at any stage of this process, whenever it was needed.

4.3.3. Data extraction and analysis

Data extraction form (Appendix 4C) was reviewed and validated by the co-authors and the information specialist from the UCL Library. From the 60 selected publications, 5 full-text original articles were randomly selected and were reviewed by FP and myself independently. If there were no disagreements, I would proceed to review the other fifty-five articles.

Data extracted from the articles can be categorised into three topics:

i. **General characteristics of the ITS studies**: first author, journal, country, study design (e.g. ITS, controlled ITS), participants, type of intervention, level of intervention (e.g. country, hospital), most granulated cluster available (e.g. hospital, individual-level) and longitudinal follow-up (e.g. prospective cohort or panel).
ii. **Data management and statistical analysis**: data source, linked data, outcome type (e.g. continuous, proportion), number of time points, time point unit, averaging-step (yes/no), statistical model (e.g. ARIMA, mixed-effects models), confounder reported (yes/no) and confounder adjusted for (yes/no), autocorrelation (considered, tested with, concluded by test, controlled by), seasonality (considered, tested with, concluded by test, controlled by), time-dependent variable (considered, handled by) and other methodological issues (considered, handled by). It is important to clarify the averaging-step definition, which is "the step from which the outcome analysed at the population level is the average of all more granulated outcome data (e.g. individual-level data) at each time-point defined for the ITS (e.g. one average outcome for each week)".
iii. **Missing data reporting and handling**: missing data considered (yes/no), proportion reported, the missing data mechanism (considered and reported), the method for handling missing data (considered, reported) and sensitivity analysis (considered, reported).

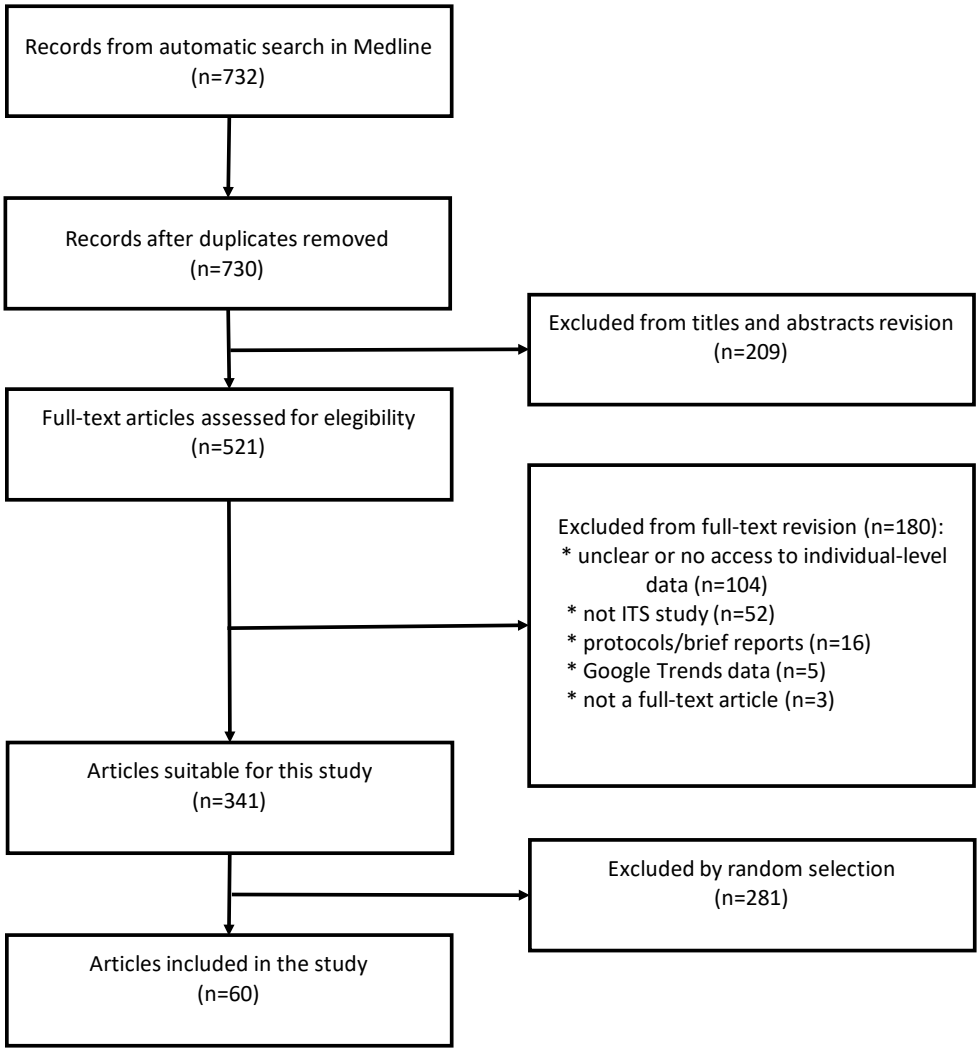
I based the data extraction on the primary outcome, or the first outcome mentioned if the authors did not set a primary outcome.

I summarised data using descriptive statistics (numbers and percentages or median, inter-quartile range, minimum and maximum values). Some cross-tabulations of frequencies were used when needed, and these are reported as supplementary material.

4.4. Results

The search strategy identified 732 titles and abstracts from Medline Ovid (Figure 4.1). After removing two duplicates, I excluded 209 titles and abstracts that did not meet the inclusion criteria, leaving 521 full-text studies to be checked for eligibility. From this full-text selection, I excluded 180, most of them having unclear or no access to individual-level data (104). After exclusion, 341 articles were suitable for the final screening; thus, I randomly selected 60 of them, and the list of these studies is provided in the Appendix 4D. Since there were no disagreements during the screening and data extraction process, FP and I only double assessed 10% of the full-text copies.

Figure 4.1. PRISMA diagram for the scoping review



Most of the 60 studies were from the US (n=28, 47%), the UK (n=7, 12%) or Canada (n=4, 7%) (Table 4.1). Only two studies (3%) were not labelled as interrupted time series studies, although they used an ITS design (they described themselves as segmented regression analyses). 48 (80%) studies using a single-ITS approach (i.e. no control group, see Chapter 2 for a more detailed description), whereas 10 (17%) applied a more sophisticated Controlled-ITS design [3]. Patients were the most prevalent participants (n=38, 63%), followed by health personnel (n=6, 10%) or general population (n=4, 7%). Policies (n=16, 27%), programs (n=14, 23%) and focused interventions (n=15, 25%) were the most common interventions evaluated in the ITS studies. Frequently, these interventions were applied at a hospital (n=18, 30%) or country level (n=17, 28%). Although all these studies (n=60, 100%) had access to individual-level data, due to the

nature of the studied ITS outcome (e.g. number of new patients before and after the intervention), only 32 (53%) could have followed/analysed the ITS outcome at an individual-level (i.e. repeated measures of the ITS outcome within individuals). However, other less granulated clusters could have been followed over time, for example, hospitals (n=13, 22%), hospital units (n=3, 5%), health facilities (n=3, 5%) or GPs (n=3, 5%) (i.e. ITS with multiple groups). A cross-table between the level of intervention and more granulated clusters available is in Appendix 4E. This provides an approximation of how often a researcher could move from modelling population ITS trajectories with population-level data points (i.e. only one ITS outcome average at each time-point), to model same trajectories with more granulated data, which was also available (e.g. individual- or hospital-level data for being modelled with mixed effects or GEE models). The longitudinal follow-up of the data collected for all these ITS studies was mostly retrospective and was available at individual-level (n=25, 42%) or other less granulated cluster levels (n=21, 35%).

Table 4.1. Characteristics of the included interrupted time series studies (N=60)

	n	(%)
Country of Study		
Australia	1	(1.7)
Bangladesh	1	(1.7)
Brazil	2	(3.3)
Cambodia	1	(1.7)
Canada	4	(6.7)
China	2	(3.3)
France	2	(3.3)
Germany	1	(1.7)
Israel	1	(1.7)
Italy	1	(1.7)
Japan	1	(1.7)
Malawi	1	(1.7)
Netherlands	1	(1.7)
Rwanda	1	(1.7)
Saudi Arabia	1	(1.7)
South Korea	1	(1.7)
Spain	2	(3.3)
Switzerland	1	(1.7)
UK	7	(11.7)
USA	28	(46.7)
Study Design		
CITS	10	(16.7)
ITS	48	(80)
SR	2	(3.3)
Participants		
Children	3	(5)

Firefighters	1	(1.7)
general population	4	(6.7)
health personnel	6	(10)
health personnel & patients	6	(10)
insured women	1	(1.7)
Medications	1	(1.7)
Patients	38	(63.3)
Type of Intervention		
guideline/protocol/sound publication or evidence	9	(15)
focused intervention	15	(25)
Policy	16	(26.7)
Program	14	(23.3)
relevant or historic event	3	(5)
Treatment	3	(5)
Level of Intervention		
cities, group of	1	(1.7)
city/district	3	(5)
Country	17	(28.3)
Hospital	18	(30)
hospitals, group of	8	(13.3)
individual-level	2	(3.3)
state/province/county	10	(16.7)
fire departments	1	(1.7)
Most Granulated Cluster Available		
GP	3	(5)
District	2	(3.3)
fire department	1	(1.7)
group of patients (by Dx)	1	(1.7)
health facility	3	(5)
Hospital	13	(21.7)
hospital unit	3	(5)
Household	1	(1.7)
individual-level	32	(53.3)
Medications	1	(1.7)
Longitudinal Follow-up		
prospective cohort (individuals)	8	(13.3)
prospective panel (cluster)	6	(10)
retrospective cohort (individuals)	25	(41.7)
retrospective panel (cluster)	21	(35)

SR: segmented regression; ITS: interrupted time series: CITS: controlled interrupted time series; GP: general practice; Dx: diagnostic

In most studies, data were routinely collected (n=46, 77%) which is often includes missing data, since the data collection procedures were not designed for an ITS study or even for any research (e.g. ITS outcomes were not collected at similar intervals across patients; such as weight measurement may have been taken at different times) (Table 4.2). Data collected were not usually linked to external data (n=10, 17%). The most common ITS outcome type was proportions (n=39, 65%) and the most usual unit of the follow-up time was a month (n=36, 60%). The median number of time-points used in the ITS analysis was 38 (IQR=55). The averaging-step was performed in 47 (78%) of the

studies. One example of the averaging step was performed by Close et al [69]. They averaged the number of stroke admissions per practice per month (73 practices observed from 2011 to 2018), for modelling the ITS on the aggregated data (i.e. one average point at each month). The most typical statistical models were the segmented regression (SR) fitted with ordinary least square estimators (SR-OLS, n=23, 38%) or with maximum likelihood type estimators (SR with generalised linear models or SR-GLM, n=15, 25%). A cross-table between averaging-step and statistical model is available in Appendix 4F, showing how researchers combine them in standard ITS studies. Confounding was reported in 41 (68%) of studies, but researchers adjusted for confounding in just 33 (55%) studies.

Table 4.2. Data and statistical analyses of the included interrupted time series studies (N=60)

	n	(%)
Data Source		
collected for the study (prospective)	14	(23.3)
routinely collected (retrospective)	46	(76.7)
Linked Data		
No	50	(83.3)
Yes	10	(16.7)
Outcome Type		
Continuous	10	(16.7)
Count	11	(18.3)
Proportion	39	(65)
Time Points Number		
median (IQR)	38	(55)
Minimum	6	
Maximum	1217	
Time Points Unit		
Day	3	(5)
half-year	1	(1.7)
Month	36	(60)
quarter year	8	(13.3)
two-month	1	(1.7)
Week	5	(8.3)
Year	6	(10)
Averaging-step		
No	11	(18.3)
Yes	47	(78.4)
Unclear	2	(3.3)
Statistical Model		
ARIMA	7	(11.7)
Joint-point (Exploratory Method)	1	(1.7)
SR-GEE	7	(11.6)
SR-GLM	15	(25)
SR-GLS	1	(1.7)
SR-OLS	23	(38.3)
Mixed Effects (random intercept only)	4	(6.7)

Mixed Effects (random intercept & slopes)	2	(3.3)
Confounder Reported		
No	19	(31.7)
Yes	41	(68.3)
Confounder Adjusted		
No	27	(45)
Yes	33	(55)

Many researchers considered the autocorrelation problem (n=41/60, 68%) (Table 4.3). However, descriptions about how they tested and handled autocorrelation were sporadic. For example, 1/3 of them did not report the test they applied -if so- for evaluating autocorrelation (n=13/41, 32%). This was different for those who worked with individual-level data and fitted GEE or mixed-effects models, reporting within-individual correlation by design (n=11/41, 27%), since they did not usually address the autocorrelation problem at the population-level. Among those who identified autocorrelation issues in their data (n=36/41, 88%), the use of Newey-West standard errors (n=7/36, 19%) or autoregressive errors terms (n=8/36, 22%) was preferred.

The seasonality issue was considered in about 1/3 of the studies (n=19/60, 32%) and, in most cases, it was not formally tested (n=14/19, 74%). Even in those studies with observation periods >1 year (n=52/60, 87%), seasonality was considered in a similar proportion (n=17/52, 33%). Regardless of the use of a formal test - graphical inspection may have been used, but not described - 18/19 (95%) concluded seasonality effects and did something to control for it. The most popular way to control for seasonality (n=12/18, 67%) was to include covariates of time (e.g. dummy variables of months) in the ITS models.

Most studies (n=49/60, 82%) considered time-dependent confounding could not be handled by a single ITS design [3]. However, more than 2/3 of the studies only reported the problem as a limitation (n=34/49, 70%), whereas less than 1/4 (n=10/49, 20%) used a control group to address the limitation [11]. Other methodological issues related to the ITS design were also considered (n=25/60, 42%), using sensitivity analyses to evaluate the impact of these issues on the results (n=6/25, 24%). For example, sensitivity analyses were used by extracting groups of patients in order to understand whether

[11] I explain how this limitation is addressed by including a control group in Chapter 2.

unmeasured events - potentially experienced by some groups - could affect the ITS outcome trajectories (n=3/25, 12%). Likewise, sensitivity analyses were applied to contrast the pre-selected ITS impact model [6,8] against other feasible models (n=3/25, 12%).

Table 4.3. Reporting and handling of methodological issues in the included interrupted time series studies (N=60)

	n	(%)
Autocorrelation - Considered (n=60)		
No	19	(31.7)
Yes	41	(68.3)
Autocorrelation - Tested With (n=41)		
Breusch-Godfrey	2	(4.9)
Cumby-Huizinga	1	(2.4)
Durbin-Watson	8	(19.5)
within-individual correlation by design	11	(26.8)
autocorrelation function	3	(7.3)
autocorrelation probability	2	(4.9)
not specified	13	(31.7)
residuals examination	1	(2.4)
Autocorrelation - Concluded by Test (n=41)		
No	5	(12.2)
Yes	36	(87.8)
Autocorrelation - Controlled by (n=36)		
Cochrane- Orcutt	1	(2.8)
GEE models	6	(16.7)
Newey-West standard errors	7	(19.4)
Prais-Winsten	2	(5.6)
Autoregressive error term	8	(22.2)
Mixed-models	5	(13.9)
Not specified	7	(19.4)
Seasonality - Considered (n=60)		
No	41	(68.3)
Yes	19	(31.7)
Seasonality - Tested With (n=19)		
Dickey-Fuller	1	(5.3)
autocorrelation/partial autocorrelation function	2	(10.5)
no formal test	14	(73.7)
not possible (short period)	1	(5.3)
regression diagnosis test	1	(5.3)
Seasonality - Concluded by Test (n=19)		
No	1	(5.3)
Yes	18	(94.7)
Seasonality - Controlled by (n=18)		
ARIMA parameter	1	(5.6)
covariate in the model	12	(66.7)
Decomposition	1	(5.6)
not handled (reported as a limitation)	2	(11.1)
seasonal ARIMA	2	(11.1)
Time-Dependent Variable - Considered (n=60)		
No	11	(18.3)

Yes	49	(81.7)
Time-Dependent Variable - Handled by (n=49)		
control group	10	(20.4)
control outcome	1	(2)
covariate (exploration)	1	(2)
covariate in the model	3	(6.1)
reported as a limitation, not handled	34	(69.4)
Other Issues - Considered (n=60)		
No	35	(58.3)
Yes	25	(41.7)
Other Issues - Handled by (n=25)		
Bonferroni adjustment (p values)	1	(4)
adjusted for survey design	1	(4)
aggregate ecological design (reported as a limitation)	1	(4)
confounders not controlled (reported as a limitation)	1	(4)
minimize immortal time bias	1	(4)
non-stationary (ARIMA controlled)	2	(8)
overdispersion evaluation (Poisson models)	2	(8)
secular trends (reported as a limitation)	2	(8)
sensitivity analysis (extracting patients)	3	(12)
sensitivity analysis (impact model)	3	(12)
sensitivity analysis (various)	6	(24)
sub-group analysis	2	(8)

Sensitivity analyses were used to evaluate missing data in two studies only (n=2/60, 3%) (Table 4.4). In one study, they compared results from multiple imputation by chained equations against results from complete case analysis (CCA). In the other study, they compared results from using a missing data category -for variables with missing data – against results from CCA. In general, only 13/60 studies (22%) reported issues related to missing data, with considerable variability across the proportion of missing values reported. Although many studies worked on retrospectively collected data (n=46/60, 77%) (Table 4.2), with an irregular recording expected for any outcome at the individual level (n=32/60, 53%) (Table 4.1), only one study (n=1/13, 8%) explicitly reported this problem (missing data on the ITS outcome <60%) (Table 4.4).

Only two studies (n=2/60, 3%) considered the missing data mechanisms and their implications on the analysis, reporting missing at random (MAR) and missing not at random (MNAR) as potential mechanisms behind their missing values (Table 4.4). About ¼ of the studies reported the method used for handling missing data, CCA being the most popular (n=14/16, 87%). Interestingly, in 2/6 investigations using mixed-effects models, researchers recognised the boundaries of these models in handling missing data on the ITS outcome.

Table 4.4. Reporting and handling of missing data issues in the included interrupted time series studies (N=60)

	n	(%)
Missing Data - Considered (n=60)		
No	47	(78.3)
Yes	13	(21.7)
Missing Data - % Reported (n=13)		
% not reported, but declared as an issue to be solved	2	(15.4)
covariates <30% / outcome <50%	1	(7.7)
covariates at baseline (<1% each, not combined)	1	(7.7)
covariates at baseline (<10% each, not combined)	2	(15.4)
covariates at baseline (<2%, flow chart)	1	(7.7)
covariates at baseline (<25% each, not combined)	1	(7.7)
covariates at baseline (<25%, flow chart)	1	(7.7)
covariates at baseline (<30% each, not combined)	1	(7.7)
covariates at baseline (<5%, flow chart)	1	(7.7)
outcome <60%	1	(7.7)
smoking (one case), but reporting outcome irregular recording as a problem	1	(7.7)
Missing Data Mechanism - Considered (n=60)		
No	58	(96.7)
Yes	2	(3.3)
Missing Data Mechanism - Reported (n=2)		
MAR	1	(50)
MNAR	1	(50)
Method for Handling Missing Data - Considered (n=60)		
No	44	(73.3)
Yes	16	(26.7)
Method for Handling Missing Data - Reported (n=16)		
Complete Case Analysis (CCA)	14	(87.4)
Mixed intercept model for handling missing outcomes	1	(6.3)
Mixed intercept and slope model for handling missing outcomes	1	(6.3)
Sensitivity Analysis for Missing Data Mechanism - Considered (n=60)		
No	58	(96.7)
Yes	2	(3.3)
Sensitivity Analysis for Missing Data Mechanism - Reported (n=2)		
comparing results from MICE versus CCA	1	(50)
comparing results from using a 'missing data category' versus CCA	1	(50)

MICE: multiple imputation by chained equations; MAR: missing at random; MNAR: missing no at random

4.5. Discussion

I identified at least five methodological issues directly associated with missing data handling in ITS studies. First, many studies have been using the averaging-step for summarising the ITS outcome at each time point, even if they had the opportunity of modelling the outcome with longitudinal individual-level data directly (or at least with data more granulated than the population level). Second, analysis tools for population-level data (e.g. aggregate-level SR) were preferred over analysis tools for individual-level data (e.g. mixed-effects models). The first and second have implications regarding bias when data are missing, as I will explain in the next paragraphs and throughout the thesis. Third, missing data on covariates at baseline are commonly handled with CCA, losing valuable information and potentially leading to bias if ITS estimates are adjusted for these covariates. Fourth, seasonality and other time-dependent confounders are barely controlled and, when they are, missing data implications were typically ignored. Finally, missing data reporting is absent in most of ITS studies and further reflections on the potential consequences of missing data mechanisms, and the subsequent selection of best methods to handle missing values, are very rare in this type of studies.

The averaging step forces the data missing at the individual level to -artificially - disappear at the population level, generating the false impression of an issue controlled. For example, if the outcome data are missing at random conditional on a fully-observed covariate at the individual level, and we calculate a simple average across subjects at each time point, the covariate that explains the outcome missingness will become unobserved at the aggregate level (e.g. if there are more missing data on the outcome on men than on women and a general outcome average –not weighted by sex- is calculated). Therefore, the mechanism will be missing *not* at random at the aggregate level. This is a potential source of bias that none of the 60 studies I reviewed has mentioned as a limitation. The data used in ITS studies are mostly retrospective (i.e. routinely collected); often, the expected amount of missing data on the outcome is high. This is particularly important for ITS designs, for which it is expected to have the outcome regularly measured at each time point [2]. With many outcome gaps due to irregular recording, researchers commonly select convenient periods as units of time (e.g. months) and average all the available records to set a unique outcome value for each time point/unit. All this with no significant reflection on the missing not at

random consequence explained above. Modelling the ITS at the individual level could be a better option (e.g. using mixed-effects models) but, as I confirmed in this review, this is not a common practice. I demonstrate how to use individual-level models with missing outcomes in Chapter 5 and provide a more detailed explanation of the issues associated with the averaging-step in Chapter 6.

The frequent use of the averaging-step seems to be related to the standard use of fixed-effect models which is abundant in the ITS literature and guidelines [6,7]. Before becoming popular in health research, ITS approach became very common in the evaluation of policies or national level interventions, for example, in economic or education research [2]. In these interventions, the ITS outcome at the population level (e.g. published annual national prevalence over long periods) was enough to model the time series; thus, the individual-level data behind the national figures were not required or even considered. At the population level, fixed-effects models such as segmented regression with ordinary least square or generalised linear models took place as the standard practice recommended by the ITS guidelines [6]. At the same time, other statistical tools for handling issues related to time series (e.g. autocorrelation or seasonality) were designed as complement or extension of these models [16]. In recent decades, access to massive individual-level data -of good quality - started to be more universal [37]. Regardless of this progress, the standard ITS recommendations are still the same: they are based on modelling population-level outcomes ignoring any potential missing data at the individual level. Researchers would have seen the averaging-step like an intuitive way to adapt the -now more available - individual-level data into the population-level data that guidelines teach to model [6]. With no methodological studies on the consequences of the averaging-step or recommendations from the ITS guidelines on how to handle missing data [61], researchers do not have sufficient information of the existence of the issue such that they would be motivated to improve practice.

Most researchers handle missing data on confounders at baseline by using complete case analysis (CCA), taking similar actions when modelling interaction terms in CITS studies, but again without major reflections on the implications. For example, if an observation was dropped by the CCA, the final ITS estimates could reduce precision or be biased. Every time an individual record is omitted due to missing data on a covariate,

we are losing not only sample size but also some outcome records as well. If the records with missing covariates contain a systematically high or low range of outcome values, and we drop these observations, then the average from CCA at each time point could be biased. Further, if this bias differs across time-points, then estimators of trajectories will themselves be biased. I have confirmed that the most popular method for handling missing data in the reviewed studies was CCA. Even among those studies not reporting the method used, but in which the ITS effect estimated was controlled for baseline confounders at the aggregate level, it seems likely that CCA was the selected procedure. Only one study reported the use of multiple imputation and used the standard chained equation method (MICE) [48]. This method brings congeniality problems when it is applied to multilevel data (e.g. individual follow up, and adjusting for a dummy month variable at an aggregate level to control for seasonality) [55]. Reaching a congenial imputation model is more complicated for MICE when researchers need to introduce time-variant confounders or interaction terms in the models, both expected steps in many CITS studies [3]. For these more complex scenarios, multilevel multiple imputation methods are preferred [56] due to their flexibility to introduce interaction terms and confounders at different levels. I demonstrate the advantage of using multilevel multiple imputation in Chapter 6 and provide an example of its application for interaction models in Chapter 7.

Seasonality and other time-dependent confounders are barely controlled in ITS studies and, when they are, missing data implications are typically ignored. Following an averaging-step procedure, the missing data at the individual level can affect the way the seasonality is observed at the aggregate level, for example, if the missingness proportion is not equal across seasons. Since the preferred method to control for seasonality seems to be to include a dummy variable of time in the models (e.g. month), the control will still incorporate noise from the points describing the seasonality at the aggregate level. Using mixed-effects to model the ITS with individual-level data, the seasonality could be controlled by specifying the structure of residual errors (i.e. when the intervention is applied at the population level). Under missing at random assumption, these mixed effect models provide unbiased estimates [70]. However, this or similar control alternatives have not been reported by any of the studies included in this review.

In ITS studies, there is a lack of missing data reporting, and further reflections on the potential consequences of missing data mechanisms and on the best methods to handle missing values are needed. Previous reviews have found an even lower proportion of missing data reporting [7], which indicates this gap is in the ITS literature. Only one study among all reviewed made some reflection about how a potential MNAR mechanism could affect its results [71]. However, no sensitivity analysis was performed by any study - to consider the impact that a possible MNAR mechanism could have on the final estimates [72], at least when individual-level data was available. Considering that most of the ITS studies are performed retrospectively (routinely collected data), the control that researchers can have on missing data is minimal; thus, a thorough evaluation/reflection on the missingness mechanisms is the only action that is viable in practice. After such an evaluation, the selection of the optimal method to handle missing values can be best informed, leading to more appropriate alternatives than CCA, which is almost never supported by a rationale in the ITS studies reviewed.

In general, study findings are consistent with those reported by other previous reviews. As I did, Polus et al. [65] identified that some ITS studies do not use the ITS label (e.g. one study could report using a segmented regression without mentioned they applied an ITS design). Polus et al. [65] concluded that not having that label could be an indicator of ignoring the study design characteristics of the ITS approach, leading researchers to a deviation from the standard ITS practice. From sixteen ITS studies published between 1976 and 2011, they found five with no statistical models and zero studies using mixed-effect models. My review focused on sixty publications from 2019, finding only six studies using mixed-effects models. This combined evidence tells us that researchers' preferences have not changed dramatically during the last decade. Jandoc et al. [61], who reviewed ITS studies in drug utilization published between 1984 and 2013, found that >92% of data sources were administrative data (routinely collected). As they reported a similar proportion of routinely collected data as me, I externally confirmed that the missing data prevalence for ITS is still a problem. They also found that, in two hundred studies reviewed, segmented regression (e.g. OLS or GLM) and ARIMA models were the most used by researchers, also reporting no use of mixed effect models. Hudson et al. [7] reported that continuous ITS outcomes were more frequent, whereas I found that proportion was the most common. Differences seem to come from the way Hudson would have typified the outcomes. For them, the outcome type would have been defined by the model used (e.g. if researchers fitted and

ordinary least square (OLS) model, then the outcome should be typified as continuous). However, taking into account that an averaging-step typically preceded the outcome modelling, the ITS outcome should be typified considering this step in order to detect other methodological problems. For example, when researchers calculate a proportion by using the number of events at each time point and then model the time series with this longitudinal set of aggregate-level proportions, they could be still using OLS models. Observations from my study confirm this approach is not atypical. Even if researchers use aggregate-level data - which should be avoided when outcome data are missing - the use of a model appropriate for the outcome type is always strongly recommended. More recently, Turner et al. [4] identified that individual-level outcomes of one type (e.g. binary) are often aggregated in population-level outcomes of another type (e.g. proportion, counts, rates and continuous), underusing individual-level analysis options (e.g. mixed-effects models). These findings confirm mine about data aggregation and analysis choices but also unveil how data manipulation before the statistical analysis occurs in many ways that can be affected by missing data at the individual level.

I recognise some strengths and limitations in this scoping review. On the strengths side, I followed standard recommendations for performing and reporting scoping reviews [66]. It is the first time that the missing data handling in ITS studies has been revised and analysed as the main aim of a review. The selected studies come from diverse countries and journals, and the results of this review are consistent with others reported in close studies [4,7,61,65], what is a good indicator of external validity. Among limitations, I only included data for one year (2019) but, considering outputs from previous investigations, there is no reason to believe that including years would have changed the conclusions. I could have missed some publications in the search process due to inconsistency in ITS studies reporting or the use of just one database (Medline). However, there are no strong reasons to believe that the review representativeness has been compromised due to these factors. I analysed a random sample from the population of publications (60/341) meaning my estimates vary about the population value. Then, although some information about missing data handling in ITS studies could have been omitted, the overall image of the problem studied here is still consistent.

In conclusion, I have confirmed that missing data reporting and handling in the most recent interrupted time series studies performed for health research was rarely performed. Researchers do not tend to evaluate the potential consequences of missing data mechanisms on their ITS estimates; their selection of proper methods for missing data handling is thus poorly reflected and informed. The complete case analysis is the most commonly applied method, but the control for confounders or interaction terms can be severely affected by its use. The averaging-step is also a widespread practice that can affect the validity of the ITS estimates when data are missing. To overcome these actual limitations, it is essential to include recommendations for missing data handling in ITS guidelines, as well as the exploration of mixed-effects models and multilevel multiple imputation as more efficient analysis alternatives in new methodological and applicative studies.

4.6. Summary

Interrupted time series (ITS) is a widely used quasi-experimental approach that is becoming more widespread in health research to answer causal questions with observational data. Current recommendations in ITS guidelines focus the attention on autocorrelation, seasonality and sample size as potential sources of bias, whereas little is advised on missing data reporting and handling. Therefore, issues associated with aggregate-level analysis and complete case analysis –both potential sources of bias when data are MAR- seem to be common practices in ITS studies. The study aimed to review recent ITS investigations on health topics for determining 1) the data management strategies and statistical analysis performed in these ITS studies; and 2) how often missing data were considered and, if so, how they were evaluated, reported and handled.

I performed a scoping review following standard recommendations from the *PRISMA Extension for Scoping Reviews*. I included all studies of ITS that assessed any intervention relevant to health care (e.g. policies or programs) published in 2019 with abstracts visible in Medline. The initial search detected 732 publications, which I systematically debugged and from which I finally selected 60 to collect data about 1) general characteristics of the ITS studies, 2) data management and statistical analysis and 3) missing data reporting and handling.

In the 60 ITS studies, the averaging-step and the subsequent aggregate-level SR were a widespread practice, bringing potential issues when data are missing. They applied the averaging-step, even if they had the opportunity of modelling the outcome with longitudinal individual-level data directly (or at least with data more granulated than the population level). As I will discuss through the next thesis chapters, the averaging-step transform data MAR into data MNAR (see Chapter 6 especially). This is a potential source of bias that none of the reviewed studies reported as a limitation, showing that researchers are in general not aware of the MNAR issue associated with this process. The data used in ITS studies are mostly retrospective (i.e. routinely collected); therefore, missing data on the outcome is often high. This is particularly important for ITS designs, for which it is expected to have the outcome regularly measured at each time point. With many outcome gaps due to irregular recording, researchers select time

periods as units of time (e.g. months) and average all the available records to set a unique outcome value for each time point/unit. All this is done with no significant reflection on the MNAR issue that can induce bias. The common use of the averaging-step with individual-level data MAR motivated my exploration of better analysis alternatives for handling missing data in ITS studies (i.e. MEM in the next chapter).

Missing data are poorly evaluated and reported in ITS studies, and statistical methods applied as standard are not robust against missing data issues. This study and other independent reviews have confirmed that missing data reporting is rare, suggesting that all the potential consequences of missing data on ITS estimates are unknown or at least undervalue by researchers. For example, the complete case analysis (CCA) is the most used method for handling missing data, but it can lead to biased and/or less precise estimates. Every time an individual record is erased due to missing data on a covariate, the researcher is losing not only sample size but also some outcome records as well. If the records with missing covariates contain a systematically high or low range of outcome values, and they drop these observations, then the average from CCA at each time point could be biased. Further, if this bias differs across time-points, then estimators of trajectories will themselves be biased. Better options for handling missing covariates such as multiple imputation (MI) has rarely been applied, and the only study that used MI did not consider a multilevel approach. I will evaluate MI-JOMO against other methods in Chapter 6 and show its potential for handling missing values on interaction terms in Chapter 7.

Before evaluating these alternatives on missing data handling for ITS analysis performed with individual-level data, in Chapter 5, I will start exploring the applicability of MEM. MEM can help to handle missing outcomes at the individual-level, avoiding any averaging-step and the potential bias issues associated with this when data are MAR. I will build most of the examples in the rest of the thesis based on the next application in Chapter 5.

5

An application of Interrupted Time Series with Mixed Effects Models

5.1. Introduction

Overweight and obesity is a worldwide problem that impacts severely on population health [73]. Since the prevalence of overweight and obesity is higher in individuals with severe mental illnesses than in the general population [74,75], their risk of harmful consequences is also higher [76,77]. Individuals with severe mental illnesses are more susceptible to developing metabolic syndrome, type-2 diabetes mellitus [78] and cardiovascular diseases [76,79], leading to a higher risk of death. Adults with schizophrenia have 3.5 times the mortality risk than the general population, with cardiovascular diseases as the most common cause [76,80]. Notably, Lahti et al. demonstrated that the risk of death is higher in women than in men with schizophrenia [81], suggesting that differences between sexes need to be further investigated.

Second-generation antipsychotics (AP) are a known cause of weight gain [5,82]. Some evidence suggests that women gain more weight than men during AP treatment [83]. One study suggested that women have five times the odds of increasing body mass index (BMI) compared with men after two years or more [84]. Twenty-nine per cent of women treated with clozapine gained ≥20% of their baseline body weight after two

years of follow up, in contrast to 13% of men [85]. Other studies have demonstrated similar differences between men and women [86,87]. However, most of these studies are based on small sample sizes of less than 200 individuals, and most do not distinguish between short and long-term weight gain associated with antipsychotic treatment.

Weight gain after initiation of antipsychotic treatment may also depend on body weight when treatment is initiated. Thus, Gebhardt et al. found that low body mass index (BMI) before first AP treatment predicted a faster increment of BMI after treatment initiation [87], and a similar conclusion was reached by Najar et al. [86]. On the other hand, there is limited information about how doses of antipsychotic treatment are associated with weight gain [5].

Interrupted time series (ITS) design provides a flexible framework for analysing the weight gain associated with the initiation of antipsychotic treatment in observational data [2,6]. By using electronic health records (EHR), it is possible to follow patients weight trajectories before and after antipsychotic treatment initiation and detect changes in these trajectories. This is possible for short and relatively long periods of observation, which is an advantage when clinical trials are not feasible [1]. ITS can also be modelled for analysing weight change given a specific dose or sex. Despite these advantages, EHR bring missing data issues that standard ITS tools (e.g. aggregate-level segmented regression) are not designed to address (see Chapters 2 and 3). However, modelling ITS with mixed effect models (MEM) is a flexible approach that can efficiently overcome common missing data problems in EHR.

5.2. Objectives

The study aim was to investigate the change in body weight of patients initiated with high or low doses of the three most commonly prescribed second-generation antipsychotics (olanzapine, risperidone and quetiapine) by using the ITS approach. The specific objectives were:

1) Clinically: to evaluate the short- and long-term change in body weight in men and women upon initiation of AP; whether this is different for low and high doses; and whether low body weight at treatment initiation had the highest weight gain.
2) Methodologically: to apply the ITS approach with mixed effect models (MEM) for handling missing data in longitudinal weight records.

5.3. Methods

5.3.1 Data source

I used anonymised, longitudinal patient records from The Health Improvement Network (THIN), a database that comprises information from UK primary care electronic health records from general practices as described in Chapter 2.

5.3.2 Study population

At the individual level, I included all patients aged between 18 and 99 years at the date they started their first treatment with one of the following three antipsychotics olanzapine, risperidone or quetiapine; between 1 January 2005 and 31 December 2015. I included patients with a diagnosed psychiatric disorder (schizophrenia, bipolar disorder, other non-affective psychoses, borderline personality disorder, anxiety, depression or dementia) who had at least one further prescription of the same AP within three months after the first prescription. I judged that these individuals were more likely to have initiated treatment than those with a single prescription. Patients who had been initiated on more than one type of AP were excluded (including switchers). A few individuals had no records of the year of birth, sex or social deprivation records and were thus excluded from my study. Likewise, I excluded individuals with no available data 12 months before the date of initiation of antipsychotic treatments since they may have initiated antipsychotic treatment elsewhere.

5.3.3 Variables and measurements

The exposure of interest was the initiation of olanzapine, risperidone or quetiapine treatment. These are often used for the treatment of severe mental illnesses including psychoses, schizophrenia and bipolar disorders. In the Neuroscience-based Nomenclature olanzapine is a dopamine and serotonin receptor antagonist, risperidone is a dopamine, serotonin and norepinephrine receptor antagonist, and quetiapine is a dopamine and serotonin receptor antagonist and norepinephrine reuptake inhibitor [88]. The outcome was body weight, measured in kilograms. The main covariates were sex (women/men) and first prescribed dose of AP (hereafter called 'first dose'). All AP reported first doses in milligrams, but I used the dose-equivalence approach of Woods [89] for defining cut-off points of low/high first dose: ≤5 mg for olanzapine, ≤75 mg for quetiapine and ≤2 mg for risperidone. Using the "2 mg of haloperidol equals 100 mg of chlorpromazine" convention as a reference, Woods (2003) explored available evidence for identifying the minimum effective dose across olanzapine, quetiapine and risperidone, defining this dose equivalence. The first dose is a good predictor of all subsequent doses prescribed during treatment; thus, over time, patients usually stay in a dose range close to the first dose they were prescribed (see Appendix 5A). I also retrieved information on age, height, social deprivation (Townsend score 1-5, from least to most deprived), smoking and drinking status, having a type-2 diabetes mellitus diagnosis, systolic blood pressure (SBP), low-density lipoprotein cholesterol (LDL-cholesterol) and high-density lipoprotein cholesterol (HDL-cholesterol), recorded within the first year before initiation of treatment. If there were multiple measures during that year, I kept the record closest to AP treatment initiation. All the variables passed a data cleaning before the statistical analysis (e.g. removing impossible weight values). This information served mostly for sample characterization; only sex, age, type-2 diabetes mellitus diagnosis and social deprivation were fully observed.

5.3.4 Statistical analysis

I used an interrupted time series approach [6] to analyse weight change over time, with one model for each of the three AP initiation cohorts by sex (six models in total, one per drug per sex). I modelled weight change over time using continuous linear splines with random intercept and slopes models (unstructured covariance, restricted

maximum likelihood), from which three slopes of weight change were estimated for 1) –4 years to baseline (pre-treatment), 2) baseline to +6 weeks (short-term), 3) +6 weeks to +4 years (long-term). Differences between slopes served to describe weight change after AP treatment initiation, both crude and adjusted for age and social deprivation. The correlation between average weight at baseline (intercept) and short-term gradient of change (short-term slope) was estimated, as it provided an estimate for whether individuals with lower weight at baseline gain more or less weight after AP treatment initiation than individuals with higher body weight. Negative correlations mean that individuals with low weight gain more weight during the short-term period and vice versa. The primary analysis was performed after stratifying each of cohorts according to low/high first dose. This was to examine whether the gradient of weight change after treatment initiation varies between low/high first doses of AP. For all these models, the Intraclass Correlation Coefficient (ICC) was reported. I assumed weight records were missing at random within strata, conditional on observed weights so that modelling the observed data over time provides unbiased estimates [48]. I also assumed weight was missing at random on dose, so the complete case analysis performed for dose provides unbiased estimates [49]. The model assessment included evaluation of residuals and visual exploration of average and individual trajectories. Although the chosen impact model (linear splines with knots at baseline and +6weeks) was informed by both the clinical criteria and evidence [5], I also performed a sensitivity analysis following the suggestions from Lopez Bernal et al. [6]. This sensitivity analysis consisted of comparing the preferred linear spline model against another feasible impact model, a restricted cubic spline model (knots again at baseline and +6 weeks), using graphical and analytical tools (see Appendix 5B). Estimates are given with 95% confidence intervals. All the statistical analyses were performed using Stata 15 for Windows [90].

5.4. Results

In total, I included 16,559 men and 22,306 women in the study. The median number ± interquartile range of weight measurements within individual trajectories over 8 years of observation were 6±7 and 8±10 (olanzapine cohorts), and 7±8 and 8±9 (quetiapine and risperidone cohorts) for men and women respectively. Characteristics of the individuals are summarised in Table 5.1. On average, at the initiation of treatment, men were younger than women prescribed olanzapine (men=47.5 years ±17.8 SD, women=54.0 years ±19.5 SD) and risperidone (men=56.6 years ±22.1 SD, women=64.5 years ±21.8 SD), but were of similar age in the quetiapine cohort (men=56.5 years ±20.7 SD, women=56.1 years ±22.1 SD). On average, men were prescribed a higher dose of olanzapine (+1mg), quetiapine (+10mg) and risperidone (+0.3mg) than women (see Table 5.1).

Table 5.1. Baseline characteristics of patients using Olanzapine, Quetiapine or Risperidone (3 retrospective cohorts), by sex.

		OLANZAPINE (N=9499)				*QUETIAPINE (N=19965)*				*RISPERIDONE (N=9401)*			
		Women		*Men*		*Women*		*Men*		*Women*		*Men*	
		N=5004 (52.7%)		*N=4495 (47.3%)*		*N=12149 (60.9%)*		*N=7816 (39.1%)*		*N=5153 (54.8%)*		*N=4248 (45.2%)*	
		n	%	n	%	n	%	n	%	n	%	n	%
Age (years)													
	18-29	608	12.2	859	19.1	1646	13.5	919	11.8	436	8.5	628	14.8
	30-39	791	15.8	860	19.1	1842	15.2	1083	13.9	481	9.3	593	14.0
	40-49	906	18.1	948	21.1	2173	17.9	1440	18.4	642	12.5	647	15.2
	50-59	777	15.5	720	16.0	1481	12.2	1031	13.2	529	10.3	477	11.2
	60-69	669	13.4	504	11.2	1027	8.5	763	9.8	514	10.0	411	9.7
	70-79	659	13.2	382	8.5	1313	10.8	1173	15.0	818	15.9	627	14.8
	80-89	485	9.7	198	4.4	2011	16.6	1190	15.2	1257	24.4	721	17.0
	90-99	109	2.2	24	0.5	656	5.4	217	2.8	476	9.2	144	3.4
	missing	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	total (n)	*5004*	*100.0*	*4495*	*100.0*	*12149*	*100.0*	*7816*	*100.0*	*5153*	*100.0*	*4248*	*100.0*
Towsend													
	Least deprived	913	18.2	674	15.0	2067	17.0	1388	17.8	966	18.7	692	16.3
	2	970	19.4	725	16.1	2101	17.3	1394	17.8	936	18.2	746	17.6
	3	1032	20.6	888	19.8	2797	23.0	1618	20.7	1166	22.6	907	21.4
	4	1117	22.3	1070	23.8	2799	23.0	1709	21.9	1119	21.7	971	22.9
	Most deprived	972	19.4	1138	25.3	2385	19.6	1707	21.8	966	18.7	932	21.9
	missing	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0

	total (n)	*5004*	*(*	*100.0*	*)*	*4495*	*(*	*100.0*	*)*	*12149*	*(*	*100.0*	*)*	*7816*	*(*	*100.0*	*)*	*5153*	*(*	*100.0*	*)*	*4248*	*(*	*100.0*	*)*
Smoking Status																									
	ex-smoking	845	(	16.9	)	855	(	19.0	)	2339	(	19.3	)	2135	(	27.3	)	1057	(	20.5	)	1133	(	26.7	)
	non-smoking	2437	(	48.7	)	1427	(	31.7	)	5634	(	46.4	)	2738	(	35.0	)	2838	(	55.1	)	1609	(	37.9	)
	smoking	1131	(	22.6	)	1527	(	34.0	)	2790	(	23.0	)	1970	(	25.2	)	786	(	15.3	)	1016	(	23.9	)
	missing	591	(	11.8	)	686	(	15.3	)	1386	(	11.4	)	973	(	12.4	)	472	(	9.2	)	490	(	11.5	)
	total (n)	*5004*	*(*	*100.0*	*)*	*4495*	*(*	*100.0*	*)*	*12149*	*(*	*100.0*	*)*	*7816*	*(*	*100.0*	*)*	*5153*	*(*	*100.0*	*)*	*4248*	*(*	*100.0*	*)*
Drinking Status																									
	non-drinking	2492	(	49.8	)	1460	(	32.5	)	5754	(	47.4	)	2788	(	35.7	)	2911	(	56.5	)	1653	(	38.9	)
	ex-drinking	869	(	17.4	)	869	(	19.3	)	2388	(	19.7	)	2200	(	28.1	)	1064	(	20.6	)	1151	(	27.1	)
	drinking	1045	(	20.9	)	1290	(	28.7	)	2496	(	20.5	)	1692	(	21.6	)	780	(	15.1	)	926	(	21.8	)
	missing	598	(	12.0	)	876	(	19.5	)	1511	(	12.4	)	1136	(	14.5	)	398	(	7.7	)	518	(	12.2	)
	total (n)	*5004*	*(*	*100.0*	*)*	*4495*	*(*	*100.0*	*)*	*12149*	*(*	*100.0*	*)*	*7816*	*(*	*100.0*	*)*	*5153*	*(*	*100.0*	*)*	*4248*	*(*	*100.0*	*)*
Diabetes Diagnostic																									
	no	4357	(	87.1	)	3986	(	88.7	)	10276	(	84.6	)	6367	(	81.5	)	4147	(	80.5	)	3446	(	81.1	)
	yes	647	(	12.9	)	509	(	11.3	)	1873	(	15.4	)	1449	(	18.5	)	1006	(	19.5	)	802	(	18.9	)
	missing	0	(	0.0	)	0	(	0.0	)	0	(	0.0	)	0	(	0.0	)	0	(	0.0	)	0	(	0.0	)
	total (n)	*5004*	*(*	*100.0*	*)*	*4495*	*(*	*100.0*	*)*	*12149*	*(*	*100.0*	*)*	*7816*	*(*	*100.0*	*)*	*5153*	*(*	*100.0*	*)*	*4248*	*(*	*100.0*	*)*
Height (m)																									
	mean (sd)	1.6	(	0.1	)	1.8	(	0.1	)	1.6	(	0.1	)	1.7	(	0.1	)	1.6	(	0.1	)	1.7	(	0.1	)
	missing	1786	(	35.7	)	1481	(	32.9	)	4665	(	38.4	)	2832	(	36.2	)	2021	(	39.2	)	1439	(	33.9	)
	total (n)	*5004*	*(*	*100.0*	*)*	*4495*	*(*	*100.0*	*)*	*12149*	*(*	*100.0*	*)*	*7816*	*(*	*100.0*	*)*	*5153*	*(*	*100.0*	*)*	*4248*	*(*	*100.0*	*)*
SBP (mmHg)																									
	mean (sd)	129.2	(	16.2	)	130.7	(	14.6	)	129.5	(	14.7	)	131.2	(	13.3	)	131.3	(	15.8	)	131.3	(	14.4	)
	missing	1659	(	33.2	)	2099	(	46.7	)	3601	(	29.6	)	2631	(	33.7	)	1280	(	24.8	)	1484	(	34.9	)
	total (n)	*5004*	*(*	*100.0*	*)*	*4495*	*(*	*100.0*	*)*	*12149*	*(*	*100.0*	*)*	*7816*	*(*	*100.0*	*)*	*5153*	*(*	*100.0*	*)*	*4248*	*(*	*100.0*	*)*
LDL-Cholesterol (mmol/L)																									
	mean (sd)	3.3	(	1.1	)	3.3	(	1.1	)	3.3	(	1.1	)	3.1	(	1.1	)	3.3	(	1.1	)	3.1	(	1.1	)
	missing	1901	(	38.0	)	1757	(	39.1	)	4783	(	39.4	)	2797	(	35.8	)	1792	(	34.8	)	1544	(	36.3	)
	total (n)	*5004*	*(*	*100.0*	*)*	*4495*	*(*	*100.0*	*)*	*12149*	*(*	*100.0*	*)*	*7816*	*(*	*100.0*	*)*	*5153*	*(*	*100.0*	*)*	*4248*	*(*	*100.0*	*)*

HDL-Cholesterol (mmol/L)																							
	mean (sd)	1.6	(	0.5	)	1.3	(	0.4	)	1.5	(	0.5	)	1.3	(	0.5	)	1.5	(	0.6	)	1.3	(	0.5	)
	missing	1459	(	29.2	)	1385	(	30.8	)	3596	(	29.6	)	1995	(	25.5	)	1307	(	25.4	)	1089	(	25.6	)
	total (n)	*5004*	*(*	*100.0*	*)*	*4495*	*(*	*100.0*	*)*	*12149*	*(*	*100.0*	*)*	*7816*	*(*	*100.0*	*)*	*5153*	*(*	*100.0*	*)*	*4248*	*(*	*100.0*	*)*
First Dose (mg)																							
	mean (sd)	6.2	(	4.6	)	7.5	(	4.9	)	84.9	(	111.9	)	95.3	(	129.0	)	1.3	(	1.1	)	1.6	(	1.4	)
	missing	1369	(	27.4	)	1138	(	25.3	)	4865	(	40.0	)	3164	(	40.5	)	1735	(	33.7	)	1396	(	32.9	)
	total (n)	*5004*	*(*	*100.0*	*)*	*4495*	*(*	*100.0*	*)*	*12149*	*(*	*100.0*	*)*	*7816*	*(*	*100.0*	*)*	*5153*	*(*	*100.0*	*)*	*4248*	*(*	*100.0*	*)*
Body Weight (kg) *																							
	mean (sd)	69.7	(	17.2	)	81.1	(	17.5	)	73.3	(	18.3	)	82.5	(	18.0	)	70.1	(	18.2	)	81.6	(	18.7	)
	missing	2156	(	43.1	)	1858	(	41.3	)	5540	(	45.6	)	3438	(	44.0	)	2389	(	46.4	)	1771	(	41.7	)
	total (n)	*5004*	*(*	*100.0*	*)*	*4495*	*(*	*100.0*	*)*	*12149*	*(*	*100.0*	*)*	*7816*	*(*	*100.0*	*)*	*5153*	*(*	*100.0*	*)*	*4248*	*(*	*100.0*	*)*

(*) Body weight has been calculated as the average of the wieght records available up to 12 months before treatment initiation.

In the short (<6 weeks) and long term (≥6 weeks to ≤4 years), individuals treated with any of the three AP drugs gained weight, especially those patients prescribed olanzapine. Pre-treatment weight change was negligible for quetiapine (women and men) and risperidone (men only) cohorts, and slightly negative for the rest of cohorts. In the short-term after olanzapine initiation, men's weight increased by 0.569 kg/week (3.4 kg over the first six weeks) and women's weight increased by 0.382 kg/week (2.3 kg over the first six weeks) (Table 5.2 and Appendix 5C). Individuals initiated on quetiapine and risperidone also gained weight shortly after initiation of treatment, but to a lesser extent (Table 5.2 and Appendix 5C, and Figure 5.1). Individuals continued to gain weight after six weeks, but at a slower rate than the first six weeks. For example, for women initiated on olanzapine, long-term weight gain was estimated to be 0.014 kg/week (0.7 kg per year) (Table 5.2, Appendix 5C, and Figure 5.1). Women who were initiated on olanzapine were in general slightly lighter (69.7 kg) than women initiated on risperidone (73.3 kg) and quetiapine (70.1 kg), but there was not much difference for the men (weight at baseline, see Table 5.1 and Appendix 5C). Women who had a lower weight before initiation of olanzapine gained more weight in the short term than women who had a higher weight (correlation between intercept and slope=-0.068, 95%CI: -0.121 to -0.014); a similar effect was observed for men (correlation between intercept and slope=-0.050, 95%CI: -0.113 to +0.014) (Appendix 5C).

Table 5.2. Expected weight gain for an average patient prescribed a particular antipsychotic, stratified by dose and sex.

