Using Propensity Score methods for estimating real-world treatment effects in the presence of measurement error and sparse outcome data

by

Jane Burnell

A thesis submitted in partial fulfilment for the requirements for the degree of Master of Philosophy at the University of Central Lancashire

July 2022

STUDENT DECLARATION FORM

STUDENT DECLARATION FORM

Type of Award **Master of Philosophy**

School **Sport and Health Sciences**

1. Concurrent registration for two or more academic awards

I declare that while registered as a candidate for the research degree, I have not been a registered candidate or enrolled student for another award of the University or other academic or professional institution

2. Material submitted for another award

I declare that no material contained in the thesis has been used in any other submission for an academic award and is solely my own work

3. Collaboration

Where a candidate's research programme is part of a collaborative project, the thesis must indicate in addition clearly the candidate's individual contribution and the extent of the collaboration. Please state below:

Not applicable

4. Use of a Proof-reader

The following third party proof-reading service was used for this thesis *Julie Cook Academic Editing & Proofreading Services* in accordance with the Policy on Proof-reading for Research Degree Programmes and the Research Element of Professional Doctorate Programmes.
A copy of the confirmatory statement of acceptance from that service has been lodged with the Research Student Registry.

Signature of Candidate J. Burnell

Print name: JANE BURNELL

ABSTRACT

The real-world treatment effect of a novel treatment can be estimated by analysing routinely collected patient data, in the form of Electronic Health Records (EHR). Unlike a Randomised Clinical Trial, the treatment allocation in this type of data is not randomised and there may be systematic differences between the treatment groups. Propensity Score (PS) (Rosenbaum & Rubin, 1983) methods are commonly used to correct for these differences and hence reduce the bias in the treatment effect estimate. The combined impact on the treatment effect estimate of two common issues in EHRs were investigated; covariate measurement error and sparse outcome data.

A comparison was made between the performance of four PS methods: 3:1 PS matching; Inverse Probability Treatment Weighting (IPTW) for the Average Treatment Effect (ATE); IPTW for the Average Treatment Effect on the Treated (ATT); and PS stratification. Simulation experiments were run, based on a data extract from The Health Improvement Network holding UK primary care data. The impact of measurement error and outcome prevalence were investigated for different scenarios to represent real-life situations. For each simulation, PS conditioning was applied to the data to address treatment allocation bias before using Cox proportional hazards regression to estimate the treatment effect on the time-to-event outcome.

In data with higher outcome prevalence, covariate measurement error had little effect on the treatment effect estimate. For data with sparse outcomes, ≤ 1%, higher negative measurement error (corresponding to under-recording) produced treatment effect estimates with little bias, but lower precision. PS Stratification was the recommended method for estimating ATE with lower bias and higher precision over the measurement error range and over the outcome prevalence range. IPTW for ATT was the recommended method for estimating ATT with higher precision in all scenarios and lower bias, particularly when there was lower outcome prevalence.

Contents

ACKNOWLEDGEMENTS

I would like to thank the following for their contribution to this thesis:

Dr Svetlana Tishkovskaya, Dr Chris Sutton, Dr Gordon Prescott, Prof. Amitava Banerjee and Dr Helene Thygesen and my Research Degree Tutors for their guidance and advice throughout this study.

The University of Birmingham for supplying the data, which were extracted from The Health Improvement Network (THIN).

Dr Mark Lunt for his discussion on the use of PS methods and Dr Tim Morris for his guidance and tutorial on simulations.

WISER and June Thompson for their advice on academic writing style.

My colleagues in the Research Support Team (academic) and my fellow students in the Post Graduate Research Virtual Office, who have all helped me get through the lockdowns and enforced working from home in 2020 and 2021 due to the Covid-19 restrictions.

Last but not least, my husband, James, and my son, Ben, who have supported me through thick and thin during the course of this study.

TABLE OF TABLES

TABLE OF FIGURES

ABBREVIATIONS

Abbreviation	In Full
AF	Atrial Fibrillation
AIC	Akaike Information Criteria
AP	Apixaban
ATE	Average Treatment Effect
ATT	Average Treatment Effect on the Treated
ATU	Average Treatment Effect on the Untreated
BIC	Bayesian Information Criteria
CHA2DS2-VASc	Stroke risk score for patients with Atrial Fibrillation
CI	Confidence Interval
CKD	Chronic Kidney Disease
CPRD	Clinical Practice Research Datalink
CPS	Conventional PS
DGM	Data Generation Mechanism
eGFR	Estimated Glomerular Filtration Rate
EHR	Electronic Health Record
EPV	Events Per Variable
HAS-BLED	Risk score for bleeding in patients with Atrial Fibrillation
hdPS	High Dimension PS
HES	Hospital Episode Statistics
HR	Hazard Ratio
iFS	Inclusive Factor Score
IPTW	Inverse Probability of Treatment Weighting
LRT	Likelihood Ratio Test
MCS	Monte Carlo Simulations
MI	Myocardial Infarction
MI-EC	Multiple Imputation for External Calibration
MINAP	Myocardial Ischaemia National Audit Project
MLE	Maximum Likelihood Estimate
MO	Multiple Overimputation
MPS	Modified Propensity Score
MSE	Mean Squared Error
MVN	Multi-Variate Normal
NICE	National Institute for Health and Care Excellence
NOAC	Novel Oral Anti-Coagulant
OAC	Oral Anti-Coagulant
ONS	Office for National Statistics
OR	Odds Ratio
PH	Proportional Hazards
PLE	Penalized Maximum Likelihood
PS	Propensity Score
PSC	PS Calibration
RCT	Randomised Control Trial
REWARD	Performance-Based Innovation Rewards
RI	Rivaroxaban
RMSE	Root Mean Squared Error
SD	Standard Deviation

SE	Standard Error
SIMEX	Simulation Extrapolation
STUVA	Stable Unit Treatment Value Assumption
THIN	The Health Improvement Network
TIA	Transient Ischemic Attack
WA	Warfarin

Chapter 1 INTRODUCTION

1.1 Background

Although a Randomised Controlled Trial (RCT) is seen as the gold standard for estimating the effect of a novel treatment, the treatment effect in a real-world setting when it is prescribed to a more general population is likely to be different. In a RCT the inclusion criteria will often mean that participants are likely to have fewer comorbidities and be younger than the general population. There may be higher adherence to the treatment in a RCT due to additional information provided to participants and more follow up appointments. These could account for a better performance of the novel treatment in a RCT than in the real-world setting. Estimation of the real-world treatment effect of a novel treatment will add to the evidence of the performance of the novel treatment to support national prescribing guidance.

The increasing availability of Electronic Health Records (EHR) data offers the opportunity for the estimation of the real-world treatment effect from observational studies. However, the treatment allocation is not randomised so there are likely be systematic differences between the treatment groups, and if this is not accounted for the treatment effect estimate will be biased. Propensity Score (PS) methods are popular in applied medical research for adjusting for this treatment allocation bias. This study applies four commonly used PS methods in the presence of other common real-world problems associated with data from EHR, measurement error and sparse outcome data, to estimate the real-world treatment effect. The aims of this study are to demonstrate the impact of measurement error and sparse outcome data on the 'treatment effect estimate' and to recommend PS methods for use under these conditions.

1.2 Observational studies and real-world effectiveness

In applied medical research there are generally two types of studies used to determine the effectiveness of a novel treatment: RCTs and Observational Studies (Cochran & Chambers, 1965). RCTs are the gold standard for assessing the efficacy of a novel treatment (Sibbald & Roland, 1998). Randomisation will, on average, balance the treatment groups for measured and non-measured variables. Participants are randomly assigned to the novel or control treatment, with the allocation balanced by key variables, such as recruitment centre, which may be thought to affect the outcome. As observational studies are not randomised, systematic differences in the baseline characteristics of participants are likely to exist between the treatment groups. Additionally, treatment allocation may be due to participant characteristics (Austin, 2011a). Any estimation of the treatment effectiveness will be biased in the presence of either of these.

Observational studies have some advantages. They can overcome ethical issues of randomising patients to a potentially less favourable treatment in an RCT. For example it would not be ethically correct to randomise participants to a treatment which had been an earlier standard treatment and was substantially less effective to make a comparison to a new treatment (Pruzek, 2011). By using existing data these treatments can be compared in an observational study. Running a RCT is a costly and time-intensive process and the number of participants recruited relatively low. In observational studies data will generally be available for more participants so the study has access to a larger population and hence the treatment effect estimate will have lower variability (Altman, 1991, p. 167). In a RCT the inclusion and exclusion criteria for participant recruitment can mean that sub-groups of participants most likely to respond well to the treatment are recruited, so the effectiveness results in a RCT can be better than in a general population. Additionally, in a RCT participants may be more likely to adhere to the treatment regime due to additional information, ongoing follow up from the trial management teams and closer care from the clinical teams. Observational studies give an estimate of effectiveness in a real-world setting where the treatment is used by a more general population and in clinical practice. However, the 'real-world' treatment effectiveness found in an observational study is likely to differ from the estimate of the treatment effectiveness given by RCTs. These are both estimates of the true treatment effectiveness.

1.3 Sources of bias in Electronic Health Records

Sources of bias in the design and analysis of studies are well documented (B. C. Choi & Pak, 2005; Sackett, 1979). However, studies using EHR introduce specific sources of bias which should be considered when planning these studies. If these are not addressed, they will lead to a bias in the treatment effect estimate. Treatment allocation bias (Section 1.2) is one such source of bias. Unmeasured confounders are another source of bias. If variables which contribute to the treatment allocation model or the outcome model are not recorded and available for use or contain incomplete data, the treatment effect estimate will be biased. The effect is widely reported in the literature and is not investigated as part of this study. Random measurement error of baseline covariates also introduces bias. Generally, the data in EHR have been collected for medical reasons and their use for applied research is secondary. De Gil et al. (2015) and Steiner, Cook and Shadish (2011) demonstrate that measurement error does introduce bias and report that it should not be ignored, although the effect is not as strong as that introduced by missing confounders (Steiner et al., 2011). The focus of this study is the effect of measurement error but not that from missing or incomplete confounders.

1.4 Propensity Score analysis

Propensity Score (PS) methods (Rosenbaum & Rubin, 1983) comprise of a range of approaches to balancing treatment groups, thus reducing the treatment allocation bias and hence obtaining a less biased estimate for the treatment effect estimate. This may be in the form of the Average Treatment Effect estimate (ATE) (Imbens, 2004), the effect of the treatment on the whole population, or the Average Treatment Effect for the Treated (ATT) (Imbens, 2004), the treatment effect of those who are selected to receive the treatment. The PS is defined as the probability of treatment assignment based conditionally on observed baseline covariates (Austin, 2011a). The closer the PS is to one the more likely the participant is to be in the novel treatment group, and the closer the PS is to zero the more likely they are to be in the control group. The PS is a balancing score, and is a function of the observed covariates for which the conditional distribution of the covariates, given the balancing score, is the same for the control and treated (Raykov, 2012). The adjustment for the treatment allocation bias, using the PS method, is applied separately from the outcome analysis.

Other methods to adjust for the systematic differences between the treatment groups include regression-based methods. PS methods offer some advantages over regression methods. These include firstly, separating the study design from the analysis, using PS methods (matching, stratification and IPTW) the PS model is built and checked without knowledge of the outcome. With regression methods the outcome is always used. Secondly, it is easier to check for correct specification of the PS model using balancing checks than using regression where goodness-of-fit methods do not check that the model is correctly specified nor that the systematic differences between the treatment groups have been eliminated. Thirdly, with rare binary or time-to-event outcomes, PS methods have more flexibility, unlike regression which may have poorer performance with a low number of events per covariate in the model. Fourthly, PS methods allow inspection of the overlap in the distribution of the baseline covariates between the treatment groups. If the overlap is small, the researcher will be aware of this and can decide whether to proceed with the analysis or not. Using regression methods, there may be no such indication (Austin, 2011a). PS methods were therefore chosen for this study.

1.5 Measurement error

If the value of an observation does not match its true value, this is known as ‘measurement error’ (Wallace, 2020). Often in the literature ‘measurement error’ relates to a continuous variable and ‘misclassification’ refers to a categorical variable (Keogh et al., 2020). In this study measurement error will be used to mean both. There has been a lack of application of

measurement error methods within applied research, with a common approach to dealing with it being to ignore it (De Gil et al., 2015; Millimet, 2011). The focus of this study is to consider measurement error when PS methods are used to balance the treatment groups in an observational study before applying the outcome analysis.

The types of measurement error which may occur in EHR are covariate measurement error, outcome measurement error and treatment allocation measurement error. The measurement error investigated in this study is measurement error in a covariate in the treatment allocation model, which is the PS model. If such measurement error exists, the treatment groups will be balanced on the observed not the true covariates, so differences between the treatment groups will still exist, meaning that this could be a source of bias in the outcome analysis (Nguyen & Stuart, 2020). Measurement error may be differential, where the measurement error also depends on the outcome, or non-differential, where the measurement error does not depend on the outcome (Carroll, Ruppert, Stefanski & Crainiceanu, 2006, p. 36). The study data are assumed to have non-differential covariate measurement error; in this context that means that measurement error is the same across the treatment groups. Differential measurement error may occur if different data sets are combined for analysis, one for the treated group and a different one for the control, or if different methods or validated tools have been used to measure the covariate (Hong et al., 2017). This is not the case in this study, where the data were collected in the same way for both treatment groups. Additionally, they were collected prior to treatment allocation or when the primary outcome occurred. Suitable methods to correct for non-differential covariate measurement error are discussed in Section 2.4.