Drug	Sex	N*	Dose**	Weight gained during short-time (0-6 weeks) in Kilograms	95% CI	Weight gained during long-time (6weeks - 4 years) in Kilograms	95% CI	Total weight gained
OLANZAPINE (N=9,499)		5004	Overall	2.3	(1.9 to 2.7)	2.8	(2.2 to 3.5)	5.1
	Women	*2535*	Low	1.9	(1.4 to 2.4)	2.5	(1.6 to 3.3)	4.4
		1100	High	3.2	(2.4 to 4.0)	2.9	(1.6 to 4.2)	6.1
		4495	Overall	3.4	(3.0 to 3.8)	1.7	(0.9 to 2.4)	5.1
	Men	*1887*	Low	2.6	(2.0 to 3.2)	1.9	(0.8 to 3.0)	4.5
		1470	High	4.5	(3.6 to 5.3)	1.4	(0.2 to 2.7)	5.9
QUETIAPINE (N=19,965)		12149	Overall	1.2	(1.0 to 1.5)	1.1	(0.6 to 1.6)	2.3
	Women	*5372*	Low	0.7	(0.3 to 1.0)	0.9	(0.1 to 1.6)	1.6
		1912	High	2.3	(1.6 to 2.9)	1.6	(0.4 to 2.7)	3.9
		7816	Overall	0.8	(0.4 to 1.1)	0.7	(0.1 to 1.3)	1.5
	Men	*3326*	Low	0.5	(0.0 to 0.9)	-0.7	(-1.8 to 0.3)	-0.3
		1326	High	1.6	(0.9 to 2.4)	1.0	(-0.3 to 2.2)	2.6
RISPERIDONE (N=9,401)	Women	5153	Overall	0.9	(0.5 to 1.3)	0.7	(-0.1 to 1.5)	1.6
		3102	Low	1.0	(0.5 to 1.4)	0.1	(-0.9 to 1.1)	1.1

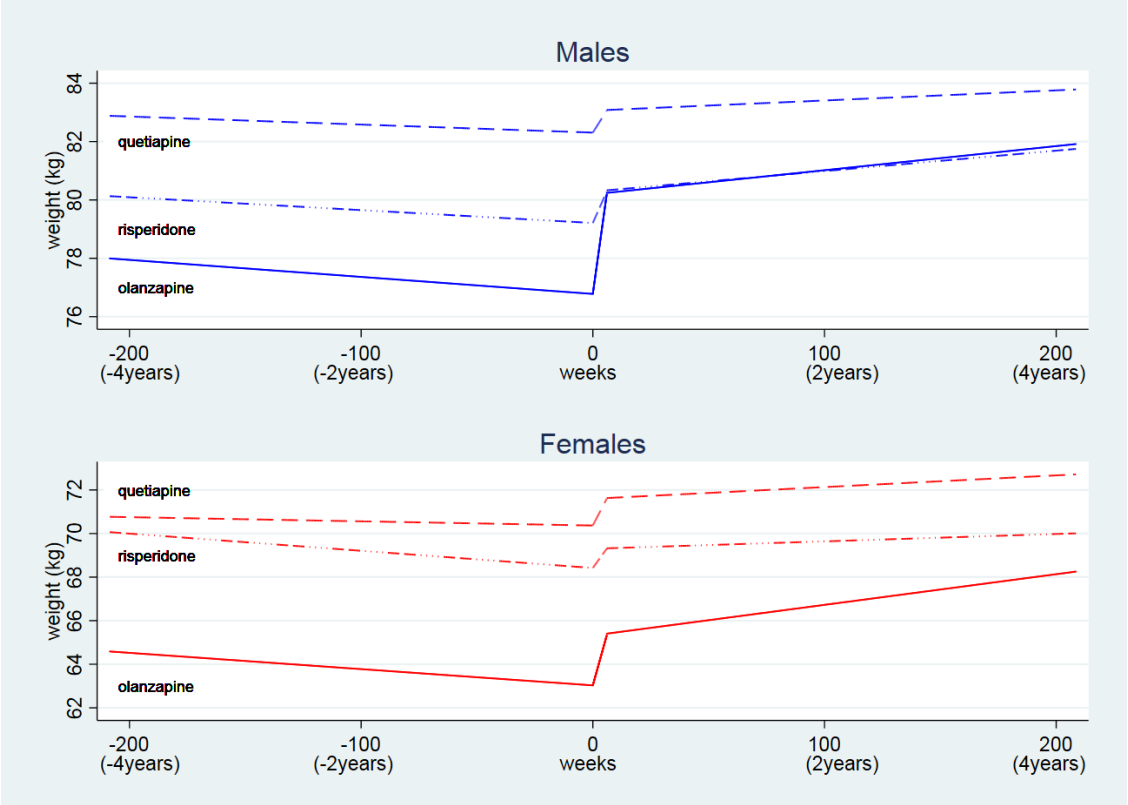
	316	High	1.1	(-0.7	to	2.9)	3.5	(1.0	to	5.9)	4.6
	4248	Overall	1.1	(0.6	to	1.5)	1.4	(0.4	to	2.4)	2.5
Men	*2411*	Low	1.0	(0.4	to	1.7)	1.1	(-0.3	to	2.6)	2.2
	441	High	1.9	(0.5	to	3.3)	1.4	(-0.7	to	3.5)	3.3

(*) Overall estimates come from Table S2 (N=38,865) and low/high dose estimates come from Table S3 (N=25,198). N from Table S2 < N from Table S3 due to missing data on dose.

(**) Cut off point for low/high dose was: ≤5 mg for Olanzapine, ≤75 mg for Quetiapine and ≤2 mg for Risperidone.

The weight gain in individuals who were initiated on a high dose of AP was greater than those initiated on a low dose. When olanzapine was initiated at high dose (>5mg), women gained +0.534 kg/week (+3.2kg over 6 weeks) and men +0.743 kg/week (+4.5kg over 6 weeks) compared with a low-dose gain of +0.314 kg/week (+1.9kg over 6 weeks) for women and +0.425 kg/week (+2.6kg over 6 weeks) for men (Table 5.2 and Appendix 5D). The short-term effect of initiation of quetiapine was also more substantial for those given high doses (>75mg) (women +2.3kg and men +1.6kg, both over 6 weeks) than given low doses (women +0.7kg and men +0.5kg, both over 6 weeks). However, there was a relatively small difference for those initiated on risperidone low doses (≤2mg) (+1.0kg over 6 weeks for both women and men) and high doses (women +1.1kg and men 1.9kg, both over 6 weeks). In the short-term, those given low doses of olanzapine tended to gain more weight as their weight at baseline was lower (women: correlation between intercept and slope=-0.155, 95%CI: -0.230 to -0.078; men: correlation between intercept and slope=-0.135, 95%CI: -0.235 to -0.033).

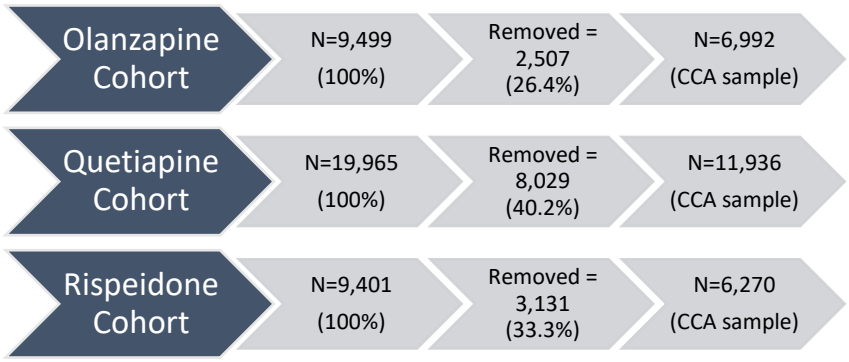
Figure 5.1. Changes in body weight over time before and after treatment initiation by drugs and sex.



Cumulative weight gain in the long-term was particularly high in patients prescribed olanzapine but, for any drug, on average, people did not lose the extra weight they gained during the short-term (Table 5.2 and Appendix 5D). For example, after four years from the first olanzapine prescription, a woman typically gained 2.3kg (short-term, 95%CI: 1.9kg-2.7kg) + 2.8kg (long-term, 95%CI: 2.2kg-3.5kg) = 5.1kg (total) and a typical man gained 3.4kg (short-term, 95%CI: 3.0kg-3.8kg) + 1.7kg (long-term, 95%CI: 0.9kg-2.4kg) = 5.1kg (total) of AP induced extra-weight. The prescribed dose of olanzapine was also critical, particularly for women in the long-term. For example, given a low dose (<5mg), women gained 1.9kg + 2.5kg = 4.4kg after four years; given a high dose (>5mg), women gained 3.2kg + 2.9kg = 6.1kg. A similar impact of higher doses was observed for quetiapine and risperidone (Table 5.2 and Appendix 5D).

MEM approach worked well in general (i.e. estimation process was not slow, there were no convergence issues), but I lost a fraction of data when included a covariate with missing values in the analysis (i.e. dose). Comparing sample sizes from Appendices 5C and 5D, Figure 5.2 shows data lost involved in the CCA performed for generating Table 5.2.

Figure 5.2. Participants removed from the olanzapine, quetiapine and risperidone cohorts due to missing data on dose, for performing complete case analysis (CCA).



5.5. Discussion

This retrospective cohort study reports data from patients seen in primary care, before and after AP treatment initiation. Pre-treatment weight change was insignificant or slightly negative for all cohorts during 4 years before baseline. Individuals starting treatment with any AP gained weight on average, especially those patients prescribed olanzapine. Weight gain was much more rapid in the short-term than in the long-term. People who were initiated on high dose AP experienced much greater absolute weight gain than those initiated on low dose AP. Cumulative weight gain during the long-term was particularly high in individuals treated with olanzapine but, for all AP's, people typically never lost the extra weight they gained during the first 6 weeks of AP treatment. MEM worked well for the ITS analyses, but I lost between 26% and 40% of cases in the ITS analysis stratified by dose due to missing values on dose (i.e. complete case analysis).

Previous studies have suggested that olanzapine is associated with a large short-term weight gain, whereas risperidone and quetiapine have a moderate effect on weight [5]. In the long-term, contrary to one previous finding [91], I found that weight gain did not stabilize during 4 years of follow up. However, my finding of the long-term effect of weight gain is consistent with previous studies by Bushe et al. [92] and Osborn et al. [82], but I can quantify the effect more accurately. Previous research has suggested women's weight is more affected by AP exposure [83]; however, I found that only olanzapine (in the long-term) and quetiapine (in the long- and short-term) induced more weight gain in women. Since my study population is a mixture of naïve and recurrent antipsychotic consumers, short and long-term weight gain in olanzapine naïve individuals and long-term weight gain in risperidone naïve individuals can be higher than the weight gain reported here [5]. Risperidone seemed to be associated with higher weight gain in men than women both in the short and long-term, and men prescribed olanzapine gained more weight in the short-term. Regarding the dosage, one recent study reanalysed results of 14 clinical trials to explore variations in weight gain across doses of olanzapine and risperidone [93]. Their conclusions about olanzapine are consistent with my results; that the excess risk of at least 7% weight gain is 16.1% for low doses (0-10 g chlorpromazine equivalent dose) and 46.8% for high doses (0-20 g chlorpromazine equivalent dose). They could not be conclusive about the effects of

risperidone as they showed only a trend in weight gain; however, this trend is in line with my findings. Some advantages from my original study are: 1) I added similar information about quetiapine, I 2) observed longer periods of weight change (4 years) and 3) analysed information at individual-level from cohorts with more than 38,000 patients in total.

MEM presented advantages over the standard ITS approach, but some disadvantages due to a missing covariate were faced. Conversely to the standard 'aggregate-level' segmented regression (Section 2.1.2), no averaging-step was needed for fitting MEM. This implies not only a more efficient way to analyse the data (i.e. I did not lose information by averaging the outcome in artificial time-windows) but also a preferable approach by modelling individual-level data instead of time-point averages. For fully observed covariates, MEM estimator (REML, Section 3.3) is unbiased under MAR assumptions [70]. Moreover, MEM model variances directly (also see Section 3.3) allowing to evaluate the association between weight at baseline (random intercept) and short-term weight gain (beta 2) in a single ITS model. Nevertheless, the solely MEM approach to handle missing covariates is the complete case analysis, reducing the precision of estimates (i.e. bigger standard errors). Assuming weight is MAR on dose -which is not a bold assumption -, the risk of bias is relatively low. However, for a more complex scenario of two or more missing covariates, the complete case analysis can be inviable in practice. That is a crucial reason for exploring alternatives that assist in the handling of missing covariates (i.e. Chapter 6).

Among its strengths, this study presents evidence from a large sample (N>38,000) of people prescribed antipsychotic medications, taken from a population which is broadly representative of the UK [27]. Patients prescribed antipsychotics are often treated for long periods, and so quantifying the risk of long-term side effects is particularly important. Clinical trials often fail to do this because of their short durations and much smaller sample size, so my study provides a necessary long-term perspective. I applied an analysis approach that has not been used previously in assessing AP induced weight gain. The approach utilises all individual weight records at their time of measurement, therefore avoiding the loss of information seen in previous studies which categorise outcomes or use period means or incidence rates as summary measures [5,82]. My longitudinal model-based approach also accounts for missing weight records –

assuming weight recording is missing at random within strata, conditional on observed weight measurements [94] – while incorporating informative pre-baseline weight data. From this method, I expect unbiased estimates if data were missing at random [50], a property that is not ensured by analyses applied elsewhere [82,92]. Following standard recommendations [6], I guaranteed good statistical power by having equal periods of observation before and after baseline, and large sample size. Additional analyses showed that my proposed linear spline models were very similar to the restricted cubic splines models (Appendix 5B), and for primary analysis, I used the former as its interpretation is more straightforward.

This study does have several potential limitations. Information on possible time-varying confounders (for example, symptoms level or illness severity) was not included; however, it is reasonable to assume limited variation from patients' baseline values for unmeasured confounders. Treatment initiation has been defined using the first prescription date in general practice; but, for some individuals, the first prescription date might occur while the individual is under the care of secondary care mental services (these data are not recorded in primary care). However, it is most likely these patients have a first prescription date very close to the one in primary care; thus, no major impact on estimates is expected. I did not control for drugs prescribed to reduce antipsychotic-induced weight gain, or for multiple prescriptions of other drugs that could potentially affect weight as well. However, I know that drugs prescribed for ameliorating weight gain would only reduce the estimate of the real weight gain of the target population; thus, I am not overestimating the weight gain effect. I did not assess weight gain associated with other antipsychotic medications as there were not enough data on them, but the three drugs included in this study are the most commonly prescribed antipsychotic medications in the UK [95] and have previously been associated with weight gain [82]. The weight gain trajectories I described are averages; thus, they should be interpreted as typical patient trajectories. In practice, individual patients' weight gain will vary from these average trajectories. However, the first weeks of treatment are critical for everyone. Finally, I did not control the number of prescriptions beyond the second prescription (treatment duration), meaning that studied patients can include those treated for long periods, those treated sporadically, just for a short period, or those who did not adhere to treatment regularly. This lack of control may reduce my long-term estimates of weight gain, but – given the evidence

about dosage – I anticipate that patients exposed to AP on a regular basis and for long periods will have larger estimates of long-term weight gain.

In conclusion, over 4 years, olanzapine treatment was associated with the highest increase in weight, with around 6 kg for those on a high dose and 4.5 kg for those on a low dose. The weight gain was less dramatic for individuals treated with quetiapine and risperidone. In general, in the long-term (i.e. up to 4 years), individuals did not lose the weight gained during the first 6 weeks of treatment. Doctors and patients may want to take the issue of a substantial weight gain into consideration when making decisions on the initiation of antipsychotic treatments, and doctors should prescribe the lowest effective dose to balance mental health benefits, weight gain and other adverse effects. MEM are useful to model ITS with missing outcomes, but a new alternative should be explored for handling missing covariates.

5.6. Summary [12]

Second-generation antipsychotics (AP) are a known cause of weight gain. However, most previous studies are based on small sample sizes, and most do not distinguish between short and long-term weight gain associated with antipsychotic treatment. On the other hand, there is limited information about how doses of antipsychotic treatment are associated with weight gain. Interrupted time series (ITS) design provides a flexible framework for analysing the antipsychotic-induced weight gain in observational data. Despite this advantage, observational data bring missing data issues that standard ITS tools (e.g. aggregate-level segmented regression) are not designed to address. However, modelling ITS with mixed effect models (MEM) is a flexible approach that could efficiently overcome the problem.

The objectives were: 1) Clinically: to investigate the change in body weight of patients initiated with high or low doses of the three most commonly prescribed second-generation antipsychotics, and 2) Methodologically: to apply MEM for handling missing data in longitudinal weight records. In total, I included 16,559 men and 22,306 women in the study.

Olanzapine was associated with the highest increase in weight, in the short and long term, and higher doses were associated with higher weight gain. For example, when women were prescribed olanzapine at high dose, they gained 3.2kg on average during the first 6 weeks and a total of 6.1kg at the end of the 4 years observation period. A low dose was associated with 1.9kg of weight gain in the first 6 weeks and a total of 4.4kg in the long-term period (4 years). For any second-generation antipsychotic, the weight gain remained in the long-term; thus, on average, patients never lost the weight they gained during the first 6 weeks of treatment.

MEM was a functional analysis tool for the ITS design, but it was not efficient for handling missing covariates. Fully observed covariates helped to inform the implicit imputation of the outcome at individual-level made by the Restricted Maximum

[12] This study has been published as an original article. Please, see Appendix 10.5 for details.

Likelihood (REML) estimator of MEM. Under MAR assumptions, MEM provide unbiased estimates. MEM can produce additional information that standard ITS methods are not able to do. For example, MEM helped to investigate the association between weight at baseline (random intercept) and short-term weight gain (beta 2) within a single ITS design. This would not be possible directly from standard ITS regression models (e.g. aggregate-level segmented regression). However, MEM can only handle missing covariates by listwise deletion, which results in a reduction of the precision and potential problems with bias.

In the next Chapter 6, I explore other choices for handling missing covariates, such as multilevel multiple imputation, and formally compare MEM and the aggregate-level segmented regression for handling missing outcomes, via simulations.

6

Evaluating methods for missing data handling in Interrupted Time Series analysis via simulation studies

6.1. Introduction

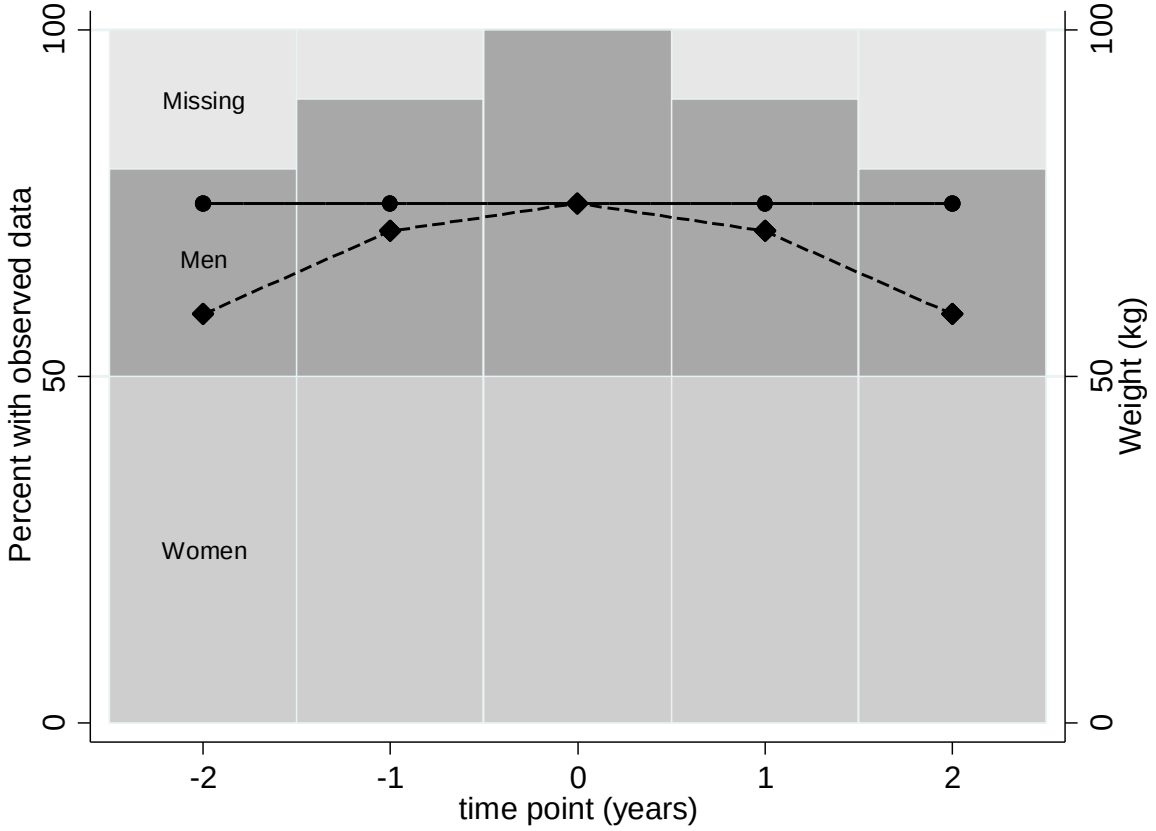
Interrupted time series (ITS) is a widely used quasi-experimental approach that evaluates the potential impact of an intervention over time, using longitudinal observational data [6]. It has frequently been used to evaluate intervention effects in longitudinal population studies; for example, to evaluate the impact of policies and social interventions on clusters, such as districts, cities and countries [96,97]. While ITS comes from social science literature, it is becoming more widespread in health research [7,61]. ITS may be used to address causal questions that are not feasible for a randomised controlled trial, but with stronger assumptions [1]. The methodology for the analysis of ITS studies is well developed [3,6,8], and typically uses segmented regression (SR) analysis [7,61]. Given a time point, for example, the initiation of treatment, we may observe a change in the values of a variable before and after that time point, and then compare the trajectories of change at the intervention. The pre-treatment trajectory is regarded as the control 'period' and the post-treatment

trajectory as the intervention 'period', so that each individual acts as their own control. The difference between mean trajectories at the intervention time is then used to estimate the effect of the intervention [6].

In SR analysis, when individual-level data are available, a typical approach is to average the data at each of the predefined time points/units (e.g. months or years) and then model the time series over these time points [4,7]. In other words, all outcome variable measurements available from individuals are averaged at each time point, and then these averages are used as population-level data for performing the SR analysis. This approach is reasonable if the same people provide data at each time point, but in observational data, this is rarely the case. For example, in clinical practice, younger women are more likely than younger men to have weight recorded when they consult their family physician (general practitioner) [32]. In other words, the distribution of missing data in weight depends on the individual's sex, so weight is missing at random (MAR) given sex. The same will apply to other partially observed outcomes that are MAR. With such data, the average points will be biased – and so will the intercept of the trajectories estimated by SR models – because they will include more measurements from women than men, and women will typically weigh less than men. Moreover, if the proportion of women and men with observed weight varies at each time point, the slope of the trajectories can also be biased.

Figure 6.1 presents a scenario where weight is constant over time for all individuals (half men, half women; men weigh 85kg, and women weigh 55kg, resulting in an overall average of 70kg). In this scenario, all individuals have a weight measurement at treatment initiation (t=0), but at different time points before and after treatment initiation the relative proportion of women and men with a weight record varies due to missing data. The average observed weight at each time point becomes biased, providing a false impression of weight change over time. Thus, the 'aggregate-level' SR analysis performed with averages calculated at predefined time points can produce biased estimates due to missing data.

Figure 6.1. Real weight trajectory (circle/solid line) and observed weight trajectory (diamond/dash line) following the averaging-step with different proportion of women and men observed at each time point in a recreated scenario.



An alternative approach to the 'aggregate-level' SR analysis is to use mixed models, which are based on individual-level data, avoiding the averaging-step described above. Formally, these mixed models are also segmented models, but they include random intercept and slopes (random effects) that cannot be included by the 'aggregate-level' SR models due to the averaging-step. Mixed models estimate identical linear trajectories to 'aggregate-level' SR models under perfect balance (when all individuals are included at each time point). However, in contrast to 'aggregate-level' SR models, the mixed model approach can provide unbiased estimates when data in the outcome variable are MAR [70]. Following the same example as before, a mixed model directly uses weight measurements taken at different time points from the same individual and models the population trajectory based on all individual trajectories, taking account of the longitudinal correlation. Thus, no initial averaging-step at each time point is needed. If individuals have missing weight records over time, the mixed model approach implicitly imputes those missing values (see Section 3.3 for details), meaning

that observations from all individuals – even those with just one record over time – contribute to the analysis.

Despite these advantages, mixed models cannot automatically handle missing data in the covariates, and individuals with covariates missing are by default omitted from regression analyses in all standard software packages. One way to address this issue is to use multilevel multiple imputation (MMI) for missing covariate data in conjunction with mixed models. MMI generates multiple datasets with missing covariate values replaced by imputed values (drawn from the conditional predictive distribution of the missing data given the observed data). Then, MMI fits the substantive model of interest in each imputed dataset and, in the final step, combines the model estimates into an overall estimate, taking into account variation within and between the imputed datasets [98] (see sections 3.2.3 to 3.2.5 for details). In this setting, the substantive model fitted at the second step is a mixed model.

In this study, I demonstrate how standard ITS analysis, based on average estimates at each predefined time point, gives biased results when data are MAR. Subsequently, I illustrate how the use of mixed models, with or without MMI of individual data, avoids this bias.

6.2. Objectives

The study objectives are 1) to examine the potential problems arising from the 'aggregate-level' SR analysis when outcome data are missing, evaluating mixed models as an alternative approach; 2) to compare the performance of mixed models with and without MMI for handling missing data on covariates.

The rest of Chapter 6 is structured as follows. In Section 6.3, I present a motivating example of ITS to estimate the effect of initiating antipsychotic drugs (olanzapine) on weight gain, showing that the standard approach of aggregating the data and then using SR gives clinically different results to using mixed models (with and without MMI). Section 6.4 presents a simulation study, which demonstrates that this difference

is because the standard 'aggregate-level' approach is biased when data are MAR. I conclude in Section 6.4 by discussing the practical and methodological implications of my findings.

6.3. Motivating example: ITS for the effect of antipsychotics on weight.

In this motivating example, as well as in the later simulation study, I focus on assessing estimators for the regression coefficients of pre- and post-treatment weight trajectories.

6.3.1 Data and first analysis

I used data from The Health Improvement Network (THIN) database, which includes electronic health records from ~12 million individuals registered with 711 UK general practices [99]. In the UK, more than 95% of people are registered with a general practice (GP), and THIN is roughly representative of the general population [27]. THIN data include demographics (e.g. sex, age, social deprivation) and clinical records (e.g. drug treatments, diagnoses, health outcomes). In this study, I only included data from general practices that met quality criteria for computer usage [35] and whose reported mortality rate is consistent with national statistics [36].

The substantive model is similar to the one I fitted in Chapter 5. I performed an ITS analysis to investigate the long-term effects of the initiation of antipsychotic drug treatment on people's body weight. It is known that specific antipsychotic treatments are likely to increase body weight substantially over a relatively short period [5], but there is less information on potential long-term effects [100]. In this study, the exposure of interest was the initiation of olanzapine (a second-generation antipsychotic), and the outcome was body weight (in kilograms). I modelled the development of weight over time using linear splines with two knots. In other words, my model estimated how weight changed in three time periods: 1) *pre-treatment:* from 4 years before treatment initiation up to treatment initiation; 2) *short-term:* from treatment initiation to 6 weeks (short-term), and 3) *long term:* from 6 weeks to 4 years post-treatment. I adjusted for sex, age (in years) and smoking (smoker vs non-smoker), all at the initiation of treatment. I included individuals who were aged between 18 and 99 years, with data available between 1st January 2005 and 31st December 2015, and who initiated their first

olanzapine treatment within this period. All had a diagnosed psychotic disorder before treatment initiation and at least one further prescription of olanzapine within three months following the first prescription. I included this criterion as there may some individuals who received just one prescription, but never used the medication. However, if they had at least two prescriptions, it seems more likely that they initiated treatment. I excluded individuals who initiated other antipsychotics than olanzapine, as well as those with no available data for 12 months before the treatment initiation.

In addition to the inclusion and exclusion criteria given above, I restricted the data to those with complete data on sex, age and smoking at treatment initiation. As this is observational data, weight measurements did not follow any fixed schedule. For example, if we look for a weight measurement every two weeks for every individual, we will find that >90% of weight measurements are missing. In other words, the weight has been irregularly recorded over the observation period (416 weeks), as it is expected for most electronic health records (see Section 2.2).

Centring each patient's follow-up time (in weeks) at their treatment initiation, I fitted the following mixed model to these data:

[Equation 6.1]

$$weight_{ij} = \beta_0 + u_{0j} + (\beta_1 + u_{1j})time_{ij} \times 1[time_{ij} < 0] + (\beta_2 + u_{2j})time_{ij} \times [0 \leq time_{ij} \leq 6] + (\beta_3 + u_{3j})time_{ij} \times [time_{ij} > 6] + \epsilon_{ij},$$

$$\begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma\right); \qquad \epsilon_{ij} \sim N(0, \sigma^2),$$

where *i* denotes the follow-up time and *j* denotes the patient, and 1[ ] is an indicator for the event in square brackets. I then fitted the same model adjusting for sex, age and smoking at treatment initiation (as fixed effects). These mixed intercept and slope models were fitted by Restricted Maximum Likelihood, and hereafter I call them just mixed-effects models (MEM).

I also fitted an 'aggregate-level' SR model by averaging available weight records at each time point (across-individuals average), and then fitting the standard regression version of [Equation 6.1] – i.e. omitting the person-specific random effects-. Because this model is fitted to data aggregated over individuals, no adjustment for sex, age or smoking was possible.

Finally, I fitted a similar model, but now weighting by the inverse of the number of bodyweight values observed at each time point. I called this model the 'aggregate-level' SR-W1, which may help to improve standard errors by including a more accurate sample size information at each time-point.

These models were used to examine the issues arising from the 'aggregate-level' SR analysis when outcome (weight measurements) data are missing, which was part of the first study objective.

6.3.2 Imposed missing data and second analysis

For my second objective, I wanted to explore the issues arising from covariate data missing at treatment initiation. Therefore, I intentionally set smoking records MAR on sex, and increased the amount of missing data on weight MAR on sex, to explore later the potential differences between estimates from complete case analysis (removing cases with smoking missing) and MMI (preserving those cases and imputing smoking). This controlled missing data generation scenario was used to evaluate all analysis methods: 'aggregate-level' SR, 'aggregate-level' SR-W1, MEM, and MMI followed by a mixed-effects model (MI-JOMO with MEM).

In detail, I set weight values MAR dependent on sex and time from treatment initiation, so that a fraction of observed data was similar to that shown in Figure 6.1. Besides, I set smoking MAR on sex, randomly removing 80% of records from men and 20% from women. Both missing mechanisms are described in detail in Appendix 6A.

In the subsequent analyses, I first fitted the same MEM [Equation 1] to the incomplete data, adjusting for covariates (complete case analysis). Then, I used a substantive-model-compatible joint-modelling multilevel multiple imputation (MI-JOMO) [56] to impute the missing smoking values and fitted the same substantive model (MEM adjusted) to each imputed data set and combined the results using Rubin's rules. I generated 20 imputed datasets with MI-JOMO, and I used a burn-in of 1000 iterations and then a further 1000 iterations between each imputation. I name this model MI-JOMO with MEM.

Lastly, I fitted the 'aggregate-level' SR and 'aggregate-level' SR-W1 models. Full details and codes for all models are given in Appendix 6A.

6.3.3 Results

Overall, there were 6,522 individuals with at least one weight measurement and complete age, sex or smoking status data. Of these 2,954 (45.3%) were men and 3,568 (54.7%) were women. On average, there were 4.8 (sd 5.5) weight records per person over the observation period. Individuals were aged 50.2 (sd 18.9) years on average, and 2,658 (40.8%) reported being current smokers.

There were substantial differences between estimates derived from MEM and SR (Table 6.1, section 'THIN: Data Fully Observed'). For example, the short-term weight change (beta2) was 0.462kg/week from MEM (adjusted) and 0.816kg/week and 0.807kg/week from SR and SR-W1, respectively. Likewise, pre-treatment and long-term periods, weight change rates from SR and SR-W1 were more than double the MEM estimates. In general, all estimates of weight change from SR analyses were higher in magnitude than those from MEM, which also implies a more substantial ITS treatment effect.

After further removal of weight records, 6,181 individuals remained with one or more weight records. There were 4.3 (sd 5.3) average weight records per person over the

observation period. The average age was 50.6 (sd 19) years, and 2,613 (42.3%) were men. After the removal of smoking records at baseline, there were only 3,379 individuals with a record of their smoking status and 1,188 (35.2%) of them were current smokers.

In general, estimates from MEM with and without MI-JOMO were similar for pre-treatment and long-term effects, and both close to those estimated under MEM with full data. However, the MI-JOMO with MEM for short-term were closest to those estimated under MEM with full data (Table 6.1). ITS estimates from SR differed substantially from the estimates from MEM with and without MI-JOMO (Table 6.1, Figure 6.2), with SRs reporting a weight pre-treatment (beta1) and long-term trajectories (beta3) closer to zero. For SR-W1, the long-term treatment effect was similar to the MEM estimates, while the estimates of the short-term effects (beta2) were much higher than MEM estimates. For both the SR and SR-W1 models, pre-treatment and long-term effects were also different when fitted to data with and without imposed missing values.

The immediate treatment effect, estimated as the difference between the negative and positive trajectories before and after olanzapine initiation, was highest for the SR approach (Table 6.1 and Figure 6.2). For example, the SR-W1 method suggested a cumulative short-term weight gain of 4.72kg, a long-term of 2.13kg, and a total of 6.85kg. In contrast, the estimates based on MEM with MI-JOMO (short-term=2.47kg, long-term=2.46kg, total=4.93kg) and without MI-JOMO (short-term=2.75kg, long-term=2.70kg, total=5.45kg) were less for the short-term and the total accumulated (see 95% CI in Appendix 6B).

In summary, an individual data model such as MEM [Equation 1] produced notably different results from SR models with 'aggregate-level' data. Further, if covariate values are MAR, use of MI-JOMO can recover information by bringing individuals with these missing covariates back into the analysis, avoiding potential bias and increasing precision. By contrast, the often-used 'aggregate-level' SR analysis cannot adjust for covariates and appears to be biased when weight data are MAR (depending on time and covariates). This may often be the case when analysing health care records.

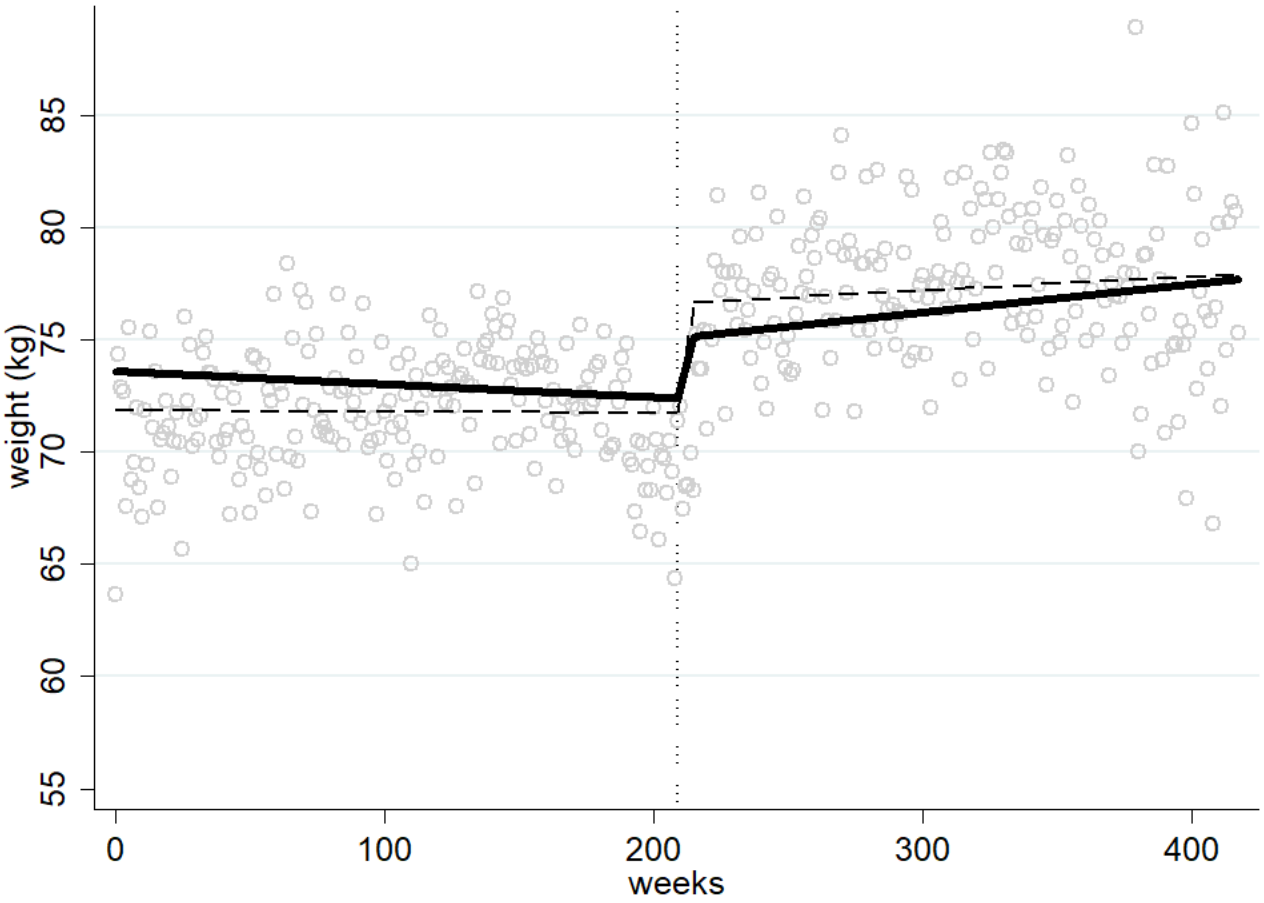
Table 6.1. Estimated weight change over time before and after olanzapine treatment initiation from the example in section 2, which also describes the various analysis methods.

	THIN: Data Fully Observed												THIN: Data with weight records MAR on sex and time, and smoking records MAR on sex and time.											
Estimate	MEM (Unadjusted) (N=6,522 - weight records=31,153)			MEM (Adjusted) (N=6,522 - weight records=31,153)			SR (N=418)			SR-W1 (N=418)			MEM (Adjusted) (N=3,379 - weight records=16,709)			MI-JOMO with MEM (Adjusted) (N=6,181 - weight records=26,880)			SR (N=418)			SR-W1 (N=418)		
	Beta	SE	p	Beta	SE	p	Beta	SE	p	Beta	SE	p	Beta	SE	p	Beta	SE	p	Beta	SE	p	Beta	SE	p
t1 [β1]	-0.0051	0.0012	<0.001	-0.0058	0.0012	<0.001	-0.0158	0.0030	<0.001	-0.0193	0.0030	<0.001	-0.0075	0.0017	<0.001	-0.0082	0.0013	<0.001	0.0008	0.0035	0.826	-0.0045	0.0032	0.158
t2 [β2]	0.4642	0.0289	<0.001	0.4617	0.0289	<0.001	0.8160	0.0875	<0.001	0.8071	0.0871	<0.001	0.4116	0.0412	<0.001	0.4576	0.0303	<0.001	0.7994	0.0998	<0.001	0.7863	0.0943	<0.001
t3 [β3]	0.0127	0.0015	<0.001	0.0125	0.0015	<0.001	0.0246	0.0032	<0.001	0.0279	0.0036	<0.001	0.0121	0.0021	<0.001	0.0133	0.0017	<0.001	0.0043	0.0036	0.234	0.0105	0.0044	0.017
sex [β4]	*	*	*	-12.9299	0.4173	<0.001	*	*	*	*	*	*	-13.7330	0.8075	<0.001	-13.1364	0.4471	<0.001	*	*	*	*	*	*
age [β5]	*	*	*	0.9498	0.0589	<0.001	*	*	*	*	*	*	0.8276	0.0832	<0.001	0.9126	0.0614	<0.001	*	*	*	*	*	*
age2 [β6]	*	*	*	-0.0096	0.0005	<0.001	*	*	*	*	*	*	-0.0085	0.0007	<0.001	-0.0093	0.0006	<0.001	*	*	*	*	*	*
smoking [β7]	*	*	*	-3.8524	0.4382	<0.001	*	*	*	*	*	*	-3.6762	0.6304	<0.001	-3.7532	0.5992	<0.001	*	*	*	*	*	*
intercept [β0]	73.5754	0.2630	<0.001	62.2883	1.5107	<0.001	75.0677	0.3688	<0.001	75.4822	0.3542	<0.001	66.4391	2.2499	<0.001	63.8191	1.6031	<0.001	71.7488	0.4207	<0.001	72.3779	0.3749	<0.001

Beta estimates are in kilograms. THIN=The Health Improvement Network Database, MAR=missing at random, SR='aggregate-level' segmented regression, SR-W1='aggregate-level' segmented regression weighted with the inverse of the number of observed weight records at each time point, MEM=random intercept and slope model with restricted maximum likelihood and unstructured covariance matrix, MI-JOMO= joint modelling multiple imputation using a similar MEM model, t1 = time before treatment initiation (209 weeks), t2 = short-term after treatment initiation (6 weeks), t3 = long-term after treatment initiation (203 weeks).

On the right side of this table, sample size between MEM and MI-JOMO differs due to missing data on smoking at baseline.

Figure 6.2. Estimated weight trajectories before and after initiation of olanzapine treatment, from the data in Section 2 (motivating example). Circles are weight averages at each time point, dashed line – SR model fitted to these averages; solid line – model [Equation 1] fitted to the raw data by MEM.



6.4. Simulation study

I now report the results of a simulation study, based on the motivating clinical example and designed to evaluate the performance of SR and MEM (with and without MMI) under controlled conditions. I am adding to this evaluation another method called Prais-Winsten regression, which is similar to SR but is recommended by ITS guidelines to account for autocorrelation at the aggregate level [6]. In particular, I wish to determine whether the differences between the various analysis methods are due to the way they handle missing data.

6.4.1 Simulation design

6.4.1.1 Study model

For the simulation study, I designed an ITS dataset where the treatment of interest was the initiation of antipsychotic treatment, and I examined the change in body weight (in kilograms) over time. The covariates were sex, age (years) and smoking status (yes/no), measured at initiation of treatment. The ITS impact model [8] is a linear weight trajectory whose slope changes only once – at treatment initiation – i.e. slightly simpler than my previous example. I included five time-units before and five after treatment initiation. I modelled the evolution of weight over time using two continuous linear splines, jointing at treatment initiation. For the simulation study design, I followed the ADEMP approach (defining aims, data-generating mechanisms, estimands, methods, and performance measures) recommended by Morris et al [101].

6.4.1.2 Data generation

Each simulated dataset with 1,000 observations was generated as follows:

1 Sex was generated as a random variable from a Bernoulli distribution with probability 0.5.

2 For each individual, weight observation times were fixed at the same 11 equally spaced times between -5 and +5, i.e. centred at treatment initiation, which is at time 0.

3 Weight was generated from the following random intercept and slopes model:

[Equation 6.2]

$$weight_{ij} = 75 + u_{0j} + \left(-0.5 + u_{1j}\right)time_{ij} \times 1\left[-5 \le time_{ij} < 0\right] + \left(3.4 + u_{2j}\right)time_{ij} \times \left[0 \le time_{ij} \le 5\right] + 10 * sex_i + \varepsilon_{ij},$$

$$\begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{pmatrix} \sim N\begin{pmatrix} 0 & 5 & 0 & 0 \\ 0, & 0 & 1.1 & -.7 \\ 0 & 0 & -.7 & 1.1 \end{pmatrix}; \qquad \varepsilon_{ij} \sim N(0,2),$$

where *i* denotes the follow-up time and *j* denotes the patient, and 1[ ] is an indicator for the event in square brackets. We referred to this as 'Data Generation Mechanism Base' (DGM-base). I also generated data from DGM-extended covariates:

[Equation 6.3]

$$weight_{ij} = 75 + u_{0j} + \left(-0.5 + u_{1j}\right)time_{ij} \times 1\left[-5 \le time_{ij} < 0\right] + \left(3.4 + u_{2j}\right)time_{ij} \times [0 \le time_{ij} \le 5] + 10 * sex_i + 0.05 * age_i - 0.0005 * age_i^2 + 2.5 * smoking_i + \varepsilon_{ij}.$$

Age was generated as a random variable from a normal distribution with mean 45 and sd 10. Smoking was binary and generated as follows:

$$logit\left(P(smoking_i = 1)\right) = -2 + 1.5 * sex_i + 0.04 * age_i - 0.0005 * age_i^2.$$

Having generated the full data, I made observations missing using two missing data mechanisms:

1. MAR-1: starting with the fully observed weight variable at treatment initiation (t_0), pre and post-treatment initiation values of weight at times $t_{0\pm j}$ were set to missing ($j = 1,2,3,4,5$) dependent on the individual's sex. For the missing sequence, pre-treatment setting of missing values was reverse-sequential ($t_{-1}, t_{-2}, t_{-3}, t_{-4}, t_{-5}$) and post-treatment setting was forward-sequential (t_1, t_2, t_3, t_4, t_5). For both directions ($\pm j$) of MAR-1 mechanism, I defined the probability of being missing by:

$$logit\left(P\left(weight_{ij} = missing\right)\right) = -2.5 + 5 * sex_i,$$

shaping the patterns of missing weight data and setting more weight records being observed for women than men. Both patterns and the proportion of missing values are available in Figures 6.3 and 6.4. MAR-1 was applied to data generated under DGM-base only.

2. MAR-2: similar to MAR-1, but now the probability of weight being missing also depends on the individual's random intercept, age and smoking. As the random intercept is unobservable (as smoking will partially be), this mechanism is a mix between MAR and MNAR (missing not at random). Moving away from treatment initiation (in both directions), the probability of weight being missing is monotonically given by:

$$\begin{aligned} logit\left(P\left(weight_{ij} = missing\right)\right) \\ = -0.25 - 2 * u_{0j} - 1.5 * sex_i - 0.05 * age_i + 0.0005 * age_i^2 \\ - 1.5 * smoking_i, \end{aligned}$$

where -0.25 helped to shape the overall proportion of missing data over time; -1.5 set more weight records to be observed for men (only for explicative purposes); -2 set more weight records to be observed for individuals who are heavier at treatment initiation; -0.05 and 0.0005 set more missing data for younger individuals, and -1.5 set more weight records to be observed for smokers. I also set about 30% of smoking values to be missing with probability:

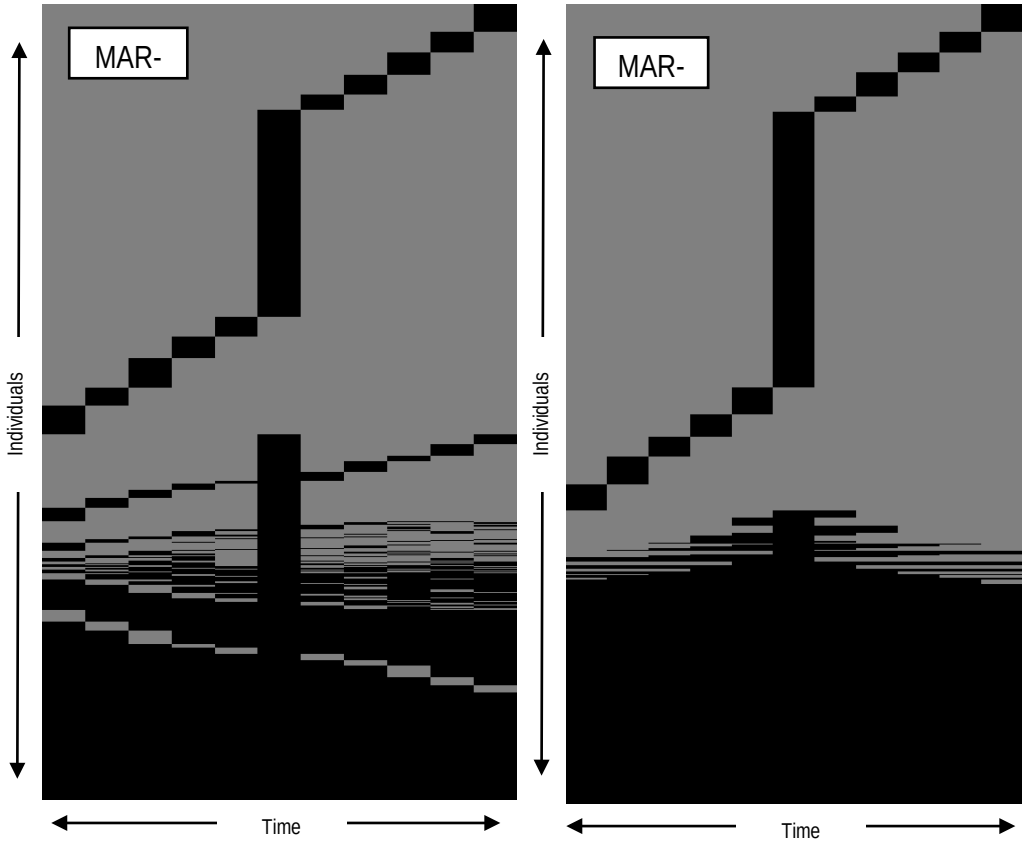
$$logit\big(P(smoking_i = missing)\big) = -3 + 3 * sex_i - 0.01 * age_i + 0.0003 * age_i^2,$$

MAR-2 was applied to data from DGM-extended-covariates only. For both described mechanisms (MAR-1 and -2), the proportion of missing weight data in the simulated sample was set to approximately 60% of individuals. In the other 40% of the data, I set only one weight record per individual at any time point, setting more individuals with only one weight record at treatment initiation (MAR dependent on the treatment initiation). This additional mechanism sought to emulate the missing data proportions and patterns seen in the clinical data used for the illustrative example (Figures 6.3 and 6.4).

I simulated 1,000 full datasets for each of the two scenarios, and then applied the missing data mechanisms to obtain the partially observed data.

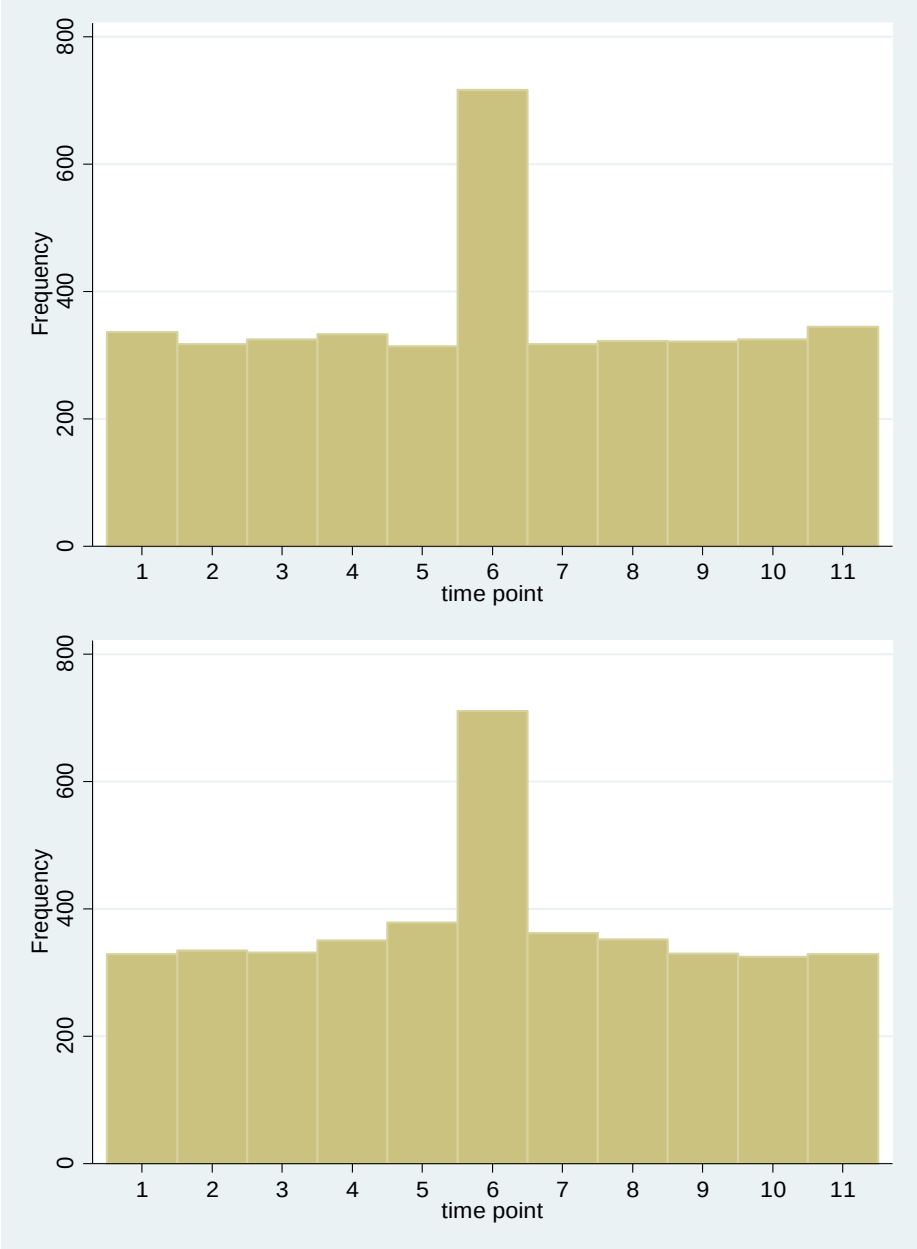
Figure 6.3 shows an example (one simulation) of the missing data patterns generated by the MAR-1 mechanism, and another example for the MAR-2 mechanism. Figure 6.4 shows the distribution of weight records over time for the same mechanisms. It is visible how these mechanisms reproduced similar characteristics of real data (i.e. THIN data, see Figure 3.1 as a reference), in terms of outcome distribution over time and missing outcome patterns.

Figure 6.3. Missing data patterns in an example of MAR-1 and another example of MAR-2 datasets.



In the grey area, each black rectangle is a weight record, each row is a person (1000 in total) followed over 11 time points (columns).

Figure 6.4. Missing data distributions in an example of MAR-1 and another example of MAR-2 datasets (N=1000).



6.4.1.3 Analysis methods evaluated

I analysed the full and partially observed data using each of the following six methods (see summary in Appendix 6C):

1) SR: this averaged observed individual weight measures at each time point and then fits a linear regression on time (maximum likelihood estimator), with a knot at zero.
2) SR-W1: (weighted SR version 1) similar to SR but weighted by the inverse of the number of observed weight records at each time point.
3) SR-W2: (weighted SR version 2) similar to SR-W1 but the number of observed weight records – used for weighting – were counted at each time point by sex and age. I categorised age using its quartiles (before averaging). When smoking data were incomplete, smoking was not included as a covariate for SR-W2.
4) Prais-Winsten: regression similar to SR but adjusted for serial correlation at the aggregate level by assuming errors that follow a first-order autoregressive process [102], an approach typically used in ITS analysis for controlling the autocorrelation issue [6].
5) MEM: I fitted the data generating model [Equations 2 and 3] using Restricted Maximum Likelihood with an unstructured covariance matrix for the random effects.
6) MI-JOMO (with MEM): I first imputed the missing covariate values, using multilevel substantive-model-compatible joint modelling multiple imputation, with the JOMO package in R. As described in [57,103] this imputes missing values consistent with the substantive model [Equation 6.1]. It does this by factorising the joint model into a joint model for the covariates and a conditional model for the outcome given the covariates. Then, the estimation and imputation process allows compatibility between the imputation and analysis models (MEM in this case), even with longitudinal data [57] (see sections 3.2.3 to 3.2.5 for details). I used 5 imputations and 1000 iterations (before the first, and between each subsequent imputation) to impute the missing covariate smoking status. I did not impute the missing weight, as (in the absence of auxiliary variables) no information can be recovered by doing this. Thus, outcome was implicitly imputed when 5 MEM were fit (one per each imputed dataset). Rubin's rules

summarise these estimates. Note that standard fully conditional specification [48] is not evaluated because it is inappropriate for handling the irregular observation times I expect in real longitudinal data. I only used MI-JOMO in the MAR-2 scenario.

6.4.1.4 Estimands and performance measures

I focused on the slope estimates (true values: $time_{before}: \beta_1 = -0.05$ and $time_{after}: \beta_2 = 3.4$) from all methods evaluated in both MAR scenarios (MAR-1 and MAR-2), by examining the bias, empirical standard error, model-based standard error and confidence interval coverage [101].

6.4.2 Simulation results

In the first scenario (DGM-base), all SR methods were biased except when data were fully observed (Table 6.2). However, the coverage of these methods was low (<61%) due to their small model-based standard errors, even the weighted methods (SR-W1 and SR-W2) and the method adjusted for serial correlation (Prais-Winsten). Conversely, MEM provided reasonably good coverage for β_1 and β_2 (>94%) for unbiased estimates.

Where weight was missing based on sex only (MAR-1), MEM showed unbiased results and the best coverage (≥95%). SR and SR-W1 produced biased estimates for both pre- and post-treatment initiation slopes, showing the highest model-based standard errors. Because the missingness mechanism depended on sex, and women weighed less than men, the preliminary data aggregation step in SR and SR-W1 biased the estimated slopes (see example in Figure 6.5, MAR-1). The SR bias was corrected using inverse-probability weights based on sex (SR-W2), but coverage was low (<74%) due to too-small model-based standard errors. The Prais-Winsten was not successful in correcting the SR bias since it does not incorporate information on missing data at the individual-level as SR-W2 does.

In the second scenario (DGM-extended-covariates), with full data, all methods were unbiased (Table 6.2). MEM provided the best coverage for β_1 and β_2 (>95%), followed by SR-W2 (>90%). Although with unbiased estimates, SR, SR-W1 and Prais-Winsten provided low coverage (<55%) due to their small model-based standard errors. SR, SR-W1 or Prais-Winsten cannot provide different averages by sex and age at each time point, which can be provided by SR-W2. Having more variability at each time point produced higher – and more realistic – standard errors from SR-W2.

On the other hand, with missing values in weight and smoking status (MAR-2), MI-JOMO had the best performance. MEM showed more reduced performance after all covariates were included in the imputation and study models and there were missing smoking data, producing slightly biased estimates and low coverage (<79%). In the same scenario, MI-JOMO performed better than MEM, providing less biased estimates, closer values of empirical and model-based standard errors, and higher coverage (>87%). For both methods, we should consider that there is some residual bias because of the dependence of observation of weights on the random intercepts. While the results in the bottom half of Table 6.2 show this resulted in a bias in the MI-JOMO analysis, this was not severe, and the resulting inferences were still usable. Conversely, SR, SR-W1, SR-W-2 and Prais-Winsten performed extremely poorly, showing large bias and low coverage (<18%).

The 'aggregate-level' SR analysis biased the slope trajectories in different directions, which I illustrated by the simulations (Figure 6.5).

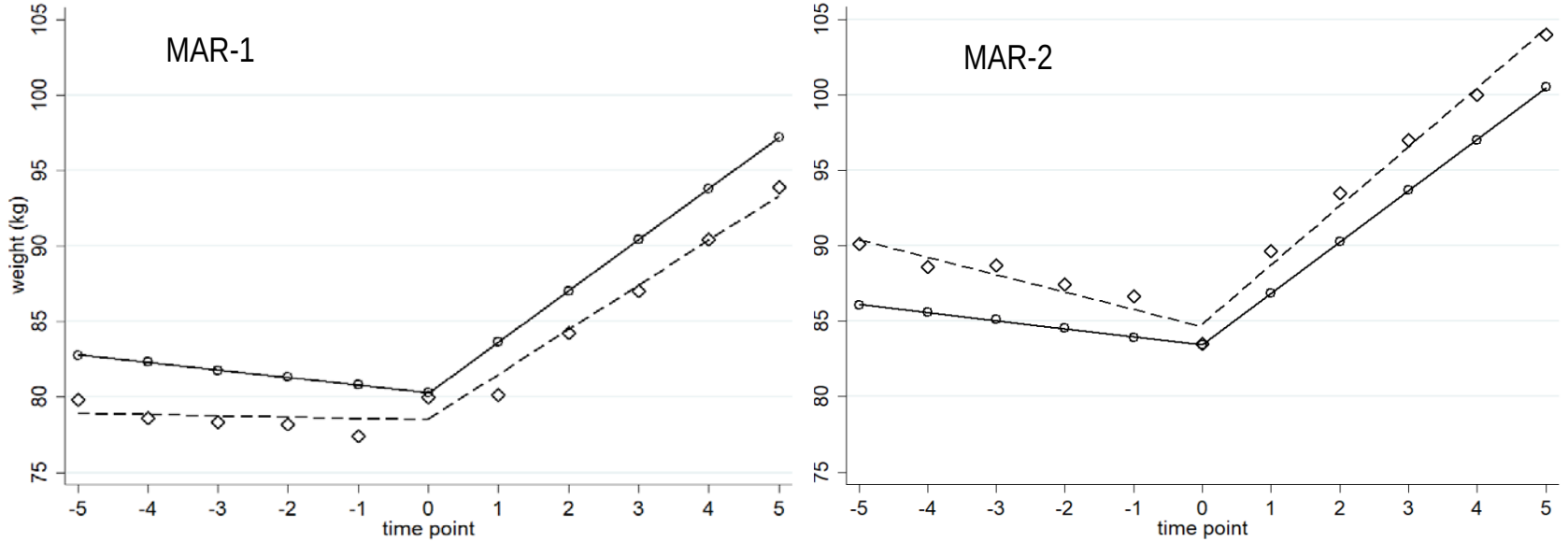
Table 6.2. Simulation results.

Data Generation Mechanism (DGM)	Missing data Mechanism	Estimand	True Parameter Value	Method	Simulated Data (time points=11; individuals=1000; replications=1000) Bias	Empirical Standard Error	Model-based Standard Error	Coverage
DGM-base	Data fully observed	β1 (before treatment initiation slope)	-0.5	MEM	-0.0004	0.0356	0.0343	94.3
				SR	-0.0004	0.0356	0.0134	52.4
				SR-W1	-0.0004	0.0356	0.0111	42.7
				SR-W2	-0.0004	0.0356	0.0165	57.5
				Prais-Winsten	-0.0004	0.0356	0.0105	41.4
		β2 (after treatment initiation slope)	3.4	MEM	0.0006	0.0345	0.0343	95.2
				SR	0.0006	0.0345	0.0134	52.6
				SR-W1	0.0006	0.0345	0.0114	44.5
				SR-W2	0.0006	0.0345	0.0164	60.5
				Prais-Winsten	0.0006	0.0345	0.0105	42.3
	MAR-1 (weight MAR on sex only)	β1 (before treatment initiation slope)	-0.5	MEM	0.0014	0.0523	0.0535	95.7
				SR	0.3761	0.0682	0.2388	88.6
				SR-W1	0.5801	0.0724	0.3113	64.9
				SR-W2	0.0010	0.0591	0.0381	72.2
				Prais-Winsten	0.3690	0.0679	0.1795	40.1
		β2 (after treatment initiation slope)	3.4	MEM	-0.0003	0.0529	0.0534	95.0
				SR	-0.3748	0.0719	0.2388	87.1
				SR-W1	-0.5786	0.0755	0.3117	66.8
				SR-W2	-0.0002	0.0606	0.0380	73.7
				Prais-Winsten	-0.3682	0.0718	0.1795	40.9
DGM-extended-covariates	Data fully observed	β1 (before treatment initiation	-0.5	MEM	0.0000	0.0340	0.0343	95.1
				SR	0.0000	0.0340	0.0134	53.6
				SR-W1	0.0000	0.0340	0.0112	47.0

	slope)		SR-W2	0.0000	0.0340	0.0316	90.3
			Prais-Winsten	0.0000	0.0340	0.0106	43.1
	β2 (after treatment initiation slope)	3.4	MEM	-0.0001	0.0336	0.0343	96.0
			SR	-0.0001	0.0336	0.0134	54.0
			SR-W1	-0.0001	0.0336	0.0113	44.7
			SR-W2	-0.0001	0.0336	0.0315	90.1
			Prais-Winsten	-0.0001	0.0336	0.0106	43.5
MAR-2 (weight MAR on sex, age, and smoking (smoking has missing data))	β1 (before treatment initiation slope)	-0.5	MEM	-0.0649	0.0884	0.0674	78.7
			MI-JOMO	-0.0335	0.0691	0.0697	88.7
			SR	-0.4535	0.0709	0.1954	17.6
			SR-W1	-0.5963	0.0765	0.2605	11.3
			SR-W2	-0.5385	0.0667	0.0872	0.0
			Prais-Winsten	-0.4515	0.0708	0.1565	1.4
	β2 (after treatment initiation slope)	3.4	MEM	0.0652	0.0883	0.0674	76.2
			MI-JOMO	0.0333	0.0698	0.0685	87.7
			SR	0.4513	0.0724	0.1954	17.9
			SR-W1	0.5941	0.0780	0.2609	13.7
			SR-W2	0.5367	0.0672	0.0874	0.0
			Prais-Winsten	0.4495	0.0724	0.1565	2.3

MAR=missing at random, SR='aggregate-level' segmented regression, SR-W1='aggregate-level' segmented regression weighted with the inverse of the number of observed weight records at each time point, SR-W2= similar to SR-W1 but the number of observed weight records were counted by each time point, sex and age group (quintiles), MEM=random intercept and slope model with restricted maximum likelihood and unstructured covariance, MI-JOMO=substantive model compatible joint modelling multiple imputation using a similar MEM model, Prais-Winsten= 'aggregate-level' Prais-Winsten regression.

Figure 6.5. Weight trajectories from a simulated dataset in which weight is fully observed (circles and solid lines) or missing at random (diamonds and dashed lines).



6.5 Discussion

ITS provides a conceptually attractive approach for assessing the impact of treatments because each individual acts as their own control. However, its innate strength, leading to its increasing use [61], raises important questions about how to handle missing data appropriately. As my example illustrates, incomplete outcome data (in this case, weight) is an intrinsic feature of this kind of study because the underlying observational data do not follow any pre-planned schedule. This means that, at any specific time, the marginal distribution of the response is unlikely to be representative of the underlying population.

The results of my studies demonstrate that the 'aggregate-level' approach will generally be biased when individual-level data are missing at random (MAR). Indeed, the motivating example shows this bias could lead to a substantial exaggeration of the actual effect of the studied intervention. In the example, the difference between pre- and immediate post-treatment weight change (biased slopes) increases the overall effect attributed to olanzapine. However, it is not always possible to determine the direction of bias. This is because the direction of the average-points bias depends on how the covariate is associated with the missingness of weight records. Even when the 'aggregate-level' SR analysis does not bring about a bias issue, these results highlight that the precision is inaccurate as the standard errors for this method are typically grossly underestimated.

When data are missing-at-random at the individual level, averaging before SR means that data are missing-not-at-random at the cluster level. This leads to the bias observed for the 'aggregate-level' SR analysis. For example, in the MAR-1 mechanism, 'aggregate-level' SR analysis loses the information about the distribution of weight records that are MAR on sex at each time point, due to the averaging-step. Thus, sex becomes unobservable at the 'aggregate-level', making weight records MNAR on sex at this level and biasing the subsequent analysis using those averages. As I demonstrated in the same simulation study, this issue could be handled by including sex in the averaging-

step (SR-W2). However, in practice, any version of SR-W2 will be hard to apply since other covariates are typically incomplete as well.

A natural alternative to the 'aggregate-level' analysis is to model the individual patient data explicitly. When the reason for outcome data being observed depends principally on time (e.g. before and after treatment initiation), underlying patient characteristics (e.g. sex, age) and observed outcomes (e.g. observed weights), the unseen values are plausibly MAR. In this setting, the simulation results demonstrate that a carefully formulated longitudinal model provides a practical approach for improved inference.

Longitudinal models should be specified carefully to include covariates predictive of both the outcome and the chance of observing it, which are vital for avoiding bias. Where it is not desired – or appropriate – to include some such variables in the substantive analysis, an MMI approach should be considered, where these variables are included as auxiliary variables. Care should also be taken to model the longitudinal correlation of the outcome appropriately, as this is particularly important for missing data, as well as to use the observed rather than expected information for likelihood-based models. In particular, having random intercepts alone, or having uncorrelated random intercepts and slopes, should be avoided (see Appendix 6D for other practical suggestions) [56]. If data at the individual level are not available, and the researcher suspects that a strong MAR mechanism affect the outcome points over time (e.g. averages or rates), the issue should be stated as a limitation as recommended in reporting guidelines [104,105].

Study results show that MMI provides a practical approach for handling missing covariates in the analysis. When performing MMI, it is essential to both use an approach that appropriately takes account of the multilevel structure and uses an approach that is compatible with the substantive model (which here includes splines for the effect of time). The JOMO package in R has the flexibility to do both.

I set the example and simulations with averages of a continuous variable, but a similar problem can happen with other types of outcomes. Rates (proportions), another

common ITS outcome [7], can also be biased when outcome data are MAR at the individual level. For example, if the numerator of the rate (the events) is higher in women than men, and the missingness process generates more missing records for women, the rate will be underestimated at the 'aggregated-level' (e.g. at time points, hospitals or districts). The ITS analysis will use those rates as consecutive points, biasing the estimated trajectories. Similar reasoning can be applied to binary and count ITS outcomes. Even using other recommended analysis methods than SR, such as ARIMA models [6], the bias problem will remain in the 'aggregate-level' used for the time series. Although I did not formally evaluate these alternative methods, some reflections can be enlightened by the study findings. In the aggregate-level approach, ARIMA models will be fitted after the averaging-step; therefore, the ITS will be based on population-level average points already biased. Other options useful for individual-level data, such as generalised estimated equations (GEE) can be applicable. However, because they are moment based estimates, precisely like the aggregate data analysis, its estimates will be biased unless data are missing completely at random [106,107].

This is the first time that this averaging-step problem for MAR data has been studied with simulations and real data. Study results will help to guide future ITS researches. I focused my study on the situation when data are missing at random. However, I am aware there may be other scenarios where data missing not at random (MNAR) could bias estimates. For example, if weight is only recorded for those with a high or low weight. This scenario goes beyond the scope of this thesis, but in practice, when a strong MNAR mechanism is suspected, a sensitivity analysis is possible using a pattern mixture approach [53,72]. Although I designed the data-generation mechanism to approximate the associations and missing data patterns that are visible in routinely collected data (i.e. THIN data), the complexity of real data is always hard to reproduce. Thus, the simulation is a simplified reproduction of reality, imposing some limits to its external validity (as with any simulation study). In practice, simulations demonstrated a serious issue with analyses that use an averaging step. Researchers can trust in the performance of MEM and MMI for handling missing data in standard ITS studies [6], assuming data are MAR at the individual level. For more sophisticated analyses to be valid (e.g. in controlled ITS designs), these methods should work well as long as the substantive models can be correctly specified as mixed-effects models, while keep similar missing data assumptions.

In conclusion, the segmented regression using averaged data points can over or underestimate the effect evaluated in interrupted time series analyses, when performed on outcome data missing at random at the individual level. However, such a problem can be addressed by using mixed models. If there are also covariates missing at random, mixed models can be combined with multilevel multiple imputation and provide unbiased results.

6.6 Summary [13]

Interrupted time series (ITS) is a widely used quasi-experimental approach that evaluates the potential impact of an intervention over time, using longitudinal observational data. In ITS on individual-level data, it is a common practice to average the outcome of interest at each time point, and then fit an 'aggregate-level' segmented regression (SR). This procedure can lead to biased estimates if data are missing at random (MAR) at the individual level. Alternative mixed effect models (MEM) can avoid average the outcome by modelling longitudinal individual-level data directly. However, MEM can be inefficient for handling missing covariates; thus, other alternatives such as multilevel multiple imputation (MI-JOMO) need further exploration. In this chapter, I compared all these methods by an illustrative example and a simulation study.

In the illustrative example, I found substantial differences between estimates derived from MEM and aggregate-level SR. For example, short-term weight change (β_2) was 0.462kg/week from MEM and 0.816kg/week from aggregate-level SR. Likewise, in pre-treatment and long-term periods, weight change rates from aggregate-level SR were more than double the MEM estimates. In general, all estimates of weight change from aggregate-level SR analyses were higher in magnitude than those from MEM, which also implies a more substantial ITS treatment effect. To investigate my hypothesis that the aggregate-level SR was producing biased results, I used a set of contextually informed simulation studies.

Simulation results confirmed that the averaging-step causes bias in ITS estimates when data are MAR at the individual level. This occurs because taking averages of individual-level data before SR means that data at the cluster level are missing not at random. I also confirmed that the aggregate-level SR can over or underestimate the ITS effect. Nevertheless, it is not always possible to determine the direction of bias, because the direction of bias at each average-point depends on how the covariate is associated with the outcome missingness. Even when the 'aggregate-level' SR analysis does not cause any bias, results from simulations highlight that the precision is smaller as the standard

[13] This study has been accepted for publication as an original article. Please, see Appendix 10.5 for details.

errors for this method are typically grossly underestimated. Avoiding the averaging-step and using MEM is recommended for handling missing values in the outcome.

For handling values MAR in covariates, MEM must be combined with MI-JOMO to obtain less biased estimates. I found that the most efficient way to do it is a two-steps process. First, missing values on covariates need to be multiply imputed, considering a model that is consistent with the substantive models. MI-JOMO allows us to complete this task efficiently. Then, MEM should be estimated using all the multiple imputed datasets, and multiple results can be summarised using Rubin´s rules. Simulation study results confirm the estimator based on MI-JOMO followed by a MEM is unbiased when data at individual-level is MAR.

Imputation of covariates with missing records could also help in the evaluation of interaction terms, useful for comparison between trajectories of different sub-groups and further CITS studies. In the next Chapter 7, I explore how MEM with MI-JOMO can help to handle missing data in ITS studies when evaluating variables that could modify outcome trajectories overt time.

7

An application of Multilevel Multiple Imputation to Interrupted Time Series analysis

7.1 Introduction

Second-generation antipsychotics increase the weight of patients in the first weeks of treatment [5]. In the long term, the evidence is more scarce, though some research has recently appeared addressing this [82,92,100]. Olanzapine is the antipsychotic that has been shown to induce more weight gain across the studies, in both the short and long term. Risperidone and quetiapine also increase weight, but less than olanzapine; and the weight gain for these seems to be more pronounced in the short than in the long term [100]. Most studies have examined people prescribed antipsychotics in general (>18 years with any diagnosis) [5] or with specific mental health conditions (e.g. >65 years with dementia) [108]. In the latter review (i.e. for people with dementia), the evidence was quite heterogeneous, reporting from no weight change to significant

weight gain (or reduced weight loss) due to second-generation antipsychotics. The designs of the reviewed studies were also heterogeneous (e.g. different sample sizes or comparison groups), and most were performed during short terms (<3 months). In older people, evidence from longer terms is needed since this population experience a natural weight loss over time that younger people typically do not experience.

There is a gap in the literature on antipsychotic-induced weight gain in older people (>65 years) with any diagnosis, alone or in comparison with younger people [91]. Since older adults are likely to have lower basal metabolism [108], it is possible that second-generation antipsychotics may have a different impact on weight in older adults. In young adults, starting treatment with low weight can lead to more rapid weight gain [86,87], but this effect has not been thoroughly studied in the older population. This knowledge has important clinical implications, especially for informing how prescriptions in older patients can affect their healthy weight [109].

We also need a better understanding of the potential modifying effect of sex and dose [14] on antipsychotic-induced weight gain. Evidence of the weight gain differences between women and men comes from short-term clinical trials (typically <1 year) [110] and some observational studies [82,100]. Evidence of the role that dosage plays in weight gain is scarcer, but I recently explored this question with observational data (Chapter 5) [100]. Nevertheless, further evaluation of the modifying effects of sex and dose will improve our knowledge in this area. In the absence of long-term randomised clinical trials, observational data can have long enough follow-up and enough patients to be powered to detect effect modification [1]. Interrupted Time Series (ITS) allow estimation of causal relationships in observational data by comparing outcome trajectories before and after treatment initiation, by using regression models, as shown in Chapter 6. Given the initiation of a particular antipsychotic treatment, the differences in weight change trajectories between men and women, or between low or high dose, can be evaluated with interaction term in ITS regression models [2,3].

[14] Although dose is not a modifying effect formally speaking (it can be seen as levels of the exposure), I am keeping this terminology through the chapter to facilitate the narrative, which is centred in the evaluation of effect modification.

Observational datasets usually contain missing values that need to be handled appropriately for ITS models to provide unbiased and efficient estimates. In Chapter 6, I showed that Mixed Effects Modelling (MEM) is a valid tool for handling ITS analysis with outcome data missing at random in individual-level data (e.g. electronic health records). I also demonstrated how MEM, in combination with multilevel multiple imputation (i.e. MI-JOMO), provides unbiased and more efficient estimates when covariates are incomplete. Now, I will apply MEM with MI-JOMO to handle missing data in the ITS regression models with interaction terms in order to evaluate the modifying effects of age, sex and dose.

7.2 Objectives

The overall aim was to investigate how the sex, age at treatment initiation and the prescribed dose may impact weight change induced by olanzapine, risperidone and quetiapine in patients aged 40 to 89 years.

The specific objectives were:

1. Clinical: To evaluate the weight trajectories before and after initiation of antipsychotic treatment in ITS analysis and examine whether sex, age and dose may independently modify these trajectories.

2. Methodological: To apply MEM in combination with MI-JOMO to handle missing data when these trajectories and the mentioned modifying effects are modelled.

7.3 Methods

7.3.1 Data source

I used anonymised, longitudinal patient records from The Health Improvement Network (THIN), a database that comprises information from UK primary care electronic health records from general practices as described in Chapter 2.

7.3.2 Study population

At the individual level, I included all patients aged between 40 and 89 years at the date they started their first treatment with olanzapine, risperidone or quetiapine; between 1 January 2007 and 31 December 2017. I did not include patients ≥90 years to ensure a plausible positivity assumption (e.g. observed low/high dose vary within all subgroups by age) [111]. Although the main interest was in people aged 60 years or more, I included younger people (40-60 years) to be able to detect differences across ages. All patients had a diagnosed psychiatric disorder (schizophrenia, bipolar disorder, other non-affective psychoses, borderline personality disorder, anxiety, depression or dementia) and at least one additional prescription of the same antipsychotic (AP) within three months after the first prescription. The idea is that people with additional prescriptions were more likely to have initiated treatment than those with a single prescription. Patients were excluded if they initiated on more than one type of AP (including switchers) at any time, as well as if they had no records of the year of birth, sex or social deprivation. Likewise, I excluded individuals with no available data at any point during the 12 months before the date of treatment initiation to avoid the potential impact of unrecorded prescription of other antipsychotics.

7.3.3 Variables and measurements

The exposure of interest was the initiation of olanzapine, risperidone or quetiapine treatment. The outcome was the change in body weight, measured in kilograms. The main covariates were sex (women/men), age (40-49, 50-59, 60-69, 70-79, 80-89 years) and first prescribed dose of AP. All AP reported the first doses in milligrams, but I used

the dose-equivalence approach of Woods [89] explained in Chapter 5 (section 5.3.3). As I explained in the same chapter, the first dose is a good predictor of subsequent doses prescribed during treatment. For descriptive purposes, I also retrieved information on height, social deprivation (quintiles of Townsend score 1-5, from least to most deprived), smoking and drinking status, having a type-2 diabetes mellitus diagnosis, systolic blood pressure (SBP), low-density lipoprotein cholesterol (LDL-cholesterol) and high-density lipoprotein cholesterol (HDL-cholesterol); recorded within the first year before initiation of treatment. If there were multiple measures during that year, I kept the record closest to AP treatment initiation. All the variables passed a data cleaning before the statistical analysis (e.g. removing impossible weight records).

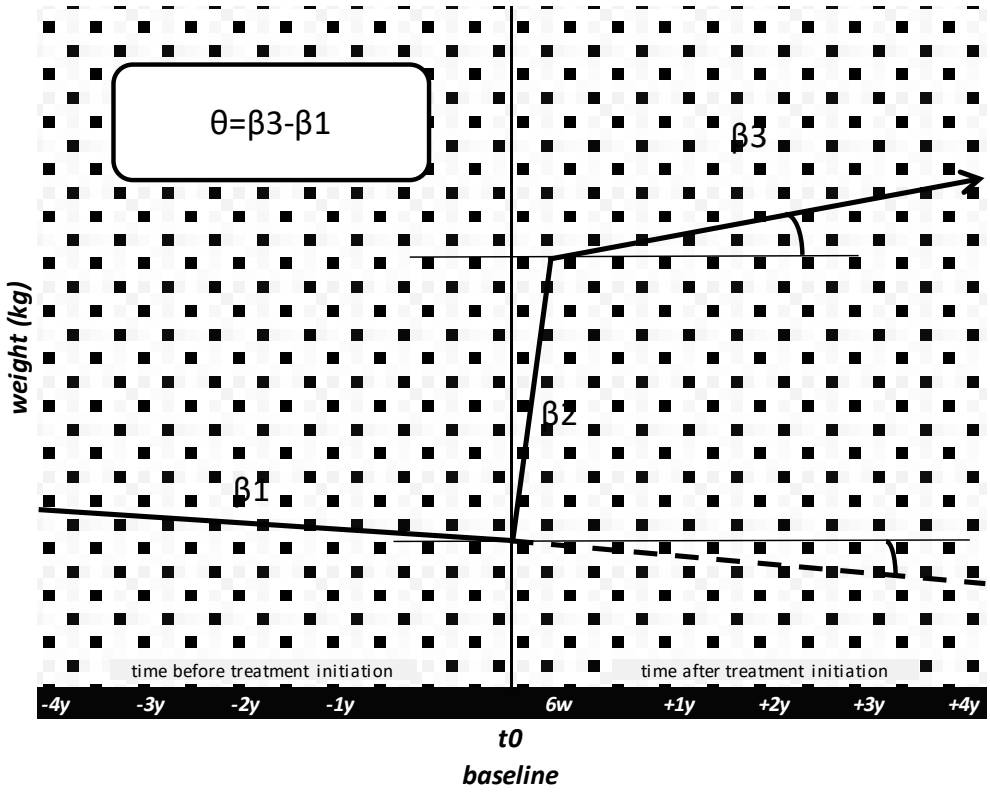
7.3.4 Statistical analysis

I used an interrupted time series approach [6] to analyse weight change over time within three AP initiation cohorts, during four years before and four years after treatment initiation. Weight trajectories over time were modelled by using continuous linear splines with random intercept [15] and slopes (unstructured covariance for the random intercept and slopes, restricted maximum likelihood), from which three slopes of weight change were estimated for: 1) −4 years to baseline (pre-treatment), 2) baseline to +6 weeks (short-term), 3) +6 weeks to +4 years (long-term).

Figure 7.1 provides a graphical representation of the main model, as well as the long-term effect size (θ) that the ITS design helps to visualise:

[15] The intercept (time zero) is at treatment initiation.

Figure 7.1. Visualization of the model estimates and the long-term ITS effect size



The model is similar to those fitted in Chapter 5 [100] although in this study, the target population is different (40-89 years) and some extra effects are evaluated [16]:

a) Interaction between sex and time (for each slope) adjusted for age and social deprivation at treatment initiation (all variables fully observed). Slopes of weight change were also estimated separately for each group by sex.
b) Interaction between age (categorised) and time (each slope) unadjusted and adjusted for sex, social deprivation and dose (unique covariate with missing values). Slopes of weight change were also estimated for each group by age category separately.
c) Interaction between dose and time (each slope) unadjusted and adjusted for sex, age and social deprivation. Slopes of weight change were also estimated for each group by dose (low/high) separately.
d) Short-term weight change (0-6 weeks) within subgroups by dose and age categories were estimated from similar MEMs.

[16] I call them 'model type' (a), (b), (c) or (d) in Section 7.4 on study results.

Data used in (a) did not involve incomplete covariates, only missing outcomes, which were handled by using MEM. For models in (b), missing dose values were handled by complete case analysis (CCA); and for models in (c), missing dose values were handled by MI-JOMO [17]. For models in (d) -and the descriptive cross-tabulation of age categories against dose -, I used CCA.

I assumed weight records were missing at random within covariate strata, conditional on observed weights. In THIN data, weight records before and after treatment initiation help to inform the implicit imputation made by MEM [18], assuming these records are associated with missing values of weight in the same trajectory. Under this assumption, modelling the observed data over time provides unbiased estimates [48]. I also assumed that weight was MAR on dose; thus, the complete case analysis I performed provides unbiased estimates [49].

Parameter estimates are reported with 95% confidence intervals. All the statistical analyses were performed using Stata 15 for Windows [90] and R for windows [56,112].

[17] I used substantive-model compatible MI-JOMO to impute consistent with the interaction with partially observed dose. For more details about how MI-JOMO is compatible with the substantive ITS models (MEM), please see Section 3.2.

[18] See Section 3.3 for more details.

7.4 Results

7.4.1 Cohort characteristics

I included data from a total of 29,039 patients during the 8 years of follow-up [19], in the cohorts of olanzapine (n=6,776), quetiapine (n=14,970) and risperidone (n=7,293). A total of 155,175 weight records were available for that period, 35,299 for olanzapine, 81,131 for quetiapine and 38,745 for risperidone cohorts. Across the cohorts, patients were mainly women (>55%), younger if they were prescribed olanzapine, and older if they were women and prescribed risperidone (Table 7.1). Non-drinkers and non-smokers were more prevalent among women (≈52%) than men (≈37%), and diagnoses with type-2 diabetes mellitus were slightly less prevalent in women (≈19%) than men (≈22%). Systolic blood pressure was similar across cohorts (mean≈130, standard deviation(sd)≈15 mmHg). Low-density lipoprotein (LDL) cholesterol (mean≈3.3, sd≈1.1 mmol/L) and high-density lipoprotein (HDL) cholesterol (mean≈1.5, sd≈0.6 mmol/L) were slightly higher in women for all cohorts. For men, the first dose of olanzapine (mean=7.1, sd=4.8 mg), quetiapine (mean≈81.5, sd≈112.7 mg) and risperidone (mean≈1.4, sd≈1.3 mg) was on average higher than for women. Weight reported at treatment initiation is similar across cohorts, but this is just due to missing weight records (>42% across cohorts), which is corrected in latter tables (e.g. when MEM and MI-JOMO are used).

In general, high doses [20] were less prevalent in older patients, for all drugs (Table 7.2). For example, for people aged 40-49 years and who were prescribed olanzapine, 854 (43%) were on low dose and 614 (31%) on high dose, whereas people aged 80-89 years were 472 (62%) on low dose and 58 (8%) on high dose. Missing values of dose were 1853 (27%) for olanzapine, 6020 (40%) for quetiapine and 2532 (35%) for risperidone.

[19] From the index date (antipsychotic treatment initiation), 4 years before and 4 years after.
[20] For details on how high/low doses were defined, please see the dose-equivalence approach of Woods explained in Chapter 5 (Section 5.3.3).

Table 7.1. Baseline characteristics of people using olanzapine, quetiapine and risperidone treatment by sex.

Characteristic	*OLANZAPINE (N=6,776)*				*QUETIAPINE (N=14,970)*				*RISPERIDONE (N=7,293)*			
	Female		*Male*		*Female*		*Male*		*Female*		*Male*	
	N=3,803 (56.1%)		*N=2,973 (43.9%)*		*N=8,765 (58.6%)*		*N=6,205 (41.4%)*		*N=4,127 (56.6%)*		*N=3,166 (43.4%)*	
	n	%	n	%	n	%	n	%	n	%	n	%
Age (years)												
40-49	973	25.6	1007	33.9	2379	27.1	1598	25.8	668	16.2	685	21.6
50-59	847	22.3	773	26.0	1664	19.0	1174	18.9	549	13.3	516	16.3
60-69	736	19.4	540	18.2	1125	12.8	844	13.6	558	13.5	456	14.4
70-79	717	18.9	423	14.2	1429	16.3	1280	20.6	926	22.4	688	21.7
80-89	530	13.9	230	7.7	2168	24.7	1309	21.1	1426	34.6	821	25.9
Missing	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
total (n)	*3803*	*100.0*	*2973*	*100.0*	*8765*	*100.0*	*6205*	*100.0*	*4127*	*100.0*	*3166*	*100.0*
Townsend												
Least deprived	730	19.2	483	16.2	1573	17.9	1170	18.9	804	19.5	562	17.8
2	808	21.2	548	18.4	1636	18.7	1204	19.4	803	19.5	621	19.6
3	778	20.5	615	20.7	1980	22.6	1294	20.9	949	23.0	667	21.1
4	825	21.7	672	22.6	1963	22.4	1334	21.5	855	20.7	693	21.9
Most deprived	662	17.4	655	22.0	1613	18.4	1203	19.4	716	17.3	623	19.7
Missing	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
total (n)	*3803*	*100.0*	*2973*	*100.0*	*8765*	*100.0*	*6205*	*100.0*	*4127*	*100.0*	*3166*	*100.0*
Smoking Status												
ex-smoking	737	19.4	750	25.2	1927	22.0	2008	32.4	963	23.3	1036	32.7

non-smoking	1927	(	50.7	)	1036	(	34.8	)	4322	(	49.3	)	2306	(	37.2	)	2295	(	55.6	)	1236	(	39.0	)
Smoking	746	(	19.6	)	799	(	26.9	)	1596	(	18.2	)	1219	(	19.6	)	521	(	12.6	)	554	(	17.5	)
Missing	393	(	10.3	)	388	(	13.1	)	920	(	10.5	)	672	(	10.8	)	348	(	8.4	)	340	(	10.7	)
total (n)	*3803*	*(*	*100.0*	*)*	*2973*	*(*	*100.0*	*)*	*8765*	*(*	*100.0*	*)*	*6205*	*(*	*100.0*	*)*	*4127*	*(*	*100.0*	*)*	*3166*	*(*	*100.0*	*)*
Drinking Status																								
non-drinking	1975	(	51.9	)	1052	(	35.4	)	4419	(	50.4	)	2355	(	38.0	)	2349	(	56.9	)	1276	(	40.3	)
ex-drinking	744	(	19.6	)	749	(	25.2	)	1968	(	22.5	)	2042	(	32.9	)	973	(	23.6	)	1060	(	33.5	)
Drinking	660	(	17.4	)	666	(	22.4	)	1462	(	16.7	)	1032	(	16.6	)	512	(	12.4	)	500	(	15.8	)
Missing	424	(	11.1	)	506	(	17.0	)	916	(	10.5	)	776	(	12.5	)	293	(	7.1	)	330	(	10.4	)
total (n)	*3803*	*(*	*100.0*	*)*	*2973*	*(*	*100.0*	*)*	*8765*	*(*	*100.0*	*)*	*6205*	*(*	*100.0*	*)*	*4127*	*(*	*100.0*	*)*	*3166*	*(*	*100.0*	*)*
Diabetes Diagnostic																								
No	3198	(	84.1	)	2469	(	83.0	)	7082	(	80.8	)	4778	(	77.0	)	3176	(	77.0	)	2393	(	75.6	)
Yes	605	(	15.9	)	504	(	17.0	)	1683	(	19.2	)	1427	(	23.0	)	951	(	23.0	)	773	(	24.4	)
Missing	0	(	0.0	)	0	(	0.0	)	0	(	0.0	)	0	(	0.0	)	0	(	0.0	)	0	(	0.0	)
total (n)	*3803*	*(*	*100.0*	*)*	*2973*	*(*	*100.0*	*)*	*8765*	*(*	*100.0*	*)*	*6205*	*(*	*100.0*	*)*	*4127*	*(*	*100.0*	*)*	*3166*	*(*	*100.0*	*)*
Height (m)																								
mean (sd)	1.6	(	0.1	)	1.7	(	0.1	)	1.6	(	0.1	)	1.7	(	0.1	)	1.6	(	0.1	)	1.7	(	0.1	)
Missing	1428	(	37.5	)	1043	(	35.1	)	3503	(	40.0	)	2291	(	36.9	)	1622	(	39.3	)	1179	(	37.2	)
total (n)	*3803*	*(*	*100.0*	*)*	*2973*	*(*	*100.0*	*)*	*8765*	*(*	*100.0*	*)*	*6205*	*(*	*100.0*	*)*	*4127*	*(*	*100.0*	*)*	*3166*	*(*	*100.0*	*)*
SBP (mmHg)																								
mean (sd)	132.9	(	15.5	)	132.4	(	14.7	)	131.8	(	14.1	)	131.7	(	13.1	)	133.0	(	15.2	)	132.3	(	14.4	)
Missing	1157	(	30.4	)	1009	(	33.9	)	2302	(	26.3	)	1556	(	25.1	)	895	(	21.7	)	769	(	24.3	)
total (n)	*3803*	*(*	*100.0*	*)*	*2973*	*(*	*100.0*	*)*	*8765*	*(*	*100.0*	*)*	*6205*	*(*	*100.0*	*)*	*4127*	*(*	*100.0*	*)*	*3166*	*(*	*100.0*	*)*
LDL-Cholesterol (mmol/L)																								
mean (sd)	3.4	(	1.2	)	3.3	(	1.1	)	3.3	(	1.1	)	3.2	(	1.1	)	3.3	(	1.1	)	3.1	(	1.1	)
Missing	1071	(	28.2	)	787	(	26.5	)	2505	(	28.6	)	1620	(	26.1	)	1146	(	27.8	)	833	(	26.3	)
total (n)	*3803*	*(*	*100.0*	*)*	*2973*	*(*	*100.0*	*)*	*8765*	*(*	*100.0*	*)*	*6205*	*(*	*100.0*	*)*	*4127*	*(*	*100.0*	*)*	*3166*	*(*	*100.0*	*)*
HDL-Cholesterol (mmol/L)																								
mean (sd)	1.6	(	0.5	)	1.3	(	0.5	)	1.5	(	0.6	)	1.3	(	0.6	)	1.5	(	0.6	)	1.3	(	0.5	)

Missing	706	(	18.6	)	492	(	16.5	)	1575	(	18.0	)	956	(	15.4	)	738	(	17.9	)	503	(	15.9	)
total (n)	*3803*	*(*	*100.0*	*)*	*2973*	*(*	*100.0*	*)*	*8765*	*(*	*100.0*	*)*	*6205*	*(*	*100.0*	*)*	*4127*	*(*	*100.0*	*)*	*3166*	*(*	*100.0*	*)*
First Dose (mg)																								
mean (sd)	6.0	(	4.5	)	7.1	(	4.8	)	76.9	(	103.6	)	81.5	(	112.7	)	1.2	(	1.0	)	1.4	(	1.3	)
Missing	1058	(	27.8	)	795	(	26.7	)	3479	(	39.7	)	2541	(	41.0	)	1448	(	35.1	)	1084	(	34.2	)
total (n)	*3803*	*(*	*100.0*	*)*	*2973*	*(*	*100.0*	*)*	*8765*	*(*	*100.0*	*)*	*6205*	*(*	*100.0*	*)*	*4127*	*(*	*100.0*	*)*	*3166*	*(*	*100.0*	*)*
Body Weight (kg)																								
mean (sd)	70.2	(	16.6	)	82.1	(	16.6	)	73.5	(	17.9	)	82.9	(	17.5	)	70.4	(	17.6	)	82.0	(	17.9	)
Missing	1699	(	44.7	)	1269	(	42.7	)	4093	(	46.7	)	2822	(	45.5	)	1938	(	47.0	)	1432	(	45.2	)
total (n)	*3803*	*(*	*100.0*	*)*	*2973*	*(*	*100.0*	*)*	*8765*	*(*	*100.0*	*)*	*6205*	*(*	*100.0*	*)*	*4127*	*(*	*100.0*	*)*	*3166*	*(*	*100.0*	*)*

Table 7.2. Low/High dose and age category in olanzapine, quetiapine and risperidone cohorts.

Age versus Dose		Low Dose		High Dose		Missing Dose		Overall	
		n	%	n	%	n	%	n	%
Olanzapine	Age								
	40-49	854	43.1%	614	31.0%	512	25.9%	1980	100.0%
	50-59	668	41.2%	497	30.7%	455	28.1%	1620	100.0%
	60-69	621	48.7%	305	23.9%	350	27.4%	1276	100.0%
	70-79	665	58.3%	169	14.8%	306	26.8%	1140	100.0%
	80-89	472	62.1%	58	7.6%	230	30.3%	760	100.0%
	Overall	3280	48.4%	1643	24.2%	1853	27.3%	6776	100.0%
Quetiapine	Age								
	40-49	1377	34.6%	948	23.8%	1652	41.5%	3977	100.0%
	50-59	1057	37.2%	637	22.4%	1144	40.3%	2838	100.0%
	60-69	924	46.9%	303	15.4%	742	37.7%	1969	100.0%
	70-79	1519	56.1%	137	5.1%	1053	38.9%	2709	100.0%
	80-89	1979	56.9%	69	2.0%	1429	41.1%	3477	100.0%
	Overall	6856	45.8%	2094	14.0%	6020	40.2%	14970	100.0%
Risperidone	Age								
	40-49	749	55.4%	212	15.7%	392	29.0%	1353	100.0%
	50-59	572	53.7%	144	13.5%	349	32.8%	1065	100.0%
	60-69	625	61.6%	71	7.0%	318	31.4%	1014	100.0%
	70-79	989	61.3%	27	1.7%	598	37.1%	1614	100.0%
	80-89	1364	60.7%	8	0.4%	875	38.9%	2247	100.0%
	Overall	4299	58.9%	462	6.3%	2532	34.7%	7293	100.0%

7.4.2 Differences by sex

Given a treatment of olanzapine, short-term weight gain was less pronounced in women (model type (a), see Table 7.3). This is seen in the interaction term between sexes estimated as $\beta_{women*t_2}$=-0.1440 kg/week (95%CI:-0.2510 to -0.0371). On the other hand, there was no difference between men and women for short term weight gain for quetiapine ($\beta_{women*t_2}$=-0.0187 kg/week; 95%CI:-0.0553 to 0.0926) or risperidone ($\beta_{women*t_2}$=-0.0301 kg/week; 95%CI:-0.1355 to 0.0753). There were minor differences between sexes on the long-term weight gain for olanzapine ($\beta_{women*t_3}$=0.0056 kg/week; 95%CI: 0.0003 to 0.0108) and quetiapine ($\beta_{women*t_3}$= 0.0040 kg/week; 95%CI: -0.0002 to 0.0082). Estimates of weight trajectories within women and men are provided separately in Table S3 (Appendix 7A).

Table 7.3. ITS models with interaction between sex and time for the olanzapine, quetiapine and risperidone cohorts.

Weight Trajectories by SEX	Olanzapine (women=3,803; men=2,973; combined=6,776)					Quetiapine (women=8,765; men=6,205; combined=14,970)					Risperidone (women=4,127; men=3,166; combined=7,293)				
	Beta	95% CI			p	Beta	95% CI			p	Beta	95% CI			p
t1 (-4 years to 0 weeks)	-0.0120	(-0.0153	to	-0.0088)	<0.001	-0.0098	(-0.0120	to	-0.0077)	<0.001	-0.0096	(-0.0124	to	-0.0067)	<0.001
t2 (0 weeks to 6 weeks)	0.4455	(0.3646	to	0.5264)	<0.001	0.1014	(0.0441	to	0.1588)	0.001	0.0974	(0.0170	to	0.1779)	0.018
t3 (6 weeks to 4 years)	0.0067	(0.0027	to	0.0107)	0.001	-0.0002	(-0.0035	to	0.0032)	0.917	0.0024	(-0.0023	to	0.0071)	0.315
Women	-12.08	(-13.02	to	-11.14)	<0.001	-12.97	(-13.62	to	-12.32)	<0.001	-11.71	(-12.60	to	-10.82)	<0.001
women*t1	-0.0029	(-0.0071	to	0.0014)	0.187	0.0013	(-0.0015	to	0.0041)	0.372	-0.0036	(-0.0073	to	0.0002)	0.062
women*t2	-0.1440	(-0.2510	to	-0.0371)	0.008	0.0187	(-0.0553	to	0.0926)	0.621	-0.0301	(-0.1355	to	0.0753)	0.575
women*t3	0.0056	(0.0003	to	0.0108)	0.039	0.0040	(-0.0002	to	0.0082)	0.064	-0.0012	(-0.0073	to	0.0049)	0.696
Intercept	82.03	(73.35	to	90.71)	<0.001	88.76	(82.73	to	94.78)	<0.001	86.59	(77.98	to	95.19)	<0.001

ITS=interrupted time series. All estimates are adjusted for age and deprivation. Intercept was set at the very beginning of the observation period. Units: time in weeks and weight in kg (i.e. t1, t2 an t3 are in kg/weeks).

7.4.3 Differences by age

Given a treatment of olanzapine, short-term weight gain was much less pronounced for older people than younger people (model type (b), see Table 7.4). Thus, the modifying effect of age was confirmed even after adjusting for sex, social deprivation and dose. In the short term, the weight gain was highest in younger people (40-49 years: β_{t2}=0.6294 kg/week; 95%CI: 0.5120 to 0.7469) and diminished progressively for people older at treatment initiation (50-59 years: β_{t2}=0.4685 kg/week; 60-69 years: β_{t2}= 0.2467 kg/week; 70-79 years: β_{t2}= 0.0865 kg/week; 80-89 years: β_{t2}= -0.0416 kg/week) (Figure 7.2 and Appendix 7B). All the interaction terms between age and short-term time (age*t2) confirmed this trend (Appendix 7B).

Similar short-term trends were observed for quetiapine and risperidone cohorts (Appendices 7C and 7D respectively).

For the long-term, the modifying effect of age on weight gain was weak for olanzapine (Table 7.4). A difference was visible: the long-term rate of weight gain was higher for younger people of 40-49 years (β_{t2}= 0.0108 kg/week; 95%CI:0.0056 to 0.0161), 50-59 years (β_{t2}=0.0135 kg/week; 95%CI: 0.0080 to 0.0189) and 60-69 years (β_{t2}=0.0102 kg/week; 95%CI: 0.0027 to 0.0176) when were compared against older people of 70-79 years (β_{t2}=0.0006 kg/week; 95%CI: -0.0070 to 0.0082) and 80-89 years (β_{t2}=-0.0021 kg/week; 95%CI: -0.0106 to 0.0063) (Figure 7.2 and Appendix 7B).

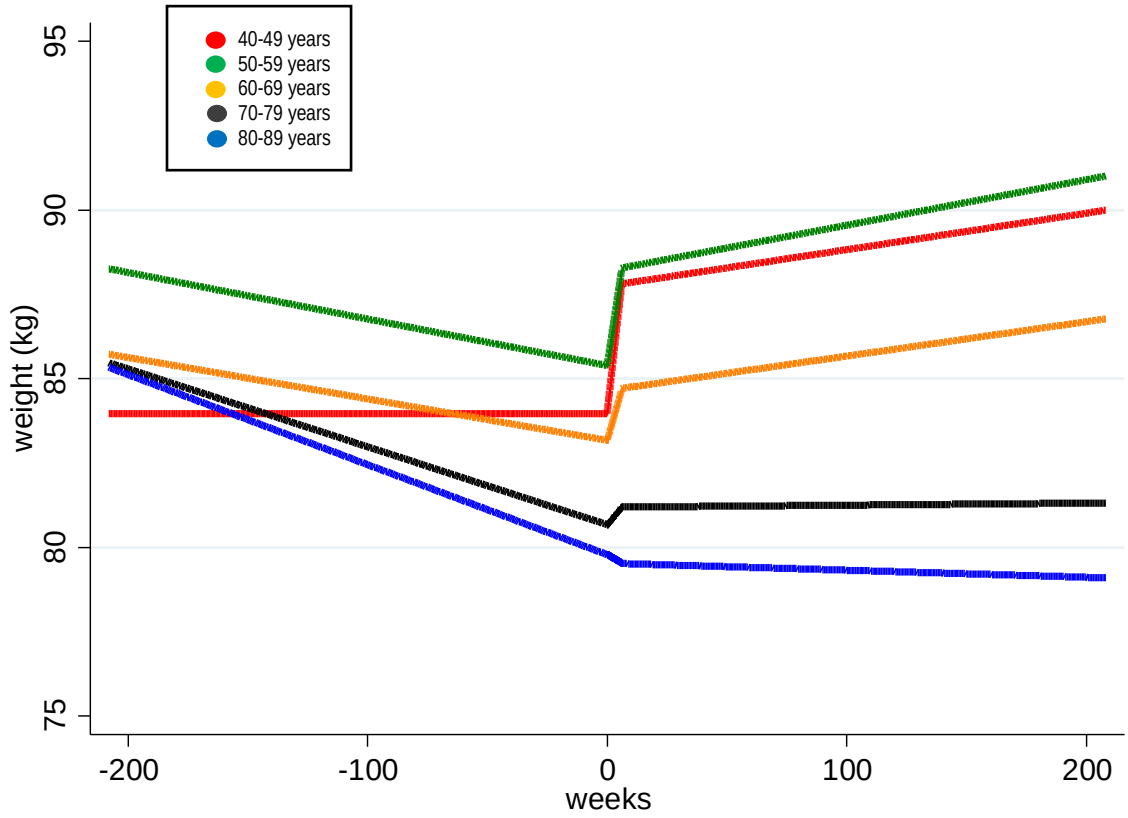
A more explicit modifying effect in the long-term was verified for users of quetiapine and risperidone (Appendices 7C and 7D, respectively).

Table 7.4. ITS models with interaction between age and time for the olanzapine cohort.