1.6 Sparse data

Sparseness in data can be caused by any of the following: small sample size (Siino, Fasola & Muggeo, 2018); rare exposure (treatment) (Hajage, Tubach, Steg, Bhatt & De Rycke, 2016); rare outcome events (Siino et al., 2018) which lead to a low number of events per variable (EPV) (Greenland, Mansournia & Altman, 2016); unbalanced or highly predictive risk factor variables (Siino et al., 2018) with narrow distributions or categories which are uncommon; variables which almost perfectly predict the outcome (Greenland et al., 2016); variables that together almost perfectly predict the exposure (treatment) (Greenland et al., 2016). Sparse outcome data are the focus of this study and it was assumed that there were rare outcome events because the prevalence of the outcome events was low and not due to missingness in the outcome data. Sparse outcomes can take the form of rare outcomes in observational studies to estimate treatment effect, or Serious Adverse Events in drug safety studies (Ross et al., 2015). Even though Serious Adverse Events in drug safety studies may be rare, they are important. An

example is Das et al. (2016) who looked at the challenges of trial design in the Neonatal Research Network under these circumstances. Rare outcomes in a large dataset are not uncommon (Chao, 1994; Franklin, Eddings, Austin, Stuart & Schneeweiss, 2017; Paul & Deng, 2000). Sparse data bias produces treatment effect estimates which are away from the null, so the treatment effect estimates are inflated (Greenland et al., 2016). Methods for working with sparse data are discussed in Section 2.5, although this study is limited to the use of PS methods in the presence of sparse outcome data.

1.7 The study dataset

The Health Improvement Network (THIN) is one of the UK primary care datasets, containing data collected for clinical purposes. The motivational example for this study is the assessment of the treatment effect of Novel Oral Anti-Coagulants (NOAC) compared with the existing treatment Warfarin, an Oral Anti-Coagulant (OAC) for the prevention of future stroke or Transient Ischaemic Attack (TIA) (this will be called 'future stroke') in patients with Atrial Fibrillation (AF). By using the data extract from THIN provided for the REWARD (Performance-Based Innovation Rewards) study (Banerjee et al., 2020), an assessment of real-world effectiveness was made. This dataset included data on the treatments prescribed (primarily anticoagulants), the outcomes (time to stroke and time to bleed event) and variables (covariates) likely to be influential to these.

A sub-set of the full REWARD data extract was used in this study with cases (used to describe study participants) with AF. The PS methods used in this study only adjust for treatment allocation bias between two treatment groups. Cases prescribed Apixaban and Dabigatran were dropped (Section 3.2) and those who had been prescribed the novel treatment Rivaroxaban (a NOAC) or the control treatment Warfarin were retained. Warfarin cases whose first NOAC/OAC prescription was before the National Institute for Health and Care Excellence (NICE) approval date for Rivaroxaban were discarded. The primary outcome was future stroke. This sub-set of the data will be referred to as 'the study dataset'.

The types of measurement error potentially present in the study dataset are covariate measurement error (covariates used in either the treatment allocation model or the outcome model) and outcome measurement error. There is unlikely to be treatment allocation measurement error as a comprehensive listing of prescribing data was provided in the original data extract, which is believed to be correct and accurate. The data are assumed to have non-differential covariate measurement error that is the same across the treatment groups.

This study will focus on the impact of measurement error in the variable for previous stroke, a covariate in the treatment allocation model, on the treatment effect estimate. Work in the REWARD study (Burnell, 2015) has shown stroke is under-recorded in primary care records when compared with linked Hospital Episode Statistics (HES) data. This is supported by Herrett et al. (2013) who showed that there was a 25% under-recording of myocardial infarction (MI) in primary care data when compared with the recording in three data sources: primary care data; hospital data; and the disease registry. However, the version of the data used to generate the dataset for the current study did not have linked HES data, so there were no external calibration data available.

Although the study dataset had 21,259 cases, it only had 232 future strokes (the primary outcome) recorded. The dataset therefore had an outcome prevalence of approximately 1.1% and is an example of a large dataset with sparse outcome data and will be used to investigate sparse data bias.

1.8 Summary of this study

This study will assess the impact of the combination of measurement error and rare outcomes when using PS methods to adjust for treatment effect bias when using EHR to estimate real-world treatment effect. This is a novel approach and to date has not been reported in the literature. The aims of the study are to compare the performance of four selected PS methods in the estimation of the treatment effect estimate in the presence of covariate measurement error and sparse data. This may be used to inform which PS methods perform best in the estimation of real-world treatment effect in situations where there are problems commonly seen in EHR datasets: covariate measurement error and sparse outcome data.

In the REWARD study, the study dataset was extracted to investigate the performance of a group of NOACs compared to Warfarin in the prevention of future stroke for patients with AF. However, in the current study it was used to generate simulated datasets to compare the performance of four different PS methods when used in the estimation of the treatment effect in the presence of covariate measurement error and sparse outcome data.

1.9 Introduction to the thesis

Chapter 2 reviews the literature relating to PS methods, measurement error and sparse data. Details of the literature searches are given in Appendix A. It considers methods which are used to adjust for measurement error when PS methods are used and the use of PS methods in the presence of sparse outcome data. These methods are assessed to see if they are compatible with the characteristics of the study data. Chapter 3 describes the selection of the study dataset

from the original data extract. The methods are presented which applied the PS methods, 3:1 PS matching, IPTW for ATE, IPTW for ATT and PS stratification, to the study dataset to correct for systematic differences between the treatment groups. It then describes how Cox proportional hazards regression was applied for the outcome analysis to estimate the treatment effect estimate. The original characteristics of the dataset are retained, before applying measurement error and changes to the outcome prevalence in the following chapters. Chapter 4 develops the simulations framework which runs the method from Chapter 3. The simulations method also allowed parameters to be introduced to vary the amount of introduced measurement error and the sparseness in the outcomes to assess their impact on the treatment effect estimate's performance. By running simulations, performance measures of the treatment effect are generated to allow for the comparison of the different PS methods. Chapter 5 presents the results of the simulations. It compares the performance of the PS methods with the original data characteristics. Measurement error is introduced into the baseline variable for previous stroke, a covariate in the treatment model. The effect size in the treatment model of the variable with measurement error (previous stroke) is varied, as well as the measurement error, to investigate the impact of a stronger predictor of treatment allocation and to make the work more generalisable. The prevalence of future stroke, the primary outcome, is varied in addition to the introduced measurement error, to demonstrate the impact of sparse data bias. Finally, the measurement error in previous stroke (a covariate in the PS model), its effect size in the PS model and the outcome prevalence (future stroke) are jointly varied. Recommendations for the PS method to use in the estimation of the ATE and in the estimation of the ATT are made in these scenarios. Chapter 6 discusses the study's findings in the context of the literature.

Chapter 2 LITERATURE REVIEW

2.1 Introduction

This study compares the performance of four commonly used PS methods when using EHRs to estimate the real-world treatment effect of a novel product. It also investigates the effect of other real-world problems associated with EHR data: covariate measurement error and sparse outcome data. This chapter reviews the literature regarding PS methods, their implementation and their use with measurement error and sparse data in order to understand the current knowledge gaps and research needs. The potential (or counterfactual) outcomes framework is introduced, as PS methods work within this. The steps to run an analysis using PS methods are presented, including details of the four main categories of PS methods and a review of the comparison between their performance. An overview of measurement error models is given and PS analysis methods which correct for measurement error are compared for their applicability to the study data. The problems relating to sparse data in the estimation of treatment effects are introduced. Methods to address sparse data bias are presented from the literature and studies which used PS methods in sparse data settings and are compared with the current study's requirements.

2.2 The Potential Outcomes Framework and assumptions

In the Potential Outcomes Framework (or Counterfactual Framework) every participant can have two potential outcomes. For participant *i* they are $Y_i(0)$ if the control treatment were received and $Y_i(1)$ if the novel treatment were received. The treatment effect for participant *i* would be $Y_i(1) - Y_i(0)$. Each participant will only receive one of the treatments (the other is counterfactual) so this cannot be calculated. The observed outcome is $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$ where $Z = 0$ for the control treatment and $Z = 1$ for the novel treatment. Using all participants in the study population $E[Y(1) - Y(0)]$ will give the Average Treatment Effect (ATE) (Imbens, 2004). This could be described as moving the whole population from untreated to treated (Austin, 2011a). The Average Treatment Effect of the Treated (ATT) is given by $E[Y(1) - Y(0) \mid Z = 1]$ (Imbens, 2004). This is the effect of the treatment over the control for the sub-population of those treated. The ATT could be thought of as the treatment effect for those for whom the treatment was intended (Caliendo & Kopeinig, 2008). Occasionally the Average Treatment Effect of the Untreated (ATU) is used, given by $E[Y(1) - Y(0) \mid Z = 0]$ (Williamson, Morley, Lucas & Carpenter, 2012a). The research question to answer will determine whether the ATE, ATT or ATU should be calculated (Williamson et al., 2012a).

In a RCT the ATE would on average equal the ATT due to randomisation as the treated population should not systematically differ from the whole population (Austin, 2011a). In observational studies there are likely to be systematic differences between the populations from which the groups have been sampled, so the ATE and ATT are likely to differ.

In terms of time-to-event data there is the idea of two survival curves, one where all participants received the treatment and the other where they all received the control. The treatment effect could be calculated using 1) an estimate of the absolute difference between the groups or 2) an estimate of the relative treatment effect. These are known as 'measures of ATE' and by considering only those who received the treatment this becomes 'measures of ATT' (Austin, 2014b).

The Potential Outcomes Framework makes the assumptions of *positivity*, *ignorable treatment assignment assumption* and *Stable Unit Treatment Value Assumption* (STUVA). Positivity means that each participant has the potential to receive either treatment. The ignorable treatment assignment assumption is that conditional on a set of covariates, for the participants the assignment to the treatment is independent of their potential outcomes Y_0 and Y_1. This is also known as 'unconfoundedness', 'selection on observables' and 'exogeneity' (S. Guo & Fraser, 2015, p. 29). STUVA is that the two potential outcomes for a participant are independent from any other participant's exposure.

2.3 Propensity Score methods

Propensity Score methods were first presented by Rosenbaum and Rubin (1983) as a way to give an unbiased estimate of the causal effect in non-randomised studies. They use the idea of treatment assignment being strongly ignorable, and then show that any balancing score will remove systematic differences and gives an unbiased estimate of the treatment effect at that value of the balancing score. They propose the PS and show that it is the coarsest of all balancing scores. They then apply the PS to adjust for confounding in three existing techniques: matched sampling; subclassification; and covariance adjustment.

2.3.1 Overview of Propensity Score analysis

The literature suggests the steps to run a PS analysis are: generate the PS; check for balance; apply the PS method; check for balance; estimate treatment effect; run sensitivity analyses (Austin, 2009a; Austin, 2011b; Garrido et al., 2014; Li, 2013). Table 1 summarises these and the following sections describe each step in more detail. If any of the tests fail, then the PS model should be redefined and the process starts again. This can be an iterative process. PS methods are applied in addition to the main outcome analysis. This is regarded as a two-step approach;

firstly the PS analysis is applied to adjust for treatment allocation bias (Table 1 Steps 1 to 4) and secondly the outcome analysis is performed (Table 1, Step 5).

Table 1: The steps in Propensity Score analysis.

Step #	Step Title	Step Description
1	Choose the PS model and Generate the PS values	Select the covariates to be used in the PS model. Choose the modelling method. Generate the value of the PS.
2	Checks for balance	Run basic checks, check for common support, that is overlap of the PS distributions between groups. Check for balance of the PS and covariate balance between the treatment groups.
3	PS Method	Select and Apply the PS method (also known as PS conditioning).
4	Checks for balance after PS conditioning	Check for covariate balance, appropriate to the PS conditioning method.
5	Estimate Treatment Effect	Apply the method appropriate for the outcome analysis to estimate treatment effect.
6	Sensitivity tests	Compare the treatment effect estimate with an unbiased value.

2.3.2 Step 1 - Choose and generate the Propensity Score

The literature is divided over the selection of the covariates to include in the PS model. The options are variables which influence treatment allocation and outcome, variables which influence only treatment allocation, or variables which influence only outcome. Including variables which influence treatment allocation and outcome is more likely to include any (measured) confounders than the other options. The best model includes only true confounders, makes the model more parsimonious and gives greater precision with no increase in bias (Austin, Grootendorst & Anderson, 2007). However omitting a true confounder can lead to imbalance between the treatment groups and a biased estimation of the treatment effect, (Austin et al., 2007). In small studies, variables strongly related to treatment allocation and weakly related to outcome can give a treatment estimate with a higher Mean Squared Error (MSE) (Brookhart et al., 2006). Including variables which influence only treatment allocation can decrease precision without decreasing bias (Brookhart et al., 2006). Additionally, including variables which influence only treatment allocation may not improve balance and can reduce the number of matched pairs (where a treated participant is matched to an untreated participant with a similar PS) when using PS matching (Austin et al., 2007). Previously the PS was seen as a treatment allocation model but now there is a strong suggestion to include at least some variables which affect the outcome, (Austin et al., 2007). Brookhart et al. (2006) suggest, at least in a large dataset, including all the variables affecting the outcome as they may be related to treatment

assignment and hence reduce unmeasured confounders and reduce bias (Brookhart et al., 2006; Garrido et al., 2014). However, in smaller datasets including all variables may cause too much 'noise', counteracting any benefits of reduction of bias from their inclusion. Covariates which are weakly associated with outcome and not associated with treatment allocation should be excluded (Garrido et al., 2014). Additionally, clinical knowledge should be used to choose the variables to include in the PS model combined with one of the above suggestions. Caliendo and Kopeinig (2008) summarise variable choice as based on theory and previous empirical findings.

Logistic regression is the most popular method used to model the PS (Austin, 2011a; Cham & West, 2016). Although interaction and polynomial terms can be used in the PS model, logistic regression assumes linearity between the terms and the logit of the PS. Machine learning methods, classification trees, random forests and generalised boosted modelling, can also be used to generate the PS. They estimate arbitrary nonlinear relationships between the covariates and the assignment to the treatment or control group (Cham & West, 2016). Random Forests with weighting method is recommended over Classification Trees as it has a lower bias for both ATE and ATT (Cham & West, 2016). When using generalised boosted modelling, selecting the correct number of iterations can be difficult. If it is too high, the generalised boosted model will be overfitted and the estimated PS values are biased towards 0 or 1 (Cham & West, 2016). Boosted logistic regression, another machine learning method, can lead to better covariate balance, but due to the highly flexible model it can produce high variance when there are rare outcomes (Franklin et al., 2017). In high dimensional covariate space, Lasso regression and Bayesian logistic regression can be used to model the PS (Franklin et al., 2017). Lasso regression can generate imprecise coefficients and some covariates are dropped from the PS model, so Bayesian logistic regression, which shrinks estimated coefficient towards zero will retain all covariates in the PS model, is recommended (Franklin et al., 2017).

Regardless of the method chosen, the PS model should be built without knowledge of the outcome (Garrido et al., 2014). This reduces any bias introduced by the analyst and is in keeping the method similar to a RCT where study design and randomisation are complete before an outcome is measured. If several outcomes are to be analysed, a different PS model could be used for each outcome (Austin et al., 2007).

2.3.3 Step 2 - Checks for balance

Once the value of the PS is calculated a visual check for overlap also known as 'common support' of PS distributions between the treatment groups should be made (Garrido et al., 2014). If there is insufficient common support, in PS matching the treated participants would be matched with

only a small selection of the comparison group. In this case the PS model should be redefined and the process restarted. If the lack of common support is only evident at the lower or higher strata (using the PS values), then untreated subjects with very low or very high PS could be dropped, but this changes the study population (Austin, 2011b).

Not all authors suggest performing balance checking at this stage, although (Garrido et al., 2014) recommends creating strata or blocks by the PS and checking for balance of the PS values between treatment groups and then checking for balance of covariates within PS blocks. Step 4 (Section 2.3.5) describes these balance checks. If imbalance is found the PS model will need to be redefined and the PS analysis re-run.

2.3.4 Step 3 - Select and apply the PS method

There are four general categories of methods for using PS to remove the effect of confounding: PS matching, stratification on the PS, inverse probability treatment weighting (IPTW) on the PS and covariate adjustment on the PS (Austin, 2011a). In their original paper, Rosenbaum and Rubin (1983) presented PS adjustment to the existing methods of matched sampling, subclassification (stratification) and covariance adjustment (covariate adjustment). IPTW was first presented by Rosenbaum (1987). PS matching will estimate the ATT (Imbens, 2004) and the other methods can be implemented to estimate the ATE or ATT.

PS matching is when matched pairs or groups are created by matching each treated participant to an untreated participant with a similar PS. The estimate of treatment effect is generated from the matched sample or dataset, where only cases for which a match is made are retained. The variance of the treatment effect can also be estimated from the matched sample, but there is discussion as to whether the treated and untreated participants are independent or not (Austin, 2011a). If the data in the matched sample are to be regarded as not independent, then statistical tests which account for the matched nature of the data will need to be applied in the outcome analysis. Analysis using PS methods can mimic that of a RCT, using the same reporting metrics as used for RCTs, which will depend on the type of outcome.

There are different matching algorithms which can be used. Austin (2014a) compares 12 matching algorithms. The following choices will help to determine the matching algorithm used:

- With or without replacement?
    - With replacement – the matched control case is returned to the pool for the next (and subsequent) matches

    - Without replacement - the matched control case is not returned to the pool for subsequent matches
- Greed or optimal?
    - Greedy matching - the best match is made for each treated participant
    - Optimal matching - minimises the within pair difference for the whole dataset
- How close?
    - Nearest neighbour (NN) - a treated case is matched to the untreated case with the closest PS
    - NN with caliper - imposes a maximum difference between the PS score for a match to be allowed. There is discussion about the best size of caliper to use. For example, matching is often on the logit of the PS with a caliper of 0.2 of pooled standard deviation (Austin, 2011a)
- How many matches?
    - The most common is 1:1 matching
    - Many:1 matching, for example 3:1 matching matches 3 untreated to 1 treated
    - Full-matching is where matched sets are formed of either one treated case and at least one untreated case or one untreated case and at least one treated case

PS matching can be combined with additional matching on a variable thought to have a strong influence on treatment allocation or outcome. Both PS matching and PS stratification within pair/stratum regression analysis can be performed to account for residual differences between the treated and untreated (Austin, 2011a).

PS stratification is developed from subclassification which Cochran (1968) developed to balance data. Subclassifying on a covariate which is known to affect treatment allocation removes the bias due to this covariate. Removing bias in this way to account for additional covariates generates large numbers of strata. Combining all such covariates into the PS and stratifying by the PS reduces the number of strata needed. This was first implemented by Rosenbaum and Rubin (1984) to reduce bias in the treatment effect estimate.

Stratification on the PS is performed by ordering the records of all participants by PS and then stratified into groups, with five equal-sized strata commonly used. The treatment effect is estimated within each stratum and then these stratum-specific results pooled to generate the ATE and the Standard Error (SE) of the estimate (S. Guo & Fraser, 2015, p. 205). These estimates are weighted by the number of cases in each stratum, to generate the ATE from *k* strata each

stratum is weighted as $1/k$, to generate the ATT the weights will be the number of treated cases in each stratum (Austin, 2011a). PS Stratification can be combined with multivariate analysis, including survival analysis.

The number of strata used needs consideration. The PS should be constant within a stratum or the number of strata sufficiently large and the differences in the values of the PS between strata small (Imbens & Wooldridge, 2009). Five strata were suggested by Cochran (1968) and Rosenbaum and Rubin (1984) and are widely used in studies. Lunceford and Davidian (2004) showed that five strata may not remove bias and that the number of strata should be based on the number which can give covariate balance. The number of strata used is a compromise. A larger number of strata gives better homogeneity within each strata and reduces the bias in the treatment effect estimate, while a smaller number of strata will have more observations in each strata giving lower variance in treatment estimates (S. Guo & Fraser, 2015, p. 208).

Inverse Probability of Treatment Weighting (IPTW) on the PS uses weights, based on the PS, to generate a synthetic dataset. The weight is defined as the inverse probability of receiving the treatment the participant actually received. The ATE, ATT and ATU can be calculated but each will use a different formula for the weights. Variance estimates must take account of the weighted nature of the data and robust variance estimation is commonly used (Joffe, Ten Have, Feldman & Kimmel, 2004).

IPTW was first proposed by Rosenbaum (1987). Joffe et al. (2004) report that weighting removes the covariate imbalance across treatment groups. IPTW has a doubly-robust property so will lead to unbiased estimate of the treatment effects even if the regression models do not represent the true models, hence IPTW is more robust to model misspecification (Lunceford & Davidian, 2004). Morgan and Todd (2008) presented a 9-step routine when using weighted regression, to estimate an average causal effect and to assess its bias. The method in the current study (Section 2.3) is not dissimilar to theirs. Hirano, Imbens and Ridder (2003) implemented IPTW and showed that using a nonparametric estimate of the propensity score, rather than the true propensity score, led to efficient average treatment effects estimates.

In Covariate Adjustment using the PS, the outcome is regressed on the treatment (as an indicator) and the PS. This method is the only one that requires access to the outcome at this stage.

The four main PS methods, described above, have different implementations. PS matching and PS stratification use the PS to group the data but do not use the PS directly when estimating the

treatment effect. As IPTW and covariate adjustment both directly use the PS in the analysis, Austin (2014b) suggests that these methods are more sensitive to the accuracy of the PS estimate. However, weighting is a 'doubly-robust' property so is more stable to model misspecification (Lunceford & Davidian, 2004). PS matching, stratification and IPTW use a 2-step method: model the PS then perform the outcome analysis to estimate the treatment effect. Once the PS model has been correctly specified the modelling for the outcome requires no further adjustment. However when using covariate adjustment, the form of the outcome model needs to be considered, such as if it is linear or non-linear. Covariate adjustment is the only method where the outcome needs to be 'visible' so the temptation is to model towards the known outcome (Austin, 2011a).

Most papers which compared the performance of different PS methods used data with binary outcomes. Different PS methods address systematic bias by different amounts (Austin, 2011b). PS matching and IPTW remove more systematic differences between treatment groups than does PS stratification and covariate adjustment, with PS matching and IPTW removing the systematic differences to a similar extent (Austin, 2009b, 2011a, 2011b). When the common support is not good, matching performs better (Busso, DiNardo & McCrary, 2014). So properties of the dataset may guide the choice of PS conditioning.

Using PS stratification, the bias increases as the sample size does. This can be offset by increasing the number of strata used, however quintile stratification was commonly used (Lunceford & Davidian, 2004). Weighting methods give unbiased estimates for 'realistic' sample sizes. Stratification and unadjusted methods can outperform weighting.

Greater balance is achieved by matching compared to stratification, but matching was working with a smaller sample size (Austin & Mamdani, 2006). This may be reflected in stratification having greater bias and matching having reduced precision. This is reported as 'the variance versus bias trade off' (Carroll et al., 2006, p. 60). A comparison between 12 PS matching algorithms drew a similar conclusion, reporting that matching algorithms without caliper have larger sample, and hence greater precision, whereas matching with caliper gives less bias (Austin, 2014a). In large samples (asymptotically) all PS matching methods should give the same results, although for smaller samples the choice of matching is usually a trade-off between bias and efficiency (Caliendo & Kopeinig, 2008). Ultimately, the choice of PS method will depend on the data properties.

When using time-to-event outcomes, like the study data, Austin (2014b) recommends using matching and IPTW. Neither method was seen as superior as they both reduced treatment

allocation bias to the same extent. Each method had advantages and limitations; the situation may dictate which to use. With PS matching the method was more 'transparent', and less sensitive to model misspecification, but only estimates ATT, and needed a pool of controls at least as large of that of the treated. Weighting (IPTW for ATE and IPTW for ATT) was applicable to more complex situations. Austin (2013) ran simulations on time-to-event data to investigate the performance of different PS methods, different methods to account for the matched or weighted nature of the data and varied the exposure prevalence. They reported that PS stratification and covariate adjustment gave biased estimates of the marginal hazard ratio (HR) and recommend PS matching and IPTW (for ATE & ATT). Both had minimal bias but IPTW had lower Mean Squared Error (MSE). The limitations were that they only used greedy NN matching and 1:1 matching not many:1 matching. The recommended PS methods for use with time-to-event data to estimate the marginal HR are PS matching and PS weighting (IPTW), (Austin, 2013). The circumstances will dictate which method to use (Austin, 2014b).

A further recommendation is to use a number of different PS methods and present the most promising (Caliendo & Kopeinig, 2008; Garrido et al., 2014), which would be the method achieving the best balance. So although PS matching or IPTW would appear to be the methods to consider for analysis of the study data, comparisons using PS stratification and covariate adjustment may be useful.

2.3.5 Step 4 - Balance of covariates after Propensity Score conditioning

Balance of the more influential variables in the PS model is important. If the data are not shown to be balanced, the variable selection for the PS model will have to be adjusted (Austin, 2011a). If PS matching or IPTW have been used, then these checks should be applied to the generated matched or weighted sample respectively. The literature agrees that checks for balance should be carried out at this stage (Austin, 2011a; Garrido et al., 2014; Rubin, 2004; Williamson et al., 2012a).

Conditional on the *true* PS, treatment allocation is independent of the measured covariates. This means treated and untreated cases with the same *true* PS will have the same covariate distribution (Rosenbaum & Rubin, 1983). If the distribution of the covariates is similar for the matched cases with the same PS, then the PS is sufficiently well defined (Ho, Imai, King & Stuart, 2007). As the *estimated* PS is being generated, tests on the difference of the covariate distributions will indicate if the *estimated* PS is sufficiently close to the *true* PS (Austin, 2009a).

These checks are applicable to the PS matched dataset and can be applied for use with PS Stratification by applying the checks within each strata (Austin, 2011a). For PS matching which

uses many-to-one matching, Austin (2008a) recommends weighting the matched control cases before performing the balance checks.