Weight Trajectories by AGE (Complete Case Analysis)		Olanzapine											
		Unadjusted (combined=6,776) (40-49 years=1,980; 50-59 years=1,620; 60-69 years=1,276; 70-79 years=1,140; 80-89 years=760)						Adjusted (combined=4,923) (40-49 years=1,468; 50-59 years=1,165; 60-69 years=926; 70-79 years=834; 80-89 years=530)					
		Beta	95% CI				p	Beta	95% CI				p
t1 (-4 years to 0 weeks)		-0.0017	(-0.0059	to	0.0025	)	0.424	-0.0002	(-0.0051	to	0.0046	)	0.930
t2 (0 weeks to 6 weeks)		0.6399	(0.5405	to	0.7394	)	0.000	0.6478	(0.5346	to	0.7610	)	0.000
t3 (6 weeks to 4 years)		0.0096	(0.0050	to	0.0143	)	0.000	0.0101	(0.0048	to	0.0154	)	0.000
age			ref						ref				
	50-59	-0.1296	(-1.4932	to	1.2339	)	0.852	0.6778	(-0.8439	to	2.1996	)	0.383
	60-69	-2.8110	(-4.2496	to	-1.3723	)	0.000	-0.9582	(-2.5668	to	0.6503	)	0.243
	70-79	-4.1645	(-5.6312	to	-2.6978	)	0.000	-2.1086	(-3.7563	to	-0.4609	)	0.012
	80-89	-8.9634	(-10.6556	to	-7.2711	)	0.000	-6.0138	(-7.9581	to	-4.0695	)	0.000
age*t1			ref						ref				
	50-59	-0.0089	(-0.0150	to	-0.0029	)	0.004	-0.0133	(-0.0204	to	-0.0063	)	0.000
	60-69	-0.0102	(-0.0164	to	-0.0040	)	0.001	-0.0120	(-0.0192	to	-0.0048	)	0.001
	70-79	-0.0241	(-0.0303	to	-0.0178	)	0.000	-0.0230	(-0.0302	to	-0.0158	)	0.000
	80-89	-0.0238	(-0.0312	to	-0.0164	)	0.000	-0.0261	(-0.0347	to	-0.0174	)	0.000
age*t2			ref						ref				
	50-59	-0.1843	(-0.3292	to	-0.0394	)	0.013	-0.1789	(-0.3443	to	-0.0135	)	0.034
	60-69	-0.4248	(-0.5767	to	-0.2729	)	0.000	-0.4153	(-0.5879	to	-0.2427	)	0.000
	70-79	-0.5284	(-0.6877	to	-0.3692	)	0.000	-0.5617	(-0.7440	to	-0.3794	)	0.000
	80-89	-0.7440	(-0.9483	to	-0.5397	)	0.000	-0.7545	(-0.9852	to	-0.5238	)	0.000
age*t3			ref						ref				
	50-59	0.0027	(-0.0042	to	0.0095	)	0.446	0.0036	(-0.0042	to	0.0114	)	0.367
	60-69	0.0017	(-0.0056	to	0.0091	)	0.643	-0.0004	(-0.0088	to	0.0081	)	0.934
	70-79	-0.0047	(-0.0129	to	0.0035	)	0.262	-0.0099	(-0.0192	to	-0.0006	)	0.038
	80-89	-0.0122	(-0.0237	to	-0.0008	)	0.037	-0.0113	(-0.0242	to	0.0016	)	0.086
intercept		77.8	(76.9	to	78.7	)	0.000	84.09	(82.51	to	85.67	)	0.000

ref=reference group. ITS=interrupted time series. All estimates come from mixed-effects models (restricted maximum likelihood and unstructured covariance matrix), adjusted for sex, social deprivation (Townsend) and first dose when indicated. The intercept was set at the very beginning of the observation period. Units: time in weeks and weight in kg (i.e. t1, t2 and t3 are in kg/weeks).

Figure 7.2. Pre-treatment, short and long-term weight trajectory by age group for the olanzapine cohort.



7.4.4 Differences by dose

Given a treatment of olanzapine, short-term weight gain was more pronounced for people prescribed a high dose (model type (c), see Table 7.5). Thus, the modifying effect of dose was confirmed even after adjusting for sex, age and social deprivation. In the short term, the weight gained by those prescribed low dose was 0.3113 kg/week (95%CI: 0.2433 to 0.3793); whereas the weight gained by patients prescribed high dose was 0.4517 kg/week (95%CI: 0.3548 to 0.5487) (Appendix 7E). Hence, over six weeks, those on low dose gained 1.9kg, whereas those on high dose gained 2.7kg.

Similar short-term trends were observed for quetiapine but not for risperidone cohorts (Appendices 7F and 7G respectively).

For the long-term, the modifying effect of dose was confirmed in the quetiapine group (Appendix 7F). For low dose, the average weight change was minimal (0.0006 kg/week; 95%CI: -0.0020 to 0.0032), equivalent to 0.1kg over 202 weeks (from week 6 to 4 years). For high dose, the weight change was 0.0061 kg/week (95%CI: 0.0020 to 0.0102), equivalent to 1.2kg over 202 weeks (Figure 7.3 and Appendix 7F).

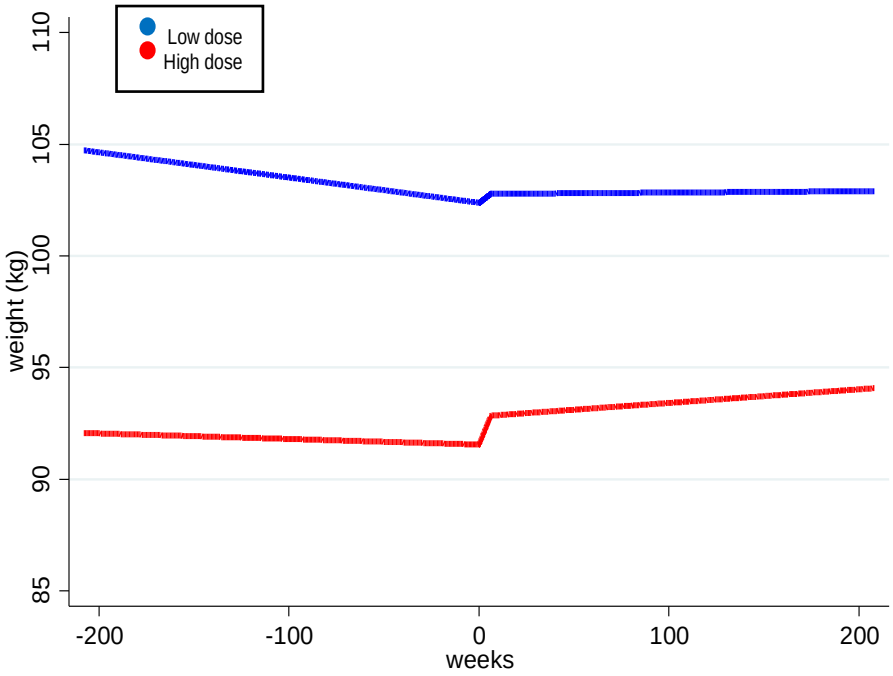
A lack of long-term modifying effect of dose was observed for olanzapine (Table 7.5) and risperidone (Appendix 7G).

Table 7.5. ITS models with interaction between dose and time for the olanzapine cohort.

Weight Trajectories by DOSE	Olanzapine																	
	Unadjusted (total=6,776) (low-dose=3,280; high-dose=1,643; missing dose=1,853; combined=4,923)						Adjusted (CCA) (total=6,776) (low-dose=3,280; high-dose=1,643; missing dose=1,853; combined=4,923)						Adjusted (MI-JOMO) (total=6,776) (low-dose=4,403; high-dose=2,323; combined=6,776)					
	Beta	95% CI				p	Beta	95% CI				p	Beta	95% CI				p
t1 (-4 years to 0 weeks)	-0.0133	(	-0.0163	to	-0.0104	) 0.000	-0.0138	(	-0.0167	to	-0.0108	) 0.000	-0.0144	(	-0.0171	to	-0.0117	) 0.000
t2 (0 weeks to 6 weeks)	0.2896	(	0.2152	to	0.3639	) 0.000	0.2914	(	0.2172	to	0.3656	) 0.000	0.3100	(	0.2415	to	0.3786	) 0.000
t3 (6 weeks to 4 years)	0.0095	(	0.0057	to	0.0132	) 0.000	0.0093	(	0.0055	to	0.0130	) 0.000	0.0097	(	0.0062	to	0.0131	) 0.000
high-dose	4.0634	(	2.8449	to	5.2820	) 0.000	0.8732	(	-0.3106	to	2.0570	) 0.148	0.8529	(	-0.6017	to	2.3075	) 0.246
high-dose*t1	0.0014	(	-0.0039	to	0.0066	) 0.615	0.0012	(	-0.0041	to	0.0065	) 0.650	0.0018	(	-0.0033	to	0.0070	) 0.481
high-dose*t2	0.2182	(	0.0900	to	0.3463	) 0.001	0.2117	(	0.0837	to	0.3396	) 0.001	0.1496	(	0.0250	to	0.2742	) 0.019
high-dose*t3	0.0006	(	-0.0056	to	0.0068	) 0.858	0.0007	(	-0.0055	to	0.0069	) 0.834	0.0005	(	-0.0056	to	0.0065	) 0.883
intercept	74.3	(	73.6	to	75.0	) 0.000	80.39	(	70.28	to	90.50	) 0.000	96.47	(	86.31	*to*	106.62	) 0.000

ITS=interrupted time series. All estimates come from mixed-effects models (restricted maximum likelihood and unstructured covariance matrix), and were adjusted for sex, social deprivation (Townsend) and age when indicated. Missing dose was handled with complete case analysis (CCA) and multilevel multiple imputation (MI-JOMO). Results from both methods are visible in the table. The intercept was set at the very beginning of the observation period. Units: time in weeks and weight in kg (i.e. t1, t2 and t3 are in kg/weeks).

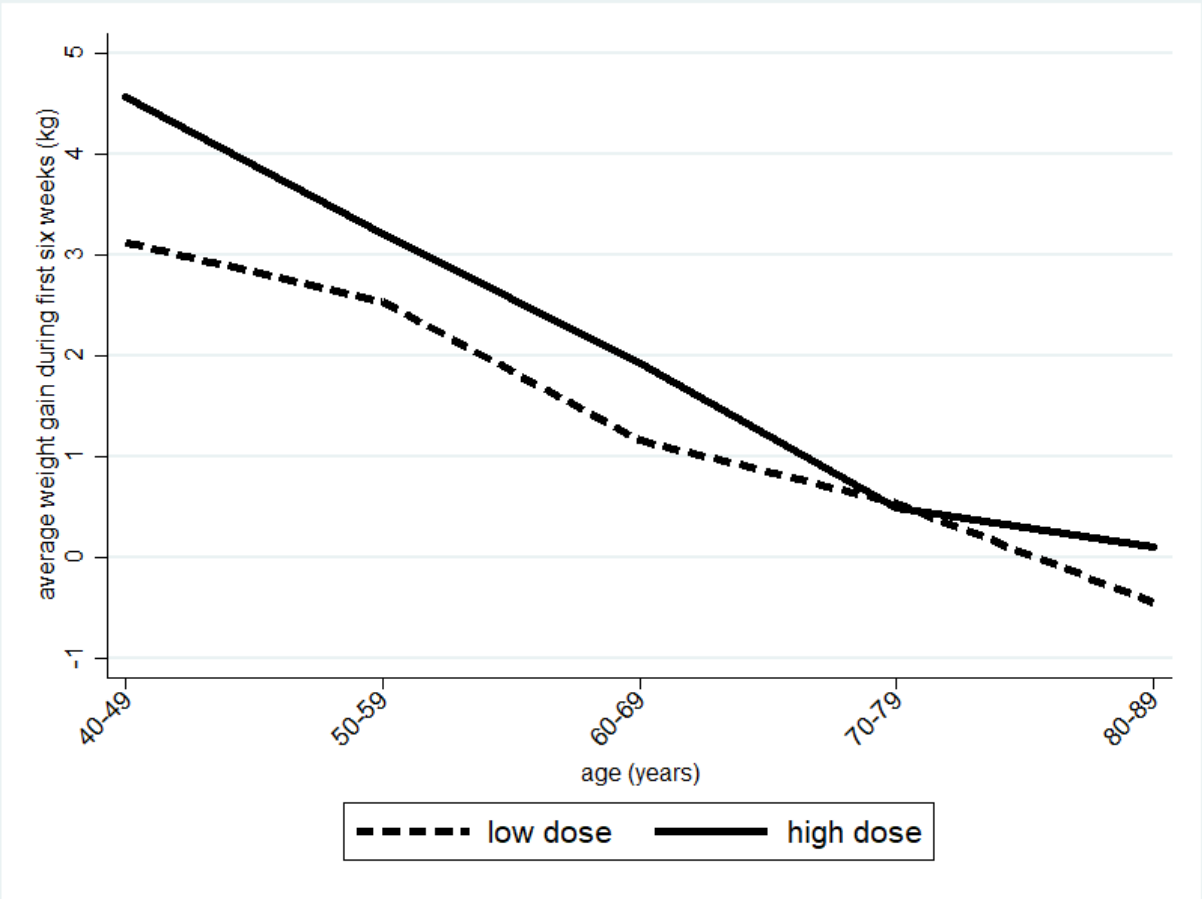
Figure 7.3. Pre-treatment, short and long-term weight trajectory by low/high dose for the quetiapine cohort.



7.4.5 Differences by age and dose

Figure 7.4 (drawn by using estimates from model type (d)) shows that short-term weight gain is almost zero or even negative in the oldest group (80-89 years) prescribed olanzapine, which is very different from the weight gain experienced by younger people (40-49 years). Likewise, differences in short-term weight gain between low/high doses are more evident in younger people.

Figure 7.4. Short-term (6 weeks) weight gain (kg) across subgroups by age and dose for the olanzapine cohort (Note: for this figure lines do not represent individual follow up).



7.4.6 ITS effect sizes

Figure 7.5 (drawn by using estimates from Appendix 7C) helps to illustrate how the expected or ′natural′ weight gain (β_1) can be contrasted against the observed weight gain after treatment initiation (β_3) to estimate the effect size in the long-term (attribute of the ITS design exposed in Figure 7.2). Figure 7.5 shows that, on average, -0.0231 kg/week of weight change would be expected in people aged 80-89 years if no quetiapine treatment would have been prescribed (dashed line). The observed weight change after treatment initiation was -0.0157 kg/week; thus, the long-term effect size of quetiapine is -0.0157 +0.0231=0.0074 kg/weeks.

Figure 7.5. Visualization of the model estimates and the long-term ITS effect size for persons aged 80-89 years prescribed quetiapine

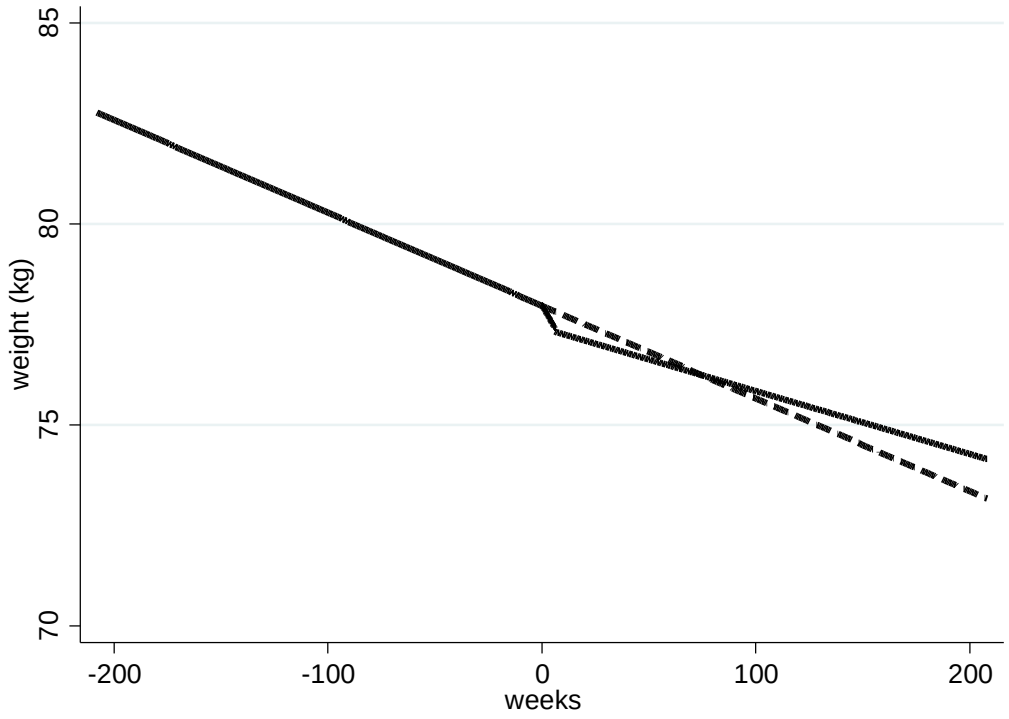


Table 7.6 shows the ITS effect size of the same second-generation antipsychotics (olanzapine, quetiapine, risperidone) on weight change in the long-term (β_3-β_1) modified by age. The table in Appendix 7I shows similar ITS effect sizes for subgroups by age and dose (for olanzapine only).

Table 7.6. ITS effect size of second-generation antipsychotics on weight change in the LONG-TERM, modified by age.

Antipsychotic	Age (years)	*pre-treatment (β1)*	*long-term (β3)*	*ITS effect size (β3-β1)*	*modifying effect*										
					*age*t1 (pre-treatment)*						*age*t3 (long-term)*				
					B	*95% CI*			*p*		**β**	*95% CI*			*p*
Olanzapine	40-49	0.0000	0.0108	0.0109	ref						ref				
	50-59	-0.0138	0.0135	0.0272	-0.0133	(-0.0204	to	-0.0063)	0.000		0.0036	(-0.0042	to	0.0114)	0.367
	60-69	-0.0123	0.0102	0.0225	-0.0120	(-0.0192	to	-0.0048)	0.001		-0.0004	(-0.0088	to	0.0081)	0.934
	70-79	-0.0231	0.0006	0.0237	-0.0230	(-0.0302	to	-0.0158)	0.000		-0.0099	(-0.0192	to	-0.0006)	0.038
	80-89	-0.0267	-0.0021	0.0245	-0.0261	(-0.0347	to	-0.0174)	0.000		-0.0113	(-0.0242	to	0.0016)	0.086
Quetiapine	40-49	0.0096	0.0025	-0.0071	ref						ref				
	50-59	-0.0022	0.0049	0.0071	-0.0115	(-0.0170	to	-0.0060)	0.000		0.0027	(-0.0044	to	0.0098)	0.464
	60-69	-0.0089	0.0005	0.0095	-0.0183	(-0.0243	to	-0.0124)	0.000		-0.0012	(-0.0090	to	0.0066)	0.755
	70-79	-0.0178	-0.0078	0.0100	-0.0268	(-0.0322	to	-0.0213)	0.000		-0.0085	(-0.0169	to	0.0000)	0.050
	80-89	-0.0231	-0.0157	0.0074	-0.0322	(-0.0375	to	-0.0269)	0.000		-0.0190	(-0.0281	to	-0.0099)	0.000
Risperidone	40-49	0.0094	0.0097	0.0003	ref						ref				
	50-59	-0.0051	0.0056	0.0107	-0.0140	(-0.0222	to	-0.0059)	0.001		-0.0038	(-0.0140	to	0.0064)	0.466
	60-69	-0.0085	0.0010	0.0094	-0.0173	(-0.0253	to	-0.0094)	0.000		-0.0068	(-0.0173	to	0.0036)	0.201
	70-79	-0.0177	-0.0032	0.0144	-0.0262	(-0.0334	to	-0.0190)	0.000		-0.0120	(-0.0231	to	-0.0009)	0.033
	80-89	-0.0212	-0.0259	-0.0047	-0.0300	(-0.0370	to	-0.0231)	0.000		-0.0361	(-0.0482	to	-0.0240)	0.000

*ref=reference group. ITS = interrupted time series. In this approach, the **long-term effect size** can be estimated by the difference between pre-treatment trajectory (beta 1) and long-term post-treatment trajectory (beta 3). Units: time in weeks and weight in kg (i.e. betas are in kg/weeks).*

7.4.7 Missing data

The consequences of missing data were different, depending on the analysis performed. For analyses with fully observed covariates, for example, in Table 7.5 (model type (c)), it was visible how the estimates of weight trajectories (t1, t2, t3) and their interaction with dose (dose*t1, dose*t2, dose*t3) changed whether the model was unadjusted or adjusted (both with CCA). Part of this difference could be due to the fact that the missing weights are more plausibly MAR when we condition/adjust for the additional covariates (i.e. sex, social deprivation and age). For analyses with partially observed covariates (i.e. dose), analysis after imputation with MI-JOMO gave different estimates to CCA (both with MEM) and narrowed the 95% confidence intervals (or produced smaller standard errors). Here the difference can be due to the inclusion of the imputed covariate values and an intrinsic limitation of multiple imputation I discuss in the next chapter.

7.4.8 Summary of key results

A summary of the results on modifying effects analysis reported in tables 7.3, 7.4 and 7.5 (model types (a), (b) and (c)) is available in Table 7.7. This summary is useful for starting the next discussion on key study findings. For clinical purposes, the Table in Appendix 7H summarises the cumulative weight change (in kg) for short (6 weeks) and long-term periods (from 6 weeks to 4 years).

Table 7.7. Summary table of the modification effects of sex, age and dose on antipsychotic induced weight change

Modifying effect of	Antipsychotic	*Does it modify the antipsychotic-induced weight change?*		*Type of modifying effect*	
		short-term	long-term	short-term	long-term
Sex	olanzapine	yes	yes, but weakly	weight gain is more severe in men	weight gain is more severe in women
	quetiapine	no	yes, but weakly	no effect	weight gain is more severe in women
	risperidone	no	no	no effect	no effect
Age	olanzapine	yes	yes, but weakly	weight gain is more severe in younger people	weight gain is more severe in younger people
	quetiapine	yes	yes	weight gain is more severe in younger people	weight gain is more severe in younger people
	risperidone	yes	yes	weight gain is more severe in younger people	weight gain is more severe in younger people
Dose	olanzapine	yes	no	weight gain is more severe when a high dose is prescribed	no effect
	quetiapine	yes	yes	weight gain is more severe when a high dose is prescribed	weight gain is more severe when a high dose is prescribed
	risperidone	no	no	no effect	no effect

7.5 Discussion

Sex, age and dose modify the weight change associated with specific second-generation antipsychotics in patients ≥40 years (see a summary in Table 7.7). In general, women gain weight with a higher rate than men in the long-term, but with a lower rate in the short-term. In the short and long-term, a higher dose of any antipsychotic is mostly associated with a higher rate of weight gain. Antipsychotic-induced weight gain has a lower rate in older people, in the short a long-term. In the short term, differences in weight gain between low/high dose are more evident in younger people. For any age, the long-term ITS effect size usually reflects a consistent weigh gain that would not occur in the absence of an antipsychotic prescription. Interaction terms in MEM facilitated the evaluation of differences in weight trajectories over time between groups by age, sex and dose. MEM was combined with MI-JOMO for handling missing values of dose, but with some limitations discussed close to the end of this section.

Sex and dose have been studied before as modifiers of antipsychotic-induced weight gain, but not via a formal evaluation of the interaction effect in ITS models. It has been reported before that men tend to gain more weight than women, in the short-term, when prescribed olanzapine or risperidone [91]. I found in a previous study (Chapter 5) [100] an apparent difference between men and women for olanzapine, which was less evident for risperidone. Both differences indicated more weight gain for men. In this study, the evaluation of the interaction term allowed to formally compare weight change trajectories of women and men ≥40 years, for which olanzapine is the only drug that produces a weight gain that is modified by sex. In the long term, the rate of weight gain after treatment initiation was minimal [100]; thus, the comparison between sexes is not straightforward. Here, I identified a weak interaction effect suggesting that women prescribed olanzapine (on average, 2.5kg for women versus 1.3kg for men) and quetiapine (on average, 0.8kg for women versus -0.1kg for men) may gain more weight than men in the long-term. Regarding dose, a higher short-term weight gain for patients consuming high doses of olanzapine has been found elsewhere [93,100] and confirmed here, using a formal interaction evaluation. This exploits an advantage of using ITS models on electronic health records for evaluating modifying effects, namely that they have a markedly higher number of weight records over long-term periods (e.g. 4 years), relative to typical randomised clinical trials.

The modifying effect of age in the antipsychotic weight trajectories is evident in the short-term, but in the long-term should be analysed considering the ITS effect size (i.e. pre versus post-treatment initiation weigh trajectory). An early study with olanzapine and risperidone reported that younger patients gained more weight in the short-term [110]. More recently, Yeung et al. [113] reported weight gain in older people treated with olanzapine, but a weight loss in older people treated risperidone, both in short- to medium-term. In the present study, I found an effect modification of age in the short-term. For example, patients aged between 60-69 years gain 1.5kg on average during the first 6 weeks of treatment, but people aged between 70-89 gain almost no weight during the same period with olanzapine treatment. I also found that long-term weight gain is less pronounced in older people compared to younger people. Nevertheless, to better understand the real long-term impact of these antipsychotics, the natural age-related loss of weight should also be considered. In ITS designs, the pre-treatment trajectory provides an approximation of the natural weight loss [21]; thus, the difference between this trajectory and the post-treatment long-term trajectory is informative as a measure of the effect size (see Figures 7.2 and 7.5) [6]. As is visible in Table 7.6, the effect size is positive (weight gain) and similar across many age groups, and only quetiapine for younger people and risperidone for older people led to weight loss. The result of this analysis suggests the antipsychotic treatment may prevent weight loss for elderly patients. In older people, natural weight loss can be a risk factor; thus, some doctors have suggested prescribing small doses of olanzapine in short periods to avoid this weight loss in patients with severe mental illness [109]. The results from the present study give an approximation of the potential impact of this recommendation [22]. The results in this chapter suggest that, even with low doses of olanzapine, the natural weight loss in older people can be controlled for long-term periods (Appendix 7I).

[21] In the long-term, some minor difference could be expected between the pre-treatment and the natural trajectory that is inferred for the treatment period (as an extension of the pre-treatment trajectory). For example, pre-treatment weight trajectory of a woman from her 76 years to her 80 years is not exactly the same than the natural weight trajectory she would have had between her 80 years to her 84 years (counterfactual). However, using as reference the pre-treatment trajectories of the next more aged sub-group (see Table 7.6), it is easy to approximate that the (potential) difference would slightly increase the effect size estimated.

[22] This is a controversial topic among clinicians. In the US, antipsychotic medication in older people has a black box warning due to increased risks of death [141]. In the UK, antipsychotics in dementia should also be used with caution [142], although many people are prescribed them.

MEM models are useful for handling irregularly observed outcomes (which may be viewed as missing as they do not follow any regular measurement schedule) for evaluating ITS effects. In the evaluation of dose, it is notable how the estimated weight trajectories and the interaction between dose and time — both using a complete records analysis — changed when the MEM model was adjusted for covariates (with CCA). There are two explanations for the observed difference: (i) that covariate adjustment (even if there are no missing outcomes) could change effect estimates (this could even happen if it were a randomised study) and (ii) that covariate adjustment could change estimates because it makes more plausible the assumption that outcomes are MAR given covariates in the model (i.e. covariates inform the outcome implicit imputation). Both are likely to be happening to some extent, and it is not easily possible to say which is the greater effect. Although in ITS the time-invariant covariates (e.g. sex) are expected to be controlled by design [2], in observational studies like this one, these covariates also need to be included in the estimation of ITS effects with MEM. I have previously demonstrated that MEM with fully observed covariates makes bias correction to CCA when the probability of being missing depends on observed values of covariates (conclusion from the previous simulation study, Chapter 6).

MI-JOMO is useful for handling missing covariates in models with interaction terms, but the approach has a minor limitation. For models with dose missing, MI-JOMO reported different estimates than MEM and narrowed the 95% confidence intervals (or produced smaller standard errors). Interestingly, the estimates of long-term weight change varied just slightly across unadjusted and adjusted models (with CCA or MI-JOMO). As the simulation study showed in Chapter 6, we expect that MI-JOMO will provide less biased and more efficient estimates when adjusting for covariates with missing values. However, for evaluating interactions with imputed categorical covariates (i.e. dose: low/high), each imputation from MI-JOMO can produce a different distribution of new values per each dataset (i.e. the number of patients in each dose group varies across multiply imputed datasets). This difference can bias the variance estimates from Rubin′s formula (Section 3.2.2). This limitation is minor in the sense that it is the variance and not the point estimate that can be biased. Indeed, the limitation is inherent to any multiple imputation method applied for evaluating this type of interaction.

Some strength and limitations are inherent to this study. It is the first time that the modifying effect of age on the antipsychotic-induced weight change has been studied, which contributes to the evidence base informing prescription guidance, particularly for older patients. The present study improves on previous shorter-term studies [110,113], by using larger sample size and including many more weight records over time to draw weight trajectories. However, in contrast to randomised controlled trials, I was unable to perform formal comparisons among antipsychotics. The comparisons would be valid if age, sex and deprivation were the only variables affecting treatment decisions [23], but it is likely that a number of additional confounders remain unadjusted for. Nevertheless, since the relative weight gain for the different treatments in the short term is similar to randomised trials', there is a reasonable justification for saying that longer-term differences between treatments, estimated from the ITS analysis, have a strong causal basis. Long-term trajectories of weight change could be affected by other factors such as competing risk [24], treatment duration [91] or other medications [114] not included in this study. However, these factors are most likely to have a greater impact on long-term estimates and less impact on the short-term effects. Not everyone had treatment during all the four years of follow up. Nevertheless, my goal was to show doctors the long-term consequences of the pragmatic treatment policy 'starting treatment with second-generation antipsychotics', whether patients stop or continue this treatment during the next four years. Finally, potential deviations from the MAR assumption of weight are possible; for example, that GPs are more likely to monitor and record weight more frequently if it there is a visible (or patient-reported) change (i.e. MNAR mechanism). MNAR evaluation and handling goes beyond the scope of this thesis; thus, this remains an open question for future research (I will return to this point in Section 8.5).

The study findings have implications for future clinical research and practice. From previous studies, we know that second-generation antipsychotics (SGA) are associated with mortality in older people [115], and randomised clinical trials show them to be associated with cerebrovascular events in dementia [116]. Generally, it is considered that

[23] For example, olanzapine prescription is usually avoided if the patient is overweight. Instead, these patients are commonly prescribed low doses of quetiapine or risperidone. This selection criterion adds heterogeneity between drug cohorts (i.e. outcome at baseline and unmeasured confounders), making a formal comparison hard to achieve. Therefore, all my conclusions about interaction effects are conditioned to a "given prescription" for people as described in each cohort.

[24] Especially mortality (i.e. patient must survive to have a weight). I will come back to the attrition problem latter in Section 8.4.

part of the risk of SGA is on cardiovascular risk via weight gain. However, this cannot be the case in older people. There must be other direct pathways to cardiovascular risk increases which do not act via weight gain, which is a relevant topic for future research. For future clinical practice, dietary recommendations that are valid for people prescribe SGA at any age must be different for older people. New versions of dietary guidelines for older patients can be initially informed by the observed weight trajectories I have shown here.

In conclusion, I found that sex, age and dose do modify the weight change induced by specific second-generation antipsychotics in patients aged ≥40 years. In general, weight gain is more severe for women than men in the long-term but less severe in the short-term. In the short and long-term, a higher dose is mostly associated with a higher rate of weight gain. Antipsychotic-induced weight gain is less severe in older people, in the short a long-term. In particular, the effect modification due to age has important clinical implications for informing prescriptions in older patients. MEM combined with substantive-compatible model MI-JOMO has been useful for handling missing covariates in models with interaction terms.

7.6 Summary

Second-generation antipsychotics increase the weight of patients in the first weeks of treatment, but their long-term effects on patient´s weight have not been thoroughly studied. There is also a gap in the literature on antipsychotic-induced weight gain in older people, alone or in comparison with younger people. A better understanding of how sex and dose can modify the antipsychotic-induced weight gain is needed as well. Interrupted Time Series (ITS) design applied to routinely collected data is a powerful approach for answering these questions if individual-level missing data (i.e. the irregular weight measurement schedule and missing covariates) are appropriately handled.

The study had two objectives: 1) (clinical) to evaluate the weight trajectories four years before and four years after treatment initiation of antipsychotic treatment, and how age, sex and dose independently modify these trajectories; 2) (methodological) to show how mixed-effect modelling (MEM), in conjunction with multilevel multiple imputation (MI-JOMO), can be used to handle missing data when these trajectories and the mentioned modifying effects are modelled.

I observed a total of 29,039 patients during the 8 years of follow-up, in the cohorts of olanzapine (n=6,776), quetiapine (n=14,970) and risperidone (n=7,293). Patients were mainly women (>55%) aged between 40 and 89 years. In the short term (6 weeks), older patients' weight was less affected by the first antipsychotic prescription compared to younger people. Given a specific antipsychotic prescription, the weight of women and men are affected differently in the short and long-term. High doses are also associated with more weight gain in people >40 years. Moreover, there is strong evidence of how antipsychotic-induced weight gain is different between younger and older persons. These findings need to be considered in future guidelines for prescribing second-generation antipsychotics in clinical practice. From the methodological side, interaction terms in MEM facilitated the evaluation of differences in weight trajectories over time between groups by age, sex or dose; handling missing data efficiently when combined with MI-JOMO.

In conclusion, MEM alone or combined with MI-JOMO are useful tools for handling missing data in ITS studies evaluating modifying effects with individual-level data. Clinically, age, sex and dose modify the weight change induced by specific second-generation antipsychotics in patients aged 40-89 years. My findings about differences by age in the antipsychotic-induced weight gain will inform treatment decision making and support dietary recommendations, especially for the elderly population.

8

Discussion

8.1. Summary of thesis

The motivation for this PhD stems from the need to address specific clinical problems, such as the impact of initiation of antipsychotic medication on short and long-term changes in body weight. While it has been demonstrated in randomised clinical trials that initiation of antipsychotic medication can substantially increase body weight over a short period, far less is known about the (arguably more important) long-term effects of antipsychotic treatment initiation. I chose to examine this question by using electronic health records from UK primary care. The records take the form of longitudinal data from clinical care. After some consideration, I chose to use the interrupted time series (ITS) design, which is a quasi-experimental design evaluating the effect of an intervention or treatment by comparing the outcome trajectory over time before and after initiation of the intervention [6]. Initially, ITS were used for evaluating interventions at the population level (e.g. policies); thus, the development of

statistical methods was mainly orientated to model population-level data [2]. However, extensive routinely collected data such as electronic health records allow applications of the ITS modelling of individual-level data, with all the advantages this brings. Furthermore, the modelling of individual-level data provides an opportunity to address the issue of missing data which traditional ITS methods are not designed to address.

In Chapter 4, I showed that the issues of missing data have rarely been addressed in most recent ITS studies with individual-level data. Despite its limitations, complete case analysis (CCA) is the most used method for handling missing data. Furthermore, I found that individual-level data are usually transformed into population-level data before fitting ITS models; for example, averaging the outcome at each time point ('averaging-step') before fitting a segmented regression ('aggregate-level' SR). I also confirmed that, to date, very few studies had applied mixed-effect modelling (MEM) for individual-level ITS data. Multilevel multiple imputation (i.e. MI-JOMO) was developed in 2015 and has not been used to impute missing data in this context.

I applied MEM to study antipsychotic-induced weight gain using the ITS design on electronic health records (Chapter 5). MEM allowed fully observed covariates and partially observed outcomes to inform the implicit imputation of missing outcomes at the individual level. This assumes that missing outcome data are missing at random (MAR). However, I found that relying on MEM is not ideal when covariates are missing (e.g. dose of medication). Therefore, I evaluated MEM combined with MI-JOMO for handling missing covariates, in the next study. From the clinical point of view, this ITS study facilitated new clinical evidence in the long-term. Thus, I could demonstrate that typical patients do not lose the weight they gained during the first six weeks of antipsychotic treatment.

Thereafter, I evaluated - via simulation studies - the performance of different types of methods for handling missing outcome or covariate data in ITS, assuming the data are MAR (Chapter 6). Specifically, I explored aggregate-level SR, MI-JOMO and MEM. I showed that the averaging-step biases ITS estimates when data were MAR at the individual level. The aggregate-level SR can over- or underestimate the ITS effect, depending on how the missingness mechanism is operating. MEMs are efficient and valid for handling outcome data that are MAR but must be combined with MI-JOMO when covariates are also MAR.

Finally, I applied MEM with MI-JOMO to assess how dose and age modify the antipsychotic-induced weight gain (Chapter 7). Interaction terms in MEM helped to evaluate differences in weight trajectories over time between groups by dose or age. MEM was combined with MI-JOMO for handling missing values of dose. Again, the ITS approach provided new clinical evidence; for example, that older patients' weight is less affected by first olanzapine prescription compared to younger people.

Findings from the thesis are described in more detail in the next sub-sections. Then, I provide a broader discussion of my findings (Section 8.2) and explain their major methodological and clinical implications (Section 8.3). I also consider the strengths and limitations inherent to the thesis (Section 8.4) and the direction of future research (Section 8.5). Finally, I present my overall conclusions (Section 8.6).

8.1.1. Current practices in missing data handling for interrupted time series studies: a scoping review

In Chapter 4, I reviewed recent ITS investigations of health topics to understand 1) the data management strategies and statistical analysis performed in these ITS studies; and 2) how often missing data were considered and, if so, how they were reported and handled.

Many studies have been using the averaging-step for summarising the ITS outcome at each time point, preferring analysis tools designed for modelling population-level data (e.g. aggregate-level SR) over analysis tools for individual-level data (e.g. mixed-effects models). The averaging-step was utilised in 47/60 (78%) of the studies. The most typical statistical models were the segmented regression (SR) fitted with ordinary least square estimators (SR-OLS, n=23, 38%) or with maximum likelihood type estimators (SR with generalised linear models or SR-GLM, n=15, 25%). Although all these studies (n=60, 100%) had access to individual-level data, due to the nature of the studied ITS outcome (e.g. number of new patients before and after the intervention), only 32 (53%) could have followed/analysed the ITS outcome at an individual-level (i.e. repeated measures of the ITS outcome within individuals). However, other less granulated clusters could have been used for the analyses, for example, hospitals (n=13, 22%), hospital units (n=3,

5%), health facilities (n=3, 5%) or GPs (n=3, 5%) (i.e. ITS with multiple groups). This provides an approximation of how often a researcher could move from modelling ITS trajectories with population-level data points (i.e. only one ITS outcome average at each time-point) to model same trajectories with more granulated data, which were also available (e.g. individual- or hospital-level data for being modelled with MEM).

In ITS studies, the aggregate-level SR analysis was bringing potential issues when data are missing. As I have discussed throughout the thesis, the averaging-step transforms data that are MAR into data MNAR (see Chapter 6 for a detailed discussion). This is a potential source of bias that none of the reviewed studies reported as a limitation, showing that researchers are in general, not aware of the implications of cross-averaging the longitudinal data. The data used in ITS studies are mostly retrospective (i.e. routinely collected); therefore, the expected proportion of missing data on the outcome is usually high (i.e. due to irregular recording). This is particularly important for ITS designs, for which it is expected to have the outcome regularly measured at each time point. With many outcome gaps due to irregular recording, researchers select time windows as units of time (e.g. months) and average all the available records to set a unique outcome value for each time point/unit. All this is done with no significant reflection on how the MNAR issue related to data aggregation can induce bias. The widespread and uncritical use of the averaging-step with individual-level data MAR motivated my exploration of better analysis alternatives for handling missing data in ITS studies (i.e. MEM in chapter 5).

Missing data are poorly evaluated and reported in ITS studies, and statistical methods applied as standard are not designed for handling missing data issues. This study and other independent reviews have confirmed that missing data are rarely reported, which is consistent with researchers being unaware of the potential consequences of missing data and the statistical methods that could help to address this. For example, the complete case analysis (CCA) is the most used method for handling missing data, but it can lead to biased and less precise estimates. Every time an individual record is omitted due to missing data on a covariate, the researcher is reducing not only sample size but also losing some outcome records as well. If the records with missing covariates contain a systematically high or low range of outcome values, and they drop these observations, then the average from CCA at each time point could be biased. Further, if this bias

differs across time-points, then estimators of trajectories will themselves be biased. Alternative options for handling missing covariates such as multiple imputation (MI) have rarely been applied. Only relatively recently multilevel MI methods have become widely available. With multilevel methods such as MI-JOMO, it is possible not only to address the missing data problem on covariates but also to do this in a way that is consistent with interaction terms. Models with interaction terms are common in Controlled ITS (CITS) studies; thus, any broad MI solution should consider including these terms in the imputation models in order to ensure consistency with substantive models. MI-JOMO has flexibility for allowing interaction terms in MEM, as well as for handling missing covariates. That is why I evaluated MI-JOMO against other methods in Chapter 6 and showed its potential for handling missing values on interaction terms when examining impacts of initiation antipsychotic treatments in Chapter 7.

8.1.2. An application of interrupted time series with mixed-effects models

In Chapter 5, I applied MEM as an alternative to the aggregate-level SR to evaluate weight trajectories before and after antipsychotic treatment initiation when applying the ITS approach. My aims were: 1) Clinically: to investigate the change in body weight of patients initiated with high or low doses of the three most commonly prescribed second-generation antipsychotics, and 2) Methodologically: to apply MEM for handling missing data in longitudinal weight records.

The results showed olanzapine was associated with the highest increase in weight, in the short and long term, and higher doses were associated with an increased rate of higher weight gain. For example, the models showed a typical woman with a first prescription of olanzapine is expected to gain 2.3kg (95% CI: 1.9-2.7) in the first 6 weeks and 2.8kg (95% CI: 2.2-3.5) in the remaining 4 years of the follow-up period. When women were prescribed olanzapine at high dose (>5 mg), they were expected to gain 3.2kg (95% CI: 2.4-4.0) on average during the first 6 weeks and 2.9kg (95% CI: 1.6-4.2) at the rest of the 4 years observation period. A low dose was associated with 1.9kg (95% CI: 1.4-2.4) of weight gain in the first 6 weeks and 2.5kg (95% CI: 1.6-3.3) in the rest of the long-term period (up to 4 years). For all second-generation antipsychotics, weight gain remained in the long-term; thus, on average, in the follow-up period, patients never lost the weight they gained during the first 6 weeks of treatment. This evidence has direct

implications for clinical practice. It suggests doctors and patients may want to take the issue of a substantial weight gain into consideration when making decisions on the initiation of antipsychotic treatments, and doctors should prescribe the lowest effective dose to balance mental health benefits, weight gain and other adverse effects.

Mixed-effects models brought several advantages when used for ITS analysis. First, fully observed covariates helped to increase the plausibility of the MAR assumption for the irregularly observed weight measurements (unobserved weights can be viewed as 'missing' when compared with the ideal, regular weight measurement schedule). MAR assumes that, given covariates, the distribution of observed and missing weights over time is the same. Therefore, assuming the model is correct (or practically so), every time the outcome is MAR on these fully observed covariates, MEM will provide unbiased estimates [25]. As seen in Chapter 6, this is an advantage over the standard methods for ITS analysis, which make much less plausible assumptions about the 'missing' weight data. Moreover, MEM can produce additional information that standard ITS methods are not able to do. For example, MEM helped to investigate the association between weight at baseline (random intercept) and short-term weight gain (beta 2) within a single ITS design. This would not be possible directly using standard ITS regression models. Instead, we would need the estimation of more than one ITS of weight (i.e. one ITS per each group by weight at baseline, which is hard to achieve in EHR due to the high proportion of missing weight at baseline). All the advantages of MEM can be preserved even when the proposed ITS model is contrasted against other plausible ITS impact models (e.g. sensitivity analysis against non-linear trends over time).

However, MEM can only handle missing covariates by listwise deletion, which results in a reduction of the precision and potential problems with bias, mainly when the probability of missing covariates is associated with the outcome. In other words, MEMs handle missing covariates by CCA only, which, in routinely collected data, can also severely reduce the sample size affecting the standard errors. I recognised this limitation here, during the study of MEM as an alternative ITS method, motivating me to explore other choices like multiple imputation in Chapter 6. Based on my findings, in

[25] Indeed, this property justifies adjust for covariates in ITS analysis with individual-level data, even if the covariate is time-invariant.

Chapter 7, I reported a further study showing the potential of MEM and MI-JOMO for handling missing data in a more complex ITS design.

8.1.3. Evaluating methods for missing data handling in interrupted time series analysis via simulation studies

In Chapter 6, I performed an illustrative example where I contrasted the aggregate-level segmented regression (SR), MEM and MI-JOMO for the first time, showing their different behaviour, and the different results they give when handling missing values in same ITS study. I then executed a formal comparison of these methods via simulation studies. My aims were: 1) to examine the potential problems arising from the aggregate-level SR analysis when outcome data are missing, evaluating mixed models as an alternative approach; 2) to compare the performance of MEM with and without MI-JOMO for handling missing data on covariates.

In the illustrative example, I found substantial differences between estimates derived from MEM and aggregate-level SR. For example, short-term weight change (β_2) was 0.462kg/week from MEM and 0.816kg/week from aggregate-level SR. Likewise, in pre-treatment and long-term periods, weight change rates from aggregate-level SR were more than double the MEM estimates. In general, all estimates of weight change from aggregate-level SR analyses were higher in magnitude than those from MEM, which also implies a more substantial ITS treatment effect. To investigate my hypothesis that the aggregate-level SR was producing biased results, I used a set of contextually informed simulation studies.

Simulation results confirmed that the averaging-step causes bias in ITS estimates when data are MAR at the individual level. This occurs because taking averages of individual-level data before SR means that data at the cluster level are missing not at random. I also confirmed that the aggregate-level SR can over or underestimate the ITS effect. Nevertheless, it is not always possible to determine the direction of bias, because the direction of bias at each average-point depends on how the covariate is associated with the weight records missingness. Even when the 'aggregate-level' SR analysis does not cause any bias, results from simulations highlight that the precision is smaller as the standard errors for this method are typically grossly underestimated. Avoiding the

averaging-step and using MEM is recommended for handling missing values in the outcome.

For handling values MAR in covariates, MEM must be combined with MI-JOMO to obtain less biased estimates. I found that the most efficient way to do it is a two-steps process. First, missing values on covariates need to be multiply imputed, considering a model that is consistent with the substantive models. MI-JOMO allows us to complete this task efficiently. Then, MEM should be estimated using all the multiple imputed datasets, and multiple results can be summarised using Rubin´s rules. Simulation study results confirm the estimator based on MI-JOMO followed by a MEM is fairly unbiased when data at individual-level is MAR.

Imputation of covariates with missing records could also help in the evaluation of interaction terms, useful for comparison between trajectories of different sub-groups and further CITS studies. In Chapter 7, I explored how MEM with MI-JOMO can help to handle missing data in ITS studies, evaluating variables that could modify outcome trajectories overt time.

8.1.4. An application of multilevel multiple imputation to interrupted time series analysis

In Chapter 7, I applied MEM combined with MI-JOMO to handle missing data in an ITS study that evaluates how certain variables modify the antipsychotic-induced weight gain in people aged ≥40 years. The main aims were: 1) Clinical: to evaluate the weight trajectories before and after treatment initiation of antipsychotic treatment in ITS analysis and examine how sex, age and dose may independently modify these trajectories; and 2) Methodological: to apply MEM in combination with MI-JOMO to handle missing data when these trajectories and the mentioned modifying effects are modelled.

The results showed that age, sex and dose modify the weight change induced by specific second-generation antipsychotics in patients aged ≥40 years. Given an olanzapine

prescription, short-term weight gain is lower in women, but long-term weight gain is higher. Olanzapine-induced short-term weight gain is less severe for those older at baseline. A similar trend was observed for quetiapine and risperidone. All these ITS estimates have been adjusted for sex, deprivation and dose. When evaluating the interaction between dose and time, I found that higher doses of olanzapine and quetiapine are associated with more short-term weight gain. Higher doses of quetiapine are also associated with more long-term weight gain.

MEM alone or combined with MI-JOMO are useful tools for handling missing data in ITS studies evaluating modifying effects with individual-level data. Interaction terms in MEM facilitated the evaluation of differences in weight trajectories over time between groups by age, sex or dose. MEM was combined with MI-JOMO for handling missing values of dose, improving precision by recovering cases that are lost with MEM alone.

8.2. General Discussion

Other authors have researched missing data methods associated with general time series. Early work by Dunsmuir and Robinson [117] proposed a method for estimating time series models when the outcome has been collected irregularly (missing outcome at some time points). The method was focused on modelling population-level data; for example, levels of air pollution in a city. Later on, different methods to handle missing data in time series were compared via simulation studies: deletion, mean imputation, mean imputation using adjacent observations, and maximum likelihood (ML) estimation [118]. They found ML to be the most accurate method [26], but again when modelling data at the population level only. In 2010, Honaker and King [119] proposed a method for multiple imputation of population-level data that allowed smooth time trends and accounted for correlations within populations (e.g. countries) and shared same time points. Other newer methods included a single imputation of missing values in time series by using information from previous similar time series [120]; a machine learning algorithm (Generative Adversarial Networks) that works on multivariate time series for predicting most accurate imputation values [121]; and another method with a similar multivariate purpose but using a Bayesian framework [122]. These useful investigations demonstrate that attention to the missing data problem in time series analysis was, as expected, on the population-level data. No previous investigation has covered the specific topic of missing data handling of individual-level data for ITS studies.

There are other methods than MI-JOMO that could be applied in ITS analysis with individual-level data, but with different limitations. Multiple imputation by full conditional specification (FCS) [48] could be a good option for relatively short observation periods. However, it could fail to converge in long-term data (e.g. due to the large number of parameters to be estimated per each iteration). The two-fold conditional specification algorithm can reduce convergence issues by using shorter cycles (i.e. fewer parameters per iteration by setting consecutive short time windows) [123,124]. However, in general, FCS-type algorithms only allow imputation of two-level multilevel data with no non-linear effects or interactions (i.e. in ITS, data must be

[26] They generated 100 time points with 50 different conditions (including % missing data, levels of slope and levels of autocorrelation) as time-series datasets for the evaluation. They did not replicate these samples.

reshaped into a wide format with time points as variables). Whenever the substantive model needs to be more complex (e.g. adding more levels or interaction terms), achieving congeniality with FCS becomes harder. Moreover, FCS and two-fold conditional specification need to create artificial blocks of time (i.e. that will be variables in the wide-format dataset) for which the outcome value will be typically represented by a summary measure (e.g. an average). This averaging-step can introduce bias even with individual-level data MAR (Chapter 6). MI-JOMO overcomes these limitations and has shown to be as good as -or better than - other methods for handling missing values in the longitudinal analysis [125,126].

Some researchers have paid particular attention to the consequences of using aggregate data in ITS analyses. Gebski et al. [127] identified that standard aggregate-level SR could ignore informative heterogeneity; for example, when hospital units have different intercept and slopes but their data are aggregated at hospital level and analysed with fixed-effect models. They proposed a method where the intercept and slopes are pooled from these units, and then the overall effect is calculated by using inverse variance weights in a two-stage meta-analysis approach [27]. Taljaard et al. [128] recognised that aggregating data could reduce statistical power, but focused the attention on the cluster levels (e.g. hospital units) and not on the individual level (e.g. patients). Motivated by similar concerns about heterogeneity, Fretheim et al. [129] performed a sensitivity analysis comparing their standard ITS estimates against those obtained from a GEE model using individual-level data and controlling for clusters (i.e. sites). Since they only focused on controlling for heterogeneity, they were not aware on the missing data issue that GEE is not able to address [28] (major differences between GEE and aggregate-level SR estimators would not be expected because both methods are similarly biased when data are MAR). In an extensive piece of work, Ewusie [130] proposed a method that incorporates sites and individual weights in the estimator used to fit segmented regressions, but again focusing on solving the heterogeneity problem while omitting any reflection on missing data. Results from Chapter 6 (Simulation Study) showed that a similar weighting procedure could reduce bias only when all covariates are fully observed, which is rare in observational data. Recently, Beard et al [131] recognised the

[27] However, they concluded that "Where feasible… if there is strong evidence of heterogeneity between the units, an analysis incorporating a random effect for units may be appropriate", giving an early recognition to the potential of MEMs.

[28] I explain this in the discussion of Chapter 6.

link between aggregate data and missing data, and how it can bias ITS estimates, but provided very general recommendations on missing data handling and did not perform any formal evaluation of the methods suggested.

The above paragraphs bring out the novelty of the approach developed in this thesis. Lessons learned from Chapter 4 (Scoping Review) confirm that these issues represent a real problem in the ITS literature, ensuring the relevance of the solutions I have identified.

8.3. Implications

This thesis involved a mix of methodological and applied research. Although all studies have a methodological component, chapters 4 (Scoping Review) and 6 (Simulation Study) are more methodologically orientated, whereas chapters 5 (Application of MEM) and 7 (Application of MI-JOMO) show how the studied methods may be applied to answer clinical questions using ITS designs. In the following sections, I explain the methodological and clinical implications of the central thesis findings.

8.3.1. Methodological Implications

The critical methodological findings from this thesis are (a) avoid aggregate-level ITS analysis of individual patient data; (b) use mixed effect models instead, and (c) handle missing values on covariates by multiple imputation, taking care that the imputation is consistent with the mixed effect model being used for the primary analysis.

Beginning with (a), observational ITS studies of the type focussed on in this thesis do not collect the outcome (here, weight) according to any regular schedule. Therefore, the averaging step (ignoring any covariates) can introduce bias (more weights may be observed when weight is high or low) and lead to misleading standard errors (because averages of different numbers of observations have different standard errors). Thus, averaging-step should be avoided. Unfortunately, the aggregate-level analysis that follows the averaging-step is widespread in ITS studies. This aggregate-level analysis

assumes that the average at each time point includes outcome records from same - or exchangeable – people. However, that is a bold assumption in many studies, which violation can bias ITS intercepts and slopes if the proportion of MAR covariates changes across time points. The recommendation to avoid the averaging-step is relevant not only for ITS but also for time series in general.