There is no definitive set of tests for balance checking, but the tests described here are among those recommended for use. For PS matching, standardised differences should be calculated for each continuous or binary covariates. If the PS matching has adjusted for the systematic differences between the treatment groups, the standardised differences should be low. These are often compared with the standardised differences for the covariates in the original, unmatched dataset by presenting them on the same plot. Additional tests include comparison of higher order moment of baseline covariates between treatment groups (Imai, King & Stuart, 2008). An example is Variance Ratios (Austin, 2009a) which compares variances of these covariates and gives a broader comparison. Visual inspection of side-by-side box plots and Q-Q plots for continuous variables can be inspected to assess how the systematic differences have been removed by the PS matching (Austin, 2009a; Garrido et al., 2014).

Caliendo and Kopeinig (2008) also listed checks for balance of covariates as standardised bias, t-test, joint significance and pseudo R^2 and the stratification test. However, tests which were previously used to assess balance, such as comparison of distributions of the PS for each group, significance testing of baseline covariates and the c-statistic and Receiver Operating Characteristic (ROC) Curve of the PS, should no longer be used as they have been shown to not differentiate between correctly specified PS models and misspecified PS models (Austin, 2009a).

Morgan and Todd (2008) presented a 9-step method to perform an analysis using inverse weighting of the PS (or IPTW). This included checking the balance between the groups in the weighted sample, using an average of standardised mean differences across treatment and control groups, and also the higher moments of the distributions of the covariates in the PS model. Joffe et al. (2004) used graphical displays (boxplots) to show the covariate distributions in original and the weighted sample. Austin and Stuart (2015) presented graphical comparisons of distributions of continuous variables and numerical comparison of distributions of continuous variables (using the Kolmogorov-Smirnov test). If the weights and hence the PS model fail to balance the groups, the PS model should be re-specified.

When using covariate adjustment, Austin (2008b) proposed two diagnostic methods to check for balance. The first is weighted conditional standardised difference and can be used for continuous or dichotomous variables. The standardised difference is the difference in means of that variable in the two treatment groups divided by the common standard deviation of the variable in the two treatment groups. It is the number of standard deviations by which the two

treatment groups differ. The conditional standardised absolute difference is integrated over the distribution of the estimated PS to produce the weighted conditional standardised absolute difference.

The second method is quantile regression, which compares conditional distributions of continuous variables. It compares the distributions of measured baseline covariate(s) in the different treatment groups with the same PS. A quantile of the dependent variable is regressed on baseline covariates, this can be repeated with several quantiles, and Austin (2008b) proposes using 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th}. The distribution of the outcome at specific values of the PS for the treatment groups can be investigated by plotting the estimated regression quantiles against the estimated PS for the participants in each treatment group.

Weighted conditional standardised differences will show if there is a difference between the means of a covariate between the treatment groups, conditional on the PS. Quantile regression to compare conditional distributions of continuous variables will show more about the distribution of the conditional difference between the two treatment groups. However, under the following conditions quantile regression does not provide any extra information and weighted conditional standardised differences should be used. Firstly, the distribution of the baseline covariate, conditional on the PS, is symmetrical for each treatment group and the conditional distribution takes the same shape for each treatment group. Secondly, if the conditional distribution is the same shape for each treatment group and only the location is shifted.

2.3.6 Step 5 - Treatment effect estimates

To obtain the estimated treatment effect, standard statistical analysis should be run. The methods must take account of the nature of the data: following PS matching the matched nature of the data must be accounted for in the analysis; following IPTW robust variance estimation is often used to account for the weighted nature of the data (Austin, 2011a; Li, 2013).

The estimate of the variance of the treatment effect when using standard analyses following the use of PS methods does not take into account that the PS is itself estimated from the data (rather than being the true PS). This approach gives estimates of the variance which produce CIs of the treatment effect that are too wide (Williamson, Morley, Lucas & Carpenter, 2012b). Formulae are available which estimate the analogous marginal variance, the variance accounting for the uncertainty in estimating the PS, only for PS stratification (Williamson et al., 2012b) and IPTW (Lunceford & Davidian, 2004). When using PS stratification, commonly used variance estimation produced CIs which are too wide when the PS model includes variables which predict the

outcome but only weakly predict the treatment allocation, in comparison to using the analogous marginal variance method. For IPTW, commonly used variance estimation produces CIs which are too wide in all circumstances, when compared to using the analogous marginal variance method (Williamson et al., 2012b).

2.3.7 Step 6 - Sensitivity tests

Li (2013) and Caliendo and Kopeinig (2008) suggested sensitivity tests at this stage. They are from the fields of economics and management respectively, and sensitivity testing has not been seen following the use of PS methods to correct for treatment allocation bias in applied medical research guidance papers. The reasons for this are unclear as sensitivity analysis is widely used in other in areas of applied medical research, such as the analysis of RCT data.

Li (2013) suggests that sensitivity analysis on the treatment (or causal) effect estimate should be carried out to assess if there are any unmeasured confounders. Preferably this would be by comparison with an unbiased estimate, but usually this is not available. Alternate comparisons can be made by using an alternative control group, changing the specification of the PS equation, or measuring the effect of unobserved variables using instrumental variables method or Rosenbaum Bounds. (The Rosenbaum Bounds approach assesses the size of the impact of an unmeasured confounding variable needed to bias causal effects from a matching analysis (DiPrete & Gangl, 2004)).

Caliendo and Kopeinig (2008) also suggest investigation of the effect of unmeasured confounders and give methods for the assessment of how strong the effect of unobserved confounders must be to alter the treatment effect. Additionally, Caliendo and Kopeinig (2008) suggest sensitivity analysis by including any cases dropped to ensure common support.

2.4 Measurement error methods

2.4.1 Overview

Measurement error, the difference between the observed value of a variable and its true value, can be modelled. Some common examples are: the classical measurement error model, where the observed value equals the true value plus random noise; the linear measurement error model, where the observed value equals the true value plus random noise and systematic error; the Berkson model, where the true value equals the observed value plus error which is independent of the observed value (Keogh et al., 2020). Measurement error can be 'differential' and 'non-differential' (Carroll et al., 2006, p. 36). If the measurement error is differential, the outcome affects the measurement error. Generally, measurement error is non-differential. The

study data are assumed to have non-differential covariate measurement error; in this context that means that measurement error is the same across the treatment groups. The data were collected in the same way for both treatment groups, prior to treatment allocation or when the primary outcome occurred (Section 1.5).

If the measurement error model is not known, ancillary data can be used to fit the model. The data for calibration or validation can be internal, using a subset of the main data, or external, using an independent dataset or study. For each of these, the data can be validation data, where the true value is observed directly, replication data where repeated measures of the same observed variable are taken, or Instrumental data where an additional variable is measured. Carroll et al. (2006, p. 33) recommend using internal data, as direct examination of the data structure is possible and it gives greater precision of estimation and inference. External data can be used but assumptions are always being made when transporting models. Replicate data can be used if the replicate mean is thought to be better than the single observation. An instrumental variable, when used externally, should only be used if it is unbiased for the true value.

In Chapter 1 the effect of measurement error in PS analysis was introduced. The literature reported that measurement error was previously ignored, but now there is more application of methods to correct for measurement error. The author of this thesis believes that the extent of measurement error on a variable may not be obvious on first inspection, unlike that of missing data. This means that any effect caused by measurement error may go unnoticed and the treatment effect estimate may be biased.

For a continuous variable, $\boldsymbol{W} = \boldsymbol{X} + e$ where $\boldsymbol{W}$ are the observed covariate(s), $\boldsymbol{X}$ are the true covariates and $\boldsymbol{e}$ is the measurement error, with an assumed distribution. For a binary variable, $\boldsymbol{W}$ and $\boldsymbol{X}$ will take the value 0 or 1, so the misclassification model can be expressed as misclassification probabilities $pr(\boldsymbol{W} = 1|\boldsymbol{X} = 0)$ and $pr(\boldsymbol{W} = 0|\boldsymbol{X} = 1)$. For both continuous and categorical variables, if only $\boldsymbol{W}$ (the mis-measured $\boldsymbol{X}$) is known then the strongly ignorable treatment assignment assumption required for PS conditioning is not met and using $\boldsymbol{W}$ in place of $\boldsymbol{X}$ fails to control for all confounders (Rudolph & Stuart, 2018). If a covariate in the PS model is measured with error, the PS does not remove the systematic differences between the treatment groups and hence a biased treatment effect will be estimated (Hong, Rudolph & Stuart, 2017).

2.4.2 Generic methods to address measurement error

Measurement error adjustment methods can be categorised as: generic methods, where the measurement error adjustment is applied to the data then the standard PS analysis is carried out; and specific to PS analysis, where the measurement error adjustment is combined with the PS analysis.

A brief summary of generic methods which may be suitable for this study are given. Multiple Over-imputation (MO) (Blackwell, Honaker & King, 2017), regards measurement error as partially missing information, and then completely missing values as an extreme form of measurement error. So, this approach corrects for both measurement error and missing data. It imputes the missing or mismeasured values and then 'over-imputes' them, so in other words overwrites them. Simulation Extrapolation (SIMEX) (Cook & Stefanski, 1994) adds additional measurement error to a single mismeasured covariate and extrapolates it back to a situation with no measurement error using simulations. Minimal Assumption Bounds (Black, Berger & Scott, 2000) specifies a range of parameter values which meet certain set assumptions on the error model. This method only gives bounds; details within the bounds are not available, which may not be enough detail for this study. If calibration data are available, Regression Calibration (Carroll & Stefanski, 1990) could be used, which replaces the mismeasured variable with an estimate of the underlying unobserved variable and then performs the analysis on these calibrated data.

2.4.3 PS-specific methods to address measurement error

2.4.3.1 Comparison of PS methods when covariate measurement error exists

The effect of covariate measurement error when using PS conditioning was demonstrated by Conover et al. (2021), De Gil et al. (2015) and Hong, Aaby, Siddique and Stuart (2019). De Gil et al. (2015) ran simulations varying several parameters including covariate measurement error and compared different PS methods. Their findings included covariate measurement error affected bias, Type I error control and Confidence Interval (CI) convergence. It did not affect common support, balance, Root Mean Squared Error (RMSE) and CI width. De Gil et al. (2015) also varied both covariate measurement error (covariate reliability) and effect size (strength of relationship between covariates and treatment assignment). Both of these, as individual parameters and their interaction, were among parameters which affected CI coverage.

Conover et al. (2021) used plasmode simulations (Vaughan et al., 2009) to assess the impact of a misclassified binary covariate in the PS model and compared the performance of IPTW, IPTW following trimming, and 1:1 PS matching. The misclassification was introduced in several

scenarios and they focused on misclassification which was differential by outcome status as this generated higher bias. They found that when the variable with misclassification was a strong indicator for treatment the bias increased with increasing exposure prevalence, but when the variable with misclassification was a strong contraindicator for treatment the bias increased with the inverse of exposure prevalence. The direction of the bias depended on whether the variable with misclassification was an indicator or contraindicator of treatment allocation. Scenarios with only false positive misclassifications produced higher bias than scenarios with only false negative misclassifications. This is how the measurement error was implemented in the current study. In Conover et al. (2021) generally in the scenarios they covered, 1:1 PS matching had lower bias and higher precision than IPTW using the untrimmed dataset (this was seen for a strong contra indication of a rare exposure and a strong indication of a common exposure), but following trimming generally IPTW performed better than 1:1 PS matching . They report that 'modest' amounts of measurement error, in around ≤5% of observations, can introduce bias.

Hong et al. (2019) ran simulations which showed: that the bias and MSE reduced as the reliability of mismeasured confounders approached 1; the bias and MSE is lower when the true covariates are correlated even when the mismeasured variables are used (observed); correlation the measurement error (of the observed, mismeasured variables) increased bias and MSE.

2.4.3.2 Study dataset requirements

The measurement error of interest in this study's data was covariate misclassification. There was no reason to suggest that the measurement error was differential across treatment groups (Section 1.5). There may be outcome measurement error in the dataset but this was not the current focus. Several PS conditioning methods were used in this study and no calibration dataset was available. These requirements guide which of the methods are applicable to this study.

2.4.3.3 Methods to address covariate measurement error

If a covariate in the PS model is measured with error, the treatment groups are not balanced because the PS is based on the observed not true covariate values and hence a biased treatment effect will be estimated (Hong et al., 2017). There are methods which perform measurement adjustment specifically in PS analysis (Braun et al., 2017; Dong & Millimet, 2020; Hong et al., 2017; Raykov, 2012; Rudolph & Stuart, 2018; Webb-Vargas, Rudolph, Lenis, Murakami & Stuart, 2017). These types of method were more suited to the study data and are evaluated and discussed in Section 2.4.3.

Braun et al. (2017) presents a method to adjust for treatment allocation measurement error which is unlikely in the study dataset. Shu and Yi (2019a) present methods to address outcome measurement error using validation data, replicant data or a doubly-robust method. Shu and Yi (2019c) developed an Augmented SIMEX (ASIMEX) method to account for both covariate measurement error and misclassified outcomes. Gravel and Platt (2018) used modified Inverse Probability weighting and MLEs of misclassified parameters derived from internal validation when there was outcome misclassification. These methods relate either to treatment measurement error, which is unlikely in the study data, or outcome measurement error, which was not the focus of this study. These methods were not considered further.