Therefore, moving to (b), it follows that in ITS studies with routinely collected data, individual-level data analysis should be preferred over population-level data analysis. Most popular statistical methods in ITS have been designed for modelling population-level data, mainly ignoring the issues of having missing data at the individual level. Conversely, statistical tools such as MEM can model individual-level data directly but under a more plausible MAR assumption. In particular, MEMs are recommended over other individual-level analysis such as GEEs, because they are unbiased under MAR assumption. By its implicit outcome imputation, MEMs allow to easily handle the problem of irregular recording (which is typically expected in routinely collected data), avoiding the generation of any artificial time window. Moreover, the generalised linear mixed models (GLMM) facilitates the study of other ITS outcomes such as counts or proportions. By defining an appropriate correlation structure, MEM also allows controlling for autocorrelation, which is one of the critical issues in ITS analysis. Another advantage of MEMs is that we can explicitly model the variance (heterogeneity). It allows exploring questions like how the ITS outcome variability change with age, sex, and other variables, or how the intercept at treatment initiation (e.g. weight at baseline) covariates with the magnitude of immediate change (e.g. β_2 as the rate of short-term weight change). In sum, MEMs bring an excellent platform for modelling ITS trajectories on individual-level data directly.

Turning to (c), MI-JOMO is an excellent choice for handling covariates MAR when it is used consistently with the substantive model that, in ITS with individual-level data, should be a MEM. In practice, this is not a hard task because of the flexibility of MI-JOMO to incorporate any MEM structure (e.g. more than two levels, interaction or non-linear terms). Moreover, MI-JOMO can include auxiliary variables (i.e. not included in the substantive model) to improve the precision and reduce bias. Both MEM and MI-JOMO are also good choices for handling missing data in more sophisticated ITS designs (e.g. Controlled ITS). For example, Controlled ITS can be analysed by fitting

MEM with interaction terms (substantive model), while MI-JOMO allows imputation of covariates by using an imputation model that is consistent with the substantive model. In summary, for ITS studies applied on observational data at the individual level – that usually bring missing values – the modern preference among analysts should be orientated towards the use of alternative methods such as MEM and MI-JOMO.

Therefore, supported by the results presented in this thesis, I recommend that in future for ITS analyses when working with individual-level data with missing data issues:

- Avoid the averaging-step in ITS analyses
- Consider fit MEM with this type of data, especially for handling outcome values due to irregular recording.
- Consider using MI-JOMO for imputing covariates data MAR to reduce the risk of bias.

These recommendations should be given consideration in future guidelines about ITS studies. Currently, there is a significant effort to boost better practices in ITS studies for health research. Differences between the ideal and actual practices in analysis and reporting have recently been criticised [4,7,61,65], emphasising the lack of attention to the missing data problem [7,61]. This gap remains in modern ITS studies because most popular tutorials and guidelines [2,3,6,132] do not pay sufficient attention to the scopes of the missing data issue, or bring special recommendations on how to solve it. This thesis has generated a new understanding of an essential issue in ITS analysis based on individual-level data, a methodological knowledge that will be very useful for future health research.

8.3.2. Clinical Implications

The clinical implications of my research related to how second-generation antipsychotics plays a role in the long-term effects on weight. First, typical patients never lose the weight they gained during the first 6 weeks of treatment, which makes the decision to initiate treatment for doctors and patients complex. Currently, when a patient is overweight or obese at treatment initiation, doctors try to avoid the prescription of olanzapine due to its known effect on weight [95]. Nevertheless,

risperidone and quetiapine also increase the patient´s weight – although less severely –. However, the induced weight gain varies across dose as well, with higher doses associated with more weight gain. It means that the weight gain accumulated in the long-term can be similar for a woman prescribed a low dose of olanzapine, and for another woman prescribed a high dose of risperidone. The clinical prescribing decision should, therefore, consider both dose and drug type. Doctors should balance the long-term benefits and harms of both drug type and dose when deciding the best treatment together with their patients.

Before this thesis, it was unclear how the antipsychotic-induced weight gain was also an issue in older people. In this thesis, I demonstrated that the effect of antipsychotic initiation varies by age and was more pronounced in younger than older people. Specifically, in people aged 80 years or more, weight tends to stabilise after antipsychotic prescription. Thus, the natural weight loss that characterises this age in life –and brings some additional risks to older people – is less likely for elderly patients that need an antipsychotic prescription. Enlightened by this evidence, nutrition recommendations that usually come with the initiation of second-generation antipsychotics for the general population should be adapted for older people.

All these new clinical findings have been possible due to the use of ITS design. However, there is still further potential by the more widespread adoption of the approach I have used to model patient-level data. Most ITS researches in health topics focus on evaluating population-level interventions [61], but the ITS approach is also beneficial for assessing individual-level interventions (e.g. Chapters 5 and 7) when randomisation is not a feasible option [2,133]. Future clinical research can find in the ITS design with routinely data a fruitful approach to investigate a new type of treatments, therapies and programs applied at the patient level.

8.4. Strengths and Limitations

To the best of my knowledge, in the ITS literature, it is the first time that the issues about missing data have been formally assessed with simulation studies. I have demonstrated alternative methods of analysis that are not only better at handling missing data than the popular aggregate-level SR but also easy to use. I have started the conversation on why researchers should prefer statistical methods for individual-level data analysis instead of traditional ITS analysis tools designed for modelling population-level data, when the former data is available and face missing data issues. In a world where production and access to routinely collected data are ever-growing, the more general use of such as methods seems to be the future.

Nevertheless, these methods are not free of some limitations, but more applied and methodological research can be promoted to understand these better.

Lack of detailed information around prescription initiation can affect the consistency of the ITS treatment initiation point and data across patients. It happens because primary care databases only report the primary care prescribing, while some patients might be initiated treatment from secondary care. For example, at the specialist level (e.g. additional prescription or nutritional indications), in the hospital (e.g. special medication for emergency episodes) or pharmacy (i.e. when patients get the medication effectively). The impact of this is difficult to evaluate, but it is particularly crucial for short-term estimates. In the future, it may be possible to link primary and secondary prescribing data, but currently, there are few such data sources available.

There is some possible residual confounding in the single ITS design, especially for long-term estimates. Time-variant confounders are not controlled by single ITS, but a Controlled ITS (CITS) design [3] can help to control for observed confounders. For example, multiple prescriptions in older patients can also be explicative of weight change (e.g. drugs other than antipsychotics). The application of a CITS design can be the next step for research on antipsychotic-induced weight gain in the long-term.

Selective attrition is another potential issue in ITS with individual-level data, especially in older people. For example, attrition due to death (i.e. no imputation should be done after the date of death), leaving the GP or moving to a care home (i.e. the assumption of no association between these changes and the studied outcome could be substantial in some cases). When a MEM is fitted, attrition can be a source of bias due to the implicit imputation made by REML (see Chapter 3, Section 3.3). This is an issue that can be more or less severe, depending on the population and outcome studied. Although some sensitivity analysis could be performed, for example comparing estimates from those with and without attrition in more restrictive observation periods, further methodological research may be needed in order to find more efficient ways to tackle the problem.

Finally, I have some concerns about the accuracy of Rubin's variance formulas when multiple imputation (MI) is applied on covariates that define groups for interaction models (e.g. dose in Chapter 7). MI will provide an unequal number of groups across imputations, potentially leading to bias in Rubin´s variance formula. Although there are other alternative approaches (e.g. Reiter's rules, Robins and Wang's rules, or Bootstrap MI), this is a problem that certainly needs further research [134].

8.5. Future work

In the future, it would be relevant to adapt analysis tools for evaluating the impact of MNAR scenarios in ITS analysis with individual-level data. A version of the Delta Method applied for RCTs [53,72] could be applied for ITS studies to evaluate potential deviations from the MAR assumption (i.e. towards MNAR), but other alternatives such as mean score are also promising [135]. The main challenge is to propose a sensitivity analysis procedure that can be valid and efficient for a two-step method which applies explicit multiple imputation of covariates (MI-JOMO) followed by implicit imputation of the missing outcomes (MEM). There is no standard procedure for doing this kind of sensitivity analysis, then further methodological research is needed.

Solutions for the selective attrition problem and for improving variance estimates in ITS performed with individual-level data are also needed. I suggest to start exploring the sensitivity analysis approach described in the previous Section 8.4, in an applicative paper, in order to understand -with more detail - the impact of attrition due to death in clinical data (e.g. weight change due to antipsychotics in later life). For the variance estimates problem (i.e. from Rubin's formula), I recommend starting with a simulation study evaluating under which conditions the combined MI-JOMO plus MEM approach could lead to bias in ITS studies. In practice, some conditions could be more protective against biased variances (e.g. small differences across imputed datasets). After the first explorations, both problems will likely need further theoretical research (e.g. how an alternative ML type estimator can overcome the selective attrition problem), but the exploration itself can provide a more sharped and manageable image of these problems.

New applications of the approach I studied in my thesis can be performed by exploring new clinical and more complex questions. For example, the question of how weight trajectories induced by antipsychotics explain future cardiovascular (CVD) events in the long-term can be answered by an extension of my applied research but using joint modelling instead. Thus, it would be possible to estimate the association between antipsychotic treatment initiation and weigh change over time (what I have done already) by fitting a MEM jointly with a time-to-event model that explains the association between weight trajectories and CVD outcomes. Since this type of study

requires very long-term periods of observation (i.e. 10 years), key time-dependent confounders, and the use of a more sophisticated CITS design should be considered. In such as investigation, the missing data issue becomes more complicated since data may not only be incomplete for covariates at treatment initiation and the first outcome (weight change) but also on the final outcome (CVD event) and time-dependent confounders [29]. It means that the substantive model becomes more complex, and the imputation model must, therefore, reflect this complexity to be congenial [30].

A future work that I consider essential is to write a tutorial for handling missing values in ITS analyses applied to individual-level data. The tutorial could include details from my PhD study as well as from the future work I described above. Some topics should be central in this tutorial; for example, (i) the use of MEM and MI-JOMO on different ITS impact models and outcomes (e.g. counts and proportions), (ii) sensitivity analysis for plausible MNAR scenarios in ITS modelling with MI-JOMO plus MEM, and practical warnings about (iii) selective attrition and MEM estimator, and (iv) accuracy of variance estimates when MI-JOMO is applied. It could be written in an easy-to-read manner to be published in a journal of epidemiology or medical statistics and linked to an interactive website (e.g. with videos and forums) for promoting active learning [31].

Finally, an alternative version of the ITS approach proposed in this thesis can be developed on the Bayesian framework. Informative priors could help to simultaneously improve imputations, inferences and predictions if the MMI-plus-MEM analysis is performed as a fully Bayesian procedure. For example, explicit imputation of covariates by MMI (congenial with MEM) can be affected by severe non-normality [126]. In some scenarios, the informative priors combined with the observed likelihood (i.e. posterior distribution) can provide more realistic parameters estimates for the imputation models, improving imputations in consequence. If the substantive MEM model is also Bayesian, the posterior distribution can be more accurate for the estimation of the ITS parameters. Moreover, fitting Bayesian MEM can help to improve individual-level predictions, which is an expected use for many ITS models (e.g. classical forecasting,

[29] Handling missing data in time-dependent confounding in ITS studies is a further thesis in itself; thus, it will probably be interesting for a future doctoral or post-doctoral researcher.
[30] I have recently obtained funding from NIHR to explore how the methods I studied in my thesis can be extended to ensure proper imputations as well as other alternatives designed for handling missing data in joint modelling [50].
[31] The team of www.missingdata.org.uk are kindly willing to host this website when ready.

but now improved with individual-level information). In principle, MEM overcome the limitation of aggregate-level SR models for producing individual-level predictions, by including random intercepts and slopes. In the frequentist approach, MEM standard errors can be improved by bootstrapping [136], but this resampling technique is computationally inefficient with large datasets (e.g. primary care data). Alternatively, standard errors can be improved with Bayesian inference [137], and the uncertainty of predictions quantified. Model error -also called structural error - defines most of prediction uncertainty [138], and structural errors can be represented and quantified (e.g. generating probabilistic weights [139]), together with MEM parameters, in the Bayesian framework. Finally, the Generalised Linear Mixed Models (GLMM) can also be fitted as Bayesian models [140]. GLMM handle other types of outcomes which are typical in ITS studies (e.g. proportions or counts) [7], avoiding data aggregation while keeping the advantages of using random intercepts and slopes. Most of these Bayesian tools need further development or formal evaluation for informing future applications in ITS studies with individual-level missing data.

8.6. Conclusions

Antipsychotic dose and type play a key role in patient's weight gain, in the short and long term. The clinical prescribing decision should, therefore, consider both dose and drug type when initiating treatment. Doctors should balance the long-term benefits and harms of both drug type and dose when deciding the best treatment together with their patients. Antipsychotic-induced weight gain varies substantially with age. Enlightened by this evidence, nutrition recommendations that usually come with the initiation of second-generation antipsychotics for the general population should be adapted for older people.

For ITS performed with missing data at individual-level, the averaging-step in the ITS analysis should be avoided. MEM provides unbiased estimates if the outcome is MAR, but its validity and precision can be reduced if covariates are also missing. In such cases, MEM can be combined with MI-JOMO in a two-step process to handle data MAR effectively. These methods have shown to be effective for the evaluation of clinical treatment at individual-level and can be easily extended to other types of ITS studies. These findings show clearly that statistical methods for modelling individual-level data should be the standard approach, and that population-level approaches (especially when applied to individual-level data) should no longer be used.

9. REFERENCES

[1] Kontopantelis E, Doran T, Springate DA, Buchan I, Reeves D. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. Bmj 2015;350:h2750. https://doi.org/10.1136/bmj.h2750.

[2] Cook TD, Campbell DT, Shadish W. Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin Boston; 2002.

[3] Lopez Bernal J, Cummins S, Gasparrini A. The use of controls in interrupted time series studies of public health interventions. Int J Epidemiol 2018;47:2082–93. https://doi.org/10.1093/ije/dyy135.

[4] Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, Cheng AC, et al. Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: a review. J Clin Epidemiol 2020;122:1–11. https://doi.org/10.1016/j.jclinepi.2020.02.006.

[5] Bak M, Fransen A, Janssen J, van Os J, Drukker M. Almost all antipsychotics result in weight gain: a meta-analysis. PLoS One 2014;9:e94112. https://doi.org/10.1371/journal.pone.0094112.

[6] Lopez Bernal J, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol 2017;46:348–55. https://doi.org/10.1093/ije/dyw098.

[7] Hudson J, Fielding S, Ramsay CR. Methodology and reporting characteristics of studies using interrupted time series design in healthcare. BMC Med Res Methodol 2019;19:137. https://doi.org/10.1186/s12874-019-0777-x.

[8] Lopez Bernal J, Soumerai S, Gasparrini A. A methodological framework for model selection in interrupted time series studies. J Clin Epidemiol 2018;103:82–91. https://doi.org/10.1016/J.JCLINEPI.2018.05.026.

[9] Zhang F, Wagner AK, Ross-Degnan D. Simulation-based power calculation for designing interrupted time series analyses of health policy interventions. J Clin Epidemiol 2011;64:1252–61. https://doi.org/10.1016/j.jclinepi.2011.02.007.

[10] Gilmour S, Degenhardt L, Hall W, Day C. Using intervention time series analyses to assess the effects of imperfectly identifiable natural events: A general method and example. BMC Med Res Methodol 2006;6. https://doi.org/10.1186/1471-2288-6-16.

[11] Cruz M, Bender M, Ombao H. A robust interrupted time series model for analyzing complex health care intervention data. Stat Med 2017;36:4660–76. https://doi.org/10.1002/sim.7443.

[12] Cook TD, Campbell DT, Day A. Quasi-experimentation: Design & analysis issues for field settings. vol. 351. Boston: Houghton Mifflin Boston; 1979.

[13] Box GEP, Jenkins GM, Reinsel GC. Time Series Analysis: forecasting and Control. Fourth. New Jersey: John Wiley & Sons; 2008.

[14] Bhaskaran K, Gasparrini A, Hajat S, Smeeth L, Armstrong B. Time series regression studies in environmental epidemiology. Int J Epidemiol 2013;42:1187–95. https://doi.org/10.1093/ije/dyt092.

[15] Godfrey LG. Misspecification tests and their uses in econometrics. J Stat Plan Inference

1996;49:241–60. https://doi.org/10.1016/0378-3758(95)00039-9.

[16] Nelson BK. Time Series Analysis Using Autoregressive Integrated Moving Average (ARIMA) Models. Acad Emerg Med 1998;5:739–44. https://doi.org/10.1111/j.1553-2712.1998.tb02493.x.

[17] Kimura M, Croll P, Li B, Wong CP, Gogia S, Faud A, et al. Survey on medical records and EHR in Asia-Pacific Region. Methods Inf Med 2011;50:386–91. https://doi.org/10.3414/ME11-02-0002.

[18] Hoerbst A, Ammenwerth E. Quality and certification of electronic health records: An overview of current approaches from the US and Europe. Appl Clin Inform 2010;1:149–64. https://doi.org/10.4338/ACI-2010-02-R-0009.

[19] PAHO. Electronic Medical Records in Latin America and the Caribbean: An Analysis of the current situation and recommendations for the Region. Washington DC: 2016.

[20] Akanbi MO, Ocheke AN, Agaba PA, Daniyam CA, Agaba EI, Okeke EN, et al. Use of Electronic Health Records in sub-Saharan Africa: Progress and challenges. J Med Trop 2012;14:1–6.

[21] NHS-Digital. Patients Registered at a GP Practice. Lastest Version 2020. https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice (accessed January 31, 2020).

[22] NHS England. GP Online Services n.d. https://www.england.nhs.uk/gp/review-of-access/infrastructure/patient-online/ (accessed November 6, 2019).

[23] University of Oxford, Emisnug EH. QResearch n.d. https://www.qresearch.org/ (accessed November 6, 2019).

[24] NIHR. Clinical Practice Research Datalink n.d. https://www.cprd.com/home/ (accessed November 6, 2019).

[25] THIN Database Research Group. The Health Improvement Network (THIN) n.d. https://www.ucl.ac.uk/epidemiology-health-care/node/7035/database (accessed November 6, 2019).

[26] Vision. The Health Improvement Network (THIN) n.d. https://www.visionhealth.co.uk/portfolio-items/the-health-improvement-network-thin/ (accessed November 6, 2019).

[27] Blak B, Thompson M, Dattani H, Bourke A. Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. J Innov Heal Informatics 2011;19:251–5. https://doi.org/10.14236/jhi.v19i4.820.

[28] Chisholm J. The Read clinical classification. BMJ Br Med J 1990;300:1092. https://doi.org/10.1136/bmj.301.6742.45.

[29] Davé S, Petersen I. Creating medical and drug code lists to identify cases in primary care databases. Pharmacoepidemiol Drug Saf 2009;18:704–7. https://doi.org/10.1002/pds.1770.

[30] Scott T, Maynard A. Will the new GP contract lead to cost effective medical practice? York: 1991.

[31] NHS-Digital. Quality and Outcomes Framework (QOF), enhanced services and core contract extraction specifications (business rules) n.d. https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/quality-and-outcomes-

framework-qof (accessed November 6, 2019).

[32] Petersen I, Welch CA, Nazareth I, Walters K, Marston L, Morris RW, et al. Health indicator recording in UK primary care electronic health records: key implications for handling missing data. Clin Epidemiol 2019;11:157–67. https://doi.org/10.2147/CLEP.S191437.

[33] Parkin E. A paperless NHS: electronic health records. London: 2016.

[34] QuintilesIMS. THIN Data Guide for Researchers. London: 2017.

[35] Horsfall L, Walters K, Petersen I. Identifying periods of acceptable computer usage in primary care research databases. Pharmacoepidemiol Drug Saf 2013;22:64–9. https://doi.org/10.1002/pds.3368.

[36] Maguire A, Blak BT, Thompson M. The importance of defining periods of complete mortality reporting for research using automated data from primary care. Pharmacoepidemiol Drug Saf 2009;18:76–83. https://doi.org/10.1002/pds.1688.

[37] Dinov ID. Volume and value of big healthcare data. J Med Stat Informatics 2016;4. https://doi.org/10.7243/2053-7662-4-3.

[38] Liu W, Ye S, Barton BA, Fischer MA, Lawrence C, Rahn EJ, et al. Simulation-based power and sample size calculation for designing interrupted time series analyses of count outcomes in evaluation of health policy interventions. Contemp Clin Trials Commun 2020;17:100474. https://doi.org/10.1016/j.conctc.2019.100474.

[39] Welch C, Petersen I, Walters K, Morris RW, Nazareth I, Kalaitzaki E, et al. Two-stage method to remove population- and individual-level outliers from longitudinal data in a primary care database. Pharmacoepidemiol Drug Saf 2012;21:725–32. https://doi.org/10.1002/pds.2270.

[40] O'Keeffe AG, Nazareth I, Petersen I. Time trends in the prescription of statins for the primary prevention of cardiovascular disease in the United Kingdom: a cohort study using The Health Improvement Network primary care data. Clin Epidemiol 2016;8:123. https://doi.org/10.2147/CLEP.S104258.

[41] Langley TE, Szatkowski LC, Wythe S, Lewis SA. Can primary care data be used to monitor regional smoking prevalence? An analysis of the Health Improvement Network primary care data. BMC Public Health 2011;11:1–6. https://doi.org/10.1186/1471-2458-11-773.

[42] Khadjesari Z, Marston L, Petersen I, Nazareth I, Walters K. Alcohol consumption screening of newly-registered patients in primary care: A cross-sectional analysis. Br J Gen Pract 2013;63:e706–12. https://doi.org/10.3399/bjgp13X673720.

[43] Peng M, Chen G, Kaplan GG, Lix LM, Drummond N, Lucyk K, et al. Methods of defining hypertension in electronic medical records: validation against national survey data. J Public Health (Bangkok) 2016;38:e392–9. https://doi.org/10.1093/pubmed/fdv155.

[44] Sharma M, Petersen I, Nazareth I, Coton SJ. An algorithm for identification and classification of individuals with type 1 and type 2 diabetes mellitus in a large primary care database. Clin Epidemiol 2016;8:373–80. https://doi.org/10.2147/CLEP.S113415.

[45] Blackburn R, Osborn D, Walters K, Falcaro M, Nazareth I, Petersen I. Statin prescribing for people with severe mental illnesses: A staggered cohort study of "real-world" impacts. BMJ Open 2017;7:e013154. https://doi.org/10.1136/bmjopen-2016-013154.

[46] Gao S, Juhaeri J, Reshef S, Dai WS. Association between body mass index and suicide,

and suicide attempt among british adults: The health improvement network database. Obesity 2013;21:E334–42. https://doi.org/10.1002/oby.20143.

[47] Rubin DB. Inference and missing data. Biometrika 1976;63:581–92. https://doi.org/10.1093/biomet/63.3.581.

[48] Van Buuren S. Flexible imputation of missing data. Boca Raton: Chapman and Hall/CRC; 2012.

[49] White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat Med 2010;29:2920–31. https://doi.org/10.1002/sim.3944.

[50] Tsiatis AA, Kenward MG, Fitzmaurice G, Verbeke G, Molenberghs G. Handbook of missing data methodology. Boca Raton: Chapman and Hall/CRC; 2014.

[51] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. BMJ 2009;339:157–60. https://doi.org/10.1136/bmj.b2393.

[52] Rubin DB. Formalizing subjective notions about the effect of nonrespondents in sample surveys. J Am Stat Assoc 1977;72:538–43. https://doi.org/10.1080/01621459.1977.10480610.

[53] Carpenter JR, Kenward MG. Multiple imputation and its application. West Sussex: John Wiley & Sons; 2013.

[54] Meng X-L. Missing Data: Dial M for ??? J Am Stat Assoc 2000;95:1325. https://doi.org/10.2307/2669781.

[55] White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med 2011;30:377–99. https://doi.org/10.1002/sim.4067.

[56] Quartagno M, Grund S, Carpenter J. jomo: A Flexible Package for Two-level Joint Modelling Multiple Imputation. R J 2019;9.

[57] Quartagno M, Carpenter JR. Multiple imputation for discrete data: Evaluation of the joint latent normal model. Biometrical J 2019;61:bimj.201800222. https://doi.org/10.1002/bimj.201800222.

[58] Quartagno M, Carpenter JR. Multilevel multiple imputation in presence of interactions, non-linearities and random slopes. Stud. Theor. Appl. Stat. Meet. Ital. Stat. Soc. Palermo, 2018, p. 20–2.

[59] Enders CK. Analyzing longitudinal data with missing values. Rehabil Psychol 2011;56:267. https://doi.org/10.1037/a0025579.

[60] StataCorp. Stata 13 Base Reference Manual. College Station, TX: Stata Press; 2013.

[61] Jandoc R, Burden AM, Mamdani M, Lévesque LE, Cadarette SM. Interrupted time series analysis in drug utilization research is increasing: systematic review and recommendations. J Clin Epidemiol 2015;68:950–6. https://doi.org/10.1016/J.JCLINEPI.2014.12.018.

[62] Bao L, Peng R, Wang Y, Ma R, Ren X, Meng W, et al. Significant Reduction of Antibiotic Consumption and Patients' Costs after an Action Plan in China, 2010–2014. PLoS One 2015;10:e0118868. https://doi.org/10.1371/journal.pone.0118868.

[63] Yoo KB, Lee SG, Park S, Kim TH, Ahn J, Cho MH, et al. Effects of drug price reduction

and prescribing restrictions on expenditures and utilisation of antihypertensive drugs in Korea. BMJ Open 2015;5. https://doi.org/10.1136/bmjopen-2014-006940.

[64] Uijtendaal E V., Zwart-van Rijkom JEF, de Lange DW, Lalmohamed A, van Solinge WW, Egberts TCG. Influence of a strict glucose protocol on serum potassium and glucose concentrations and their association with mortality in intensive care patients. Crit Care 2015;19:270. https://doi.org/10.1186/s13054-015-0959-9.

[65] Polus S, Pieper D, Burns J, Fretheim A, Ramsay C, Higgins JPT, et al. Heterogeneity in application, design, and analysis characteristics was found for controlled before-after and interrupted time series studies included in Cochrane reviews. J Clin Epidemiol 2017;91:56–69. https://doi.org/10.1016/J.JCLINEPI.2017.07.008.

[66] Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169:467–73. https://doi.org/10.7326/M18-0850.

[67] Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018;18:143. https://doi.org/10.1186/s12874-018-0611-x.

[68] Grant MJ, Booth A. A typology of reviews: An analysis of 14 review types and associated methodologies. Health Info Libr J 2009;26:91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x.

[69] Close J, Fosh B, Wheat H, Horrell J, Lee W, Byng R, et al. Longitudinal evaluation of a countywide alternative to the Quality and Outcomes Framework in UK General Practice aimed at improving Person Centred Coordinated Care. BMJ Open 2019;9:29721. https://doi.org/10.1136/bmjopen-2019-029721.

[70] Rabe-Hesketh S, Skrondal A, Skrondal A. Multilevel and longitudinal modeling using stata. Texas: Stata Press Publication; 2008.

[71] Miller RL, Chiaramonte D, Strzyzykowski T, Sharma D, Anderson-Carpenter K, Fortenberry JD. Improving Timely Linkage to Care among Newly Diagnosed HIV-Infected Youth: Results of SMILE. J Urban Heal 2019;96:845–55. https://doi.org/10.1007/s11524-019-00391-z.

[72] Cro S, Carpenter JR, Kenward MG. Information-anchored sensitivity analysis: theory and application. J R Stat Soc 2019;182:623–45. https://doi.org/10.1111/rssa.12423.

[73] Newcomer JW, Haupt DW. The metabolic effects of antipsychotic medications. Can J Psychiatry 2006;51:480–91. https://doi.org/10.1177/070674370605100803 Article information.

[74] Holt RIG, Peveler RC. Obesity, serious mental illness and antipsychotic drugs. Diabetes, Obes Metab 2009;11:665–79. https://doi.org/10.1111/j.1463-1326.2009.01038.x.

[75] Elmslie JL, Silverstone JT, Mann JI, Williams SM, Romans SE. Prevalence of overweight and obesity in bipolar patients. J Clin Psychiatry 2000;61:179–84. https://doi.org/10.4088/jcp.v61n0306.

[76] Osborn DPJ, Levy G, Nazareth I, Petersen I, Islam A, King MB. Relative risk of cardiovascular and cancer mortality in people with severe mental illness from the United Kingdom's General Practice Research Database. Arch Gen Psychiatry 2007;64:242–9. https://doi.org/10.1001/archpsyc.64.2.242.

[77] Hayes JF, Marston L, Walters K, King MB, Osborn DPJ. Mortality gap for people with

bipolar disorder and schizophrenia: UK-based cohort study 2000–2014. Br J Psychiatry 2017;211:175–81. https://doi.org/10.1192/bjp.bp.117.202606.

[78] De Hert M, van Winkel R, Van Eyck D, Hanssens L, Wampers M, Scheen A, et al. Prevalence of diabetes, metabolic syndrome and metabolic abnormalities in schizophrenia over the course of the illness: a cross-sectional study. Clin Pract Epidemiol Ment Heal 2006;2:14. https://doi.org/10.1186/1745-0179-2-14.

[79] Emul M, Kalelioglu T. Etiology of cardiovascular disease in patients with schizophrenia: current perspectives. Neuropsychiatr Dis Treat 2015;11:2493. https://doi.org/10.2147/NDT.S50006.

[80] Olfson M, Gerhard T, Huang C, Crystal S, Stroup TS. Premature mortality among adults with schizophrenia in the United States. JAMA Psychiatry 2015;72:1172–81. https://doi.org/10.1001/jamapsychiatry.2015.1737.

[81] Lahti M, Tiihonen J, Wildgust H, Beary M, Hodgson R, Kajantie E, et al. Cardiovascular morbidity, mortality and pharmacotherapy in patients with schizophrenia. Psychol Med 2012;42:2275–85. https://doi.org/10.1017/S0033291712000396.

[82] Osborn DPJ, Petersen I, Beckley N, Walters K, Nazareth I, Hayes J. Weight change over two years in people prescribed olanzapine, quetiapine and risperidone in UK primary care: Cohort study in THIN, a UK primary care database. J Psychopharmacol 2018;32:1098–103. https://doi.org/10.1177/0269881118780011.

[83] Seeman M V. Secondary effects of antipsychotics: women at greater risk than men. Schizophr Bull 2008;35:937–48. https://doi.org/10.1093/schbul/sbn023.

[84] Koga M. Clinical factors related to gains in body mass index (BMI) among patients under long-term antipsychotic treatment. Psychiatr Neurol Jpn 2003;105:473–88.

[85] Covell NH, Weissman EM, Essock SM. Weight gain with clozapine compared to first generation antipsychotic medications. Schizophr Bull 2004;30:229–40. https://doi.org/10.1093/oxfordjournals.schbul.a007074.

[86] Najar H, Joas E, Kardell M, Pålsson E, Landén M. Weight gain with add-on second-generation antipsychotics in bipolar disorder: a naturalistic study. Acta Psychiatr Scand 2017;135:606–11. https://doi.org/10.1111/acps.12737.

[87] Gebhardt S, Haberhausen M, Heinzel-Gutenbrunner M, Gebhardt N, Remschmidt H, Krieg J-C, et al. Antipsychotic-induced body weight gain: predictors and a systematic categorization of the long-term weight course. J Psychiatr Res 2009;43:620–6. https://doi.org/10.1016/j.jpsychires.2008.11.001.

[88] Nutt DJ, Blier P. Neuroscience-based Nomenclature (NbN) for *Journal of Psychopharmacology*. J Psychopharmacol 2016;30:413–5. https://doi.org/10.1177/0269881116642903.

[89] Woods SW. Chlorpromazine equivalent doses for the newer atypical antipsychotics. J Clin Psychiatry 2003. https://doi.org/10.4088/jcp.v64n0607.

[90] Stata. Stata Statistical Software: Release 15. 2017.

[91] Haddad P. Weight change with atypical antipsychotics in the treatment of schizophrenia. J Psychopharmacol 2005;19:16–27. https://doi.org/10.1177/0269881105058378.

[92] Bushe CJ, Slooff CJ, Haddad PM, Karagianis JL. Weight change from 3-year observational data: findings from the worldwide schizophrenia outpatient health

outcomes database. J Clin Psychiatry 2012;73:e749-55. https://doi.org/10.4088/jcp.11m07246.

[93] Spertus J, Horvitz-Lennon M, Abing H, Normand S-L. Risk of weight gain for specific antipsychotic drugs: a meta-analysis. NPJ Schizophr 2018;4:12. https://doi.org/10.1038/s41537-018-0053-9.

[94] Haneuse S, Bogart A, Jazic I, Westbrook EO, Boudreau D, Theis MK, et al. Learning about missing data mechanisms in electronic health records-based research: a survey-based approach. Epidemiology 2016;27:82. https://doi.org/10.1097/EDE.0000000000000393.

[95] Marston L, Nazareth I, Petersen I, Walters K, Osborn DPJ. Prescribing of antipsychotics in UK primary care: a cohort study. BMJ Open 2014;4:e006135. https://doi.org/10.1136/bmjopen-2014-006135.

[96] Myer AJ, Belisle L. Highs and Lows: An Interrupted Time-Series Evaluation of the Impact of North America's Only Supervised Injection Facility on Crime. J Drug Issues 2018;48:36–49. https://doi.org/10.1177/0022042617727513.

[97] Ejlerskov KT, Sharp SJ, Stead M, Adamson AJ, White M, Adams J. Supermarket policies on less-healthy food at checkouts: Natural experimental evaluation using interrupted time series analyses of purchases. PLOS Med 2018;15:e1002712. https://doi.org/10.1371/journal.pmed.1002712.

[98] Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, et al. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol 2017;9:157–66. https://doi.org/10.2147/CLEP.S129785.

[99] Roland M, Guthrie B, Thomé DC. Primary medical care in the United Kingdom. J Am Board Fam Med 2012;25:S6–11.

[100] Bazo-Alvarez JC, Morris TP, Carpenter JR, Hayes JF, Petersen I. Effects of long-term antipsychotics treatment on body weight: A population-based cohort study. J Psychopharmacol 2019:026988111988591. https://doi.org/10.1177/0269881119885918.

[101] Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med 2019;38:2074–102. https://doi.org/10.1002/sim.8086.

[102] Prais SJ, Winsten CB. Trend estimators and serial correlation. Cowles Commission discussion paper Chicago; 1954.

[103] Goldstein H, Carpenter JR, Browne WJ. Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. J R Stat Soc Ser A 2014;177:553–64. https://doi.org/10.1111/rssa.12022.

[104] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008;61:344–9. https://doi.org/10.1016/J.JCLINEPI.2007.11.008.

[105] Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLOS Med 2015;12:e1001885. https://doi.org/10.1371/journal.pmed.1001885.

[106] Funatogawa I, Funatogawa T. Longitudinal data analysis: autoregressive linear mixed effects models. Singapore: Springer; 2018.

[107] Copas AJ, Seaman SR. Bias from the use of generalized estimating equations to analyze incomplete longitudinal binary data. J Appl Stat 2010;37:911–22. https://doi.org/10.1080/02664760902939604.

[108] Atti AR, Gozzi BF, Zuliani G, Bernabei V, Scudellari P, Berardi D, et al. A systematic review of metabolic side effects related to the use of antipsychotic drugs in dementia. Int Psychogeriatrics 2014;26:19–37. https://doi.org/10.1017/S1041610213001658.

[109] Qureshi NA, Al-Dossari DS, Salem SO, Alkhamees OA, Alsanad SM. Psychotropic medication-induced weight gain or loss looked through the lens of age and psychiatric diagnoses: A narrative review. Int Neuropsychiatr Dis J 2018:1–30. https://doi.org/10.9734/INDJ/2018/42585.

[110] Basson BR, Kinon BJ, Taylor CC, Szymanski KA, Gilmore JA, Tollefson GD. Factors influencing acute weight change in patients with schizophrenia treated with olanzapine, haloperidol, or risperidone. J Clin Psychiatry 2001;62:231–8. https://doi.org/10.4088/jcp.v62n0404.

[111] Petersen ML, Porter KE, Gruber S, Wang Y, Van Der Laan MJ. Diagnosing and responding to violations in the positivity assumption. Stat Methods Med Res 2012;21:31–54. https://doi.org/10.1177/0962280210386207.

[112] R Core Team. R: A language and environment for statistical computing. 2019.

[113] Yeung EYH, Chun S, Douglass A, Lau TE. Effect of atypical antipsychotics on body weight in geriatric psychiatric inpatients. SAGE Open Med 2017;5:2050312117708711. https://doi.org/10.1177/2050312117708711.

[114] Casey DE, Daniel DG, Wassef AA, Tracy KA, Wozniak P, Sommerville KW. Effect of divalproex combined with olanzapine or risperidone in patients with an acute exacerbation of schizophrenia. Neuropsychopharmacology 2003;28:182–92. https://doi.org/10.1038/sj.npp.1300023.

[115] Ralph SJ, Espinet AJ. Increased All-Cause Mortality by Antipsychotic Drugs: Updated Review and Meta-Analysis in Dementia and General Mental Health Care. J Alzheimer's Dis Reports 2018;2:1–26. https://doi.org/10.3233/adr-170042.

[116] Steinberg M, Lyketsos CG. Atypical Antipsychotic Use in Patients With Dementia: Managing Safety Concerns. Am J Psychiatry 2012;169:900–6. https://doi.org/10.1176/appi.ajp.2012.12030342.

[117] Dunsmuir W, Robinson PM. Estimation of time series models in the presence of missing data. J Am Stat Assoc 1981;76:560–8. https://doi.org/10.1080/01621459.1981.10477687.

[118] Velicer WF, Colby SM. A Comparison of Missing-Data Procedures for Arima Time-Series Analysis. Educ Psychol Meas 2005;65:596–615. https://doi.org/10.1177/0013164404272502.

[119] Honaker J, King G. What to Do about Missing Values in Time-Series Cross-Section Data. Am J Pol Sci 2010;54:561–81. https://doi.org/10.1111/j.1540-5907.2010.00447.x.

[120] Wellenzohn, Kevin; Böhlen, Michael H; Dignös, Anton; Gamper, Johann; Mitterer H. Continuous Imputation of Missing Values in Streams of Pattern-Determining Time Series. Proc. 20th Int. Conf. Extending Database Technol. EDBT 2017, Venecia, Italy: 2017.

[121] Luo Y, Cai X, Zhang Y, Xu J, Yuan X. Multivariate Time Series Imputation with Generative Adversarial Networks. NIPS Proc., 2018, p. 1596–607.

[122] Boaz RM, Lawson AB, Pearce JL. Multivariate air pollution prediction modeling with partial missingness. Environmetrics 2019;30. https://doi.org/10.1002/env.2592.

[123] Welch C, Bartlett J, Petersen I. Application of multiple imputation using the two-fold fully conditional specification algorithm in longitudinal clinical data. Stata J 2014;14:418–31. https://doi.org/10.1177/1536867x1401400213.

[124] Welch CA, Petersen I, Bartlett JW, White IR, Marston L, Morris RW, et al. Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data. Stat Med 2014;33:3725–37. https://doi.org/10.1002/sim.6184.

[125] Huque MH, Carlin JB, Simpson JA, Lee KJ. A comparison of multiple imputation methods for missing data in longitudinal studies. BMC Med Res Methodol 2018;18:168. https://doi.org/10.1186/s12874-018-0615-6.

[126] Enders CK, Hayes T, Du H. A Comparison of Multilevel Imputation Schemes for Random Coefficient Models: Fully Conditional Specification and Joint Model Imputation with Random Covariance Matrices. Multivariate Behav Res 2018;53:695–713. https://doi.org/10.1080/00273171.2018.1477040.

[127] Gebski V, Ellingson K, Edwards J, Jernigan J, Kleinbaum D. Modelling interrupted time series to evaluate prevention and control of infection in healthcare. Epidemiol Infect 2012;140:2131–41. https://doi.org/10.1017/S0950268812000179.

[128] Taljaard M, McKenzie JE, Ramsay CR, Grimshaw JM. The use of segmented regression in analysing interrupted time series studies: An example in pre-hospital ambulance care. Implement Sci 2014;9:1–4. https://doi.org/10.1186/1748-5908-9-77.

[129] Fretheim A, Soumerai SB, Zhang F, Oxman AD, Ross-Degnan D. Interrupted time-series analysis yielded an effect estimate concordant with the cluster-randomized controlled trial result. J Clin Epidemiol 2013;66:883–7. https://doi.org/10.1016/j.jclinepi.2013.03.016.

[130] Ewusie J. Improved Methods for Interrupted Time Series Analysis Useful When Outcomes are Aggregated: Accounting for heterogeneity across patients and healthcare settings. McMaster University, 2018.

[131] Beard E, Marsden J, Brown J, Tombor I, Stapleton J, Michie S, et al. Understanding and using time series analyses in addiction research. Addiction 2019;114:1866–84. https://doi.org/10.1111/add.14643.

[132] Bastuji-Garin S, Sbidian E, Gaudy-Marqueste C, Ferrat E, Roujeau JC, Richard MA, et al. Impact of STROBE Statement Publication on Quality of Observational Study Reporting: Interrupted Time Series versus Before-After Analysis. PLoS One 2013;8. https://doi.org/10.1371/journal.pone.0064733.

[133] Kontopantelis E, Doran T, Springate D, Bmj IB-, 2015 U. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. BmjCom n.d. https://doi.org/10.1136/bmj.h2750.

[134] Reiter JP. Multiple imputation when records used for imputation are not used or disseminated for analysis. Biometrika 2008;95:933–46. https://doi.org/10.1093/biomet/asn042.

[135] White IR, Carpenter J, Horton NJ. A mean score method for sensitivity analysis to departures from the missing at random assumption in randomised trials. Stat Sin 2018;28:1985–2003. https://doi.org/10.5705/ss.202016.0308.

[136] Bates D, Maechler M, Bolker B, Walker S, Christensen RHB, Singmann H, et al. Package

'lme4.' Version 2018;1:17.

[137] Bürkner P-C. Advanced Bayesian Multilevel Modeling with the R Package brms. R J 2018;10:395–411.

[138] Sargsyan K, Huan X, Najm HN. Embedded model error representation for Bayesian model calibration. Int J Uncertain Quantif 2019;9.

[139] Tasdighi A, Arabi M, Harmel D, Line D. A Bayesian total uncertainty analysis framework for assessment of management practices using watershed models. Environ Model Softw 2018;108:240–52.

[140] Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, et al. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol 2009;24:127–35.

[141] Gareri P, Segura-García C, Manfredi VGL, Bruni A, Ciambrone P, Cerminara G, et al. Use of atypical antipsychotics in the elderly: A clinical review. Clin Interv Aging 2014;9:1363–73. https://doi.org/10.2147/CIA.S63942.

[142] National Institute for Health Care Excellence. Dementia: assessment, management and support for people living with dementia and their carers. NICE Guidel [NG97] 2018. https://www.nice.org.uk/guidance/ng97/chapter/Recommendations (accessed June 5, 2020).

10. APPENDICES

10.1. Appendices of Chapter 4

10.1.1. Appendix 4A: Scoping review protocol (PROSPERO format)

1. Title: Current practices in missing data handling for interrupted time series: a scoping review
2. General Language: English
3. Start: Feb 05, 2020
4. Early Termination Date: Mar 05, 2020
5. Review Stage: In progress
6. Named Contact
   Juan Carlos Bazo-Alvarez
7. Named Contact Email
   juan.alvarez.16@ucl.ac.uk
   juan.bazo.a@gmail.com
8. Address
   Research Department of Primary Care and Population Health
   Upper Third Floor
   UCL Medical School (Royal Free Campus)
   Rowland Hill Street
   London
   NW3 2PF
9. Contact Telephone Number
   +44 07376076260
10. Organizational Affiliation of the Review
    1. Department of Primary Care and Population Health, University College London (UCL), London, United Kingdom.
    2. Instituto de Investigación, Universidad Católica Los Ángeles de Chimbote, Chimbote, Peru.
    3. MRC Clinical Trials Unit at UCL, London, United Kingdom.
    4. Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, United Kingdom.
    5. Department of Clinical Epidemiology, Aarhus University, Denmark.
11. Members
    Juan Carlos Bazo-Alvarez[1,2], Tim P Morris[3], James R Carpenter[3,4], Irene Petersen[1,5]
12. Financing
    JCB is sponsored by FONDECYT-CONCYTEC (grant contract number 231-2015-FONDECYT). TPM and JRC are supported by the Medical Research Council (grant numbers MC_UU_12023/21 and MC_UU_12023/29). The study sponsors only had a funding role in this research. Thus, researchers will work with total independence from their sponsors.
13. Conflicts of Interest
    None
14. Collaborators
    Frank Peralta
15. Review Questions
    In health care studies using the interrupted time series (ITS) approach:
    1. How are the data management and statistical analysis?
    2. How researchers report and handle the main methodological issues of ITS statistical analysis?
    3. If so, how is missing data reported and handled?
16. Search
    We will use the MEDLINE(R) and Epub Ahead of Print, In-Process & Other Non-Indexed Citations and Daily database (Ovid version) in February 2020 for finding ITS studies published in 2019. Search strategies will use a combination of free text terms and subject headings, and authors will be consulted for inclusion of appropriate terminology. Where appropriate, the validated filters will be used for limiting searches to interrupted time series designs. The search strategy will be reviewed by an information specialist using the Peer Review checklist of electronic search strategies. Studies whose full-text is not available will not be included. Current version of the search strategy is in the appendix.
17. URL to Search for the Strategy

See section 27 and Appendix 4For details on the search strategy.

18. Study Condition or Domain
    Evidence of reporting missing data handling procedures in healthcare studies using the interrupted time series approach.
19. Participants
    We will include all original ITS studies with a minimum of two data points before and after the initiation of a healthcare intervention (e.g. programs, policies, or educational interventions). Systematic reviews, meta-analysis, RCTs or studies that did not use and ITS-type analysis will be excluded. Studies with no access to individual-level data will be excluded as well. There will be no restrictions on participants, language of study, or the type of outcome.
20. Intervention
    This is not applicable to the present systematic review as it focuses on missing data handling in interrupted time series studies that evaluate healthcare interventions.
21. Comparator
    This is not applicable to the present systematic review as it focuses on missing data handling in interrupted time series studies that evaluate healthcare interventions.
22. Types of Studies to be Included
    See section 19.
23. Context
    None
24. Outcomes
    We will extract data from the studies consisted of:
    - General Information:
        - Author
        - Year
        - DOI
        - Definition of study design (e.g. ITS, before-and-after) [study label]
        - *Country of study*
        - Study objectives (population, intervention and the outcomes of interest)
        - *Type of intervention (e.g. policy, program, treatment)*
        - Level of intervention (e.g. individual, hospital, district, country)
        - *Participants (e.g. patients, doctors)*
    - Data handling and statistical analysis (research question #1):
        - Data source (e.g. prospective, routinely collected)
        - Type of outcome (e.g. continuous, count, binary)
        - The number of data points collected pre-and post-intervention and the unit (e.g. week, month, year)
        - Averaging-step at each time point (for individual level data only: yes/no)
        - Main statistical model/tool (e.g. segmented regression, mixed models, other)
    - Methodological issues -reporting and handling- (research question #2):
        - Autocorrelation
        - Seasonality
        - Time-varying confounders
        - Others
    - Missing data handling reported (yes/no) (research question #3): If yes, we will extract:
        - Missing data proportion reported (yes/no; if yes, % reported)
        - Missing data mechanism reported (yes/no; if yes, which one was declared)
        - Missing data handling method applied (yes/no; if yes, which method was applied)
        - Sensitivity analysis for missing data assumption (yes/no; if yes, which analysis was performed)

    We will base data extraction on the primary outcome and if no defined primary outcome is reported, we will use the first reported outcome.
25. Timing and Effect Measures
    This is not applicable to the present systematic review as it focuses on missing data handling in interrupted time series studies that evaluate healthcare interventions.
26. Additional Result
    None.
27. Data Extraction (Selection and Coding)
    After the search and selection of articles, all titles will be treated by the Rayyan program; a list will be created and duplicates removed. Review process will be carried out by two reviewers A and B, with the support of a third person C for disagreements (names to be defined). Reviewer A will screen titles and

abstracts identified by the search for inclusion. Reviewer B will assess 10% of the titles and abstracts and, if there are no disagreements, then reviewer A would proceed to single screening. Full-text copies for all the potential studies will be obtained and assessed for inclusion by A, with B double assessing 10% of them. Full-text review and data extraction will be done using an Excel template. In this systematic review, data on outcomes (section 24) will be extracted using a data collection sheet. Studies will be collected in any language, excluding those studies that do not have the full-text. All the excluded studies will be listed and enumerated indicating the reason for their exclusion.

28. Assessment of Risk of Bias
    As a methodological study, risk of bias assessment was not performed on individual studies.
29. Strategy for Data Synthesis
    We will summarize data using descriptive statistics (numbers and percentages or median, 25th, and 75th centile). Some graphs could be included to facilitate the communication of specific results.
30. Subgroup or Set Analysis
    A sub-group analysis will be performed between studies with or without access to individual level data. In particular, we are interested in i) missing data methods applied (if so); ii) statistical models applied; iii) whether an averaging-step was performed before statistical models fitted (when individual level data were available).
31. Type or Method of Revision
    Scoping review
32. Language
    English
33. Country
    UK
34. Other Registration Details
    None
35. Protocol URL
    (to be defined)
36. Dissemination Plans
    The results will be incorporated in a PhD thesis and later published in a scientific journal. Presentations will be made at an epidemiology congress.
37. Keywords
    Interrupted Time Series Analysis; Segmented Regression; Missing Data; Multiple Imputation.
38. Details of any other Existing Revisions
    No other existing reviews were reported in PROSPERO with similar goals. Hudson et al [7] published a similar study, but using 2015 data and with different objectives. We are using a similar research strategy.
39. Status of the Current Revision
    Ongoing
40. Additional Information
    None
41. Details
    None
42. References

Hudson J, Fielding S, Ramsay CR. Methodology and reporting characteristics of studies using interrupted time series design in healthcare. BMC Med Res Methodol 2019;19:137. https://doi.org/10.1186/s12874-019-0777-x.

Appendix: Search strategy

1. Interrupted Time Series Analysis/
2. interrupted time series.tw,kw.
3. (segmented adj3 regression).tw,kw.
4. arima.tw,kw.
5. autoregressive integrated moving average.tw,kw.
6. 1 or 2 or 3 or 4 or 5
7. limit 6 to yr="2019"

10.1.2. Appendix 4B: Search strategy

Interrupted Time Series Analysis/

interrupted time series.tw,kw.

(segmented adj3 regression).tw,kw.

arima.tw,kw.

autoregressive integrated moving average.tw,kw.