Hong et al. (2017) present two Bayesian models for use with PS methods, 1) a Joint approach where the PS, measurement error and outcome are modelled by using all the information jointly and 2) a Two-Step approach where the PS and the measurement error are modelled in a Bayesian framework then the outcome is modelled separately. They used a 'subjective' Bayesian approach where the prior reflects expert knowledge before the data are collected, although other approaches could be used. They worked with the assumptions that there are no validation data, neither internal nor external and that there is some prior knowledge about the extent of the differential measurement error, which gives the 'prior' distribution. Their simulations show that when a covariate which is a strong predictor of outcome is mismeasured, the joint model performs best. For measurement error in weaker predictors neither model is better than the naïve model (but they do give better coverage). This was reflected in their case study which showed little advantage of their model(s) over the naïve model; this could be because the covariate with measurement error was only weakly related to the outcome.

The use of Bayesian models with PS methods is relevant for psychology or education research where individuals often 'self-select' their treatment (Hong et al., 2017). The current study uses medical data where the treatment allocation is decided by a clinician, although the covariates which inform this decision may be misrecorded. Methods presented in the next paragraph can be used for both differential and non-differential measurement error. Methods which are specifically for differential covariate measurement error, such as (Hong et al., 2017), can be discarded as they are not relevant to this study.

The methods presented which address covariate measurement error, either for single or multiple covariates, are all relevant to non-differential measurement error (Dong & Millimet, 2020; Raykov, 2012; Rudolph & Stuart, 2018; Webb-Vargas et al., 2017), but some may also suit differential measurement error. Some of the methods require a calibration/validation dataset.

Using latent variables in the PS model may off-set covariate measurement error (Whittaker, 2020). However, Sengewald, Steiner and Pohl (2019) report that balance checking for latent covariates is not possible as latent covariates are unobserved, so it is "not possible to evaluate a PS model with latent covariates". Sengewald et al. (2019) signpost to other methods including Raykov (2012)'s Modified PS using factor score estimates of latent covariates. An alternative is the inclusive factor score (iFS), (Nguyen & Stuart, 2020). Traditionally a latent variable is generated from several error-prone covariates. There are different proxies for the latent variable which can be used in PS analysis but these often do not balance the latent variable, and so lead to a biased estimate of the treatment effect. The iFS uses a structural equation model to predict the latent variable, using a joint distribution of the latent variable, the error-prone covariates and the exposure, given the observed covariates. Their simulations showed that iFS reduces the bias of the treatment effect estimate.

Raykov (2012) uses the term Conventional PS (CPS) for a PS which is modelled using at least one fallible variable, a variable measured with error. If balancing is performed using this CPS a biased estimate of the treatment effect will be obtained. They then present the Modified Propensity Score (MPS) using covariates with measurement error. The MPS uses two indicator variables for each fallible covariate, which inform the latent dimensions, that are the true values. The MPS can be applied to multiple fallible covariates providing each has two indicator variables to inform them. The true covariate values were generated using latent variable software Mplus then the MPS modelled using logistic regression based on the 'true' values of the covariates. They ran both the CPS and the MPS methods on simulated data with fallible covariates and a significant treatment difference. The CPS failed to identify the significant treatment difference whereas the MPS did.

The MPS (Raykov, 2012) remains in the spirit of PS as it does not access the outcome to build the MPS. It needs a large sample size as it uses logistic regression to build the model and sufficient indicator variables to inform the fallible variable(s). MPS has not yet been used in applied medical research, only in social sciences. Raykov (2012) uses covariate adjustment, but the MPS can be used with stratification (Kaplan, 1999) and matching (Peikes, Moreno & Orzol, 2008).

Webb-Vargas et al. (2017) present Multiple Imputation for External Calibration (MI-EC) for use with measurement error in a single covariate when the measurement error is non-differential. It uses a calibration data set, which is regarded as a gold standard. The main dataset contains Y, the outcome, T, treatment, Z, covariates without measurement error, and W, the observed value

of the true confounder X. The calibration dataset should contain only X and W. They made various assumptions: the $joint\ conditional\ distribution\ (X, Z, T, Y|W)$ is multivariate normal (MVN); the distribution is the same for the main and calibration datasets; the mean of the joint conditional distribution is linear in W and the covariance matrix is constant; and the measurement error is non-differential. If these are met, the posterior distribution of $f(X|Z, T, Y, W)$ can be generated.

They ran simulations based on the work by Y. Guo, Little and McConnell (2012) using the Naïve method (unadjusted), True method (using X), Uncongenial MI-EC (not using variables later used in the outcome model) and Congenial MI-EC (includes outcome, treatment and all confounders). Their findings were that the Congenial MI-EC was shown to be the best performing model and can be used to correct for measurement error. As some of the covariates which predict Y, the outcome, are only in the main dataset and not the calibration one, so MI-EC has the advantage over Regression Calibration which needs access to all the confounders in the calibration dataset as well.

This method can be used for a single mismeasured covariate when the measurement error is non-differential. A variable used in the analysis must be used in the imputation, which is standard practice, but removes the advantage of PS modelling being independent from the outcome which is a limitation of MI-EC. MI-EC worked well with any model linear in X and even when the joint MVN assumption was violated because the treatment allocation was binary. A small calibration dataset limits the bias correction for the Congenial MI-EC model, but relative size of the calibration to main ratio is not important. Following MI-EC, they used PS IPTW, but reported that PS matching or PS stratification could have been used. The current study did not have access to a calibration dataset, so MI-EC was not applicable.

Rudolph and Stuart (2018) show the equivalence of covariate measurement error and unobserved confounding and hence apply three methods previously used for unmeasured confounding for use in measurement error correction. These are PS Calibration, Vanderweele and Arah's bias formulas, and Rosenbaum's sensitivity analysis. They use simulations to show how these methods correct for the following, 1) classical measurement error, which is non-differential and homoscedastic, 2) systematic differential where the measurement error is different in the treatment groups, 3) heteroscedastic measurement error.

PS Calibration (PSC) (Sturmer, Schneeweiss, Avorn & Glynn, 2005) uses a calibration dataset. The naïve PS, using mismeasured covariates, and the true PS, using correctly measured covariates, are generated in the calibration dataset then extrapolated to the main dataset to calibrate the

naïve PS stored there. There are a number of assumptions associated with PSC, and PSC is violated if the measurement error is differential. PSC accounts for measurement error in multiple covariates and PSC can be used with all PS methods. It works well for PS matching, stratification and covariate adjustment, but not IPTW. PSC does not work well when the measurement error is large. It reduces bias and works best when the ATE is close to 0.

Vanderweele and Arah's bias formulas (VanderWeele & Arah, 2011) are formulae for calculating bias caused by unobserved confounders. An unobserved confounder can be related to the treatment allocation and/or to the outcome. Reasonable combinations of the values of the coefficients of these covariates are tested to find the ones which lead to a different result. They can be used for classical measurement error and differential by treatment group. They correct fully for measurement error bias in all three scenarios if the correct sensitivity parameters are used and the assumptions met. They can also be applied for any PS estimation method and can be used for classical measurement error and differential by treatment group.

Rosenbaum's sensitivity analysis (Gastwirth, Krieger & Rosenbaum, 1998) assumes that the data are PS matched pairs and the treatment groups are balanced on the observed confounders. There are different versions of this method and here the 'simultaneous' sensitivity analysis version is used, where two parameters are varied simultaneously. When the outcome, Y, is binary the analysis is used to set the upper and lower bound for McNemar's test, and when Y is continuous the parameters are varied to set the upper and lower bounds for the normalised Wilcoxon Signed Rank test statistic. It may be seen as a limitation that this method can only be used with PS matching. This method corrects for non-differential (or classical) and differential measurement error. It is difficult to interpret for continuous outcomes and it reduces bias, but not by a large amount.

Rudolph and Stuart (2018) recommended Vanderweele and Arah's bias formulas, which reduced bias by 100% if the correct sensitivity parameters were used in all three simulation scenarios, and PS Calibration for use as they worked well for a variety of PS methods and measurement error scenarios. However, Rosenbaum's sensitivity analysis, which performed least well, may still be applicable in this study if the focus were on PS matching.

Dong and Millimet (2020) presented a semi-parametric estimator to address measurement error in more than one covariate. They focus on using IPTW when the PS was of an unknown functional form. First, they estimate the functional form of the PS and second, they estimate the moment of the known form of mismeasured covariates. Their work was applied to assess the performance of financial literacy programmes for micro-entrepreneurs.

A discussion of the suitability of these methods for use in this study is given in Section 2.6.1.

2.5 Sparse data methods

2.5.1 Introduction

This section of the literature review focused on sparse data in observational studies using Electronic Healthcare Record (EHR) data and the most appropriate methods used to analyse them. Sparseness in data can be caused by any of the following: small sample size (Siino et al., 2018); rare exposure (treatment) (Hajage et al., 2016); rare outcome events (Siino et al., 2018) which lead to a low number of events per variable (EPV) (Greenland et al., 2016); unbalanced or highly predictive risk factor variables (Siino et al., 2018) with narrow distributions or categories which are uncommon (Greenland et al., 2016); variables which almost perfectly predict the outcome (Greenland et al., 2016); variables that together almost perfectly predict the exposure (treatment) (Greenland et al., 2016).

Sparse data bias produces treatment effect estimates away from the null, so inflated treatment effect estimates are produced. Estimates should be compared with existing knowledge and any previous studies, and if they strongly differ this could be an indication of bias (Greenland, Schwartzbaum & Finkle, 2000). The effects of treatment on outcome measured by risk, rate and odds and adjusted versions of these such as logistic, Poisson or Cox modelling can all be subject to bias if there are small numbers in any of the treatment/outcome combinations (Greenland et al., 2016). These estimation methods assume sufficient events at all treatment levels or categories, but when this is not met the estimate for the regression coefficients is away from the null. Greenland et al. (2016) refers to this as ‘sparse data bias’, as it can occur in quite large datasets not just due to small sample size.

2.5.2 Sparse data methods

Methods for working with sparse data include Propensity Score methods (Section 2.5.3), Penalised Likelihood Estimation (PLE), Data Augmentation and Bayesian methods. Some of the methods are closely related, for example Data Augmentation for sparse data can be used as a form of Bayesian analysis, also Data Augmentation is a form of PLE where the prior data forces the program to generate a penalty function, which imposes the prior constraints (Sullivan & Greenland, 2013).

When outcomes are binary, logistic regression is a common form of analysis. However when the data are sparse the asymptotic properties that maximum likelihood estimates (MLE) are based on no longer hold and the treatment effect estimate may therefore be infinite or heavily biased.

Firth's penalized maximum likelihood (PLE) (Firth, 1995) provides a finite treatment effect estimate even in such sparse data settings (Siino et al., 2018).

In studies with binary outcomes and two treatments (exposed or not exposed to the active treatment) the results can be reported in a 2x2 table. In studies where the data are stratified, such as by centre, the data for each stratum are represented in a separate 2x2 table. There are different ways of dealing with sparse data in this setting by combining categories, deleting cells or tables containing zero values, or Data Augmentation where a constant is added to each cell (Subbiah & Srinivasan, 2008). Data Augmentation is the most popular method in the literature.

Bayesian methods are well suited to observational studies and they perform well compared to frequentist methods particularly when the data are sparse. This is particularly the case when the number of covariates approach the number of outcomes (Sullivan & Greenland, 2013). Data Augmentation may be regarded as a semi-Bayes or partial-Bayes analysis as it does not require priors on all coefficients. The literature reports that a weakly defined prior gives better performance than frequentist methods, particularly when the data are sparse. Greenland et al. (2000) commented that like all Bayesian methods, Data Augmentation needed suitable background information.

PS methods to address treatment allocation modelling improve the handling of sparse data (Greenland et al., 2016). PS methods combine the information from several variables into one, making the EPV lower in the outcome model. The current study is limited to the use of PS methods in the presence of sparse outcome data.

2.5.3 Studies investigating Propensity Score methods and sparse data

A number of papers conducted studies using PS methods on data with some of the characteristics of the study data: observational data using EHR; large dataset with few outcomes (sparse data); outcome analysis using time-to-event data. Hajage et al. (2016), L. Choi et al. (2018), Patorno, Glynn, Hernandez-Diaz, Liu and Schneeweiss (2014) and Franklin et al. (2017) compare PS methods in their studies.

The analysis on the current study's data is a 2-step approach – firstly to adjust for any treatment allocation bias by using PS methods, secondly to perform the outcome analysis suitable for time-to-event format data. The sparseness in the data takes the form of rare outcomes, so will affect the second step. An assessment of different PS methods will be useful to determine which best prepares the data for the outcome analysis with sparse data.