1 or 2 or 3 or 4 or 5

limit 6 to yr="2019"

10.1.3. Appendix 4C: Data extraction form

DATA COLLECTION FORM – SCOPING REVIEW
(only headers, no boxes)

1. General Information
   - 1.1. Publication ID
   - 1.2. EndNote Record Number
   - 1.3. Author
   - 1.4. Year
   - 1.5. Journal
   - 1.6. DOI
   - 1.7. Definition of study design (e.g. ITS, before-and-after)
   - 1.8. Country of study
   - 1.9. Study objectives
     - 1.9.1. Study population
     - 1.9.2. Study intervention
     - 1.9.3. Study outcome
   - 1.10. Type of intervention (e.g. policy, program, treatment)
   - 1.11. Level of intervention (e.g. individual, hospital, district, country)
   - 1.12. Participants (e.g. patients, doctors)
     - 1.12.1. Study unit of analysis
     - 1.12.2. Study minimum available cluster
     - 1.12.3. Study longitudinal follow up

2. Data handling and statistical analysis (research question #1)
   - 2.1. Data source (e.g. prospective, routinely collected)
   - 2.2. Data linked? (yes/no)
   - 2.3. Type of outcome (e.g. continuous, count, binary)
   - 2.4. Data points over time
     - 2.4.1. Number of data points
     - 2.4.2. Data points unit (e.g. week, month, year)
   - 2.5. Averaging-step at each time point (for individual level data only: yes/no)
   - 2.6. Main statistical model/tool (e.g. segmented regression, mixed models, other)
   - 2.7. Confounders
     - 2.7.1. Confounders reported (yes/no)
     - 2.7.2. Confounders adjusted for
   - 2.8. Autocorrelation
     - 2.8.1. Autocorrelation reported (yes/no)
     - 2.8.2. Autocorrelation handled with
   - 2.9. Seasonality
     - 2.9.1. Seasonality reported (yes/no)
     - 2.9.2. Seasonality handled with
   - 2.10. Time-varying confounders
     - 2.10.1. Time varying confounder reported (yes/no)
     - 2.10.2. Time varying confounder handled with
   - 2.11. Others methodological issues
     - 2.11.1. Others reported (yes/no)
     - 2.11.2. Others handled with

3. Missing data handling reported (yes/no) If yes, we will extract:
   - 3.1. Missing data proportion reported
     - 3.1.1. Missing data reported (yes/no)
     - 3.1.2. Missing data proportion (% reported)
   - 3.2. Missing data mechanism reported
     - 3.2.1. Missing data mechanism reported (yes/no)
     - 3.2.2. Which missing data mechanism (MAR, MCAR, MNAR)
   - 3.3. Missing data handling method applied
     - 3.3.1. Missing data method applied (yes/no)
     - 3.3.2. Which missing data method (CCA, MI, others)
   - 3.4. Sensitivity analysis for missing data assumption
     - 3.4.1. Missing data sensitivity analysis reported (yes/no)
     - 3.4.2. Which missing data sensitivity analysis (delta method, others)

4. Especial notes

10.1.4. Appendix 4D: List of the 60 selected publications

Table 4D. List of the finally selected publications from 2019 for the scoping review (N=60)

#	1st Author	Journal	Country	Study Type	Population	Intervention Type
1	Acquisto	American Journal of Health-System Pharmacy	USA	ITS	health personnel	program
2	Adeleke	Open Heart	UK	ITS	patients	program
3	Agarwal	JAMA Network Open	USA	ITS	patients	policy
4	Akazawa	Open Forum Infectious Diseases	Japan	ITS	patients	program
5	Annear	BMJ Global Health	Cambodia	ITS	patients	policy
6	Arruda	American Journal of Infection Control	Brazil	ITS	patients	intervention
7	Barnes	American Journal of Cardiology	USA	ITS	patients	guideline/protocol/sound published evidence
8	Barrio	International Journal of Drug Policy	Spain	CITS	patients	policy
9	Belle	Clinical Nutrition	Switzerland	SR	patients	treatment
10	Berdahl	Journal of General Internal Medicine	USA	ITS	patients	policy
11	Besen	Revista Brasileira de Terapia Intensiva	Brazil	ITS	patients	guideline/protocol/sound published evidence
12	Bruckner	International Journal of Epidemiology	France	CITS	patients	relevant or historic event
13	Bui	Journal of Safety Research	USA	ITS	firefighters	intervention
14	Carlos	Journal of the American College of Radiology	USA	ITS	insured women	policy
15	Close	BMJ Open	UK	CITS	health personnel	policy
16	Flick	AIDS	Malawi	ITS	health personnel	intervention
17	Garriga	BMJ Open	UK	ITS	patients	program
18	Gould	Surgery	USA	CITS	patients	policy
19	Grout	Academic pediatrics	USA	ITS	patients	intervention
20	Guan	BMJ Open	China	CITS	medications	policy
21	Haakenstad	Preventive Medicine	USA	ITS	patients	policy
22	Hallgren	Behavior Therapy	USA	ITS	patients	treatment
23	Hecker	American Journal of Infection Control	USA	ITS	patients	treatment
24	Heinsbroek	BMJ Open	UK	ITS	patients	intervention
25	Holmberg	Resuscitation	USA	CITS	patients	guideline/protocol/sound published evidence
26	Holroyd	BMC Geriatrics	Canada	ITS	patients	program
27	Horng	JAMA Network Open	USA	ITS	patients	intervention
28	Ismail	BMC Nephrology	Saudi Arabia	ITS	patients	intervention

29	Jhuang	Scientific Reports	China	ITS	general population	policy
30	Kim	Environment International	South Korea	CITS	general population	policy
31	Lichtl	BMJ Open	Germany	SR	patients	intervention
32	Liu	Canadian Journal of Cardiology	Canada	ITS	patients	program
33	Lu	Preventive Medicine	USA	ITS	patients	policy
34	Majka	American Journal of Medical Quality	USA	ITS	health personnel & patients	intervention
35	Mamun	Canadian Journal of Public Health	Canada	ITS	health personnel & patients	program
36	Marincowitz	BMJ Open	UK	ITS	patients	guideline/protocol/sound published evidence
37	Martin	European Journal of Human Genetics	UK	ITS	general population	guideline/protocol/sound published evidence
38	Merola	Hospital Pharmacy	USA	ITS	health personnel	intervention
39	Miller	Postgraduate Medical Journal	UK	ITS	health personnel	intervention
40	Miller	Journal of Urban Health	USA	ITS	patients	program
41	Ouldali	JAMA Pediatrics	France	ITS	children	program
42	Parchman	Annals of Family Medicine	USA	CITS	health personnel & patients	guideline/protocol/sound published evidence
43	Parekh	Value in Health	USA	ITS	patients	policy
44	Petruzzo	Multiple Sclerosis and Related Disorders	Italy	ITS	patients	guideline/protocol/sound published evidence
45	Qato	JAMA Network Open	USA	CITS	patients	relevant or historic event
46	Ramaswamy	JAMA Network Open	USA	ITS	patients	program
47	Ranapurwala	Pain Medicine	USA	ITS	patients	program
48	Ribera	BMJ Open	Spain	ITS	patients	guideline/protocol/sound published evidence
49	Ross	Journal of the International AIDS Society	Rwanda	ITS	patients	policy
50	Roychoudhury	Pediatric Neurology	Canada	ITS	patients	program
51	Sadan	Injury	Israel	ITS	health personnel & patients	intervention
52	Schwartz	Clinical Infectious Diseases	Bangladesh	CITS	children	intervention
53	Shah	Journal of Surgical Education	USA	ITS	health personnel	intervention
54	Singer	JAMA Network Open	USA	ITS	patients	policy
55	Smith	Journal of Pediatric Health Care	USA	ITS	health personnel & patients	program
56	Sugg	Preventive Medicine Reports	USA	ITS	general population	relevant or historic event
57	Tufts	HPB	USA	ITS	patients	intervention
58	Tyler	Hospital Pediatrics	USA	ITS	patients	guideline/protocol/sound published evidence
59	Van Seben	BMJ Open	Netherlands	ITS	health personnel & patients	program
60	Wassie	Public Health Nutrition	Australia	ITS	children	policy

10.1.5. Appendix 4E: Cross-table between level of intervention and more granulated clusters

Table 4E. Cross-table between level of intervention and more granulated clusters available (N=60)

	Minimal Available Cluster (More Granulated)										
Level of Intervention	GP	district	fire department	group of patients (by Dx)	health facility	hospital	hospital unit	household	individual-level	medications	Total
cities, group of	0	0	0	0	1	0	0	0	0	0	1
city/district	1	1	0	0	0	1	0	0	0	0	3
country	0	0	0	0	2	4	0	1	9	1	17
hospital	0	0	0	1	0	2	2	0	13	0	18
hospitals, group of	1	0	0	0	0	1	1	0	5	0	8
individual-level	0	0	0	0	0	0	0	0	2	0	2
state/province/county	1	1	0	0	0	5	0	0	3	0	10
three fire departments	0	0	1	0	0	0	0	0	0	0	1
Total	3	2	1	1	3	13	3	1	32	1	60

10.1.6. Appendix 4F: Cross-table between averaging-step and statistical model

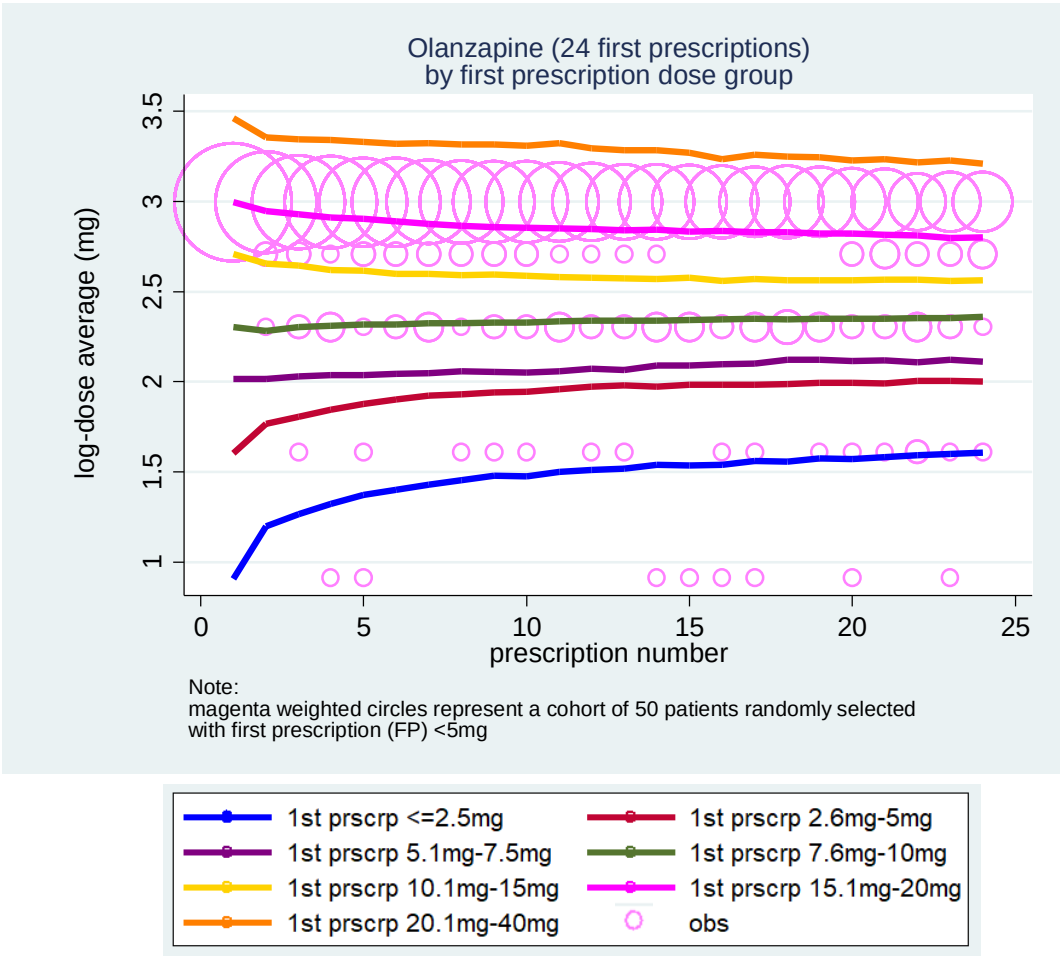
Table 4F. Cross-table between main statistical model and averaging-step (N=60)

Main Statistical Model	Averaging-step			
	no	unclear	yes	Total
ARIMA	0	0	7	7
Jointpoint (Exploratory Method)	0	1	0	1
SR-GEE	5	0	2	7
SR-GLM	1	0	14	15
SR-GLS	0	0	1	1
SR-OLS	0	1	22	23
mixed effects (random intercept only)	1	0	1	2
mixed effects (random intercept & slopes)	4	0	0	4
Total	11	2	47	60

10.2. Appendices of Chapter 5

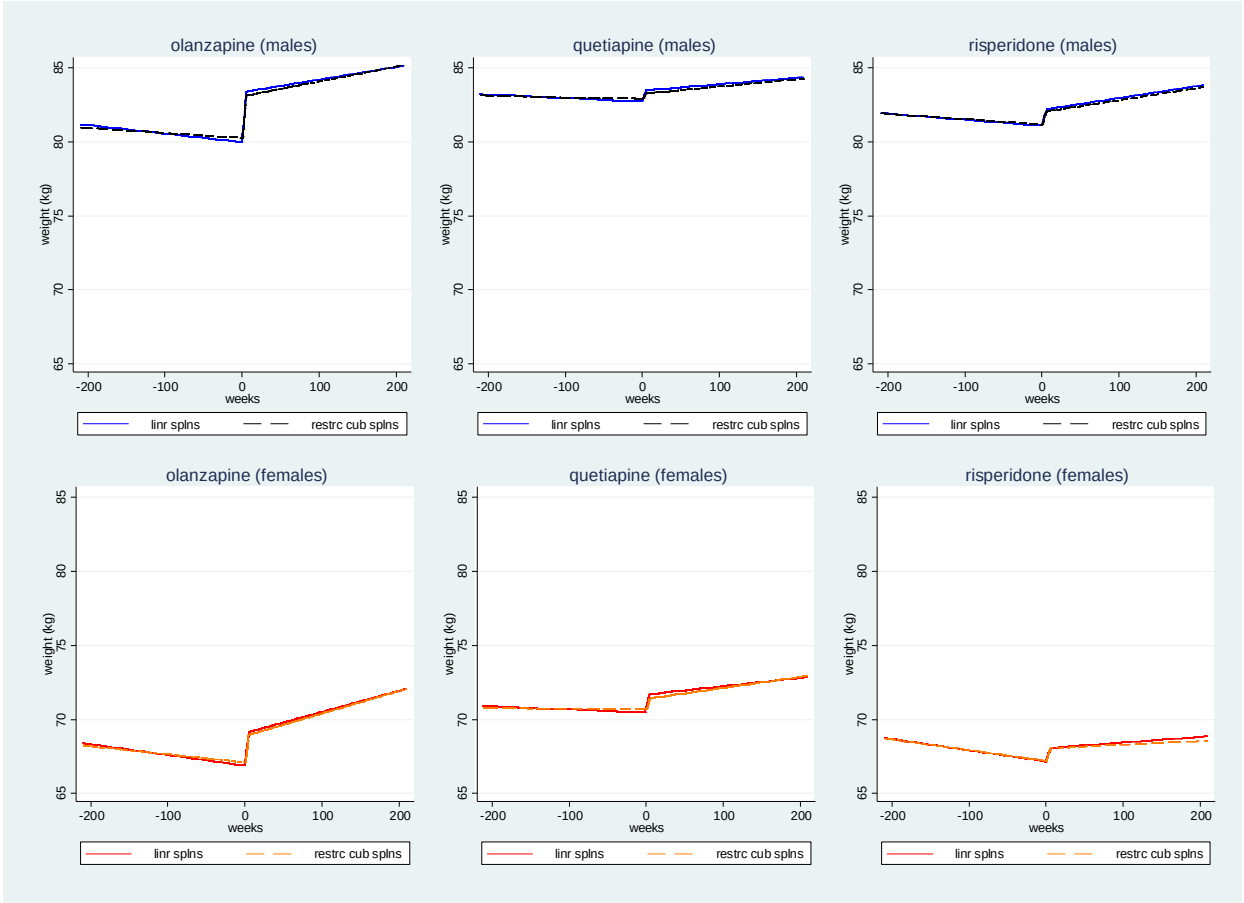
10.2.1 Appendix 5A: Visualization of antipsychotic dose over time

The next figure provides a visual approximation on how stable a first prescription of antipsychotic can be over time. I modelled trajectories of log-dose of olanzapine over 24 consecutive prescriptions for each of 7 groups differentiated by first dose (from the lowest in blue to the highest in orange, see legend below). Solid lines are average trajectories predicted from mixed effects models. Magenta weighted circles represent a cohort of 50 patients randomly selected from the magenta group (15.1mg-20mg). Bigger circles represent more persons from these 50 patients. It is visible that dose prescriptions of most of these patients stayed close to the first dose prescription that defined the magenta group, during the 24 first prescriptions. A similar pattern was confirmed across different doses and antipsychotic drugs.



10.2.2 Appendix 5B: Visual comparison between linear splines and restricted cubic models

Figure 5B. Visual comparison off linear splines models with restricted cubic models of changes in body weight over time before and after treatment initiation, by drug and sex.



10.2.3 Appendix 5C: Changes in weight by sex

Table 5C. Changes in body weight over time before and after treatment initiation, by sex

		Women						*Men*					
		Crude			*Adjusted**			*Crude*			*Adjusted**		
		weight change (kg/week)	95% CI	p	weight change (kg/week)	95% CI	p	weight change (kg/week)	95% CI	p	weight change (kg/week)	95% CI	p
OLANZAPINE (N=9499)	pre-baseline	-0.007	(-0.010 to -0.005)	<0.001	-0.007	(-0.010 to -0.005)	<0.001	-0.006	(-0.009 to -0.003)	<0.001	-0.006	(-0.009 to -0.003)	<0.001
	post short-term	0.383	(0.321 to 0.446)	<0.001	0.382	(0.319 to 0.444)	<0.001	0.570	(0.499 to 0.641)	<0.001	0.569	(0.498 to 0.640)	<0.001
	post long-term	0.014	(0.011 to 0.017)	<0.001	0.014	(0.011 to 0.017)	<0.001	0.009	(0.005 to 0.012)	<0.001	0.008	(0.005 to 0.012)	<0.001
	weight at baseline (kg)	68.4	(67.8 to 69.0)	<0.001	64.6	(62.9 to 66.3)	<0.001	81.2	(80.6 to 81.8)	<0.001	78.0	(76.3 to 79.7)	<0.001
	correlation**	-0.040	(-0.092 to 0.013)		-0.068	(-0.121 to -0.014)		-0.047	(-0.110 to 0.016)		-0.050	(-0.113 to 0.014)	
	N (%)			N=	5004	(52.7%)				N=	4495	(47.3%)	
QUETIAPINE (N=19965)	pre-baseline	-0.002	(-0.004 to 0.000)	0.025	-0.002	(-0.004 to 0.000)	0.030	-0.002	(-0.004 to 0.000)	0.022	-0.003	(-0.005 to -0.001)	0.008
	post short-term	0.204	(0.161 to 0.247)	<0.001	0.205	(0.162 to 0.248)	<0.001	0.127	(0.076 to 0.179)	<0.001	0.126	(0.074 to 0.177)	<0.001
	post long-term	0.006	(0.003 to 0.008)	<0.001	0.005	(0.003 to 0.008)	<0.001	0.004	(0.001 to 0.007)	0.006	0.003	(0.000 to 0.007)	0.025
	weight at baseline (kg)	70.9	(70.5 to 71.3)	<0.001	70.8	(69.6 to 71.9)	<0.001	83.2	(82.7 to 83.7)	<0.001	82.9	(81.5 to 84.3)	<0.001
	correlation**	0.013	(-0.020 to 0.046)		-0.022	(-0.055 to 0.011)		0.042	(-0.005 to 0.088)		0.012	(-0.035 to 0.060)	
	N (%)			N=	12149	(60.9%)				N=	7816	(39.1%)	
RISPERIDONE (N=9401)	pre-baseline	-0.008	(-0.010 to -0.005)	<0.001	-0.008	(-0.010 to -0.005)	<0.001	-0.004	(-0.007 to -0.001)	0.004	-0.004	(-0.007 to -0.002)	0.002
	post short-term	0.155	(0.091 to 0.220)	<0.001	0.147	(0.083 to 0.211)	<0.001	0.186	(0.110 to 0.262)	<0.001	0.180	(0.104 to 0.256)	<0.001
	post long-term	0.004	(0.000 to 0.008)	0.060	0.003	(-0.001 to 0.007)	0.089	0.008	(0.003 to 0.013)	0.001	0.007	(0.002 to 0.012)	0.004
	weight at baseline (kg)	68.8	(68.2 to 69.3)	<0.001	70.1	(68.2 to 72.0)	<0.001	81.9	(81.3 to 82.6)	<0.001	80.1	(78.2 to 82.1)	<0.001
	correlation**	0.001	(-0.050 to 0.052)		-0.025	(-0.077 to 0.027)		0.040	(-0.024 to 0.104)		0.014	(-0.051 to 0.079)	
	N (%)			N=	5153	(54.8%)				N=	4248	(45.2%)	

(*) Adjusted for age and deprivation (Townsend). Estimates come from linear splines random intercept and slope models. Body weight has been measured in kilograms and time in weeks. All crude and adjusted models reported ICC>0.95 as well as a p<0.001 when contrasted against null models (using Log-likelihood Ratio tests).

(**) Correlation between weight at baseline and weight change in the post short-term.

10.2.4Appendix 5D: Changes in weight by dose and sex

Table 5D. Changes in body weight over time before and after treatment initiation, by drug dose and sex

		Women						*Men*					
		Low-dose			*High-dose*			*Low-dose*			*High-dose*		
		weight change (kg/week)	95% CI	p	weight change (kg/week)	95% CI	p	weight change (kg/week)	95% CI	p	weight change (kg/week)	95% CI	p
OLANZAPINE (N=6,992)	pre-baseline	-0.010	-0.013 to -0.006	<0.001	-0.002	-0.009 to 0.004	0.425	-0.007	-0.011 to -0.003	0.001	-0.002	-0.007 to 0.004	0.552
	post short-term	0.314	0.231 to 0.397	<0.001	0.534	0.401 to 0.667	<0.001	0.425	0.326 to 0.525	<0.001	0.743	0.608 to 0.879	<0.001
	post long-term	0.012	0.008 to 0.017	<0.001	0.015	0.008 to 0.021	<0.001	0.009	0.004 to 0.015	0.001	0.007	0.001 to 0.013	0.020
	weight at baseline (kg)	65.4	63.1 to 67.8	<0.001	65.2	61.7 to 68.7	<0.001	78.1	75.5 to 80.7	<0.001	79.3	76.4 to 82.3	<0.001
	correlation*	-0.155	-0.230 to -0.078		-0.004	-0.123 to 0.115		-0.135	-0.235 to -0.033		0.022	-0.082 to 0.125	
	N	N = 2535			N = 1100			N = 1887			N = 1470		
QUETIAPINE (N=11,936)	pre-baseline	-0.005	-0.007 to -0.002	<0.001	0.010	0.005 to 0.015	<0.001	-0.009	-0.011 to -0.006	<0.001	0.010	0.004 to 0.015	0.001
	post short-term	0.110	0.046 to 0.174	0.001	0.376	0.260 to 0.492	<0.001	0.079	0.000 to 0.157	0.049	0.272	0.147 to 0.397	<0.001
	post long-term	0.004	0.000 to 0.008	0.028	0.008	0.002 to 0.014	0.009	-0.004	-0.009 to 0.001	0.153	0.005	-0.001 to 0.011	0.127
	weight at baseline (kg)	71.8	70.1 to 73.6	<0.001	69.8	66.9 to 72.7	<0.001	81.6	79.3 to 83.8	<0.001	82.9	79.5 to 86.3	<0.001
	correlation*	0.000	-0.052 to 0.052		0.001	-0.077 to 0.079		-0.012	-0.086 to 0.061		-0.005	-0.118 to 0.107	
	N	N = 5372			N = 1912			N = 3326			N = 1326		
RISPERIDONE (N=6,270)	pre-baseline	-0.009	-0.012 to -0.006	<0.001	0.018	0.005 to 0.030	0.005	-0.006	-0.009 to -0.002	0.002	0.001	-0.009 to 0.012	0.779
	post short-term	0.162	0.083 to 0.240	<0.001	0.188	-0.115 to 0.491	0.223	0.174	0.069 to 0.279	0.001	0.319	0.086 to 0.552	0.007
	post long-term	0.000	-0.004 to 0.005	0.868	0.017	0.005 to 0.030	0.005	0.006	-0.001 to 0.013	0.119	0.007	-0.004 to 0.018	0.192
	weight at baseline (kg)	70.2	67.8 to 72.6	<0.001	64.6	56.4 to 72.9	<0.001	80.0	77.4 to 82.5	<0.001	83.7	77.0 to 90.4	<0.001
	correlation*	-0.015	-0.083 to 0.054		0.120	-0.070 to 0.302		0.072	-0.017 to 0.160		0.126	-0.050 to 0.294	
	N	N = 3102			N = 316			N = 2411			N = 441		

All estimates were adjusted for age and deprivation (Townsend), coming from linear splines random intercept and slope models. Body weight has been measured in kilograms and time in weeks.Cut off point for low/high dose was: 5 mg for Olanzapine, 75 mg for Quetiapine and 2 mg for Risperidone. Missing data from drug dose are 26.4% for Olanzapine, 40.2% for Quetiapine and 33.3% for Risperidone.

(*) Correlation between weight at baseline and weight change in the post short-term.

10.3 Appendices of Chapter 6

10.3.1 Appendix 6A: Stata and R codes for the motivating example

1 I started with a THIN fully containing sex, age and smoking fully observed, but weight and date of weight as recorded (irregular recording overtime).

2 Time goes from 0 to 417 weeks, and week 209 marks the treatment initiation of olanzapine. The second knot is 215 weeks (+6 weeks after treatment initiation).

3 I fitted crude and adjusted mixed effects models (MEM) with this full data, as follows:

```
mkspline ls1 209 ls2 215 ls3 = time
mixed weight ls1 ls2 ls3 || gppatid: ls1 ls2 ls3, cov(un) stddev reml
mixed weight ls1 ls2 ls3 sex age age_2 smoking || gppatid: ls1 ls2 ls3, cov(un)
stddev reml
```

4 I also fitted SR and SR-W1 models with same full data, as follows:

```
preserve
    egen mean_sw1_weight = mean(weight), by(time)
    egen count_sw1_weight = count(weight), by(time)
    duplicates drop time mean_sw1_weight count_sw1_weight
    glm mean_sw1_weight ls1 ls2 ls3 /*SR analysis*/
    glm mean_sw1_weight ls1 ls2 ls3 [pw=count_sw1_weight] /*SR-W1 analysis*/
restore
```

5 Then, I set weight values MAR on sex, reproducing approximately the mechanism shown in Figure 6.1 (sex=0=men):

```
gen weight2 = weight
forvalues t=0(1)208 {
                local p = (209-`t')*0.004
                randomselect if (sex==0), gen(rs`t') prop(`p') seed(3956412)
                replace weight2 =. if (rs`t'==1) & (time==`t')
}

randomselect if (sex==0), gen(rs209) prop(.20)
replace weight2 =. if (rs209==1) & (time==209)
                randomselect if (sex==0), gen(rs210) prop(.30)
                replace weight2 =. if (rs210==1) & (time==210)
randomselect if (sex==0), gen(rs211) prop(.40)
replace weight2 =. if (rs211==1) & (time==211)
                randomselect if (sex==0), gen(rs212) prop(.50)
                replace weight2 =. if (rs212==1) & (time==212)
randomselect if (sex==0), gen(rs213) prop(.60)
replace weight2 =. if (rs213==1) & (time==213)
                randomselect if (sex==0), gen(rs214) prop(.70)
                replace weight2 =. if (rs214==1) & (time==214)
randomselect if (sex==0), gen(rs215) prop(.80)
replace weight2 =. if (rs215==1) & (time==215)

forvalues t=216(1)417 {
                local p = (`t'-215)*0.004
                randomselect if (sex==0), gen(rs`t') prop(`p')
                replace weight2 =. if (rs`t'==1) & (time==`t')
}
```

6 Then, I set smoking MAR on sex as follows:

```
drop rs* ls1 ls2 ls3 weight
reshape wide weight2 , i(gppatid) j(time)
randomselect if (sex==0), gen(smk_m) prop(.80)
```

```
replace smoking =. if (smk_m==1)
randomselect if (sex==1), gen(smk_f) prop(.20)
replace smoking =. if (smk_f==1)
reshape long
drop if weight2==.
mkspline ls1 209 ls2 215 ls3 = time
saveold phd_paper2_altana.dta, replace
```

7 I fitted crude and adjusted mixed effects models (MEM) again, but now on incomplete data:

```
mixed weight2 ls1 ls2 ls3 || gppatid: ls1 ls2 ls3, cov(un) stddev reml

mixed weight2 ls1 ls2 ls3 sex age age_2 smoking || gppatid: ls1 ls2 ls3, cov(un)
stddev reml
```

8 I imputed smoking status only by using MI-JOMO (using R and Stata):

```
###Preparing data
mydata <- read.dta13("phd_paper2_altana.dta")
mydata<-within(mydata, smoking<-factor(smoking))
mydata<-within(mydata, sex<-factor(sex))
mdata<-
mydata[,c("gppatid","weight2","time","ls1","ls2","ls3","sex","age","age_2","
smoking")]

###Running JOMO on missing data (MAR)
jdata<-
mydata[,c("gppatid","weight2","ls1","ls2","ls3","sex","age","age_2","smoking
")]
mylevel<-c(1,1,1,1,1,2,2,2,2)
formula<-as.formula(weight2 ~ sex + age + age_2 + smoking + ls1 + ls2 + ls3
+ (1 + ls1 + ls2 +ls3|gppatid))
jomo.imp<-jomo.lmer(formula,jdata, level=mylevel, nimp=20, nburn=1000,
nbetween=1000)

###Saving imputed data in Stata
write.dta(jomo.imp,  "jomo_appexample_2.dta")
```

```
set more off
use "\\ad.ucl.ac.uk\homez\rmjlbaz\DesktopSettings\Desktop\phd paper 2 app
exmpl\jomo_appexample_2.dta", clear
drop id
gen time = ls1 + ls2 + ls3
recode sex 1=0 2=1
recode smoking 1=0 2=1
mi import flong, id(clus time) m(Imputation) clear
mi estimate: mixed weight ls1 ls2 ls3 sex age age_2 smoking || clus: ls1 ls2 ls3,
cov(un) stddev reml
```

9 Finally, I performed the SR and SR-W1 analyses with a previous averaging-step:

```
/*using same data than in step 7*/

preserve
    egen mean_sw1_weight = mean(weight2), by(time)
    egen count_sw1_weight = count(weight2), by(time)
    sort time
    duplicates drop time mean_sw1_weight count_sw1_weight
    glm mean_sw1_weight ls1 ls2 ls3 /*SR analysis*/
    glm mean_sw1_weight ls1 ls2 ls3 [pw=count_sw1_weight] /*SR-W1 analysis*/
restore
```

10.3.2 Appendix 6B: Motivating Example: 95% confidence intervals of cumulative weight

Model	N	Weight gained during short-time (0-6 weeks), in kilograms.	95% CI	Weight gained during long-time (6 weeks-4 years), in kilograms.	95% CI	Total weight gained, in kilograms
MEM	3,379	2.47	(1.99 to 2.95)	2.46	(1.62 to 3.29)	4.93
MI-JOMO with MEM	6,181	2.75	(2.39 to 3.10)	2.70	(2.03 to 3.35)	5.45
SR	418	4.80	(3.62 to 5.97)	0.87	(-0.57 to 2.34)	5.67
SR-W1	418	4.72	(3.61 to 5.83)	2.13	(0.39 to 3.88)	6.85

Cumulative weight gain was calculated using estimates from Table 6.1 (data with missing records).CI=confidence interval, MEM=mixed effects model, MI-JOMO with MEM=multilevel multiple imputation followed by a mixed-effects model, SR=segmented regression, SR-W1=segmented regression weighted with the inverse of the number of weight records observed at each time-point

10.3.3 Appendix 6C: Summary of evaluated methods and scenarios

Study Section	Scenario	Missingness	Analysis methods evaluated	Results are visible in
Motivating Example (section 2)	Scenario 1	Incomplete weight, all covariates complete	MEM unadjusted, MEM adjusted, SR, SR-W1	Table 6.1, left section
	Scenario 2	Further incomplete weight (MAR), incomplete smoking (MAR), other covariates complete	MEM adjusted, MI-JOMO with MEM adjusted, SR, SR-W1	Table 6.1, right section
Simulation Study (section 3)	Scenario 1a	Weight and all covariates complete	MEM adjusted, SR, SR-W1, SR-W2	Table 2, top section
	Scenario 1b	Incomplete weight (MAR on sex), all covariates complete		
	Scenario 2a	Weight and all covariates complete		Table 2, bottom section
	Scenario 2b	Incomplete weight (MAR/MNAR on sex, age, smoking and intercept), smoking (MAR on sex and age), other covariates complete	MEM adjusted, MI-JOMO with MEM adjusted, SR, SR-W1, SR-W2	

Scenarios 1a and 1b have the same data generation mechanism= DGM-base. Scenarios 2a and 2b = DGM-extended-covariates. MAR=missing at random, SR='aggregate-level' segmented regression, SR-W1='aggregate-level' segmented regression weighted with the inverse of the number of observed weight records at each time point, SR-W2= similar to SR-W1 but the number of observed weight records were counted by each time point, sex and age group (quintiles), MEM=random intercept and slope model with restricted maximum likelihood and unstructured covariance, MI-JOMO=substantive model compatible joint modelling multiple imputation using a similar MEM model.

10.3.4 Appendix 6D: Comments for applicative analyses

Researchers could eventually trust in MAR as a mechanism that is feasible to handle with traditional ITS tools. However, here we showed that MCAR is the only mechanism for which SR (with averaging-step) and mixed-effects models (no averages at all) produce equally unbiased estimates. MCAR is rarely present in observational data.

In practice, two potential averaging-steps can generate bias when outcome data are MAR at the individual level. In the first step, the period for establishing temporal units is defined (e.g. a monthly unit for a follow-up period of 48 months) and all the outcome records from the same patient are averaged – or one of them could also be randomly selected within that period-. The average represents the outcome value for that patient in that month. If some days within the month, inform the missing mechanism of the outcome, and the observed values are systematically lower or higher than missing values, then the average will be biased. The second step implies to average all patients' weight averages within each month (e.g. population weight average for January, then for February, March, and so on). Now, if there is any covariate that can explain the outcome missingness and values at the individual level, that covariate will typically be lost after the second step (it becomes unobserved) and the ITS analysis will not be able to correct for bias. Considering the averaging issue studied by us, analysts should avoid any of these steps unless the individuals in the study are constant and fully observed over time (which is rarely the case in observational data).

In real data analysis, MEM and MI-JOMO present both advantages and disadvantages. MEM is computationally less demanding, does not raise congeniality issues and can include individuals with just one outcome observation in the analysis. However, MEM must omit any individual with missing covariates and, as we demonstrated, if these covariates explain the outcome missingness, we should expect biased estimates. Although in ITS designs we should not adjust for time-invariant confounders, with MEM we are forced to do it at least for those associated with the outcome missingness since MEM uses covariates information when its implicit imputation operates. Here is when multiple imputation becomes relevant. MI-JOMO can handle missing data in the covariates and can do it while keeping the congeniality issue controlled. The way we used MI-JOMO in our motivating example allows it to impute covariates values (smoking status) and not the outcome of interest (bodyweight in this case). After that, the implicit imputation of weight records by MEM operates similarly. This way to perform the analysis prevent many issues related to imputing outcomes as well (e.g. impute so many weight records missed). For example, whenever the outcome would also be imputed, MI-JOMO needs

more outcome records to converge (number of observations>number of imputation parameters), and thus individuals with just a few outcome records must be removed from the analysis to overcome the problem. That kind of operation reduces the boundaries of MI-JOMO notably, so in practice, it is usually better to proceed as suggested in our study (only imputing covariates at baseline).

10.4 Appendices of Chapter 7

10.4.1 Appendix 7A: Table 7A (for all drugs)

Table 7A. Trajectories of weight change by sex and interaction between sex and time for the olanzapine, quetiapine and risperidone cohorts.

Weight Trajectories by SEX		Olanzapine (women=3,803; men=2,973; combined=6,776)						Quetiapine (women=8,765; men=6,205; combined=14,970)						Risperidone (women=4,127; men=3,166; combined=7,293)					
		Beta	95% CI				p	Beta	95% CI				p	Beta	95% CI				p
Women	t1 (-4 years to 0 weeks)	-0.0150	(	-0.0178	to	-0.0122	) <0.001	-0.0086	(	-0.0105	to	-0.0067	) <0.001	-0.0132	(	-0.0157	to	-0.0107	) <0.001
	t2 (0 weeks to 6 weeks)	0.3021	(	0.2303	to	0.3738	) <0.001	0.1247	(	0.0760	to	0.1733	) <0.001	0.0666	(	-0.0026	to	0.1357	) 0.059
	t3 (6 weeks to 4 years)	0.0123	(	0.0087	to	0.0158	) <0.001	0.0038	(	0.0012	to	0.0065	) 0.005	0.0014	(	-0.0027	to	0.0055	) 0.500
	intercept	73.85	(	62.26	to	85.43	) <0.001	76.40	(	68.65	to	84.14	) <0.001	77.58	(	66.12	to	89.04	) <0.001
Men	t1 (-4 years to 0 weeks)	-0.0120	(	-0.0151	to	-0.0088	) <0.001	-0.0098	(	-0.0118	to	-0.0077	) <0.001	-0.0095	(	-0.0123	to	-0.0068	) <0.001
	t2 (0 weeks to 6 weeks)	0.4469	(	0.3687	to	0.5251	) <0.001	0.0976	(	0.0440	to	0.1513	) <0.001	0.0986	(	0.0198	to	0.1774	) 0.014
	t3 (6 weeks to 4 years)	0.0066	(	0.0028	to	0.0105	) 0.001	-0.0003	(	-0.0034	to	0.0028	) 0.855	0.0022	(	-0.0023	to	0.0066	) 0.343
	intercept	76.87	(	63.59	to	90.16	) <0.001	88.98	(	79.51	to	98.46	) <0.001	85.88	(	72.80	to	98.95	) <0.001
Combined	t1 (-4 years to 0 weeks)	-0.0120	(	-0.0153	to	-0.0088	) <0.001	-0.0098	(	-0.0120	to	-0.0077	) <0.001	-0.0096	(	-0.0124	to	-0.0067	) <0.001
	t2 (0 weeks to 6 weeks)	0.4455	(	0.3646	to	0.5264	) <0.001	0.1014	(	0.0441	to	0.1588	) 0.001	0.0974	(	0.0170	to	0.1779	) 0.018
	t3 (6 weeks to 4 years)	0.0067	(	0.0027	to	0.0107	) 0.001	-0.0002	(	-0.0035	to	0.0032	) 0.917	0.0024	(	-0.0023	to	0.0071	) 0.315
	women	-12.08	(	-13.02	to	-11.14	) <0.001	-12.97	(	-13.62	to	-12.32	) <0.001	-11.71	(	-12.60	to	-10.82	) <0.001
	women*t1	-0.0029	(	-0.0071	to	0.0014	) 0.187	0.0013	(	-0.0015	to	0.0041	) 0.372	-0.0036	(	-0.0073	to	0.0002	) 0.062
	women*t2	-0.1440	(	-0.2510	to	-0.0371	) 0.008	0.0187	(	-0.0553	to	0.0926	) 0.621	-0.0301	(	-0.1355	to	0.0753	) 0.575
	women*t3	0.0056	(	0.0003	to	0.0108	) 0.039	0.0040	(	-0.0002	to	0.0082	) 0.064	-0.0012	(	-0.0073	to	0.0049	) 0.696
	intercept	82.03	(	73.35	to	90.71	) <0.001	88.76	(	82.73	to	94.78	) <0.001	86.59	(	77.98	to	95.19	) <0.001

All estimates are adjusted for age and deprivation. Intercept was set at the very beginning of the observation period.

10.4.2 Appendix 7B: Table 7B (for olanzapine)

Table 7B. Trajectories of weight change by age, and interaction between age and time for the olanzapine cohort.

Weight Trajectories by AGE (Complete Case Analysis)		Olanzapine											
		Unadjusted (combined=6,776) (40-49 years=1,980; 50-59 years=1,620; 60-69 years=1,276; 70-79 years=1,140; 80-89 years=760)						Adjusted (combined=4,923) (40-49 years=1,468; 50-59 years=1,165; 60-69 years=926; 70-79 years=834; 80-89 years=530)					
		Beta	95% CI				p	Beta	95% CI				p
40-49 years	t1 (-4 years to 0 weeks)	-0.0013	(-0.0063	to	0.0036	)	0.591	0.0000	(-0.0056	to	0.0055	)	0.989
	t2 (0 weeks to 6 weeks)	0.6317	(0.5246	to	0.7387	)	0.000	0.6294	(0.5120	to	0.7469	)	0.000
	t3 (6 weeks to 4 years)	0.0097	(0.0050	to	0.0145	)	0.000	0.0108	(0.0056	to	0.0161	)	0.000
	intercept	77.76	(76.72	to	78.80	)	0.000	83.97	(81.06	to	86.89	)	0.000
50-59 years	t1 (-4 years to 0 weeks)	-0.0108	(-0.0153	to	-0.0063	)	0.000	-0.0138	(-0.0191	to	-0.0085	)	0.000
	t2 (0 weeks to 6 weeks)	0.4512	(0.3498	to	0.5527	)	0.000	0.4685	(0.3560	to	0.5810	)	0.000
	t3 (6 weeks to 4 years)	0.0121	(0.0074	to	0.0168	)	0.000	0.0135	(0.0080	to	0.0189	)	0.000
	intercept	77.71	(76.67	to	78.75	)	0.000	85.39	(82.60	to	88.17	)	0.000
60-69 years	t1 (-4 years to 0 weeks)	-0.0119	(-0.0162	to	-0.0077	)	0.000	-0.0123	(-0.0172	to	-0.0074	)	0.000
	t2 (0 weeks to 6 weeks)	0.2202	(0.1014	to	0.3389	)	0.000	0.2467	(0.1046	to	0.3888	)	0.001
	t3 (6 weeks to 4 years)	0.0119	(0.0055	to	0.0182	)	0.000	0.0102	(0.0027	to	0.0176	)	0.007
	intercept	74.97	(73.92	to	76.02	)	0.000	83.17	(80.50	to	85.84	)	0.000
70-79 years	t1 (-4 years to 0 weeks)	-0.0258	(-0.0299	to	-0.0218	)	0.000	-0.0231	(-0.0279	to	-0.0184	)	0.000
	t2 (0 weeks to 6 weeks)	0.1148	(-0.0022	to	0.2317	)	0.055	0.0865	(-0.0459	to	0.2190	)	0.200
	t3 (6 weeks to 4 years)	0.0052	(-0.0013	to	0.0118	)	0.119	0.0006	(-0.0070	to	0.0082	)	0.873
	intercept	73.63	(72.60	to	74.66	)	0.000	80.67	(78.11	to	83.24	)	0.000

80-89 years	t1 (-4 years to 0 weeks)		-0.0256	(	-0.0306	to	-0.0207	)	0.000	-0.0267	(	-0.0328	to	-0.0205	)	0.000
	t2 (0 weeks to 6 weeks)		-0.0869	(	-0.2441	to	0.0704	)	0.279	-0.0416	(	-0.2294	to	0.1461	)	0.664
	t3 (6 weeks to 4 years)		-0.0039	(	-0.0119	to	0.0040	)	0.328	-0.0021	(	-0.0106	to	0.0063	)	0.621
	intercept		68.83	(	67.65	to	70.01	)	0.000	79.78	(	77.00	to	82.55	)	0.000
Combined	t1 (-4 years to 0 weeks)		-0.0017	(	-0.0059	to	0.0025	)	0.424	-0.0002	(	-0.0051	to	0.0046	)	0.930
	t2 (0 weeks to 6 weeks)		0.6399	(	0.5405	to	0.7394	)	0.000	0.6478	(	0.5346	to	0.7610	)	0.000
	t3 (6 weeks to 4 years)		0.0096	(	0.0050	to	0.0143	)	0.000	0.0101	(	0.0048	to	0.0154	)	0.000
	age					ref							ref			
		50-59	-0.1296	(	-1.4932	to	1.2339	)	0.852	0.6778	(	-0.8439	to	2.1996	)	0.383
		60-69	-2.8110	(	-4.2496	to	-1.3723	)	0.000	-0.9582	(	-2.5668	to	0.6503	)	0.243
		70-79	-4.1645	(	-5.6312	to	-2.6978	)	0.000	-2.1086	(	-3.7563	to	-0.4609	)	0.012
		80-89	-8.9634	(	-10.6556	to	-7.2711	)	0.000	-6.0138	(	-7.9581	to	-4.0695	)	0.000
	age*t1					ref							ref			
		50-59	-0.0089	(	-0.0150	to	-0.0029	)	0.004	-0.0133	(	-0.0204	to	-0.0063	)	0.000
		60-69	-0.0102	(	-0.0164	to	-0.0040	)	0.001	-0.0120	(	-0.0192	to	-0.0048	)	0.001
		70-79	-0.0241	(	-0.0303	to	-0.0178	)	0.000	-0.0230	(	-0.0302	to	-0.0158	)	0.000
		80-89	-0.0238	(	-0.0312	to	-0.0164	)	0.000	-0.0261	(	-0.0347	to	-0.0174	)	0.000
	age*t2					ref							ref			
		50-59	-0.1843	(	-0.3292	to	-0.0394	)	0.013	-0.1789	(	-0.3443	to	-0.0135	)	0.034
		60-69	-0.4248	(	-0.5767	to	-0.2729	)	0.000	-0.4153	(	-0.5879	to	-0.2427	)	0.000
		70-79	-0.5284	(	-0.6877	to	-0.3692	)	0.000	-0.5617	(	-0.7440	to	-0.3794	)	0.000
		80-89	-0.7440	(	-0.9483	to	-0.5397	)	0.000	-0.7545	(	-0.9852	to	-0.5238	)	0.000
	age*t3					ref							ref			
		50-59	0.0027	(	-0.0042	to	0.0095	)	0.446	0.0036	(	-0.0042	to	0.0114	)	0.367
		60-69	0.0017	(	-0.0056	to	0.0091	)	0.643	-0.0004	(	-0.0088	to	0.0081	)	0.934
		70-79	-0.0047	(	-0.0129	to	0.0035	)	0.262	-0.0099	(	-0.0192	to	-0.0006	)	0.038
		80-89	-0.0122	(	-0.0237	to	-0.0008	)	0.037	-0.0113	(	-0.0242	to	0.0016	)	0.086
	intercept		77.8	(	76.9	to	78.7	)	0.000	84.09	(	82.51	to	85.67	)	0.000

ref=reference group. All estimates come from mixed effects models (restricted maximum likelihood and unstructured covariance matrix), adjusted for sex, social deprivation (Townsend) and first dose when indicated. Intercept was set at the very beginning of the observation period.

10.4.3 Appendix 7C: Table 7C (for quetiapine)

Table 7C. Trajectories of weight change by age, and interaction between age and time for the quetiapine cohort.

Weight Trajectories by AGE (Complete Case Analysis)		Quetiapine												
		Unadjusted (combined=14,970) (40-49 years=3,977; 50-59 years=2,838; 60-69 years=1,969; 70-79 years=2,709; 80-89 years=3,477)							Adjusted (combined=8,950) (40-49 years=2,325; 50-59 years=1,694; 60-69 years=1,227; 70-79 years=1,656; 80-89 years=2,048)						
		Beta	95% CI					p	Beta	95% CI					p
40-49 years	t1 (-4 years to 0 weeks)	0.0085	(	0.0051	to	0.0119	)	0.000	0.0096	(	0.0050	to	0.0141	)	0.000
	t2 (0 weeks to 6 weeks)	0.2618	(	0.1874	to	0.3362	)	0.000	0.2626	(	0.1618	to	0.3633	)	0.000
	t3 (6 weeks to 4 years)	0.0074	(	0.0036	to	0.0112	)	0.000	0.0025	(	-0.0027	to	0.0076	)	0.349
	intercept	81.52	(	80.75	to	82.30	)	0.000	88.42	(	85.77	to	91.06	)	0.000
50-59 years	t1 (-4 years to 0 weeks)	-0.0015	(	-0.0048	to	0.0018	)	0.371	-0.0022	(	-0.0065	to	0.0021	)	0.307
	t2 (0 weeks to 6 weeks)	0.2059	(	0.1267	to	0.2852	)	0.000	0.2322	(	0.1260	to	0.3383	)	0.000
	t3 (6 weeks to 4 years)	0.0071	(	0.0032	to	0.0110	)	0.000	0.0049	(	-0.0002	to	0.0099	)	0.060
	intercept	82.46	(	81.62	to	83.31	)	0.000	88.13	(	85.35	to	90.92	)	0.000
60-69 years	t1 (-4 years to 0 weeks)	-0.0107	(	-0.0142	to	-0.0073	)	0.000	-0.0089	(	-0.0133	to	-0.0046	)	0.000
	t2 (0 weeks to 6 weeks)	0.0648	(	-0.0217	to	0.1512	)	0.142	0.0578	(	-0.0518	to	0.1673	)	0.301
	t3 (6 weeks to 4 years)	0.0027	(	-0.0018	to	0.0072	)	0.239	0.0005	(	-0.0052	to	0.0063	)	0.854
	intercept	78.28	(	77.37	to	79.18	)	0.000	84.52	(	82.12	to	86.92	)	0.000
70-79 years	t1 (-4 years to 0 weeks)	-0.0193	(	-0.0221	to	-0.0165	)	0.000	-0.0178	(	-0.0214	to	-0.0142	)	0.000
	t2 (0 weeks to 6 weeks)	-0.0897	(	-0.1734	to	-0.0060	)	0.036	-0.0603	(	-0.1656	to	0.0449	)	0.261
	t3 (6 weeks to 4 years)	-0.0078	(	-0.0132	to	-0.0023	)	0.005	-0.0078	(	-0.0146	to	-0.0010	)	0.024
	intercept	74.33	(	73.66	to	75.01	)	0.000	79.75	(	78.11	to	81.38	)	0.000

80-89 years	t1 (-4 years to 0 weeks)		-0.0230	(	-0.0254	to	-0.0207	)	0.000	-0.0231	(	-0.0262	to	-0.0201	)	0.000
	t2 (0 weeks to 6 weeks)		-0.0827	(	-0.1614	to	-0.0040	)	0.039	-0.0923	(	-0.1958	to	0.0112	)	0.081
	t3 (6 weeks to 4 years)		-0.0203	(	-0.0263	to	-0.0143	)	0.000	-0.0157	(	-0.0236	to	-0.0078	)	0.000
	intercept		68.58	(	68.03	to	69.12	)	0.000	77.97	(	76.57	to	79.37	)	0.000
Combined	t1 (-4 years to 0 weeks)		0.0079	(	0.0051	to	0.0107	)	0.000	0.0091	(	0.0054	to	0.0128	)	0.000
	t2 (0 weeks to 6 weeks)		0.2805	(	0.2123	to	0.3487	)	0.000	0.2731	(	0.1818	to	0.3643	)	0.000
	t3 (6 weeks to 4 years)		0.0070	(	0.0035	to	0.0106	)	0.000	0.0022	(	-0.0025	to	0.0069	)	0.360
	age					ref							ref			
		50-59	0.9184	(	-0.0947	to	1.9315	)	0.076	0.8097	(	-0.4503	to	2.0698	)	0.208
		60-69	-3.2989	(	-4.4273	to	-2.1706	)	0.000	-2.8048	(	-4.1967	to	-1.4129	)	0.000
		70-79	-7.2780	(	-8.2945	to	-6.2615	)	0.000	-7.8847	(	-9.1800	to	-6.5894	)	0.000
		80-89	-13.0431	(	-13.9985	to	-12.0878	)	0.000	-11.8903	(	-13.1387	to	-10.6419	)	0.000
	age*t1					ref							ref			
		50-59	-0.0095	(	-0.0137	to	-0.0053	)	0.000	-0.0115	(	-0.0170	to	-0.0060	)	0.000
		60-69	-0.0186	(	-0.0232	to	-0.0140	)	0.000	-0.0183	(	-0.0243	to	-0.0124	)	0.000
		70-79	-0.0270	(	-0.0312	to	-0.0228	)	0.000	-0.0268	(	-0.0322	to	-0.0213	)	0.000
		80-89	-0.0308	(	-0.0349	to	-0.0268	)	0.000	-0.0322	(	-0.0375	to	-0.0269	)	0.000
	age*t2					ref							ref			
		50-59	-0.0760	(	-0.1784	to	0.0264	)	0.146	-0.0453	(	-0.1807	to	0.0901	)	0.512
		60-69	-0.2258	(	-0.3384	to	-0.1133	)	0.000	-0.2176	(	-0.3649	to	-0.0703	)	0.004
		70-79	-0.3890	(	-0.4994	to	-0.2785	)	0.000	-0.3480	(	-0.4929	to	-0.2032	)	0.000
		80-89	-0.3604	(	-0.4722	to	-0.2486	)	0.000	-0.3641	(	-0.5134	to	-0.2148	)	0.000
	age*t3					ref							ref			
		50-59	0.0002	(	-0.0053	to	0.0056	)	0.953	0.0027	(	-0.0044	to	0.0098	)	0.464
		60-69	-0.0036	(	-0.0096	to	0.0024	)	0.242	-0.0012	(	-0.0090	to	0.0066	)	0.755
		70-79	-0.0137	(	-0.0202	to	-0.0072	)	0.000	-0.0085	(	-0.0169	to	0.0000	)	0.050
		80-89	-0.0266	(	-0.0335	to	-0.0197	)	0.000	-0.0190	(	-0.0281	to	-0.0099	)	0.000
	intercept		81.6	(	80.9	to	82.2	)	0.000	88.27	(	86.99	to	89.55	)	0.000

ref=reference group. All estimates come from mixed effects models (restricted maximum likelihood and unstructured covariance matrix), adjusted for sex, social deprivation (Townsend) and first dose when indicated. Intercept was set at the very beginning of the observation period.

10.4.4 Appendix 7D: Table 7D (for risperidone)

Table 7D. Trajectories of weight change by age, and interaction between age and time for the quetiapine cohort.