Several papers used PS methods with sparse data but were not directly applicable to this study. Kuss (2002) used logistic regression to build the PS model for rare exposure. The sparseness of the data in L. Choi et al. (2018) related to low exposure. Patorno et al. (2014) worked with data with frequent exposure, many potential confounders and few outcomes (particularly in the exposed group), which is similar to the current study's data, but used High Dimension Propensity Scores (hdPS). These are all different scenarios to the current study. Ali et al. (2014) gave recommendations for PS balance and PS model variable selection with rare outcomes**.** Lee (2010) reported that when the data are sparse, doubly-robust adjustment (a PS method with regression adjustment) produced more biased treatment effect estimates with higher SE than when just using PS weighting. Yoshida et al. (2017) showed that matching weights performed better than 1:1:1 matching and IPTW, particularly with rare outcomes or uneven exposure distributions, but with 3 treatment groups. Chang, Perng and Shiau (2000) highlighted the problems of using Cox PH with sparse data. If there were no events in a stratum it was non-informative. In cases with many such strata, the Cox PH modelling could become unstable. They report that this estimate was unbiased and that precision can be increased by ignoring the heterogeneity among strata, but this estimate was then biased. This is an example of the trade-off between bias and precision. They propose a compromise of these two.

Hajage et al. (2016), Fabiani et al. (2015) and Franklin et al. (2017) had stronger similarities to the current study. Hajage et al. (2016) ran simulations and applied their analysis to a real-world dataset. Their method had similarities to the current study, using EHR, comparing PS methods, the outcome data were in the form of time-to-event and they estimated the marginal HR. Their data had sparse exposure (treatment) whereas the current study has sparse outcome data. The parameters they varied included: prevalence of exposure; correlation between variables in the PS model; strength of association between covariates in the PS model and exposure (the effect size); association between exposure and outcome and censoring rate. They compared the performance of PS weighting (IPTW for ATE & IPTW for ATT) and PS matching. These methods had been recommended by the literature as the best performing PS methods (Section 2.3.4). The results from Hajage et al. (2016) of most relevance to the current study were when the exposure rates were varied: rare exposure gave a biased estimate to the marginal HR; the estimate for ATE (using IPTW) had a particularly high bias; estimates for ATT were less biased; and IPTW for ATT was recommended over PS matching for estimates of the ATT.

Fabiani et al. (2015) provided an example of using a PS method (PS stratification) when the data are sparse (both exposure and outcomes). They explored if there was a link between the influenza vaccine in pregnant women and several maternal and neonatal outcomes. The study

setting was similar to the current study: EHR were used; the outcome data were in the form of time-to-event; and Cox PH regression was used. There was low exposure prevalence (2%) and many of the outcomes had a prevalence of <1%. This is an example of using PS methods in a sparse data setting, but no comparison with other PS methods was made.

Franklin et al. (2017) ran extensive simulations, based on two datasets, in scenarios where the outcome data were sparse. There were similarities with the current study, although their outcome of interest was binary, the log RR was displayed in their plots. They used a plasmode simulations method, compared the performance of several PS methods (some were different to the current study), combined them with different outcome modelling options and applied them to several scenarios which all had low outcome prevalence, of 1% to 5%. Some of the differences to the current study were that they used four methods for generating the PS, ran each analysis on trimmed and untrimmed datasets and where feasible, the PS methods were used to estimate the ATE and the ATT. Franklin et al. (2017) used seven scenarios to present a variety of parameters. In summary, their recommendations for the best PS methods to use were guided by calculating the treatment effect within quintiles of the PS. If there was little heterogeneity, they recommended regression on the PS using a nonlinear generalised additive model fit. If there was heterogeneity, they recommended matching weights. They did not recommend 1:1 matching, IPTW and stratification when the data had few outcomes and poor common support. They also averaged the results over all their simulation scenarios to recommend PS methods to estimate the ATE and ATT. Franklin et al. (2017) confirmed that IPTW could be unstable when the common support was not good, which agreed with findings from the literature. They found that extreme weights also cause problems for stratification and full matching. In full matching in areas of poor overlap a single treated case can be matched to many (500) controls, they suggest limiting this number, but could give higher variance if the number of outcome events reduces. Their findings supported the literature that trimming reduced the bias but may increase the SE of the estimate. They reported that datasets with better common support would improve the performance of not only IPTW, but also other PS methods.

2.6 Identification of the gap in the literature

2.6.1 Measurement error

This section examines the applicability of the existing methods for correcting for measurement error when using PS methods, if they were to be applied to this study's data. The measurement error in the study data to be investigated was non-differential covariate measurement error (Section 2.4.3.1). The methods which relate to treatment allocation measurement error (Braun

et al., 2017) or outcome measurement error (Gravel & Platt, 2018; Shu & Yi, 2019a, 2019c) or only non-differential covariate measurement error (Hong et al., 2017) were discarded.

If correction for non-differential, covariate measurement error were to be applied to the study's data, the selection of methods would be as follows. MI-EC and PS Calibration require an external calibration dataset which is not available for the study dataset. Rosenbaum's sensitivity analysis has only been applied to PS matching so could not be used with all the PS methods compared in this study. The Modified PS uses latent variables each of which are generated from two indicator variables. Latent variable methods were not considered for the current study. Vanderweele and Arah's bias formulas or Dong and Millimet (2020)'s semi-parametric estimator would be the methods most likely for consideration. Additional criteria which may guide the method selection will include, the ability to correct for bias, the ease of application, such as the availability of software or code and the novel application of the method.

Methods to correct for measurement error when using PS conditioning were originally more widely applied to social science data (Dong & Millimet, 2020; Hong et al., 2019; Hong et al., 2017; Nguyen & Stuart, 2020; Raykov, 2012; Sengewald et al., 2019; Webb-Vargas et al., 2017). More recently there have been applications to health survey data or RCT data, (Braun et al., 2017; Conover et al., 2021; Shu & Yi, 2019a, 2019b, 2019c). Only one study, (Gravel & Platt, 2018) had applied their methods to routinely collected EHR, CPRD data linked to HES & MINAP, which is similar to the current study's dataset. The measurement error seen in survey data (both for social science and medical research) which rely on participant recall may be different to the measurement error in EHR which are routinely collected data. A novel aspect of this study would have been the application of one of these methods to EHR data. However, the study data does not contain validation data, such as external data from HES or repeated measures, which excludes some of the methods described.

2.6.2 Sparse data methods

This study was limited to using PS methods to minimise sparse data bias. Although PS methods are a recognised method when analysing sparse data, the different types of PS methods performed differently (L. Choi et al., 2018; Franklin et al., 2017; Hajage et al., 2016; Patorno et al., 2014). Greenland et al. (2000) summarised that other types of bias, such as misclassification and selection bias, can also add to the problem of sparse data bias. In this study additional bias could be introduced by measurement error, so the focus was to compare the performance of different PS methods in the presence of covariate measurement error and sparseness of outcome data.

2.7 Summary

Comparisons of the performance of different PS methods to correct for treatment allocation bias when the outcome data is binary is made by Austin (2009b), Austin et al. (2007), Austin and Mamdani (2006) and Busso et al. (2014). IPTW and PS matching are the PS methods which provide the best balance between the treatment groups. The majority of the literature uses data with binary outcomes however Austin (2013), Austin (2014b) and Gayat, Resche-Rigon, Mary and Porcher (2012) use time-to-event data. Austin (2013) and Austin (2014b) show IPTW and PS matching to be the best performing methods with time-to-event outcomes.

De Gil et al. (2015), Conover et al. (2021) and Hong et al. (2019) demonstrated the effect of covariate measurement error on treatment effect estimates and compare the performance of some PS conditioning methods. No method to correct for covariate measurement error when using PS analysis was found to be suitable for this study's data. So, this study focuses on the comparison of the performance of PS methods in the presence of covariate measurement error.

PS analysis is recognised as a method for working with sparse data (Greenland et al., 2016). Franklin et al. (2017) and Hajage et al. (2016) compared PS methods in simulations using sparse data. Franklin et al. (2017) investigated many scenarios including sparse outcome data and rare exposure, although the outcome data were binary. Franklin et al. (2017) compared the performance of different PS methods, and their averaged results showed for the ATE PS stratification performed slightly better than IPTW for ATE, and for the ATT 1:1 PS matching performed slightly better than IPTW for ATT. Hajage et al. (2016) used data in time-to-event format from EHR, but with sparse exposure and recommended IPTW for ATT over PS matching for estimates of the ATT.

There is less in the literature where PS methods are compared when the outcome data are in the form of time-to-event, as opposed to comparisons of PS methods using binary data. No reference to work was found which makes a comparison between PS methods with varied added covariate measurement error and varied sparseness of the outcome data. This study developed a simulations method to analyse the study dataset, which has time-to-event outcomes. Simulations were run to compare the performance of the four selected PS methods when the amount of covariate measurement error and the sparseness of outcome data were varied. Chapter 3 develops the analysis method for each PS method with no introduced measurement error and no change to the sparseness of the outcome data. Chapter 4 develops the simulations method.

Chapter 3 METHODS

3.1 Introduction

The real-world treatment effect of a novel treatment can be estimated from observational (non-randomised) data. However, as the treatment allocation is not randomised, there may be systematic differences between the treatment groups, and if this is not accounted for the treatment effect estimate will be biased. Propensity Score (PS) methods are commonly used to adjust for differences between treatment groups and are used in this study. The PS methods IPTW for ATE, IPTW for ATT, PS stratification and PS matching were chosen for use in this study as they are widely used. These PS methods are all applied to the data before the outcome analysis is conducted. In later chapters (4 and 5) the performance of the different PS methods is compared in the presence of other real-world problems: measurement error and sparse data.

Before the simulations could be run to compare the performance of the PS methods in the presence of measurement error and sparse outcome data, the analysis method was developed and run on the study dataset, keeping its original characteristics. This chapter describes how the study dataset was established from an extract of primary care data and why Rivaroxaban was chosen as the NOAC treatment to compare with Warfarin. It then explains how the PS model, the treatment allocation model, was fitted to the data and the PS conditioning methods: PS matching; IPTW for ATE; IPTW for ATT; and PS stratification were applied. As the data were in time-to-event format, a Cox proportional hazards model was fitted for use in the outcome analysis. The baseline hazard was identified for use in the simulations work, (applied in subsequent chapters). Further details of the work presented in this chapter are given in Appendix B.

3.2 Establishing the study dataset

This study used data supplied to the Performance-Based Innovation Rewards study (REWARD) (Appendix B-1.2). The data extract contained variables relating to patient demographics, factors thought to affect the prescribing of NOAC/OAC medication, factors thought to affect the main study outcomes of stroke and major bleed event, details of subsequent outcomes, full details of anti-coagulant prescriptions and summary details of other medication thought to affect NOAC/OAC prescribing. The NOACs which were available to be prescribed during the REWARD study period, (January 2011 to May 2015) were Apixaban, Dabigatran and Rivaroxaban. The traditional OAC used as the control treatment was Warfarin.

The REWARD dataset contained routinely collected primary care data, so there was no randomisation to the treatment. To correct for the possible systematic differences between treatment groups this study uses Propensity Score (PS) methods, which have been developed for use with two treatments. There may be different factors which affect the prescribing of each NOAC so one of the NOACs had to be chosen to compare with Warfarin. Rivaroxaban was chosen because its use was increasing and had the highest number of cases to whom it was prescribed (Table 2). Although this meant there were likely to be more outcome events in the dataset which would make the outcome modelling more stable, the outcome data (the number of future strokes) was still sparse. The Rivaroxaban-Warfarin dataset was used as the study dataset. Early analysis using the Apixaban-Warfarin dataset is given in Appendix B-4 for comparison.

Table 2: Number of NOAC-naive patients by year of first NOAC/OAC prescription.

Year	Apixaban	Dabigatran	Rivaroxaban	Warfarin
2011 and earlier	0	26	4	46,246
2012	0	280	99	7,409
2013	114	459	735	7,117
2014	769	389	1,395	5,333
2015 (Jan to May)	461	119	699	1,334
Total	1,344	1,273	2,932	67,439

Selecting the first NOAC/OAC prescription date after the NOAC NICE approval date (May 2012) ensured that the treatment groups were compared during the same time period and that the same healthcare policies applied. Selecting only NOAC/OAC naïve patients ensured that the effect of only this treatment is being studied. This did mean that patients who had 'crossed over' from Warfarin to a NOAC were excluded from the study dataset. The original requirement was 'first NOAC/OAC prescription after first date of AF diagnosis', but the recording of the first date of AF diagnosis appeared to be inaccurate, with many patients apparently prescribed a NOAC/OAC before a recorded diagnosis of AF. So the start date of the first NOAC/OAC prescription was used in lieu of the date of first AF diagnosis.

The Rivaroxaban vs Warfarin dataset (RI-WA) was built using the selection criteria:

- The first NOAC/OAC prescription date was after the National Institute for Health and Care Excellence (NICE) approval date for the NOAC (May 2012 for Rivaroxaban) used in the dataset.
- The patients were NOAC/OAC-naïve, that means this was the first NOAC/OAC prescription this patient was recorded as being prescribed.

3.3 Modelling the Propensity Score

The PS model is the treatment allocation model and gives the probability that the patient would be allocated the novel treatment, the NOAC. So for patients who received the NOAC, it would be expected that the PS would be closer to 1 than those who received the control, Warfarin. The PS is a probability distribution for which a truncated or an S-shaped curve suit these outcomes taking values in the range 0 to 1. Methods used to model the PS include logistic regression, general location modelling, classification trees, random forest, generalised boosted model (Cham & West, 2016). Logistic Regression is the most widely used method when the data is complete (Cham & West, 2016; Luellen, Shadish & Clark, 2005; Thoemmes & Kim, 2011) and was selected for modelling the PS model in this study (Appendix B-3.1).