Weight Trajectories by AGE (Complete Case Analysis)		Risperidone											
		Unadjusted (combined=7,293) (40-49 years=1,353; 50-59 years=1,065; 60-69 years=1,014; 70-79 years=1,614; 80-89 years=2,247)						Adjusted (combined=4,761) (40-49 years=961; 50-59 years=716; 60-69 years=696; 70-79 years=1,016; 80-89 years=1,372)					
		Beta	95% CI				p	Beta	95% CI				p
40-49 years	t1 (-4 years to 0 weeks)	0.0094	(0.0042	to	0.0146	)	0.000	0.0094	(0.0033	to	0.0154	)	0.002
	t2 (0 weeks to 6 weeks)	0.2463	(0.1214	to	0.3712	)	0.000	0.2853	(0.1404	to	0.4302	)	0.000
	t3 (6 weeks to 4 years)	0.0116	(0.0056	to	0.0176	)	0.000	0.0097	(0.0026	to	0.0168	)	0.008
	intercept	80.62	(79.31	to	81.92	)	0.000	85.25	(81.35	to	89.15	)	0.000
50-59 years	t1 (-4 years to 0 weeks)	-0.0028	(-0.0084	to	0.0028	)	0.322	-0.0051	(-0.0123	to	0.0021	)	0.165
	t2 (0 weeks to 6 weeks)	0.2575	(0.1261	to	0.3889	)	0.000	0.3121	(0.1450	to	0.4792	)	0.000
	t3 (6 weeks to 4 years)	0.0041	(-0.0021	to	0.0103	)	0.195	0.0056	(-0.0024	to	0.0135	)	0.169
	intercept	80.23	(78.82	to	81.64	)	0.000	88.10	(84.06	to	92.13	)	0.000
60-69 years	t1 (-4 years to 0 weeks)	-0.0087	(-0.0139	to	-0.0035	)	0.001	-0.0085	(-0.0145	to	-0.0024	)	0.006
	t2 (0 weeks to 6 weeks)	0.0647	(-0.0518	to	0.1812	)	0.276	0.1022	(-0.0384	to	0.2429	)	0.154
	t3 (6 weeks to 4 years)	0.0024	(-0.0040	to	0.0088	)	0.466	0.0010	(-0.0063	to	0.0082	)	0.793
	intercept	77.66	(76.40	to	78.92	)	0.000	83.55	(80.15	to	86.96	)	0.000
70-79 years	t1 (-4 years to 0 weeks)	-0.0183	(-0.0218	to	-0.0148	)	0.000	-0.0177	(-0.0223	to	-0.0131	)	0.000
	t2 (0 weeks to 6 weeks)	-0.0926	(-0.1929	to	0.0077	)	0.070	-0.1131	(-0.2425	to	0.0163	)	0.087
	t3 (6 weeks to 4 years)	-0.0060	(-0.0130	to	0.0010	)	0.092	-0.0032	(-0.0114	to	0.0050	)	0.440
	intercept	73.71	(72.85	to	74.57	)	0.000	80.00	(77.72	to	82.28	)	0.000

80-89 years	t1 (-4 years to 0 weeks)		-0.0216	(	-0.0245	to	-0.0187	)	0.000	-0.0212	(	-0.0247	to	-0.0177	)	0.000
	t2 (0 weeks to 6 weeks)		-0.1011	(	-0.2086	to	0.0064	)	0.065	-0.0712	(	-0.1942	to	0.0518	)	0.257
	t3 (6 weeks to 4 years)		-0.0247	(	-0.0333	to	-0.0161	)	0.000	-0.0259	(	-0.0363	to	-0.0155	)	0.000
	intercept		69.20	(	68.50	to	69.90	)	0.000	77.22	(	75.46	to	78.98	)	0.000
Combined	t1 (-4 years to 0 weeks)		0.0087	(	0.0041	to	0.0132	)	0.000	0.0087	(	0.0033	to	0.0142	)	0.002
	t2 (0 weeks to 6 weeks)		0.2690	(	0.1581	to	0.3799	)	0.000	0.3143	(	0.1836	to	0.4449	)	0.000
	t3 (6 weeks to 4 years)		0.0115	(	0.0059	to	0.0172	)	0.000	0.0096	(	0.0029	to	0.0163	)	0.005
	age					ref							ref			
		50-59	-0.3952	(	-2.0345	to	1.2441	)	0.637	0.6467	(	-1.2340	to	2.5274	)	0.500
		60-69	-2.9903	(	-4.6398	to	-1.3408	)	0.000	-2.2956	(	-4.1842	to	-0.4071	)	0.017
		70-79	-7.0115	(	-8.4678	to	-5.5551	)	0.000	-5.6614	(	-7.3847	to	-3.9380	)	0.000
		80-89	-11.5164	(	-12.8868	to	-10.1460	)	0.000	-8.9142	(	-10.5603	to	-7.2681	)	0.000
	age*t1					ref							ref			
		50-59	-0.0118	(	-0.0185	to	-0.0051	)	0.001	-0.0140	(	-0.0222	to	-0.0059	)	0.001
		60-69	-0.0178	(	-0.0245	to	-0.0111	)	0.000	-0.0173	(	-0.0253	to	-0.0094	)	0.000
		70-79	-0.0268	(	-0.0327	to	-0.0209	)	0.000	-0.0262	(	-0.0334	to	-0.0190	)	0.000
		80-89	-0.0302	(	-0.0358	to	-0.0245	)	0.000	-0.0300	(	-0.0370	to	-0.0231	)	0.000
	age*t2					ref							ref			
		50-59	-0.0181	(	-0.1836	to	0.1474	)	0.830	-0.0211	(	-0.2206	to	0.1784	)	0.836
		60-69	-0.1968	(	-0.3627	to	-0.0308	)	0.020	-0.2157	(	-0.4130	to	-0.0185	)	0.032
		70-79	-0.3673	(	-0.5245	to	-0.2101	)	0.000	-0.4412	(	-0.6322	to	-0.2502	)	0.000
		80-89	-0.3693	(	-0.5271	to	-0.2114	)	0.000	-0.3747	(	-0.5685	to	-0.1809	)	0.000
	age*t3					ref							ref			
		50-59	-0.0075	(	-0.0159	to	0.0010	)	0.083	-0.0038	(	-0.0140	to	0.0064	)	0.466
		60-69	-0.0080	(	-0.0169	to	0.0008	)	0.074	-0.0068	(	-0.0173	to	0.0036	)	0.201
		70-79	-0.0165	(	-0.0257	to	-0.0074	)	0.000	-0.0120	(	-0.0231	to	-0.0009	)	0.033
		80-89	-0.0359	(	-0.0458	to	-0.0261	)	0.000	-0.0361	(	-0.0482	to	-0.0240	)	0.000
	intercept		80.7	(	79.6	to	81.8	)	0.000	85.91	(	84.17	to	87.64	)	0.000

ref=reference group. All estimates come from mixed effects models (restricted maximum likelihood and unstructured covariance matrix), adjusted for sex, social deprivation (Townsend) and first dose when indicated. Intercept was set at the very beginning of the observation period.

10.4.5 Appendix 7E: Table 7E (for olanzapine)

Table 7E. Trajectories of weight change by dose, and interaction between dose and time for the olanzapine cohort.

Weight Trajectories by DOSE		Olanzapine																			
		Unadjusted (total=6,776) (low-dose=3,280; high-dose=1,643; missing dose=1,853; combined=4,923)							Adjusted (CCA) (total=6,776) (low-dose=3,280; high-dose=1,643; missing dose=1,853; combined=4,923)							Adjusted (MI-JOMO) (total=6,776) (low-dose=4,403; high-dose=2,323; combined=6,776)						
		Beta		95% CI				p	Beta		95% CI				p	Beta		95% CI				p
Low dose	t1 (-4 years to 0 weeks)	-0.0134	(	-0.0163	to	-0.0106	)	0.000	-0.0140	(	-0.0168	to	-0.0111	)	0.000	-0.0145	(	-0.0171	to	-0.0119	)	0.000
	t2 (0 weeks to 6 weeks)	0.2893	(	0.2158	to	0.3629	)	0.000	0.2932	(	0.2197	to	0.3667	)	0.000	0.3113	(	0.2433	to	0.3793	)	0.000
	t3 (6 weeks to 4 years)	0.0096	(	0.0058	to	0.0134	)	0.000	0.0095	(	0.0057	to	0.0133	)	0.000	0.0099	(	0.0064	to	0.0134	)	0.000
	intercept	74.27	(	73.60	to	74.95	)	0.000	80.60	(	68.50	to	92.71	)	0.000	96.93	(	83.98	to	109.89	)	0.000
High dose	t1 (-4 years to 0 weeks)	-0.0116	(	-0.0163	to	-0.0069	)	0.000	-0.0119	(	-0.0166	to	-0.0072	)	0.000	-0.0120	(	-0.0162	to	-0.0078	)	0.000
	t2 (0 weeks to 6 weeks)	0.5017	(	0.3950	to	0.6084	)	0.000	0.4945	(	0.3881	to	0.6009	)	0.000	0.4517	(	0.3548	to	0.5487	)	0.000
	t3 (6 weeks to 4 years)	0.0100	(	0.0051	to	0.0148	)	0.000	0.0099	(	0.0050	to	0.0147	)	0.000	0.0100	(	0.0055	to	0.0145	)	0.000
	intercept	78.29	(	77.24	to	79.35	)	0.000	82.21	(	62.49	to	101.93	)	0.000	96.01	(	75.89	to	116.12	)	0.000
Missing dose	t1 (-4 years to 0 weeks)	-0.0140	(	-0.0180	to	-0.0099	)	0.000	-0.0145	(	-0.0186	to	-0.0105	)	0.000							
	t2 (0 weeks to 6 weeks)	0.3491	(	0.2409	to	0.4574	)	0.000	0.3604	(	0.2525	to	0.4683	)	0.000							
	t3 (6 weeks to 4 years)	0.0111	(	0.0057	to	0.0165	)	0.000	0.0108	(	0.0053	to	0.0162	)	0.000							
	intercept	75.73	(	74.79	to	76.67	)	0.000	89.02	(	72.11	to	105.94	)	0.000							
Combined	t1 (-4 years to 0 weeks)	-0.0133	(	-0.0163	to	-0.0104	)	0.000	-0.0138	(	-0.0167	to	-0.0108	)	0.000	-0.0144	(	-0.0171	to	-0.0117	)	0.000
	t2 (0 weeks to 6 weeks)	0.2896	(	0.2152	to	0.3639	)	0.000	0.2914	(	0.2172	to	0.3656	)	0.000	0.3100	(	0.2415	to	0.3786	)	0.000
	t3 (6 weeks to 4 years)	0.0095	(	0.0057	to	0.0132	)	0.000	0.0093	(	0.0055	to	0.0130	)	0.000	0.0097	(	0.0062	to	0.0131	)	0.000
	high-dose	4.0634	(	2.8449	to	5.2820	)	0.000	0.8732	(	-0.3106	to	2.0570	)	0.148	0.8529	(	-0.6017	to	2.3075	)	0.246
	high-dose*t1	0.0014	(	-0.0039	to	0.0066	)	0.615	0.0012	(	-0.0041	to	0.0065	)	0.650	0.0018	(	-0.0033	to	0.0070	)	0.481
	high-dose*t2	0.2182	(	0.0900	to	0.3463	)	0.001	0.2117	(	0.0837	to	0.3396	)	0.001	0.1496	(	0.0250	to	0.2742	)	0.019
	high-dose*t3	0.0006	(	-0.0056	to	0.0068	)	0.858	0.0007	(	-0.0055	to	0.0069	)	0.834	0.0005	(	-0.0056	to	0.0065	)	0.883
	intercept	74.3	(	73.6	to	75.0	)	0.000	80.39	(	70.28	to	90.50	)	0.000	96.47	(	86.31	to	106.62	)	0.000

All estimates come from mixed effects models (restricted maximum likelihood and unstructured covariance matrix), and were adjusted for sex, social deprivation (Townsend) and age when indicated. Missing dose was handled with complete case analysis (CCA) and multilevel multiple imputation (MI-JOMO). Results from both methods are visible in the table. Intercept was set at the very beginning of the observation period.

10.4.6 Appendix 7F: Table 7F (for quetiapine)

Table 7F. Trajectories of weight change by dose, and interaction between dose and time for the quetiapine cohort.

Weight Trajectories by DOSE		Quetiapine																			
		Unadjusted (total=14,970) (low-dose=6,856; high-dose=2,094; missing dose=6,020; combined=8,950)							Adjusted (CCA) (total=14,970) (low-dose=6,856; high-dose=2,094; missing dose=6,020; combined=8,950)							Adjusted (MI-JOMO) (total=14,970) (low-dose=11,023; high-dose=3,882; combined=14,970)						
		Beta	95% CI					p	Beta	95% CI					p	Beta	95% CI				p
Low dose	t1 (-4 years to 0 weeks)	-0.0114	(	-0.0134	to	-0.0094	)	0.000	-0.0116	(	-0.0136	to	-0.0096	)	0.000	-0.0113	(	-0.0130	to	-0.0097	)	0.000
	t2 (0 weeks to 6 weeks)	0.0430	(	-0.0112	to	0.0973	)	0.120	0.0467	(	-0.0073	to	0.1008	)	0.090	0.0661	(	0.0217	to	0.1105	)	0.004
	t3 (6 weeks to 4 years)	-0.0025	(	-0.0057	to	0.0006	)	0.117	-0.0029	(	-0.0061	to	0.0002	)	0.068	0.0006	(	-0.0020	to	0.0032	)	0.649
	intercept	75.54	(	75.03	to	76.04	)	0.000	90.22	(	81.17	to	99.27	)	0.000	102.38	(	93.84	to	110.91	)	0.000
High dose	t1 (-4 years to 0 weeks)	0.0026	(	-0.0016	to	0.0068	)	0.226	0.0023	(	-0.0019	to	0.0065	)	0.277	-0.0025	(	-0.0058	to	0.0007	)	0.130
	t2 (0 weeks to 6 weeks)	0.2870	(	0.1884	to	0.3855	)	0.000	0.2852	(	0.1868	to	0.3835	)	0.000	0.2091	(	0.1311	to	0.2870	)	0.000
	t3 (6 weeks to 4 years)	0.0060	(	0.0011	to	0.0109	)	0.015	0.0059	(	0.0011	to	0.0108	)	0.017	0.0061	(	0.0020	to	0.0102	)	0.004
	intercept	80.97	(	80.00	to	81.94	)	0.000	74.09	(	55.27	to	92.91	)	0.000	91.55	(	76.26	to	106.85	)	0.000
Missing dose	t1 (-4 years to 0 weeks)	-0.0099	(	-0.0121	to	-0.0077	)	0.000	-0.0100	(	-0.0122	to	-0.0079	)	0.000							
	t2 (0 weeks to 6 weeks)	0.1051	(	0.0495	to	0.1608	)	0.000	0.1011	(	0.0456	to	0.1566	)	0.000							
	t3 (6 weeks to 4 years)	0.0061	(	0.0029	to	0.0094	)	0.000	0.0061	(	0.0028	to	0.0093	)	0.000							
	intercept	77.37	(	76.83	to	77.91	)	0.000	91.52	(	82.13	to	100.90	)	0.000							
Combined	t1 (-4 years to 0 weeks)	-0.0114	(	-0.0135	to	-0.0093	)	0.000	-0.0116	(	-0.0137	to	-0.0095	)	0.000	-0.0113	(	-0.0130	to	-0.0096	)	0.000
	t2 (0 weeks to 6 weeks)	0.0496	(	-0.0065	to	0.1057	)	0.083	0.0509	(	-0.0051	to	0.1068	)	0.075	0.0690	(	0.0238	to	0.1143	)	0.003
	t3 (6 weeks to 4 years)	-0.0026	(	-0.0058	to	0.0007	)	0.120	-0.0032	(	-0.0064	to	0.0001	)	0.055	0.0004	(	-0.0022	to	0.0031	)	0.751
	high-dose	5.4369	(	4.3763	to	6.4976	)	0.000	-0.7762	(	-1.8154	to	0.2630	)	0.143	-0.5034	(	-1.5478	to	0.5411	)	0.341
	high-dose*t1	0.0140	(	0.0097	to	0.0183	)	0.000	0.0138	(	0.0095	to	0.0181	)	0.000	0.0088	(	0.0051	to	0.0126	)	0.000
	high-dose*t2	0.2329	(	0.1265	to	0.3393	)	0.000	0.2308	(	0.1247	to	0.3369	)	0.000	0.1363	(	0.0438	to	0.2289	)	0.004
	high-dose*t3	0.0086	(	0.0029	to	0.0143	)	0.003	0.0093	(	0.0036	to	0.0149	)	0.001	0.0059	(	0.0007	to	0.0110	)	0.025
	intercept	75.5	(	75.0	to	76.0	)	0.000	85.63	(	77.76	to	93.51	)	0.000	99.27	(	92.18	to	106.37	)	0.000

All estimates come from mixed effects models (restricted maximum likelihood and unstructured covariance matrix), and were adjusted for sex, social deprivation (Townsend) and age when indicated. Missing dose was handled with complete case analysis (CCA) and multilevel multiple imputation (MI-JOMO). Results from both methods are visible in the table. Intercept was set at the very beginning of the observation period.

10.4.7 Appendix 7G: Table 7G (for risperidone)

Table 7G. Trajectories of weight change by dose, and interaction between dose and time for the risperidone cohort.

Weight Trajectories by DOSE		Risperidone																			
		Unadjusted (total=7,293) (low-dose=4,299; high-dose=462; missing dose=2,532; combined=4,761)						Adjusted (CCA) (total=7,293) (low-dose=4,299; high-dose=462; missing dose=2,532; combined=4,761)						Adjusted (MI-JOMO) (total=7,293) (low-dose=6,365; high-dose=3,964; combined=7,293)					
		Beta	95% CI				p	Beta	95% CI				p	Beta	95% CI				p
Low dose	t1 (-4 years to 0 weeks)	-0.0123	(	-0.0147	to	-0.0099	) 0.000	-0.0126	(	-0.0150	to	-0.0102	) 0.000	-0.0127	(	-0.0147	to	-0.0107	) 0.000
	t2 (0 weeks to 6 weeks)	0.1062	(	0.0390	to	0.1734	) 0.002	0.1006	(	0.0335	to	0.1676	) 0.003	0.0728	(	0.0169	to	0.1287	) 0.011
	t3 (6 weeks to 4 years)	0.0008	(	-0.0032	to	0.0048	) 0.683	0.0009	(	-0.0031	to	0.0049	) 0.655	0.0009	(	-0.0024	to	0.0042	) 0.602
	intercept	74.77	(	74.17	to	75.38	) 0.000	87.34	(	76.34	to	98.34	) 0.000	97.23	(	86.47	to	107.99	) 0.000
High dose	t1 (-4 years to 0 weeks)	0.0065	(	-0.0020	to	0.0149	) 0.132	0.0059	(	-0.0025	to	0.0144	) 0.166	-0.0012	(	-0.0087	to	0.0063	) 0.754
	t2 (0 weeks to 6 weeks)	0.1706	(	-0.0339	to	0.3751	) 0.102	0.1731	(	-0.0313	to	0.3776	) 0.097	0.1338	(	-0.0656	to	0.3332	) 0.188
	t3 (6 weeks to 4 years)	0.0070	(	-0.0020	to	0.0160	) 0.126	0.0070	(	-0.0020	to	0.0160	) 0.126	0.0069	(	-0.0023	to	0.0160	) 0.141
	intercept	80.57	(	78.61	to	82.54	) 0.000	87.89	(	45.36	to	130.41	) 0.000	85.03	(	44.46	to	125.60	) 0.000
Missing dose	t1 (-4 years to 0 weeks)	-0.0124	(	-0.0155	to	-0.0092	) 0.000	-0.0127	(	-0.0159	to	-0.0096	) 0.000						
	t2 (0 weeks to 6 weeks)	0.0257	(	-0.0642	to	0.1157	) 0.575	0.0094	(	-0.0801	to	0.0989	) 0.837						
	t3 (6 weeks to 4 years)	0.0013	(	-0.0041	to	0.0067	) 0.633	0.0014	(	-0.0040	to	0.0068	) 0.615						
	intercept	74.88	(	74.08	to	75.69	) 0.000	88.25	(	72.95	to	103.56	) 0.000						
Combined	t1 (-4 years to 0 weeks)	-0.0123	(	-0.0148	to	-0.0099	) 0.000	-0.0127	(	-0.0151	to	-0.0103	) 0.000	-0.0128	(	-0.0148	to	-0.0108	) 0.000
	t2 (0 weeks to 6 weeks)	0.1082	(	0.0400	to	0.1764	) 0.002	0.1032	(	0.0351	to	0.1713	) 0.003	0.0743	(	0.0173	to	0.1312	) 0.011
	t3 (6 weeks to 4 years)	0.0008	(	-0.0032	to	0.0047	) 0.706	0.0007	(	-0.0033	to	0.0047	) 0.729	0.0008	(	-0.0026	to	0.0041	) 0.654
	high-dose	5.7331	(	3.7512	to	7.7150	) 0.000	-1.4312	(	-3.3492	to	0.4867	) 0.144	-0.6345	(	-3.1030	to	1.8341	) 0.608
	high-dose*t1	0.0195	(	0.0115	to	0.0274	) 0.000	0.0198	(	0.0118	to	0.0278	) 0.000	0.0121	(	0.0045	to	0.0197	) 0.002
	high-dose*t2	0.0504	(	-0.1469	to	0.2477	) 0.617	0.0454	(	-0.1517	to	0.2425	) 0.652	0.0456	(	-0.1615	to	0.2527	) 0.664
	high-dose*t3	0.0061	(	-0.0039	to	0.0162	) 0.231	0.0065	(	-0.0035	to	0.0165	) 0.205	0.0064	(	-0.0038	to	0.0167	) 0.217
	intercept	74.8	(	74.2	to	75.4	) 0.000	86.40	(	75.94	to	96.86	) 0.000	95.75	(	85.74	to	105.75	) 0.000

All estimates come from mixed effects models (restricted maximum likelihood and unstructured covariance matrix), and were adjusted for sex, social deprivation (Townsend) and age when indicated. Missing dose was handled with complete case analysis (CCA) and multilevel multiple imputation (MI-JOMO). Results from both methods are visible in the table. Intercept was set at the very beginning of the observation period.

10.4.8 Appendix 7H: Tables 7H-Sex, 7H-Dose and 7H-Age (for all drugs)

Table 7H-Sex. Expected weight gain for an average patient prescribed a particular antipsychotic, stratified by sex.

Drug	Sex	N	Weight gained during short-time (0-6 weeks) in Kilograms	95% CI	Weight gained during long-time (6weeks - 4 years) in Kilograms	95% CI	Total weight gained
OLANZAPINE	Women	3,803	1.8	(1.4 to 2.2)	2.5	(1.8 to 3.2)	4.3
	Men	2,973	2.7	(2.2 to 3.2)	1.3	(0.6 to 2.1)	4
QUETIAPINE	Women	8,765	0.8	(0.5 to 1.1)	0.8	(0.2 to 1.3)	1.6
	Men	6,205	0.6	(0.3 to 0.9)	-0.1	(-0.7 to 0.6)	0.5
RISPERIDONE	Women	4,127	0.4	(0 to 0.8)	0.3	(-0.5 to 1.1)	0.7
	Men	3,166	0.6	(0.1 to 1.1)	0.4	(-0.5 to 1.3)	1

Estimates come from Table 7A

Table 7H-Dose. Expected weight gain for an average patient prescribed a particular antipsychotic, stratified by dose.

Drug	Dose*	N	Weight gained during short-time (0-6 weeks) in Kilograms	95% CI	Weight gained during long-time (6weeks - 4 years) in Kilograms	95% CI	Total weight gained
OLANZAPINE	low	4,403	1.9	(1.5 to 2.3)	2	(1.3 to 2.7)	3.9
	high	2,323	2.7	(2.1 to 3.3)	2	(1.1 to 2.9)	4.7
QUETIAPINE	low	11,023	0.4	(0.1 to 0.7)	0.1	(-0.4 to 0.6)	0.5
	high	3,882	1.3	(0.8 to 1.7)	1.2	(0.4 to 2.1)	2.5
RISPERIDONE	low	6,365	0.4	(0.1 to 0.8)	0.2	(-0.5 to 0.8)	0.6
	high	3,964	0.8	(-0.4 to 1.9)	1.4	(-0.5 to 3.2)	2.2

Estimates come from Tables 7E, 7F and 7G (adjusted, MI-JOMO)

(*) Cut off point for low/high dose was: ≤5 mg for Olanzapine, ≤75 mg for Quetiapine and ≤2 mg for Risperidone.

Table 7H-Age. Expected weight gain for an average patient prescribed a particular antipsychotic, stratified by age.

Drug	Age (years)	N	Weight gained during short-time (0-6 weeks) in Kilograms	95% CI	Weight gained during long-time (6weeks - 4 years) in Kilograms	95% CI	Total weight gained
OLANZAPINE	40-49	1,468	3.8	(3.1 to 4.5)	2.2	(1.1 to 3.3)	6
	50-59	1,165	2.8	(2.1 to 3.5)	2.7	(1.6 to 3.8)	5.5
	60-69	926	1.5	(0.6 to 2.3)	2.1	(0.5 to 3.6)	3.6
	70-79	834	0.5	(-0.3 to 1.3)	0.1	(-1.4 to 1.7)	0.6
	80-89	530	-0.3	(-1.4 to 0.9)	-0.4	(-2.1 to 1.3)	-0.7
QUETIAPINE	40-49	2,325	1.6	(1.0 to 2.2)	0.5	(-0.5 to 1.5)	2.1
	50-59	1,694	1.4	(0.8 to 2.0)	1.0	(0 to 2.0)	2.4
	60-69	1,227	0.4	(-0.3 to 1.0)	0.1	(-1.1 to 1.3)	0.5
	70-79	1,656	-0.4	(-1.0 to 0.3)	-1.6	(-2.9 to -0.2)	-2
	80-89	2,048	-0.6	(-1.2 to 0.1)	-3.2	(-4.8 to -1.6)	-3.8
RISPERIDONE	40-49	961	1.7	(0.8 to 2.6)	2.0	(0.5 to 3.4)	3.7
	50-59	716	1.9	(0.9 to 2.9)	1.1	(-0.5 to 2.7)	3
	60-69	696	0.6	(-0.2 to 1.5)	0.2	(-1.3 to 1.7)	0.8

	70-79	1,016	-0.7	(-1.5 to 0.1)	-0.6	(-2.3 to 1.0)	-1.3
	80-89	1,372	-0.4	(-1.1 to 0.3)	-5.2	(-7.3 to -3.1)	-5.6

Estimates come from Tables 7B, 7C and 7D (adjusted)

10.4.9 Appendix 7I: Table 7I (for olanzapine)

Table 7I. ITS effect size of second-generation antipsychotics on weight change in the LONG-TERM, by age and dose.

		Low Dose			*High Dose*		
Antipsychotic	Age (years)	*pre-treatment (β1)*	*long-term (β3)*	*ITS effect **size (β3-β1)***	*pre-treatment (β1)*	*long-term (β3)*	*ITS effect **size (β3-β1)***
Olanzapine	40-49	0.0020	0.0130	0.0110	-0.0005	0.0093	0.0098
	50-59	-0.0088	0.0124	0.0212	-0.0200	0.0151	0.0351
	60-69	-0.0137	0.0141	0.0278	-0.0091	0.0047	0.0138
	70-79	-0.0236	-0.0008	0.0228	-0.0216	0.0027	0.0243
	80-89	-0.0269	-0.0041	0.0228	-0.0261	-0.0024	0.0237

*ITS = interrupted time series. In this approach, the **long-term effect size** can be estimated by the difference between pre-treatment trajectory (beta 1) and long-term post-treatment trajectory (beta 3). Estimates are unadjusted.*

10.5 Conference presentations, grant funding and publications related to this thesis

10.5.1 Conference presentations

I presented different components of this thesis, at different stages of progress, in the conferences listed below:

- Bazo-Alvarez JC, Petersen I, Morris T, Carpenter J. Analysing weight after antipsychotic drug treatment: understanding missing data behaviour and its impact on estimates in longitudinal electronic health records. **Poster presented** in the Symposium of Peruvian Researchers at Europe SINAPSIS, Berlin Oct 2017. https://www.sinapsis-peru.org/

- Bazo-Alvarez JC [32], Morris TP, Carpenter JR, Petersen I. Weight change after anti-psychotic drug treatment: long-term evidence from a retrospective study using electronic health records. **Poster presented** in the International Society of Pharmacovigilance 18th Annual Meeting, Geneve Nov 2018. https://www.isop2018geneva.org/programme/poster-listing.html

- Bazo-Alvarez JC [33], Morris TP, Carpenter JR, Hayes JF, Petersen I. Antipsychotic dose and weight gain: long-term evidence from a retrospective study using electronic health records. **Poster presented** in the 35th International Conference of Pharmacoepidemiology and Risk Management, Philadelphia, Aug 2019. https://onlinelibrary.wiley.com/toc/10991557/2019/28/S2

- Bazo-Alvarez JC, Petersen I, Morris TP, Carpenter JR. Handling missing data for interrupted time series analysis in longitudinal electronic health records. **Poster presented** in the 4th Symposium of Peruvian Scientists in Europe SINAPSIS, Ghent, Oct 2019. https://www.sinapsis-peru.org/

[32] Best Poster. An early version of this poster won the Best Poster Prize in the 3rd Annual Symposium of Peruvian Scientists in Europe, SINAPSIS Oct 2018, Barcelona, Spain. The presentation of the definitive version (in Geneve) was supported by the Seedcorn Funding Award given by the Research Department of Primary Care and Population Health at UCL.

[33] Travel Scholarship. I was given this scholarship by the International Society of Pharmacoepidemiology (ISPE) for presenting this poster in Philadelphia, USA.

- Bazo-Alvarez JC [34], Morris TP, Pham TM, Carpenter JR, Petersen I. How to deal with missing data in interrupted time series analysis with electronic health records. **Oral presentation accepted** for the International Conference of Pharmacoepidemiology All Access, Sep 2020 . https://www.eventscribe.com/2020/ICPEAllAccess/

10.5.2 Grant funding

During my PhD, I got two grants for funding projects that include components directly connected to my PhD thesis:

Title:	Prescribing of Antipsychotic Drugs in Older People in the UK: a cohort study using UK primary care data
Year:	2019
Amount:	GBP 29,448.93
Funder:	Funded by NHS National Institute for Health Research, School for Primary Care Research (Funding Round - 17)
Applicants:	Elizabeth Jones (PI), Irene Petersen, **JC Bazo-Alvarez**, Kate Walters, Cini Bhanu.
Connection:	This project has a component related to antipsychotic-induced weight gain in older people, which is directly connected to the Chapter 7 of my thesis.

Title:	Statins and LDL-Cholesterol in the long-term: a retrospective cohort study using electronic health records
Year:	2020
Amount:	GBP 17,069.74
Funder:	Funded by NHS National Institute for Health Research, School for Primary Care Research (Funding Round - 19)
Applicants:	**JC Bazo-Alvarez (PI)**, Irene Petersen, Wannamethee Goya, Kingshuk Pal.
Connection:	In this project, we will apply the approach I studied in this thesis (interrupted time series with mixed effects models and multilevel multiple imputation). The study will start in Oct 2020.

[34] All Access Scholarship. I was given this scholarship by the International Society of Pharmacoepidemiology (ISPE) for oral presentation in the ICPE All Access Sep 2020

10.5.3 Publications

Up to August 2020, I got one original article published and another accepted for publication:

- Bazo-Alvarez JC, Morris TP, Carpenter JR, Hayes JF, Petersen I. Effects of long-term antipsychotics treatment on body weight: A population-based cohort study. Journal of Psychopharmacology. 2020 Jan;34(1):79-85.

- Bazo-Alvarez JC, Morris TP, Pham TM, Carpenter JR, Petersen I. Handling missing values in interrupted time series analysis of longitudinal individual-level data. Accepted for publication in Clinical Epidemiology. 2020 Aug.

The first paper is based on Chapter 5 (An application of interrupted time series with mixed effect models) and the second paper is based on Chapter 6 of this thesis (Evaluating methods for missing data handling in interrupted time series analysis via simulation studies). In the next pages, I show copies of both the published and accepted papers:

Original Paper

Effects of long-term antipsychotics treatment on body weight: A population-based cohort study

Journal of Psychopharmacology
2020, Vol. 34(1) 79–85

DOI: 10.1177/0269881119885918
journals.sagepub.com/home/jop

Juan Carlos Bazo-Alvarez[1,2], Tim P Morris[3], James R Carpenter[3,4], Joseph F Hayes[5] and Irene Petersen[1,6]

Abstract

Background: Antipsychotics are often prescribed for long-term periods, however, most evidence of their impact on body weight comes from short-term clinical trials. Particularly, impact associated with dosage has been barely studied.

Aims: The aim of this study was to describe the short- and long-term change in body weight of people initiated on high or low doses of the three most commonly prescribed second-generation antipsychotics.

Methods: Retrospective cohorts of individuals with a diagnosed psychotic disorder observed from 2005 to 2015 in the UK primary care. The exposure was the first prescription of olanzapine, quetiapine or risperidone. The main outcome was change in body weight four years before and four years after initiation of antipsychotic treatment, stratified on sex and 'low' or 'high' dose.

Results: In total, 22,306 women and 16,559 men were observed. Olanzapine treatment was associated with the highest change in weight, with higher doses resulting in more weight gain. After 4 years, given a high dose of olanzapine (>5 mg), women gained on average $+6.1$ kg; whereas given a low dose ($\leqslant 5$ mg), they gained $+4.4$ kg. During the first six weeks of olanzapine treatment, they gained on average $+3.2$ kg on high dose and $+1.9$ kg on low dose. The trends were similar for men. Individuals prescribed risperidone and quetiapine experienced less weight gain in both the short- and long-term.

Conclusions: Olanzapine treatment was associated with the highest increase in weight. Higher doses were associated with more weight gain. Doctors should prescribe the lowest effective dose to balance mental-health benefits, weight gain and other adverse effects.

Keywords

Antipsychotic agents, dopamine, serotonin, noradrenaline, weight gain, electronic health records, interrupted time series analysis

Introduction

Overweight and obesity is a worldwide problem that impacts severely on population health (Newcomer and Haupt, 2006). Since the prevalence of overweight and obesity is higher in individuals with severe mental illnesses than in the general population (Elmslie et al., 2000; Holt and Peveler, 2009), their risk of harmful consequences is also higher (Hayes et al., 2017; Osborn et al., 2007). Individuals with severe mental illnesses are more susceptible to developing metabolic syndrome, type-2 diabetes mellitus (De Hert et al., 2006) and cardiovascular diseases (Emul and Kalelioglu, 2015; Osborn et al., 2007), leading to a higher risk of death. Adults with schizophrenia have three and a half times the mortality risk than the general population, with cardiovascular diseases the most common cause (Olfson et al., 2015; Osborn et al., 2007). Particularly, Lahti et al. demonstrated that the risk of death is higher in women than in men with schizophrenia (Lahti et al., 2012), suggesting that differences between sexes need to be further investigated.

Second-generation antipsychotics (AP) are a known cause of weight gain (Bak et al., 2014; Osborn et al., 2018). Some evidence suggests that women gain more weight than men during AP treatment (Seeman, 2008). One study suggested that women have five times the odds of increasing body mass index (BMI) compared with men after a period of two years or more (Koga, 2003). Of women treated with clozapine, 29% gained $\geqslant 20\%$ of their baseline body weight after two years of follow up, in contrast to 13% of men (Covell et al., 2004). Other studies have demonstrated similar differences between men and women (Gebhardt et al., 2009; Najar et al., 2017). However, most of these studies are based on small sample sizes of less than 200 individuals, and most do not distinguish between short- and long-term weight gain associated with antipsychotic treatment.

Weight gain after initiation of antipsychotic treatment may also depend on body weight when treatment is initiated. Thus, Gebhardt et al. found that low BMI before first AP treatment

[1]Research Department of Primary Care and Population Health, University College London, London, UK
[2]Instituto de Investigación, Universidad Católica los Ángeles de Chimbote, Peru
[3]MRC Clinical Trials Unit at University College London, London, UK
[4]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK
[5]Division of Psychiatry, University College London, London, UK
[6]Department of Clinical Epidemiology, Aarhus University, Denmark

Corresponding author:
Juan Carlos Bazo-Alvarez, Research Department of Primary Care and Population Health, University College London, Rowland Hill Street, London, NW3 2PF, UK.
Email: juan.alvarez.16@ucl.ac.uk

predicted a faster increment of BMI after treatment initiation (Gebhardt et al., 2009), and a similar conclusion was reached by Najar et al. (2017). On the other hand, we have limited information about how doses of antipsychotic treatment are associated with weight gain (Bak et al., 2014).

In this study, our aim was to investigate the change in body weight of patients initiated with high or low doses of the three most commonly prescribed second-generation antipsychotics: olanzapine, risperidone and quetiapine. Our objectives were to evaluate:

(1) the short- and long-term change in body weight in men and women upon initiation of AP;
(2) whether this is different for low and high doses;
(3) whether low body weight at treatment initiation had the greatest weight gain.

Methods

Data source

We used anonymized, longitudinal patients' records from The Health Improvement Network (THIN), a database that comprises information from UK primary care electronic health records from general practices (Roland et al., 2012). THIN integrates more than 12 million patients from 711 general practices, including demographic data (sex, year of birth and indicator of social deprivation (quintiles of Townsend score)) and clinical data. The clinical data are recorded using the hierarchical Read Code system (Chisholm, 1990). In the UK, more than 95% are registered with a general practice, THIN is roughly representative of the UK population and has been previously used for reporting health indicators at the national level (Blak et al., 2011). For this study, we included data from all practices after they have been deemed to be operating at the standard of acceptable computer usage (Horsfall et al., 2013) and whose reported mortality rate was consistent with national statistics (Maguire et al., 2009).

Study population

At the individual level, we included all patients aged between 18 and 99 years at the date they started their first treatment with olanzapine, risperidone or quetiapine; between 1 January 2005 and 31 December 2015. We included patients with a diagnosed psychiatric disorder (schizophrenia, bipolar disorder, other non-affective psychoses, borderline personality disorder, anxiety, depression or dementia) who had at least one further prescription of the same AP within three months after the first prescription. We judged that these individuals were more likely to have initiated treatment than those with a single prescription. Patients who had been initiated on more than one type of AP were excluded (including switchers). A few individuals had no records of year of birth, sex or social deprivation records and were thus excluded from our study. Likewise, we excluded individuals with no available data 12 months before the date of initiation of antipsychotic treatments since they may have initiated antipsychotic treatment elsewhere.

Variables and measurements

The exposure of interest was the initiation of olanzapine, risperidone or quetiapine prescription. In the Neuroscience-based Nomenclature olanzapine is a dopamine and serotonin receptor antagonist, risperidone is a dopamine, serotonin and norepinephrine receptor antagonist, and quetiapine is a dopamine and serotonin receptor antagonist and norepinephrine reuptake inhibitor (Nutt and Blier, 2016). The outcome was body weight, measured in kilograms. The main covariates were sex (women/men) and first prescribed dose of AP (hereafter called 'first dose'). All AP reported first doses in milligrams, but we used the dose-equivalence approach of Woods (Woods, 2003) for defining cut-off points of low/high first dose: $\leqslant 5$ mg for olanzapine, $\leqslant 75$ mg for quetiapine and $\leqslant 2$ mg for risperidone. Using the '2 mg of haloperidol equals 100 mg of chlorpromazine' convention as reference, Woods (2003) explored available evidence for identifying the minimum effective dose across olanzapine, quetiapine and risperidone, defining this dose equivalence. The first dose is a good predictor of all subsequent doses prescribed during treatment; thus, over time, patients usually stay in a dose range close to the first dose they were prescribed (data not shown). We also retrieved information on age, height, social deprivation (Townsend score 1–5, from least to most deprived), smoking and drinking status, having a type-2 diabetes mellitus diagnosis, systolic blood pressure (SBP), low-density lipoprotein cholesterol (LDL-cholesterol) and high-density lipoprotein cholesterol (HDL-cholesterol); recorded within first year of initiation of treatment. This information served mostly for sample characterization; only sex, age, type-2 diabetes mellitus diagnosis and social deprivation were fully observed.

Statistical analysis

We used an interrupted time series approach (Bernal et al., 2017) to analyse weight change over time, with one model for each of the three AP initiation cohorts by sex (six models in total, one per drug per sex). We modelled weight change over time using continuous linear splines with random intercept and slopes models (unstructured covariance, restricted maximum likelihood), from which three slopes of weight change were estimated for: (a) –4 years to baseline (pre-treatment), (b) baseline to +6 weeks (short-term), (c) +6 weeks to +4 years (long-term). Differences between slopes served to describe weight change after AP treatment initiation, both crude and adjusted for age and social deprivation (objective 1). The correlation between average weight at baseline (intercept) and short-term gradient of change (short-term slope) was estimated, as it provided an estimate for whether individuals with lower weight at baseline gain more or less weight after AP treatment initiation than individuals with higher body weight. Negative correlations mean that individuals with low weight gain more weight during the short-term period and vice versa (objective 3). Our main analysis was performed after stratifying each of cohorts according to low/high first dose. This was to examine whether the gradient of weight change after treatment varies between low/high first doses of AP (objective 2). For all these models, the Intraclass Correlation Coefficient (ICC) was reported. Likelihood ratio tests were performed to compare the goodness of fit between the nested models. We assumed weight records were missing at random within strata, conditional on observed weights, so that modelling the observed data over time provides unbiased estimates (van Buuren, 2012). We also assumed missing data on dose was missing at random, so that the complete case analysis we performed provides unbiased estimates (White and Carlin, 2010). Model assessment included evaluation of residuals and a visual exploration of average and

individual trajectories. Although the chosen impact model (linear splines with knots at baseline and +6 weeks) was informed by both the clinical criteria and evidence (Bak et al., 2014), we also performed a sensitivity analysis following the suggestions from Lopez Bernal et al. (Bernal et al., 2017). This sensitivity analysis consisted of comparing our preferred linear spline model against another feasible impact model, a restricted cubic spline model (knots again at baseline and +6 weeks), using graphical and analytical tools (see Figure S1 in supplemental material). Estimates are given with 95% confidence intervals (CIs). All the statistical analyses were performed using Stata 15 for Windows (Stata, 2017).

Results

In total, we included 16,559 men and 22,306 women in the study. The median number ± interquartile range of weight measurements within individual trajectories over eight years of observation were 6 ± 7 and 8 ± 10 (olanzapine cohorts), and 7 ± 8 and 8 ± 9 (quetiapine and risperidone cohorts) for men and women respectively. Characteristics of the individuals are summarized in Figure 1 and are provided in more detail in Table S1. On average, at initiation of treatment, men were younger than women prescribed olanzapine (men = 47.5 years ±17.8 SD, women = 54.0 years ±19.5 SD) and risperidone (men = 56.6 years ±22.1 SD, women = 64.5 years ±21.8 SD), but were of similar age in the quetiapine cohort (men = 56.5 years ±20.7 SD, women = 56.1 years ±22.1 SD). On average, men were prescribed higher dose of olanzapine (+1 mg), quetiapine (+10 mg) and risperidone (+0.3 mg).

In the short (< 6 weeks) and long term (⩾ 6 weeks to ⩽ 4 years), individuals treated with any of the three AP drugs gained weight, especially those patients prescribed olanzapine. Pre-treatment weight change was negligible for quetiapine (women and men) and risperidone (men only) cohorts, and slightly negative for the rest of cohorts. In the short-term after olanzapine initiation, men's weight increased by 0.569 kg/week (3.4 kg over the first six weeks) and women's weight increased by 0.382 kg/week (2.3 kg over the first six weeks) (Tables 1 and S2). Individuals initiated on quetiapine and risperidone also gained weight shortly after initiation of treatment, but to a lesser extent (Tables 1 and S2, and Figure 2). Individuals continued to gain weight after six weeks, but at a slower rate than the first six weeks. For example, for women initiated on olanzapine, long-term weight gain was estimated to be 0.014 kg/week (0.7 kg per year) (Tables 1 and S2, and Figure 2). Women who were initiated on olanzapine were in general slightly lighter (69.7 kg) than women initiated on risperidone (73.3 kg) and quetiapine (70.1 kg), but there was not much difference for the men (weight at baseline, see Figure 1 and Table S2). Women who had a lower weight before initiation of olanzapine gained more weight in the short term than women who had a higher weight (correlation between intercept and slope = –0.068, 95% CI: –0.121 to −0.014); a similar effect was observed for men (correlation between intercept and slope = –0.050, 95% CI: –0.113 to +0.014) (Table S2).

The weight gain in individuals who were initiated on high dose of AP was greater than those initiated on low dose. When olanzapine was initiated at high dose (> 5 mg), women gained +0.534 kg/week (+3.2 kg over 6 weeks) and men +0.743 kg/week (+4.5 kg over 6 weeks) compared with low-dose gain of +0.314 kg/week (+1.9 kg over 6 weeks) for women and +0.425 kg/week (+2.6 kg over 6 weeks) for men (Tables 1 and S3). The short-term effect of initiation of quetiapine was also stronger for those given high doses (> 75 mg) (women +2.3 kg and men +1.6 kg, both over 6 weeks) than given low doses (women +0.7 kg and men +0.5 kg, both over 6 weeks). However, there was a relatively small difference for those initiated on risperidone low doses (⩽ 2 mg) (+1.0 kg over 6 weeks for both women and men) and high doses (women +1.1 kg and men 1.9 kg, both over 6 weeks). In the short-term, those given low doses of olanzapine tended to gain more weight as their weight at baseline was lower (women: correlation between intercept and slope = –0.155, 95% CI: –0.230 to −0.078; men: correlation between intercept and slope = –0.135, 95% CI: –0.235 to −0.033).

Cumulative weight gain in the long-term was particularly high in patients prescribed olanzapine but, for any drug, people did not on average lose the extra-weight they gained during the short-term (Table 1 and S3). For example, after four years from the first olanzapine prescription, a typical woman gained 2.3 kg (short-term, 95% CI: 1.9–2.7 kg) + 2.8 kg (long-term, 95% CI: 2.2–3.5 kg) = 5.1 kg (total) whereas a typical man gained 3.4 kg (short-term, 95%CI: 3.0–3.8 kg) + 1.7 kg (long-term, 95% CI: 0.9–2.4 kg) = 5.1 kg (total) of AP induced extra-weight. The prescribed dose of olanzapine was also critical, particularly for women in the long-term. For example, given a low dose (< 5 mg), women gained 1.9 + 2.5 = 4.4 kg after four years; given a high dose (> 5 mg), women gained 3.2 + 2.9 = 6.1 kg. A similar impact of higher doses was observed for quetiapine and risperidone (Table 1 and S3).

Discussion

This retrospective cohort study reports data from patients seen in primary care, before and after AP treatment initiation. Pre-treatment weight change was insignificant or slightly negative for all cohorts during four years before baseline. Individuals starting treatment with any AP gained weight on average, especially those patients prescribed olanzapine. Weight gain was much more rapid in the short-term than in the long-term. People who were initiated on high-dose AP experienced much greater absolute weight gain than those initiated on low dose AP. Cumulative weight gain during the long-term was particularly high in individuals treated with olanzapine but, for all APs, people typically never lost the extra weight they gained during the first six weeks of AP treatment.

Strengths and limitations of this study

This study presents evidence from a large sample ($n > 38{,}000$) of people prescribed antipsychotic medications, taken from a population which is broadly representative of the UK (Blak et al., 2011). Patients prescribed antipsychotics are often treated for long periods, and so quantifying the risk of long-term side effects is particularly important. Clinical trials invariably fail to do this because of their short durations and much smaller sample size, so our study provides a necessary long-term perspective. We applied an analysis approach that has not been used previously in assessing AP-induced weight gain. A major advantage of our approach is that it includes pre-treatment weight change information, so

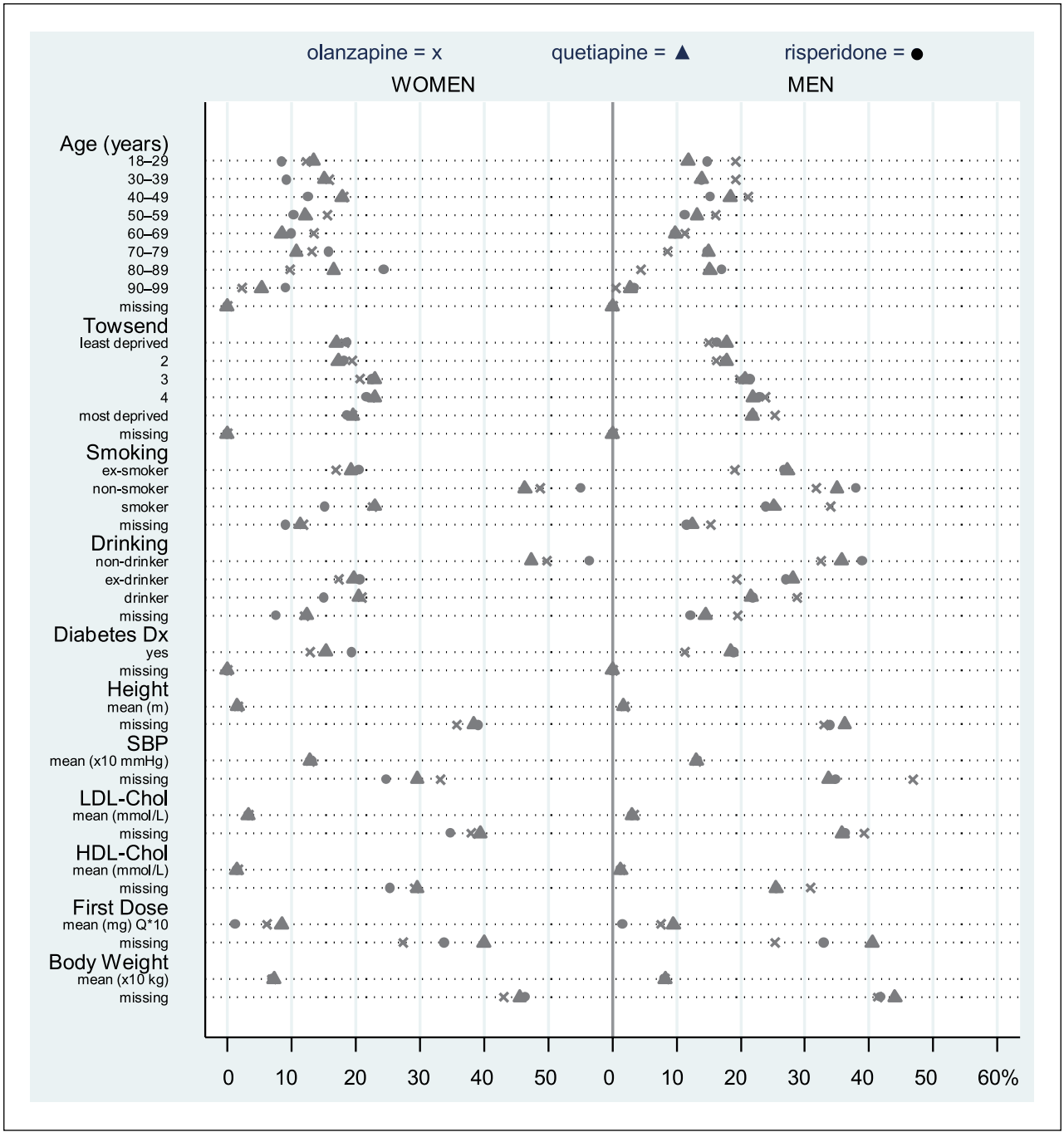


Figure 1. Baseline characteristics of patients from olanzapine, quetiapine and risperidone cohorts, stratified by sex. From height onwards, some continuous variables changed their scale as labelled below their names.

patients act as their own controls in the analysis and any additional weight change after baseline is attributable to the AP treatment. The approach utilizes all individual weight records at their time of measurement, therefore avoiding the loss of information seen in previous studies which categorize outcomes or use period means or incidence rates as summary measures (Bak et al., 2014; Osborn et al., 2018). Our longitudinal model-based approach also accounts for missing weight records – assuming weight recording is missing at random within strata, conditional on observed weight measurements (Haneuse et al., 2016) – while incorporating informative pre-baseline weight data. From this method we expect unbiased estimates if data were missing at random (Molenberghs et al., 2014), a property that is not ensured by complete case analyses applied elsewhere (Bushe et al., 2012; Osborn et al., 2018). Following standard recommendations (Bernal et al., 2017), we guaranteed good statistical power by having equal periods of observation before and after baseline, and a large sample size. Additional analyses showed that our proposed linear

Table 1. Expected weight gain for an average patient prescribed a particular antipsychotic, stratified by dose and sex.

Drug	Sex	*n*[a]	Dose[b]	Weight gained during **short-time** (0–6 weeks), kg	95% CI	Weight gained during **long-time** (6 weeks–4 years), kg	95% CI	Total weight gained
OLANZAPINE (*n* = 9499)	Women	5004	Overall	2.3	(1.9–2.7)	2.8	(2.2–3.5)	5.1
		2535	Low	1.9	(1.4–2.4)	2.5	(1.6–3.3)	4.4
		1100	High	3.2	(2.4–4.0)	2.9	(1.6–4.2)	6.1
	Men	4495	Overall	3.4	(3.0–3.8)	1.7	(0.9–2.4)	5.1
		1887	Low	2.6	(2.0–3.2)	1.9	(0.8–3.0)	4.5
		1470	High	4.5	(3.6–5.3)	1.4	(0.2–2.7)	5.9
QUETIAPINE (*n* = 19,965)	Women	12,149	Overall	1.2	(1.0–1.5)	1.1	(0.6–1.6)	2.3
		5372	Low	0.7	(0.3–1.0)	0.9	(0.1–1.6)	1.6
		1912	High	2.3	(1.6–2.9)	1.6	(0.4–2.7)	3.9
	Men	7816	Overall	0.8	(0.4–1.1)	0.7	(0.1–1.3)	1.5
		3326	Low	0.5	(0.0–0.9)	–0.7	(–1.8–0.3)	–0.3
		1326	High	1.6	(0.9–2.4)	1.0	(–0.3–2.2)	2.6
RISPERIDONE (*n* = 9401)	Women	5153	Overall	0.9	(0.5–1.3)	0.7	(–0.1–1.5)	1.6
		3102	Low	1.0	(0.5–1.4)	0.1	(–0.9–1.1)	1.1
		316	High	1.1	(–0.7–2.9)	3.5	(1.0–5.9)	4.6
	Men	4248	Overall	1.1	(0.6–1.5)	1.4	(0.4–2.4)	2.5
		2411	Low	1.0	(0.4–1.7)	1.1	(–0.3–2.6)	2.2
		441	High	1.9	(0.5–3.3)	1.4	(–0.7–3.5)	3.3

[a]Overall estimates come from Table S2 (*n* = 38,865) and low/high dose estimates come from Table S3 (*n* = 25,198). *n* from Table S2 < *n* from Table S3 due to missing data on dose.
[b]Cut off point for low/high dose was: ⩽ 5 mg for Olanzapine, ⩽ 75 mg for Quetiapine and ⩽ 2 mg for Risperidone.

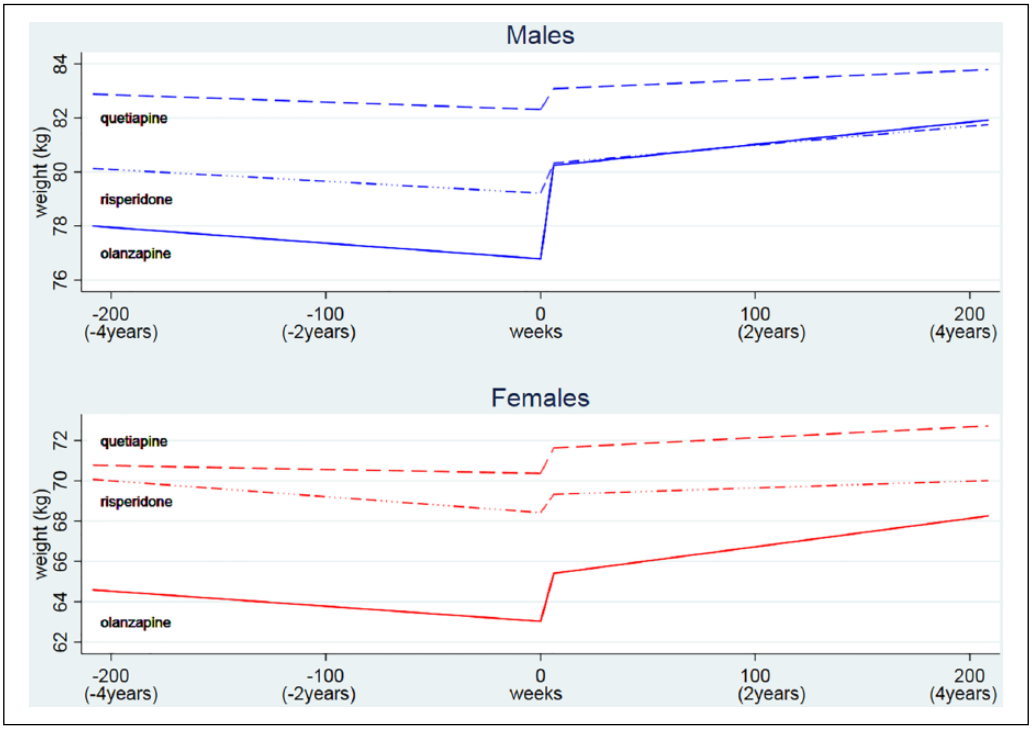


Figure 2. Changes in body weight over time before and after treatment initiation by drugs and sex.

spline models were very similar to the restricted cubic splines models (Figure S1), and for primary analysis we used the former as interpretation is more straightforward.

Our study does have a number of potential limitations. Information on possible time-varying confounders (for example, symptoms level or illness severity) was not included, however, it

is reasonable to assume limited variation from patients' baseline values for unmeasured confounders. Treatment initiation has been defined using first prescription date in general practice; but, for some individuals, the first prescription date might occur while the individual is under the care of secondary care mental services (these data are not recorded in primary care). However, it is most likely these patients have a first prescription date very close to the one in primary care, thus no major impact on estimates is expected.

We did not control for drugs prescribed to reduce antipsychotic-induced weight gain, or for multiple prescriptions of other drugs that could potentially affect weight as well. However, we know that drugs prescribed for ameliorating weight gain would only reduce the estimate of the real weight gain of the target population, thus we are not overestimating the weight gain effect. We did not assess weight gain associated with other antipsychotic medications as there were not enough data on them, but the three drugs included in this study are the most commonly prescribed antipsychotic medications in the UK (Marston et al., 2014) and have previously been associated with weight gain (Osborn et al., 2018). The weight gain trajectories we described are averages, thus they should be interpreted as typical patient trajectories. In practice, individual patients' weight gain will vary from these average trajectories. However, the first weeks of treatment are critical for everyone. Finally, we did not control the number of prescriptions beyond the second prescription (treatment duration), meaning that studied patients can include those treated for long periods, those treated sporadically, just for a short period, or those who did not adhere to treatment regularly. This lack of control may reduce our long-term estimates of weight gain, but, given the evidence about dosage, we anticipate that patients exposed to AP on a regular basis and for long periods will have larger estimates of long-term weight gain.

Comparison with other studies

Previous studies have suggested olanzapine is associated with a large short-term weight gain whereas risperidone and quetiapine have a moderate effect on weight (Bak et al., 2014). In the long-term, contrary to one previous finding (Haddad, 2005), we found that weight gain did not stabilize during four years of follow up. However, our finding of long-term effect of weight gain is consistent with previous studies by Bushe et al. (2012) and Osborn et al. (2018), but we are able to quantify the effect more accurately. Previous research has suggested women's weight is more affected by AP exposure (Seeman, 2008); however, we found that only olanzapine (in the long-term) and quetiapine (in the long and short-term) induced more weight gain in women. Since our study population is a mixture of naïve and recurrent antipsychotic consumers, short- and long-term weight gain in olanzapine naïve individuals and long-term weight gain in risperidone naïve individuals can be higher than the weight gain reported by us (Bak et al., 2014). Risperidone seemed to be associated with greater weight gain in men than women both in the short- and long-term, and men prescribed olanzapine gained more weight in the short-term. Regarding the dosage, one recent study reanalysed results of 14 clinical trials to explore variations in weight gain across doses of olanzapine and risperidone (Spertus et al., 2018). Their conclusions about olanzapine are consistent with our results; that the excess risk of at least 7% weight gain is 16.1% for low doses (0–10 g chlorpromazine equivalent dose) and 46.8% for high doses (0–20 g chlorpromazine equivalent dose). They could not be conclusive about the effects of risperidone as they showed only a trend in weight gain; however, this trend is in line with our findings. Some advantages from our original study are: (a) we added similar information about quetiapine, (b) we observed longer periods of weight change (four years) and (c) we analysed information at individual-level from cohorts with more than 38,000 patients in total.

Conclusions and policy implications

Over a four-year period, olanzapine treatment was associated with the highest increase in weight with around 6 kg for those on high dose and 4.5 kg for those on low dose. The weight gain was less dramatic for individuals treated with quetiapine and risperidone. In general, individuals did not lose the weight gained during the first six weeks of treatment. Doctors and patients may want to take the issue of a substantial weight gain into consideration when making decisions on initiation of antipsychotic treatments, and doctors should prescribe the lowest effective dose to balance mental health benefits, weight gain and other adverse effects.

Author contributions

JCB and IP wrote the study protocol. JCB performed the formal analysis and drafted the paper. JRC, TPM and IP supervised the formal analysis. JFH checked the clinical component of the study. All the authors contributed and approved the final draft of this paper.

Ethical approval

This study was approved by the THIN Scientific Review committee at IMS Health in April 2016. SRC Reference Number: 16THIN013. All data were anonymized and no participant consent was required.

Data availability

Data were analysed under THIN licence, however, they are not available for sharing.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: JCB was sponsored by FONDECYT-CONCYTEC (grant contract number 231-2015-FONDECYT). TPM and JRC are both supported by the Medical Research Council (grant numbers MC_UU_12023/21 and MC_UU_12023/29). JFH is supported by the Wellcome Trust (grant 211085/Z/18Z). The study sponsors only had a funding role in this research, thus researchers worked with total independence from their sponsors.

ORCID iDs

Juan Carlos Bazo-Alvarez https://orcid.org/0000-0002-6169-8049
Joseph F Hayes https://orcid.org/0000-0003-2286-3862

Supplemental material

Supplemental material for this article is available online.

References

Bak M, Fransen A, Janssen J, et al. (2014) Almost all antipsychotics result in weight gain: A meta-analysis. *PloS One* 9: e94112.

Bernal JL, Cummins S and Gasparrini A (2017) Interrupted time series regression for the evaluation of public health interventions: A tutorial. *Int J Epidemiol* 46: 348–355.

Blak BT, Thompson M, Dattani H, et al. (2011) Generalisability of The Health Improvement Network (THIN) database: Demographics, chronic disease prevalence and mortality rates. *Inform Prim Care* 19: 251–255.

Bushe CJ, Slooff CJ, Haddad PM, et al. (2012) Weight change from 3-year observational data: Findings from the worldwide schizophrenia outpatient health outcomes database. *J Clin Psychiatry* 73: e749–e755.

Chisholm J (1990) The read clinical classification. *BMJ* 300: 1092.

Covell NH, Weissman EM and Essock SM (2004) Weight gain with clozapine compared to first generation antipsychotic medications. *Schizophr Bull* 30: 229–240.

Dayabandara M, Hanwella R, Ratnatunga S, et al. (2017) Antipsychotic-associated weight gain: Management strategies and impact on treatment adherence. *Neuropsychiatr Dis Treat* 13: 2231.

De Hert M, van Winkel R, van Eyck D, et al. (2006) Prevalence of diabetes, metabolic syndrome and metabolic abnormalities in schizophrenia over the course of the illness: A cross-sectional study. *Clin Pract Epidemiol Ment Health* 2: 14.

Elmslie JL, Silverstone JT, Mann JI, et al. (2000) Prevalence of overweight and obesity in bipolar patients. *J Clin Psychiatry* 61: 179–184.

Emul M and Kalelioglu T (2015) Etiology of cardiovascular disease in patients with schizophrenia: Current perspectives. *Neuropsychiatr Dis Treat* 11: 2493–2503.

Gebhardt S, Haberhausen M, Heinzel-Gutenbrunner M, et al. (2009) Antipsychotic-induced body weight gain: Predictors and a systematic categorization of the long-term weight course. *J Psychiatr Res* 43: 620–626.

Haddad P (2005) Weight change with atypical antipsychotics in the treatment of schizophrenia. *J Psychopharmacol* 19: 16–27.

Haneuse S, Bogart A, Jazic I, et al. (2016) Learning about missing data mechanisms in electronic health records-based research: A survey-based approach. *Epidemiology* 27: 82–90.

Hayes JF, Marston L, Walters K, et al. (2017) Mortality gap for people with bipolar disorder and schizophrenia: UK-based cohort study 2000–2014. *Br J Psychiatry* 211: 175–181.

Holt RIG and Peveler RC (2009) Obesity, serious mental illness and antipsychotic drugs. *Diabetes Obes Metab* 11: 665–679.

Horsfall L, Walters K and Petersen I (2013) Identifying periods of acceptable computer usage in primary care research databases. *Pharmacoepidemiol Drug Saf* 22: 64–69.

Koga M (2003) Clinical factors related to gains in body mass index (BMI) among patients under long-term antipsychotic treatment. *Seishin Shinkeigaku Zasshi* 105: 473–488.

Lahti M, Tiihonen J, Wildgust H, et al. (2012) Cardiovascular morbidity, mortality and pharmacotherapy in patients with schizophrenia. *Psychol Med* 42: 2275–2285.

Maguire A, Blak BT and Thompson M (2009) The importance of defining periods of complete mortality reporting for research using automated data from primary care. *Pharmacoepidemiol Drug Saf* 18: 76–83.

Marston L, Nazareth I, Petersen I, et al. (2014) Prescribing of antipsychotics in UK primary care: A cohort study. *BMJ Open* 4: e006135.

Molenberghs G, Fitzmaurice G, Kenward MG, et al. (2014) *Handbook of Missing Data Methodology*. Boca Raton, FL: Chapman and Hall/CRC.

Najar H, Joas E, Kardell M, et al. (2017) Weight gain with add-on second-generation antipsychotics in bipolar disorder: A naturalistic study. *Acta Psychiatr Scand* 135: 606–611.

Newcomer JW and Haupt DW (2006) The metabolic effects of antipsychotic medications. *Can J Psychiatry* 51: 480–491.

Nutt DJ and Blier P (2016) Neuroscience-based Nomenclature (NbN) for Journal of Psychopharmacology. *J Psychopharmacol* 30: 413–415.

Olfson M, Gerhard T, Huang C, et al. (2015) Premature mortality among adults with schizophrenia in the United States. *JAMA Psychiatry* 72: 1172–1181.

Osborn DP, Levy G, Nazareth I, et al. (2007) Relative risk of cardiovascular and cancer mortality in people with severe mental illness from the United Kingdom's General Practice Research Database. *Arch Gen Psychiatry* 64: 242–249.

Osborn DP, Petersen I, Beckley N, et al. (2018) Weight change over two years in people prescribed olanzapine, quetiapine and risperidone in UK primary care: Cohort study in THIN, a UK primary care database. *J Psychopharmacol* 32: 1098–1103.

Roland M, Guthrie B and Thomé DC (2012) Primary medical care in the United Kingdom. *J Am Board Fam Med* 25: S6–S11.

Seeman MV (2008) Secondary effects of antipsychotics: Women at greater risk than men. *Schizophr Bull* 35: 937–948.

Spertus J, Horvitz-Lennon M, Abing H, et al. (2018) Risk of weight gain for specific antipsychotic drugs: A meta-analysis. *NPJ Schizophr* 4: 12.

Stata (2017) *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC.

van Buuren S (2012) *Flexible Imputation of Missing Data*. Boca Raton, FL: Chapman and Hall/CRC.

White IR and Carlin JB (2010) Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med* 29: 2920–2931.

Woods SW (2003) Chlorpromazine equivalent doses for the newer atypical antipsychotics. *J Clin Psychiatry* 64: 663–667.

Handling missing values in interrupted time series analysis of longitudinal individual-level data

Juan Carlos Bazo-Alvarez*[1,2], Tim P Morris[3], Tra My Pham[3], James R Carpenter[3,4], Irene Petersen[1,5]

1 Research Department of Primary Care and Population Health, University College London (UCL), London, United Kingdom.

2 Instituto de Investigación, Universidad Católica los Ángeles de Chimbote, Chimbote, Peru.

3 MRC Clinical Trials Unit at UCL, London, United Kingdom.

4 Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, United Kingdom.

5 Department of Clinical Epidemiology, Aarhus University, Denmark.

Positions

- JCB is a PhD Student at UCL and Research Fellow at Universidad Católica Los Angeles de Chimbote.
- TPM is a Senior Medical Statistician at the MRC Clinical Trial Unit at UCL.
- TMP is a Research Fellow at the MRC Clinical Trial Unit at UCL.
- JRC is a Professor of Medical Statistics at London School of Hygiene and Tropical Medicine, and Honorary Professor at the MRC Clinical Trial Unit at UCL.
- IP is a Professor of Epidemiology and Health Informatics at UCL and Professor of Biostatistics at Aarhus University.

Authors contact

Dr Tim P Morris: tim.morris@ucl.ac.uk

Dr Tra My Pham: tra.pham.09@ucl.ac.uk

Prof James R Carpenter: james.carpenter@lshtm.ac.uk

Prof Irene Petersen: i.petersen@ucl.ac.uk

Correspondent author

Juan Carlos Bazo-Alvarez

Research Department of Primary Care and Population Health

University College London

Rowland Hill Street, NW3 2PF, London, UK.

Phone: +44 07376076260, email: juan.alvarez.16@ucl.ac.uk

Abstract

Background

In the interrupted time series (ITS) approach, it is common to average the outcome of interest at each time point and then perform a segmented regression (SR) analysis. In this study, we illustrate that such 'aggregate-level' analysis is biased when data are missing at random (MAR) and provide alternative analysis methods.

Methods

Using electronic health records from the UK, we evaluated weight change over time induced by the initiation of antipsychotic treatment. We contrasted estimates from aggregate-level SR analysis against estimates from mixed models with and without multiple imputation of missing covariates, using individual-level data. Then, we conducted a simulation study for insight about the different results in a controlled environment.

Results

Aggregate-level SR analysis suggested a substantial weight gain after initiation of treatment (average short-term weight change: 0.799kg/week) compared to mixed models (0.412kg/week). Simulation studies confirmed that aggregate-level SR analysis was biased when data were MAR. In simulations, mixed models gave less biased estimates than SR analysis and, in combination with multilevel multiple imputation, provided unbiased estimates. Mixed models with multiple imputation can be used with other types of ITS outcomes (e.g. proportions). Other standard methods applied in ITS do not help to correct this bias problem.

Conclusions

Aggregate-level SR analysis can bias the ITS estimates when individual-level data are MAR, because taking averages of individual-level data before SR means that data at the cluster level are missing not at random. Avoiding the averaging-step and using mixed models with or without multilevel multiple imputation of covariates is recommended.

Keywords

Interrupted Time Series Analysis; Segmented Regression; Missing Data; Multiple Imputation; Mixed Effects Models; Electronic Health Records; Big Data.

Running title

Interrupted time series and missing data

Word Count: Abstract: 250 words / Article: 5,146 words

1. Introduction

Interrupted time series (ITS) is a widely used quasi-experimental approach that evaluates the potential impact of an intervention over time, using longitudinal observational data [1]. It has frequently been used to evaluate intervention effects in longitudinal population studies; for example, to evaluate the impact of policies and social interventions on clusters, such as districts, cities and countries [2,3]. While ITS comes from social science literature, it is becoming more widespread in health research [4,5]. ITS may be used to address causal questions that are not feasible for a randomised controlled trial, but with stronger assumptions [6]. The methodology for the analysis of ITS studies is well developed [1,7,8], and typically uses segmented regression (SR) analysis [4,5]. Given a time point, for example the initiation of treatment, we may observe a change in the values of a variable before and after that time point, and then compare the trajectories of change at the intervention. The pre-treatment trajectory is regarded as the control 'period' and the post-treatment trajectory as the intervention 'period', so that each individual acts as their own control. The difference between mean trajectories at the intervention time is then used to estimate the effect of the intervention [1].

In SR analysis, when individual-level data are available, a typical approach is to average the data at each of the predefined time points/units (e.g. months or years) and then model the time series over these time points [5,9–11]. In other words, all outcome variable measurements available from individuals are averaged at each time point, and then these averages are used as population-level data for performing the SR analysis. This approach is reasonable if the same people provide data at each time point, but in observational data this is rarely the case. For example, in clinical practice, younger women are more likely than younger men to have weight recorded when they consult their family physician (general practitioner) [12]. In other words, the distribution of missing data in weight depends on the individual's sex, so weight is missing at random (MAR) given sex. The same will apply to other partially observed outcomes that are MAR. With such data, the average points will be biased – and so will the intercept of the trajectories estimated by SR models – because they will include more measurements from women

than men, and women will typically weigh less than men. Moreover, if the proportion of women and men with observed weight varies at each time point, the slope of the trajectories can also be biased.

Figure 1 presents a scenario where weight is constant over time for all individuals (half men, half women; men weigh 85kg, and women weigh 55kg, resulting in an overall average of 70kg). In this scenario, all individuals have a weight measurement at treatment initiation (t=0), but at different time points before and after treatment initiation the relative proportion of women and men with a weight record varies due to missing data. The average observed weight at each time point becomes biased, providing a false impression of weight change over time. Thus, the 'aggregate-level' SR analysis performed with averages calculated at pre-defined time points can produce biased estimates due to missing data.

An alternative approach to the 'aggregate-level' SR analysis is to use mixed models, which are based on individual-level data, avoiding the averaging-step described above. Formally, these mixed models are also segmented models, but they include random intercept and slopes (random effects) that cannot be included by the 'aggregate-level' SR models due to the averaging-step. Mixed models estimate identical linear trajectories to 'aggregate-level' SR models under perfect balance (when all individuals are included at each time point). However, in contrast to 'aggregate-level' SR models, the mixed model approach can provide unbiased estimates when data in the outcome variable are MAR [13]. Following the same example as before, a mixed model directly uses weight measurements taken at different time points from the same individual, and models the population trajectory based on all individual trajectories, taking account of the longitudinal correlation. Thus, no initial averaging-step at each time point is needed. If individuals have missing weight records over time, the mixed model approach implicitly imputes those missing values, meaning that observations from all individuals – even those with just one record over time – contribute to the analysis.

Despite these advantages, mixed models cannot automatically handle missing data in the covariates, and individuals with covariates missing are by default omitted from regression analyses in all standard software packages. One way to address this issue is to use multilevel multiple imputation (MMI)

for missing covariate data in conjunction with mixed models. MMI generates multiple datasets with missing covariate values replaced by imputed values (drawn from the conditional predictive distribution of the missing data given the observed data). Then, MMI fits the substantive model of interest in each imputed dataset and, in the final step, combines the model estimates into an overall estimate, taking into account variation within and between the imputed datasets [14]. In our setting, the substantive model fitted at the second step is a mixed model.

In this study, we demonstrate how standard ITS analysis, based on average estimates at each predefined time point, gives biased results when data are MAR. Subsequently, we illustrate how the use of mixed models, with or without MMI of individual data, avoids this bias.

Our objectives are 1) to examine the potential problems arising from the 'aggregate-level' SR analysis when outcome data are missing, evaluating mixed models as an alternative approach; 2) to compare the performance of mixed models with and without MMI for handling missing data on covariates.

The rest of this article is structured as follows. In Section 2 we present a motivating example of ITS to estimate the effect of initiating antipsychotic drugs (olanzapine) on weight gain, showing that the standard approach of aggregating the data and then using SR gives clinically different results to using mixed models (with and without MMI). Section 3 presents a simulation study, which demonstrates that this difference is because the standard 'aggregate-level' approach is biased when data are MAR. We conclude in Section 4 by discussing the practical and methodological implications of our findings. Stata and R codes for reproducing our results are provided in the Appendix. It should be noted that this study did not cover ITS modelled on consecutive cross-sectional samples (e.g. incidence trajectories modelled with data from different individuals over time).

2. Motivating example: ITS for effect of antipsychotic drugs on weight.

In this motivating example, as well as in the later simulation study, we focus on assessing estimators for the regression coefficients of pre- and post-treatment weight trajectories.

2.1. Data and first analysis

We used data from The Health Improvement Network (THIN) database, which includes electronic health records from ~12 million individuals registered with 711 UK general practices [15]. In the UK, more than 95% of people are registered with a general practice (GP), and THIN is roughly representative of the general population [16]. THIN data include demographics (e.g. sex, age, social deprivation) and clinical records (e.g. drug treatments, diagnoses, health outcomes). In this study, we only included data from general practices that met quality criteria for computer usage [17] and whose reported mortality rate is consistent with national statistics [18].

We performed an ITS analysis to investigate the long-term effects of the initiation of antipsychotic drug treatment on people's body weight. It is known that specific antipsychotic treatments are likely to increase body weight substantially over a relatively short period [19], but we have less information on potential long-term effects [20]. In this study, the exposure of interest was the initiation of olanzapine (a second-generation antipsychotic), and the outcome was body weight (in kilograms). We modelled the development of weight over time using linear splines with two knots. In other words, our model estimated how weight changed in three time periods: 1) *pre-treatment:* from 4 years before treatment initiation up to treatment initiation; 2) *short-term:* from treatment initiation to 6 weeks (short-term), and 3) *long term:* from 6 weeks to 4 years post-treatment. We adjusted for sex, age at initiation (in years) and smoking at initiation of treatment (smoker vs non-smoker). We included individuals who were aged between 18 and 99 years, with data available between 1st January 2005 and 31st December 2015, and who initiated their first olanzapine treatment within this period. All had a diagnosed psychotic disorder before treatment initiation and at least one further prescription of olanzapine within three months following the first prescription. We included this criterion as there may some individuals who received just one prescription, but never used the medication. However, if they had at least two prescriptions it seems more likely that they initiated treatment. We excluded individuals who initiated other antipsychotics than olanzapine, as well as those with no available data for 12 months before the treatment initiation.

In addition to the inclusion and exclusion criteria given above, we restricted our data to those with complete data on sex, age and smoking at treatment initiation. As this is observational data, weight measurements did not follow any fixed schedule. For example, if we look for a weight measurement every two weeks for every individual, we will find that >90% of weight measurements are missing. In other words, the weight has been irregularly recorded over the observation period (416 weeks), as it is expected for most electronic health records.

Centring each patient's follow-up time (in weeks) at their treatment initiation, we fitted the following mixed model to these data:

$$weight_{ij} = \beta_0 + u_{0j} + (\beta_1 + u_{1j})time_{ij} \times 1[time_{ij} < 0] + (\beta_2 + u_{2j})time_{ij} \times [0 \le time_{ij} \le 6] + (\beta_3 + u_{3j})time_{ij} \times [time_{ij} > 6] + \epsilon_{ij},$$

[Equation 1]

$$\begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma\right); \qquad \epsilon_{ij} \sim N(0, \sigma^2),$$

where *i* denotes the follow-up time and *j* denotes the patient, and 1[ ] is an indicator for the event in square brackets. We then fitted the same model adjusting for sex, age and smoking at treatment initiation (as fixed effects). These mixed intercept and slope models were fitted by Restricted Maximum Likelihood, and hereafter we call them just mixed-effects models (MEM).

We also fitted an 'aggregate-level' SR model by averaging available weight records at each time point (across-individuals average), and then fitting the standard regression version of [Equation 1] – i.e. omitting the person-specific random effects-. Because this model is fitted to data aggregated over individuals, no adjustment for sex, age or smoking was possible.

Finally, we fitted a similar model, but now weighting by the inverse of the number of body weight values observed at each time point. We called this model the 'aggregate-level' SR-W1, which may help to improve standard errors by including a more accurate sample size information at each time-point.

These models were used to examine the issues arising from the 'aggregate-level' SR analysis when outcome (weight measurements) data are missing, which was part of our first study objective.

2.2. Imposed missing data and second analysis

For our second objective, we wanted to explore the issues arising from covariate data missing at treatment initiation. Therefore, we intentionally set smoking records MAR on sex, and increased the amount of missing data on weight MAR on sex, to explore later the potential differences between estimates from complete case analysis (removing cases with smoking missing) and MMI (preserving those cases and imputing smoking). This controlled missing data generation scenario was used evaluate all analysis methods: 'aggregate-level' SR, 'aggregate-level' SR-W1, MEM, and MMI followed by a mixed-effects model (MI-JOMO with MEM).

In detail, we set weight values MAR dependent on sex and time from treatment initiation, so that a fraction of observed data was similar to that shown in Figure 1. In addition, we set smoking MAR on sex, randomly removing 80% of records from men and 20% from women. Both missing mechanisms are described in detail in Appendix A.

In our subsequent analyses we first fitted the same MEM [Equation 1] to the incomplete data, adjusting for covariates (complete case analysis). Then, we used a substantive-model-compatible joint-modelling multilevel multiple imputation (MI-JOMO) [21] to impute the missing smoking values and fitted the same substantive model (MEM adjusted) to each imputed data set and combined the results using Rubin's rules. We generated 20 imputed datasets with MI-JOMO, and we used a burn-in of 1000 iterations and then a further 1000 iterations between each imputation. We name this model MI-JOMO with MEM.

Lastly, we fitted the 'aggregate-level' SR and 'aggregate-level' SR-W1 models. Full details and codes for all models are given in Appendix A.

2.3. Results

Overall, there were 6,522 individuals with at least one weight measurement and complete age, sex or smoking status data. Of these 2,954 (45.3%) were men and 3,568 (54.7%) were women. On average, there were 4.8 (sd 5.5) weight records per person over the observation period. Individuals were aged 50.2 (sd 18.9) years on average, and 2,658 (40.8%) reported being current smokers.

There were substantial differences between estimates derived from MEM and SR (Table 1, section 'THIN: Data Fully Observed'). For example, the short-term weight change (beta2) was 0.462kg/week from MEM (adjusted) and 0.816kg/week and 0.807kg/week from SR and SR-W1 respectively. Likewise, pre-treatment and long-term periods, weight change rates from SR and SR-W1 were more than double the MEM estimates. In general, all estimates of weight change from SR analyses were higher in magnitude than those from MEM, which also implies a more substantial ITS treatment effect.

After further removal of weight records, 6,181 individuals remained with one or more weight records. There were 4.3 (sd 5.3) average weight records per person over the observation period. The average age was 50.6 (sd 19) years, and 2,613 (42.3%) were men. After removal of smoking records at baseline there were only 3,379 individuals with a record of their smoking status and 1,188 (35.2%) of them were current smokers.

In general, estimates from MEM with and without MI-JOMO were similar for pre-treatment and long-term effects, and both close to those estimated under MEM with full data. However, the MI-JOMO with MEM for short-term were closest to those estimated under MEM with full data (Table 1). ITS estimates from SR differed substantially from the estimates from MEM with and without MI-JOMO (Table 1, Figure 2), with SRs reporting a weight pre-treatment (beta1) and long-term trajectories (beta3) closer to zero. For SR-W1, the long-term treatment effect was similar to the MEM estimates, while the short-term effects estimates (beta2) were much higher than MEM estimates. For both the SR and SR-W1

models, pre-treatment and long-term effects were also different when fitted to data with and without imposed missing values.

The immediate treatment effect, estimated as the difference between the negative and positive trajectories before and after olanzapine initiation, was highest for the SR approach (Table 1 and Figure 2). For example, the SR-W1 method suggested a cumulative short-term weight gain of 4.72kg, a long-term of 2.13kg, and a total of 6.85kg. In contrast, the estimates based on MEM with MI-JOMO (short-term=2.47kg, long-term=2.46kg, total=4.93kg) and without MI-JOMO (short-term=2.75kg, long-term=2.70kg, total=5.45kg) were less for the short-term and the total accumulated (see 95% CI in Appendix B).

In summary, individual-data model such as MEM [Equation 1] produced notably different results from SR models with 'aggregate-level' data. Further, if covariate values are MAR, use of MI-JOMO can recover information by bringing individuals with these missing covariates back into the analysis, avoiding potential bias and increasing precision. By contrast, the often-used SR 'aggregate-level' analysis cannot adjust for covariates and appears to be biased when weight data are MAR (depending on time and covariates). This may often be the case when analysing health care records.

3. Simulation study

We now report the results of a simulation study, based on the motivating clinical example and designed to evaluate the performance of SR and MEM (with and without MMI) under controlled conditions. We are adding to this evaluation another method called Prais-Winsten regression, which is similar to SR but is recommended by ITS guidelines to account for autocorrelation at the aggregate level [1]. In particular, we wish to determine whether the differences between the various analysis methods are due to the way they handle missing data.

3.1. Simulation design

3.1.1. Study model

For the simulation study, we designed an ITS dataset where the treatment of interest was the initiation of antipsychotic treatment, and we examined change in body weight (in kilograms) over time. The covariates were sex, age (years) and smoking status (yes/no), measured at initiation of treatment. The ITS impact model [8] is a linear weight trajectory whose slope changes only once – at treatment initiation – i.e. slightly simpler than our previous example. We included five time-units before and five after treatment initiation. We modelled the evolution of weight over time using two continuous linear splines, jointing at treatment initiation.

3.1.2. Data generation

Each simulated dataset with 1,000 observations was generated as follows:

1 Sex was generated as a random variable from a Bernoulli distribution with probability 0.5.

2 For each individual, weight observation times were fixed at the same 11 equally spaced times between -5 and +5, i.e. centred at treatment initiation, which is at time 0.

3 Weight was generated from the following random intercept and slopes model:

$$weight_{ij} = 75 + u_{0j} + (-0.5 + u_{1j})time_{ij} \times 1[-5 \leq time_{ij} < 0] + (3.4 + u_{2j})time_{ij} \times [0 \leq time_{ij} \leq 5] + 10 * sex_i + \varepsilon_{ij}, \quad \text{[Equation 2]}$$

$$\begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{pmatrix} \sim N\begin{pmatrix} 0 \\ 0, 0 \\ 0 \end{pmatrix}\begin{matrix} 5 & 0 & 0 \\ 1.1 & -.7 \\ -.7 & 1.1 \end{matrix}; \qquad \varepsilon_{ij} \sim N(0,2),$$

where *i* denotes the follow-up time and *j* denotes the patient, and 1[ ] is an indicator for the event in square brackets. We referred to this as 'Data Generation Mechanism Base' (DGM-base). We also generated data from DGM-extended covariates:

$$weight_{ij} = 75 + u_{0j} + (-0.5 + u_{1j})time_{ij} \times 1[-5 \leq time_{ij} < 0] + (3.4 + u_{2j})time_{ij} \times [0 \leq time_{ij} \leq 5] + 10 * sex_i + 0.05 * age_i - 0.0005 * age_i^2 + 2.5 * smoking_i + \varepsilon_{ij}.$$

[Equation 3]

Age was generated as a random variable from a normal distribution with mean 45 and sd 10. Smoking was binary and generated as follows:

$$logit\big(P(smoking_i = 1)\big) = -2 + 1.5 * sex_i + 0.04 * age_i - 0.0005 * age_i^2.$$

Having generated the full data, we made observations missing using two missing data mechanisms:

1. MAR-1: starting with the fully observed weight variable at treatment initiation (t_0), pre and post-treatment initiation values of weight at times $t_{0\pm j}$ were set to missing ($j = 1,2,3,4,5$) dependent on the individual's sex. For the missing sequence, pre-treatment setting of missing values was reverse-sequential ($t_{-1}, t_{-2}, t_{-3}, t_{-4}, t_{-5}$) and post-treatment setting was forward-sequential (t_1, t_2, t_3, t_4, t_5). For both directions ($\pm j$) of MAR-1 mechanism, we defined the probability of being missing by:

$$logit\left(P\big(weight_{ij} = missing\big)\right) = -2.5 + 5 * sex_i,$$

shaping the patterns of missing weight data and setting more weight records being observed for women than men. Both patterns and proportion of missing values are available in Appendix C. MAR-1 was applied on data generated under DGM-base only.

2. MAR-2: similar to MAR-1, but now the probability of weight being missing also depends on the individual's random intercept, age and smoking. As the random intercept is unobservable (as smoking will partially be), this mechanism is a mix between MAR and MNAR (missing not at random). Moving away from treatment initiation (in both directions), the probability of weight being missing is monotonically given by:

$$logit\left(P\left(weight_{ij} = missing\right)\right) = -0.25 - 2 * u_{0j} - 1.5 * sex_i - 0.05 * age_i + 0.0005 * age_i^2 - 1.5 * smoking_i,$$

where -0.25 helped to shape the overall proportion of missing data over time; -1.5 set more weight records to be observed for men (only for explicative purposes); -2 set more weight records to be observed for individuals who are heavier at treatment initiation; -0.05 and 0.0005 set more missing data for younger individuals, and -1.5 set more weight records to be observed for smokers. We also set about 30% of smoking values to be missing with probability:

$$logit\left(P(smoking_i = missing)\right) = -3 + 3 * sex_i - 0.01 * age_i + 0.0003 * age_i^2,$$

MAR-2 was applied to data from DGM-extended-covariates only. For both described mechanisms (MAR-1 and -2), the proportion of missing weight data in the simulated sample was set to approximately 60% of individuals. In the other 40% of the data, we set only one weight record per

individual at any time point, setting more individuals with only one weight record at treatment initiation (MAR dependent on the treatment initiation). This additional mechanism sought to emulate the missing data proportions and patterns seen in the clinical data used for the illustrative example (see Appendix C).

We simulated 1,000 full datasets for each of the two scenarios, and then applied the missing data mechanisms to obtain the partially observed data.

3.1.3. Analysis methods evaluated

We analysed the full and partially observed data using each of the following six methods (see summary in Appendix D):

1) SR: this averaged observed individual weight measures at each time point and then fits a linear regression on time (maximum likelihood estimator), with a knot at zero.
2) SR-W1: (weighted SR version 1) similar to SR but weighted by the inverse of the number of observed weight records at each time point.
3) SR-W2: (weighted SR version 2) similar to SR-W1 but the number of observed weight records – used for weighting – were counted at each time point by sex and age. We categorised age using its quartiles (before averaging). When smoking data were incomplete, smoking was not included as a covariate for SR-W2.
4) Prais-Winsten: regression similar to SR but adjusted for serial correlation at the aggregate level by assuming errors that follow a first-order autoregressive process [22], an approach typically used in ITS analysis for controlling the autocorrelation issue [1].
5) MEM: we fitted the data generating model [Equations 2 and 3] using Restricted Maximum Likelihood with an unstructured covariance matrix for the random effects.
6) MI-JOMO (with MEM): We first imputed the missing covariate values, using multilevel substantive-model-compatible joint modelling multiple imputation, with the JOMO package in R. As described in [23,24] this imputes missing values consistent with the substantive model

[Equation 1]. It does this by factorising the joint model into a joint model for the covariates and a conditional model for the outcome given the covariates. Then, the estimation and imputation process allows compatibility between the imputation and analysis models (MEM in this case), even with longitudinal data [24]. We used 5 imputations and 1000 iterations (before the first, and between each subsequent imputation) to impute the missing covariate smoking status. We did not impute the missing weight, as (in the absence of auxiliary variables) no information can be recovered by doing this. Note that standard fully conditional specification [25] is not evaluated because it is inappropriate for handling the irregular observation times we expect in real longitudinal data. We only used MI-JOMO in the MAR-2 scenario.

3.1.4. Estimands and performance measures

We focused on the slope estimates (true values: $time_{before}$: $\beta_1 = -0.05$ and $time_{after}$: $\beta_2 = 3.4$) from all methods evaluated in both MAR scenarios (MAR-1 and MAR-2), by examining the bias, empirical standard error, model-based standard error and confidence interval coverage [26].

3.2. Simulation results

In the first scenario (DGM-base), all SR methods were biased except from when data were fully observed (Table 2). However, the coverage of these methods was low (<61%) due to their small model-based standard errors, even the weighted methods (SR-W1 and SR-W2) and the method adjusted for serial correlation (Prais-Winsten). Conversely, MEM provided reasonably good coverage for β_1 and β_2 (>94%) for unbiased estimates.

Where weight was missing based on sex only (MAR-1), MEM showed unbiased results and the best coverage (≥95%). SR and SR-W1 produced biased estimates for both pre- and post-treatment initiation slopes, showing the highest model-based standard errors. Because the missingness mechanism depended on sex, and women weighed less than men, the preliminary data aggregation step in SR and

SR-W1 biased the estimated slopes (see example in Figure 3, MAR-1). The SR bias was corrected using inverse-probability weights based on sex (SR-W2), but coverage was low (<74%) due to too-small model-based standard errors. The Prais-Winsten model was not successful in correcting the SR bias since it does not incorporate information on missing data at the individual-level as SR-W2 does.

In the second scenario (DGM-extended-covariates), with full data, all methods were unbiased (Table 2). MEM provided the best coverage for β_1 and β_2 (>95%), followed by SR-W2 (>90%). Although with unbiased estimates, SR, SR-W1 and Prais-Winsten provided a low coverage (<55%) due to their small model-based standard errors. SR, SR-W1 or Prais-Winsten cannot provide different averages by sex and age at each time point, which can be provided by SR-W2. Having more variability at each time point produced higher – and more realistic – standard errors from SR-W2.

On the other hand, with missing values in weight and smoking status (MAR-2), MI-JOMO had the best performance. MEM showed poorer performance after all covariates were included in the imputation and study models and there were missing smoking data, producing slightly biased estimates and low coverage (<79%). In the same scenario, MI-JOMO performed better than MEM, providing less biased estimates, closer values of empirical and model-based standard errors, and higher coverage (>87%). For both methods, we should consider that there is some residual bias because of the dependence of observation of weights on the random intercepts. While the results in the bottom half of Table 2 show this resulted in a bias in the MI-JOMO analysis, this was not severe, and the resulting inferences were still usable. Conversely, SR, SR-W1, SR-W-2 and Prais-Winsten performed extremely poorly, showing large bias and low coverage (<18%).

The 'aggregate-level' SR analysis biased the slope trajectories in different directions, which we illustrated by our simulations (Figure 3).

4. Discussion

ITS provides a conceptually attractive approach for assessing the impact of treatments because each individual acts as their own control. However, its innate strength, leading to its increasing use [4], raises important questions about how to appropriately handle missing data. As our example illustrates, incomplete outcome data (in our case, weight) is an intrinsic feature of this kind of study because the underlying observational data do not follow any pre-planned schedule. This means that, at any specific time, the marginal distribution of the response is unlikely to be representative of the underlying population.

The results of our studies demonstrate that the 'aggregate-level' approach will generally be biased when individual-level data are missing at random (MAR). Indeed, the motivating example shows this bias could lead to a substantial exaggeration of the actual effect of the studied intervention. In the example, the difference between pre- and immediate post-treatment weight change (biased slopes) increases the overall effect attributed to olanzapine. However, it is not always possible to determine the direction of bias. This is because the direction of the average-points bias depends on how the covariate is associated with the missingness of weight records. Even when the 'aggregate-level' SR analysis does not bring about a bias issue, our results highlight that the precision is inaccurate as the standard errors for this method are typically grossly underestimated.

When data are missing-at-random at the individual level, averaging before SR means that data are missing-not-at-random at the cluster level. This leads to the bias observed for the 'aggregate-level' SR analysis. For example, in the MAR-1 mechanism, 'aggregate-level' SR analysis loses the information about the distribution of weight records that are MAR on sex at each time point, due to the averaging-step. Thus, sex becomes unobservable at the 'aggregate-level', making weight records MNAR on sex at this level and biasing the subsequent analysis using those averages. As we demonstrate in the same simulation study, this issue could be handled by including sex in the averaging-step (SR-W2). However, in practice, any version of SR-W2 will be hard to apply since other covariates are typically incomplete as well.

A natural alternative to the 'aggregate-level' analysis is to model the individual patient data explicitly. When the reason for outcome data being observed depends principally on time (e.g. before and after treatment initiation), underlying patient characteristics (e.g. sex, age) and observed outcomes (e.g. observed weights), the unseen values are plausibly MAR. In this setting, our simulation results demonstrate that a carefully formulated longitudinal model provides a practical approach for improved inference.

Longitudinal models should be formulated carefully to include covariates predictive of both the outcome and the chance of observing it, which are key for avoiding bias. Where it is not desired – or appropriate – to include some such variables in the substantive analysis, an MMI approach should be considered, where these variables are included as auxiliary variables. Care should also be taken to model the longitudinal correlation of the outcome appropriately, as this is particularly important for missing data, as well as to use the observed rather than expected information for likelihood-based models. In particular, having random intercepts alone, or having uncorrelated random intercepts and slopes, should be avoided (see Appendix E for other practical suggestions) [21]. If data at the individual level are not available, and the researcher suspects that a strong MAR mechanism affect the outcome points over time (e.g. averages or rates), the issue should be stated as a limitation as recommended in reporting guidelines [27,28].

Our results show that MMI provides a practical approach for handling missing covariates in the analysis. When performing MMI, it is essential to both use an approach that properly takes account of the multilevel structure, and uses an approach that is compatible with the substantive model (which here includes splines for the effect of time). The JOMO package in R has the flexibility to do both.

We set our example and simulations with averages of a continuous variable, but a similar problem can happen with other types of outcomes. Rates (proportions), another common ITS outcome [5], can also be biased when outcome data are MAR at the individual level. For example, if the numerator of the rate (the events) is higher in women than men, and the missingness process generates more missing records for women, the rate will be underestimated at the 'aggregated-level' (e.g. at time points, hospitals or

districts). The ITS analysis will use those rates as consecutive points, biasing the estimated trajectories. Similar reasoning can be applied to binary and count ITS outcomes. Even using other recommended analysis methods than SR, such as ARIMA models [1], the bias problem will remain in the 'aggregate-level' used for the time series. Although we did not formally evaluate these alternative methods, some reflections can be enlightened by the study findings. In the aggregate-level approach, ARIMA models will be fitted after the averaging-step; therefore, the ITS will be based on population-level average points already biased. Other options useful for individual-level data, such as generalised estimated equations (GEE) can be applicable. However, because they are moment-based estimates, precisely like the aggregate data analysis, its estimates will be biased unless data are missing completely at random [29,30].

This is the first time that this averaging-step problem for MAR data has been studied with simulations and real data. Our results will help to guide future ITS studies. We focused our study on the situation when data are missing at random. However, we are aware there may be other scenarios where data missing not at random (MNAR) could bias estimates. For example, if weight is only recorded for those with a high or low weight. This scenario goes beyond the scope of this study but in practice, when a strong MNAR mechanism is suspected, a sensitivity analysis is possible using a pattern mixture approach [31,32].

In conclusion, the segmented regression using averaged data points can over or underestimate the effect evaluated in interrupted time series analyses, when performed on outcome data missing at random at the individual level. However, such a problem can be addressed by using mixed models. If there are also covariates missing at random, mixed models can be combined with multilevel multiple imputation and provide unbiased results.

Acknowledgements

We want to thanks Marc Comas, Maria del Mar Garcia, Rafel Ramos and all the team in the USR Girona (https://www.idiapjgol.org/index.php/ca/girona.html), as well as the members of the Methodologists Group of the Research Department of Primary Care and Population Health at UCL, for their invaluable feedback during preliminary presentations of this study. Special thanks to Ekaterina Bordea and Andrea Gabrio for their comments on an early version of this article, and to Caroline Clark for making this support possible. We also have special thanks to Matteo Quartagno for his great support on the use of JOMO package.

Funding

This work was supported by FONDECYT-CONCYTEC (grant contract number 231-2015-FONDECYT) to JCB. TPM, TMP and JRC were supported by the Medical Research Council (grant numbers MC_UU_12023/21 and MC_UU_12023/29). The study sponsors only had a funding role in this research. Thus, researchers worked with total independence from their sponsors.

Declarations of interest

None.

Ethical Approval

This study has two components, a motivating example and a simulation study. The latter did not require any ethical approval since all data is simulated. The Scientific Review Committee (SRC) of The Health Improvement Network (THIN) approved the protocol for the motivating example in April 2016 (SRC Reference Number: 16THIN013). No further revision by another institutional review board or ethics committee was needed since all data were anonymised and THIN license includes all consents required. THIN data is not freely available.

Authors Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

ORCID iD

Juan Carlos Bazo-Alvarez: https://orcid.org/0000-0002-6169-8049

REFERENCES

1. Lopez Bernal J, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *Int J Epidemiol*. 2017;46(1):348-355.
2. Myer AJ, Belisle L. Highs and Lows: An Interrupted Time-Series Evaluation of the Impact of North America's Only Supervised Injection Facility on Crime. *J Drug Issues*. 2018;48(1):36-49. doi:10.1177/0022042617727513
3. Ejlerskov KT, Sharp SJ, Stead M, Adamson AJ, White M, Adams J. Supermarket policies on less-healthy food at checkouts: Natural experimental evaluation using interrupted time series analyses of purchases. Popkin BM, ed. *PLOS Med*. 2018;15(12):e1002712. doi:10.1371/journal.pmed.1002712
4. Jandoc R, Burden AM, Mamdani M, Lévesque LE, Cadarette SM. Interrupted time series analysis in drug utilization research is increasing: systematic review and recommendations. *J Clin Epidemiol*. 2015;68(8):950-956. doi:10.1016/J.JCLINEPI.2014.12.018
5. Hudson J, Fielding S, Ramsay CR. Methodology and reporting characteristics of studies using interrupted time series design in healthcare. *BMC Med Res Methodol*. 2019;19(1):137. doi:10.1186/s12874-019-0777-x
6. Kontopantelis E, Doran T, Springate DA, Buchan I, Reeves D. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. *bmj*. 2015;350:h2750.
7. Lopez Bernal J, Cummins S, Gasparrini A. The use of controls in interrupted time series studies of public health interventions. *Int J Epidemiol*. 2018;47(6):2082-2093.
8. Lopez Bernal J, Soumerai S, Gasparrini A. A methodological framework for model selection in interrupted time series studies. *J Clin Epidemiol*. 2018;103:82-91. doi:10.1016/J.JCLINEPI.2018.05.026
9. Bao L, Peng R, Wang Y, et al. Significant Reduction of Antibiotic Consumption and Patients' Costs after an Action Plan in China, 2010–2014. Singer AC, ed. *PLoS One*. 2015;10(3):e0118868. doi:10.1371/journal.pone.0118868
10. Yoo KB, Lee SG, Park S, et al. Effects of drug price reduction and prescribing restrictions on expenditures and utilisation of antihypertensive drugs in Korea. *BMJ Open*. 2015;5(7). doi:10.1136/bmjopen-2014-006940
11. Uijtendaal E V., Zwart-van Rijkom JEF, de Lange DW, Lalmohamed A, van Solinge WW, Egberts TCG. Influence of a strict glucose protocol on serum potassium and glucose concentrations and their association with mortality in intensive care patients. *Crit Care*. 2015;19(1):270. doi:10.1186/s13054-015-0959-9
12. Petersen I, Welch CA, Nazareth I, et al. Health indicator recording in UK primary care electronic health records: key implications for handling missing data. *Clin Epidemiol*. 2019;11:157-167. doi:10.2147/CLEP.S191437
13. Rabe-Hesketh S, Skrondal A, Skrondal A. *Multilevel and Longitudinal Modeling Using Stata*. Stata Press Publication; 2008. Accessed August 4, 2019. https://books.google.co.uk/books?hl=en&lr=&id=woi7AheOWSkC&oi=fnd&pg=PR21&dq=Multilevel+and+Longitudinal+Modeling+Using+Stata+second+edition&ots=edLw2c_PUz&sig=Hw_YmlcM7lcKV5k00uYt6WKjz5A#v=onepage&q=Multilevel and Longitudinal Modeling Using Stata second edition&f=false
14. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, et al. Missing data and multiple imputation in clinical epidemiological research. *Clin Epidemiol*. 2017;9:157-166. doi:10.2147/CLEP.S129785
15. Roland M, Guthrie B, Thomé DC. Primary medical care in the United Kingdom. *J Am Board Fam Med*. 2012;25(Suppl 1):S6-S11.

16. Blak B, Thompson M, Dattani H, Bourke A. Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. *J Innov Heal Informatics*. 2011;19(4):251-255.
17. Horsfall L, Walters K, Petersen I. Identifying periods of acceptable computer usage in primary care research databases. *Pharmacoepidemiol Drug Saf*. 2013;22(1):64-69.
18. Maguire A, Blak BT, Thompson M. The importance of defining periods of complete mortality reporting for research using automated data from primary care. *Pharmacoepidemiol Drug Saf*. 2009;18(1):76-83.
19. Bak M, Fransen A, Janssen J, van Os J, Drukker M. Almost all antipsychotics result in weight gain: a meta-analysis. *PLoS One*. 2014;9(4):e94112.
20. Bazo-Alvarez JC, Morris TP, Carpenter JR, Hayes JF, Petersen I. Effects of long-term antipsychotics treatment on body weight: A population-based cohort study. *J Psychopharmacol*. Published online November 14, 2019:026988111988591. doi:10.1177/0269881119885918
21. Quartagno M, Grund S, Carpenter J. *Jomo: A Flexible Package for Two-Level Joint Modelling Multiple Imputation*.; 2019. Accessed July 25, 2019. https://journal.r-project.org/archive/2019/RJ-2019-028/RJ-2019-028.pdf
22. Prais SJ, Winsten CB. *Trend Estimators and Serial Correlation*. Cowles Commission discussion paper Chicago; 1954.
23. Goldstein H, Carpenter JR, Browne WJ. Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *J R Stat Soc Ser A (Statistics Soc*. 2014;177(2):553-564. doi:10.1111/rssa.12022
24. Quartagno M, Carpenter JR. Multiple imputation for discrete data: Evaluation of the joint latent normal model. *Biometrical J*. 2019;61(4):bimj.201800222. doi:10.1002/bimj.201800222
25. Van Buuren S. *Flexible Imputation of Missing Data*. Chapman and Hall/CRC; 2012.
26. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074-2102. doi:10.1002/sim.8086
27. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol*. 2008;61(4):344-349. doi:10.1016/J.JCLINEPI.2007.11.008
28. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLOS Med*. 2015;12(10):e1001885. doi:10.1371/journal.pmed.1001885
29. Funatogawa I, Funatogawa T. *Longitudinal Data Analysis: Autoregressive Linear Mixed Effects Models*. Springer; 2018.
30. Copas AJ, Seaman SR. Bias from the use of generalized estimating equations to analyze incomplete longitudinal binary data. *J Appl Stat*. 2010;37(6):911-922. doi:10.1080/02664760902939604
31. Carpenter JR, Kenward MG. *Multiple Imputation and Its Application*. John Wiley & Sons; 2013. Accessed August 4, 2019. https://www.wiley.com/en-us/Multiple+Imputation+and+its+Application+-p-9780470740521
32. Cro S, Carpenter JR, Kenward MG. Information-anchored sensitivity analysis: theory and application. *J R Stat Soc Ser A (Statistics Soc*. 2019;182(2):623-645. doi:10.1111/rssa.12423