The literature presents different options for the selection of the covariates to include in the PS model: variables which influence prescribing and outcome; variables which influence only prescribing; variables which influence only outcome (Section 2.3.2). In this study, variables which influenced the prescribing, although they may also have influenced the outcome, were included in the PS model. This allowed the PS to be modelled without knowledge of the outcome. The variables were selected either by expert knowledge or by identifying non-clinical variables from the data.

All clinically relevant variables (from advice by clinicians) were kept in the PS model, regardless of their statistical significance during the model selection process. Other non-clinically relevant variables, seen to affect prescribing, were kept in the model if their p-value ≤ 0.05, showing them to be statistically significant (Appendix B-3.2). The model selection was made using the Bayesian Information Criteria (BIC) (Posada & Buckley, 2004) (Appendix B-2). Several functional forms for the continuous variables, age and date of first prescription, were selected and combinations of these added to the model containing the clinically relevant variables and the significant non-clinical variables. The 'best' model was selected, using the BIC, and used as an initial PS model (Appendix B-3.3). It was then assessed to determine if any of the variables could be dropped to simplify its use in the simulations phase (Appendix B-3.4). The refined PS model used in this study is given in Table 3.

Table 3: The refined treatment allocation model for the RI-WA dataset.

Covariate	Coefficient	SE of coeffi-cient	Z	P>|z|	[95% CI]
Previous stroke	0.123	0.061	2.03	0.042	(0.004, 0.242)
Alcohol misuse	0.098	0.128	0.76	0.446	(-0.153, 0.348)
Chronic kidney disease	0.008	0.051	0.16	0.871	(-0.093, 0.109)
Liver disease	0.033	0.437	0.07	0.941	(-0.825, 0.890)
Ischemic heart disease	-0.082	0.051	-1.61	0.108	(-0.181, 0.018)
First NOAC/OAC prescription was ≤ 28 days of first AF diagnosis?	-0.192	0.042	-4.56	<0.001	(-0.275, -0.110)
=86 if age≤86, else =age	0.077	0.013	6.13	<0.001	(0.053, 0.102)
licence_to_noac30 *	0.153	0.011	13.56	<0.001	(0.131, 0.175)
$(licence\_to\_noac30)^2$	-0.001	<0.001	-5.54	<0.001	(-0.002, -0.001)
Constant term	-10.830	1.096	-9.88	<0.001	(-12.979, -8.682)

*licence_to_noac30 is the Rivaroxaban licence date to date of first NOAC/OAC prescription, in months

Following the advice of the literature a check for common support, or overlap, was carried out once the PS model had been defined and hence the PS value calculated (Section 2.3.3). This ensured the two treatment groups had sufficient participants with similar PS values to make the PS conditioning meaningful. Figure 1 shows that there was good common support in this dataset, that is sufficient overlap of the PS distributions of the two treatment groups. A simple match on the PS showed that all NOAC cases were matched to a Warfarin case, so the PS model was sufficiently well defined to continue the analysis.

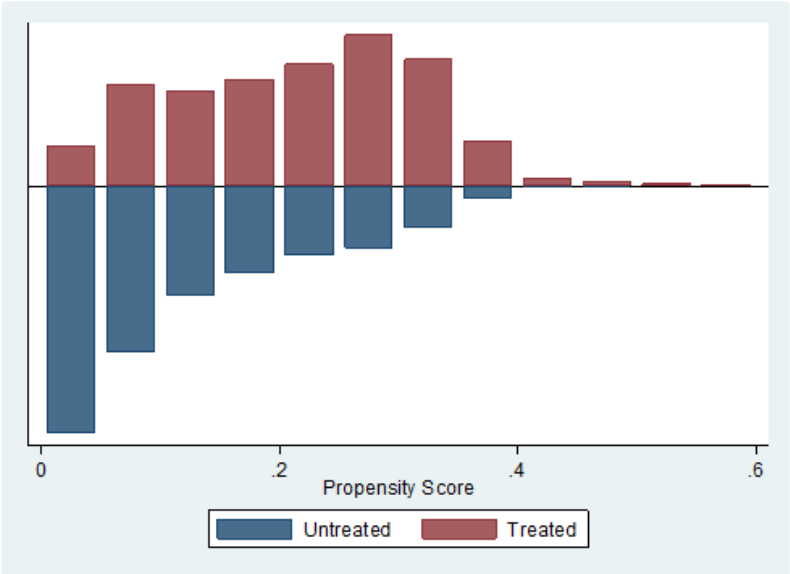


Figure 1: Histogram of Propensity Score, using Stata's -psgraph-, for Rivaroxaban (Treated) and Warfarin (Untreated) for the RI-WA dataset.

3.4 PS conditioning methods

This study compared the treatment effect estimate obtained when using four widely-used PS conditioning methods. PS matching and IPTW were recommended from the literature for use with time-to-event data. PS matching and IPTW for ATT were selected for comparison in estimating the ATT and for estimating the ATE, IPTW for ATE was compared with PS stratification. PS stratification was selected because it had performed well in sparse data settings (Section 2.7). It is acknowledged that there are many more variations of these PS methods that could have been applied in this study, but these examples of PS matching, PS stratification and IPTW were used for ease of comparison of the effect of measurement error and sparse data in the simulations phase of this study. This section describes the initial implementation of these PS conditioning methods on the original dataset with no added measurement error and no adjustment of the outcome prevalence.

3.4.1 PS matching

3.4.1.1 PS matching - method

PS matching was chosen for use as it performs well to remove systematic bias between treatment groups (Austin, 2011a). PS matching is when matched pairs, or groups, are created by matching each treated participant to one or more untreated participant with a similar PS. The estimate of treatment effect is generated from the matched sample or dataset, where only cases for whom a match is made are retained. This is particularly important for observational studies, because there may be a large proportion of cases on the control treatment, so the size of the matched dataset will be smaller than the original dataset. A number of different PS matching method were investigated for this study in order to find the most appropriate one to take forward (Appendix B-6.2). Although other Stata user written commands were considered (Appendix B-6.1), the Stata user written command -psmatch2- (Leuven & Sianesi, 2003) was used to apply these matching methods.

3.4.1.2 PS matching - balance checks

If the matched data were not shown to be balanced, the variable selection for the PS model would have to have been adjusted. The literature agrees that balance checks should be carried out after PS conditioning. Austin (2009a) describes balance checking methods; for PS matching these include reporting descriptive statistics, standardised differences, variance comparisons, use of q-q plots for important variables. Although some quantitative comparisons could be made, most of the assessment was made visually, by comparing graphs.

The balancing checks which were applied to the matched dataset were: comparison of the PS distribution between the treatment groups, standardised differences of the variables and the number of matched pairs/groups generated. Appendix B-6.3 shows the results of the balance checks from the PS matching methods listed in Appendix B-6.2 which were selected for further consideration. They showed no one PS matching method to be superior. The simplest case, *1:1 nearest neighbour no-replacement,* was selected for further use as it balanced the data equally well compared with the other PS matching methods. As the study dataset has considerably more Warfarin patients (18,348) than Rivaroxaban patients (2,911) a many:1 matched dataset could be built. In this case 3 Warfarin patients to 1 Rivaroxaban patient was chosen. The 3:1 matched dataset used more cases and contained more outcome events. So 3*:1 nearest neighbour with replacement* was also chosen for further use. Whereas the 1:1 matching created pairs of data matched on the PS, the 3:1 matching created groups of four matched on the PS.

An additional step was needed as the 3:1 matched dataset was generated using weights. The 3:1 matching used replacement (it is not possible to use the no-replacement option in -psmatch2- so a control case is available for matching to subsequent treatment cases). The weights indicated how many times a control case was matched. The working dataset was transformed to expand this weighted 3:1 dataset so that there was a separate record for each use of a weighted case. This format was needed for the outcome analysis where each matched group was considered to have its own baseline hazard function.

3.4.2 Inverse Probability of Transverse Weighting

IPTW on the PS uses weights, based on the PS, to generate a synthetic dataset or sample. The weight is defined as the inverse probability of receiving the treatment the participant actually received (Section 2.3.4). In this study, IPTW was used to estimate the ATE and ATT. A different formula was used for the calculation of the weight for ATE and ATT. IPTW was implemented by using the Stata community written command -propwt- (Lunt & Linden, n.d.) to generate the appropriate weight, for ATE or ATT. This weight was then used as Stata's *pweight* (probability weights which represent the probability of the case being used in the sample and is proportional to the probability of the case being sampled) in the outcome analysis.

Balance checking was performed on a single run on the original dataset, with no measurement error added. The standardised means were compared between the treatment groups and the continuous variables plotted to compare their distributions between the treatment groups (Appendix B-6.4). These showed the weights applied for IPTW for ATT and IPTW for ATE balanced the standardised means of each variable in the PS model between the treatment groups.

3.4.3 PS stratification

Stratification on the PS is another PS conditioning method used in this study, and is performed when the records of all participants are ordered by PS, then grouped into strata. The treatment effect is estimated within each stratum and then these stratum-specific results pooled, or similar, to generate the ATE and the SE of the estimate (Section 2.3.4). Stratification using 5, 10 and 50 strata were investigated and the balance brought by these strata compared. Five strata were traditionally used (Lunceford & Davidian, 2004), 50 strata were used as a high number of strata and 10 strata used as a compromise of these. In all cases there were both treated and untreated cases in each stratum. The standardised mean differences for each variable are shown in Appendix B-6.5. Using 5, 10 and 50 strata all reduced the standardised differences. As the number of strata increased, the standardised differences for each variable decreased. The number of strata used is a compromise between a larger number of strata reducing the bias in the treatment effect estimate, but can lead to fewer observations in each strata giving a higher variance treatment estimate. 10 strata were chosen to use in PS stratification in this study. 10 strata reduced the standardised differences more than 5 strata and each stratum was less sparse than when 50 strata were used.

3.5 Outcome modelling

The REWARD data were extracted to compare the effect of NOACs compared with Warfarin in the prevention of future stroke, the primary outcome. The outcome analysis was performed on time-to-event data, that is time to first stroke following the first NOAC/OAC prescription, using survival analysis methods. Cox regression was used for the analysis which estimated the treatment effect (Appendix B-7.1).

Different implementations of Cox regression were used to take account of the matched or weighted nature of the data, following PS conditioning used to address the systematic differences between the treatment groups (Appendix B-7.2). When using PS matching, Cox regression with stratification was chosen over frailty, a term used to represent that individuals in the population (such as members of the same family) are heterogeneous due to unobserved factors (Cleves, Gould & Marchenko, 2016, p. 327). Frailty was not suitable as the PS is not an inherent trait, it can depend on the algorithm used. Using Cox regression with stratification, each stratum was a matched pair or group in which the baseline hazard was assumed to be constant. When using IPTW, for both the ATE and ATT, the weights generated by IPTW were used directly as an option in the Cox regression. When using PS stratification, the stratified option was used, again allowing the baseline hazard to vary across strata.

The outcome model considered variables which were clinically relevant and had not been included in the treatment allocation model, thus making the outcome model a conditional model. This maximised the number of potential confounders that had been accounted for, whilst following the traditional PS 2-step approach of firstly accounting for the treatment allocation bias without sight of the outcome and secondly performing the outcome modelling. The estimated treatment effect is therefore conditional on the variables in the outcome model. The outcome model was fitted to the analysis dataset used following PS matching (Appendix B-7.3). Both 1:1 and 3:1 PS matching were considered, but the outcome modelling following 3:1 PS matching was more stable so this was the option retained for use. However, these same variables were used for the other PS methods, PS stratification, IPTW for ATE and IPTW for ATT, without refitting the model to the full dataset. This was done for consistency between the PS methods. It may be regarded as a limitation of this study that the outcome model may be regarded as misspecified. There may be confounding which is not accounted for which introduces bias into the treatment effect estimate.

The variables considered for adjustment as likely confounders in the outcome model were those in the CHA2DS2-VASc (Lip, Nieuwlaat, Pisters, Lane & Crijns, 2010) which is the stroke risk score for patients with Atrial Fibrillation, unless they had been fully accounted for in the PS model (the treatment allocation model) and those advised from expert clinical opinion. 'Univariate' models, using just one variable plus treatment, identified four variables which may be regarded as significant with p-values <0.05, being prescribed blood pressure lowering medication, statins or antiplatelets and hypercholesterolemia. Models were then fitted which included only these four variables, these four variables and the CHA2DS2-VASc score, and these four variables and the variables used in the CHA2DS2-VASc score (Appendix B-7.4). The models all included treatment. The chosen model was the four variables and the CHA2DS2-VASc score (Table 4).

Table 4: The outcome model selected for use. The model includes treatment, the 4 most significant univariate variables and the CHA2DS2-VASc score.

Covariate	HR	SE of HR	95% CI of HR	Coeffi-cient*	SE of Coeffi-cient	95% CI of Coefficient	p-value
Treatment	1.534	0.383	(0.940, 2.504)	0.428	0.250	(-0.062, 0.918)	0.087
Prescribed blood pressure lowering medication	0.339	0.110	(0.180, 0.639)	-1.081	0.323	(-1.714, -0.448)	0.001
Prescribed statins	0.677	0.245	(0.333, 1.378)	-0.390	0.362	(-1.100, 0.321)	0.282
Prescribed antiplatelets	0.646	0.225	(0.326, 1.279)	-0.437	0.349	(-1.121, 0.246)	0.210
Hypercholesterolemia	0.729	0.269	(0.354, 1.502)	-0.316	0.369	(-1.039, 0.407)	0.391
CHA2DS2-VASc score	1.360	0.165	(1.073, 1.725)	0.308	0.121	(0.070, 0.545)	0.011

*Coefficient is the log(hazard ratio)

Although the hazard ratio (HR) is often presented for time-to-event studies in applied medical research, the assessment of bias is generally carried out on the regression coefficient estimates for the Cox model covariates (the log(hazard ratio)). This method is used to assess the bias to compare different models using PS methods for time-to-event data using simulations (Austin, 2013; Gayat et al., 2012). The current study adopted this approach by using the log-hazard-ratio for the assessment of the performance measures of the estimate of the treatment effect. The Cox regression coefficients were of more interest than the hazard ratio as they in turn were used in the simulations. The selected outcome model is therefore also presented with the covariate coefficients, the log-hazard-ratios (Table 4).

When the outcome model was generated using the Cox model the baseline hazard $h_0(t)$ was not calculated. However, in the simulations, developed in Chapter 4, it is needed to generate the simulated survival time. This can be done using a parametric survival model. Empirical investigations had suggested a Weibull model would be an appropriate baseline hazard function. A Weibull distribution offered flexibility with parameters of γ, the shape parameter, and λ, the scale parameter. By varying the shape parameter, γ, the distribution of the function changes, for $\gamma = 1$ this distribution is an exponential, so the hazard is constant (Appendix B-7.5). The function of the baseline hazard was assessed empirically which gave baseline hazard parameters of λ = 0.00029933 and γ = 0.480355 (Appendix B-7.6).

3.6 Summary

The treatment allocation model (the PS model) and the outcome model were fitted to the original dataset with no added measurement error using the original outcome prevalence. The PS methods, 3:1 PS matching, IPTW for ATE, IPTW for ATT and PS stratification, were applied to remove the treatment allocation bias between two treatment groups, and the balance checking confirmed that the PS model was a close estimate of the true PS model. The outcome model

was applied to the data following each PS conditioning method, accounting for the nature of the data, such as matched or weighted. A baseline hazard function was also generated for use in the simulations work. These were all used in the framework to run the simulations which was developed in Chapter 4.

Chapter 4 DEVELOPMENT FOR THE SIMULATIONS FRAMEWORK

4.1 Introduction

In Chapter 3 the study dataset was built to compare the treatment effect estimate of Rivaroxaban, a NOAC, with Warfarin in the prevention of future stroke in patients with AF. The PS methods used were IPTW for ATE, IPTW for ATT, 3:1 PS matching, and PS stratification. The PS model, the Cox PH model (used in the outcome analysis) and the baseline hazard function were all fitted to the study dataset with the original characteristics.

This chapter describes the development of the simulations framework which implemented the methods from Chapter 3. The simulations generated performance measures of the treatment effect estimate which would be used to compare the PS methods used in this study under different conditions. Parameters were varied to assess the impact of measurement error and sparse outcomes. Measurement error was investigated in two ways: introducing measurement error in the variable for previous stroke into the PS model; and varying the effect size of this variable in the PS model. This will contribute to the guidance on the use of the different PS methods under these conditions, when using observational data (or routinely collected data) to assess the treatment effect of a novel treatment.

Section 4.4 justifies the range of measurement error used. The selection of the plasmode simulation method (Franklin, Schneeweiss, Polinski & Rassen, 2014), where the draws are made from the original data preserving the relationship between the variables for each case and its application developed for use in this study, are described in Section 4.2.2. Initial results from simulations run with the data characteristics of the study dataset are given in Section 4.3. The parameters to vary in the simulations are applied: introduced measurement error in the variable previous stroke, a variable in the PS model (Section 4.4); changes to the effect size of this variable in the treatment allocation model (PS model) (Section 4.5); and changes to the prevalence of future stroke, the primary outcome (Section 4.6). The number of simulated datasets required is calculated in Section 4.7 and a plan for the full simulation runs is given in Section 4.8. The results from these simulations are reported in Chapter 5.

4.2 Development of the simulation method

4.2.1 Examples of simulation methods from the literature

Broadly, there are two methods to create simulated datasets; resampling and Monte Carlo Simulations (MCS). In resampling methods, random draws of cases are made from the original data and saved to the generated dataset. Schafer and Kang (2008) and Franklin et al. (2014) use

resampling in their Data Generation Mechanisms (DGM). In MCS, the variables are generated from random draws of known or calculated functions. Generated variables, such as predicted treatment and outcomes, are generated from these 'new' baseline covariates. Morris, White and Crowther (2019), Tumlinson, Sass and Cano (2014) and Chu et al. (2012) report on the use of this method.

Plasmode simulation (Vaughan et al., 2009) is a resampling method, where the draws of cases, that is 'individual patient records', are made from the original data and the resulting cases copied to the generated dataset. This preserves the relationship between the baseline variables for each case. Franklin et al. (2014) used a plasmode simulation method, where a number of datasets, *J,* were created with size, n, which was less than or equal to the original dataset, from bootstrapped samples (with replacement) from the original dataset. Each of the generated datasets were analysed and the results combined for estimates of bias and variance. The current study's simulation DGM was similar to plasmode simulations. Using joint distributions in Monte Carlo simulations can also maintain the relationship between the baseline variables. In the current study's data there is a mixture of discrete and continuous variables, usually joint distributions are used when all the variables are discrete or all the variables are continuous so it is very difficult to apply to this study dataset (Thomopoulos, 2013).

4.2.2 Simulations method

The simulations method in this study was developed to compare the performance of the treatment effect estimate when using the different PS methods in the presence of measurement error in a variable in the PS model, the change of effect size of this variable in the PS model and change in outcome prevalence on the treatment effect estimate. The simulated datasets were generated by resampling (with replacement) from the original dataset. This preserved the relationship between the baseline covariates for each case. Once a dataset had been created, measurement error was introduced into the variable for previous stroke, to represent under- or over-recording of that variable, and an amended value generated for the PS value and CHA2DS2-VASc score (Appendix B-7.3). Variables for the simulated treatment allocation, simulated survival time and simulated survival outcome were created using the baseline variables, the chosen values for the effect size in the PS model of the variable with measurement error and the outcome prevalence. The treatment effect was estimated from the dataset (using the simulated variables) and recorded. Performance measures of the PS methods, mean, SD, bias, MSE, percentage change MSE and mean Model SE, were calculated from the treatment effect estimates from all the generated datasets. The calculation for all simulation runs used an assumed true mean obtained from a single simulation run using the original dataset. It was

generated as a plausible value to use in all simulations for estimations of both the ATE and the ATT, in order to investigate the variations in the treatment effect estimate due to covariate measurement error and outcome prevalence.

The simulations process is shown as a flow diagram in Figure 2.

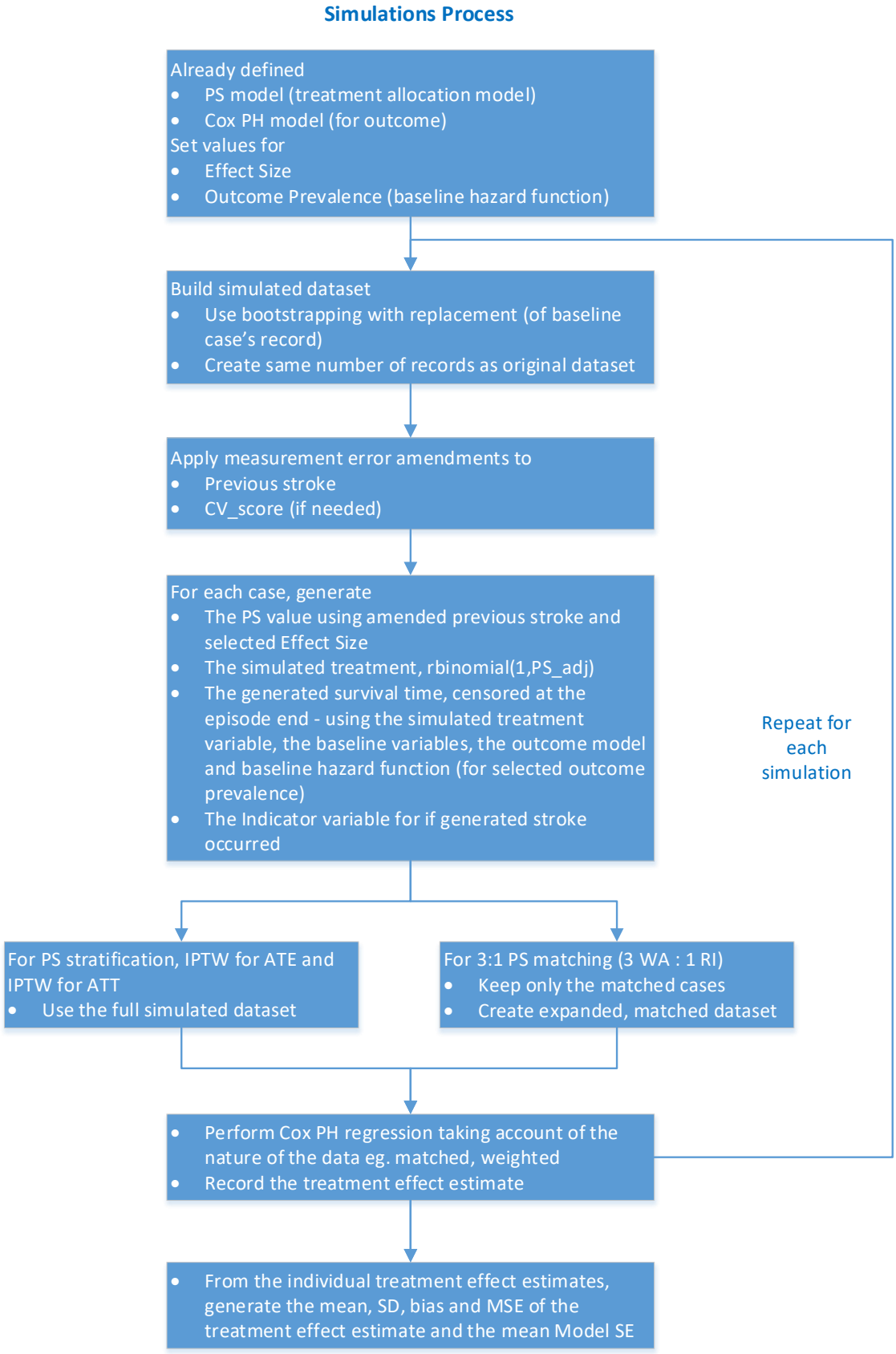


Figure 2: Flow diagram of the simulations process (CV_score is CHA2DS2-VASc score).

4.2.3 Performance measures used

Table 5 gives the formulae used to calculate the performance measures of the treatment effect estimate in the simulations, Standard Deviation (SD), Bias, the absolute Mean Squared Error (MSE) and Mean Squared Error percentage change. This terminology is used in this study.

Table 5: Definition of the performance measures used.

Performance Measure	Definition
Standard Deviation (SD) (Efron & Tibshirani, 1993, p. 47)	$SD = \sqrt{\frac{\sum_{i=1}^{n}(\theta_i - \hat{\theta})^2}{(N-1)}}$
Bias (Austin, 2013)	$bias = \left(\frac{1}{N}\right) \Sigma_{i=1}^{n}(\theta_i - \theta)$
Mean Squared Error (MSE) (Austin, 2013)	$MSE = \left(\frac{1}{N}\right) \Sigma_{i=1}^{n}(\theta_i - \theta)^2$
MSE Percentage Change	$MSE\ \%\ change = \left(\frac{MSE_0 - MSE_k}{MSE_0}\right) * 100$
Model SE Mean (Morris et al., 2019)	Mean of the SE of the treatment effect estimate collected from the outcome analysis from each generated dataset

where

HR is the hazard ratio, the increased hazard when prescribed Rivaroxaban compared to when prescribed Warfarin

θ is the true marginal log(HR)

θ_i is the estimate of log(HR) from the ith dataset

$\hat{\theta}$ is the sample mean of θ_i

MSE_0 is the MSE at no measurement error

MSE_k is the MSE at k measurement error

N is number of cases in the dataset

4.3 No added measurement error

Using the methods described in Section 4.2.2, the simulations were run (Table 6). 100 simulated datasets were used to verify the functionality of the simulations method. The simulated datasets had the characteristic of the original dataset. The performance measures given in Section 4.2.3 were recorded. The mean estimated treatment effect was displayed as the log of the hazard ratio (Section 3.5) and SD, bias and MSE all relate to this. *Number Valid* recorded the number of non-missing values which were used in these calculations. The coding of the simulations captured the return code of the outcome analysis. If the Cox regression failed, for example a Cox model could not be fitted to the outcome data in the dataset, the results from that dataset were recorded as missing and this allowed the simulations to continue by moving to the next simulated dataset. In these simulations there was no missing data. *Number Events* was the

number of outcome events (future stroke), *Number WA* was the number of cases with a generated treatment of Warfarin (the control treatment) and *Number RI* was the number of cases with a generated treatment of Rivaroxaban (the novel treatment). For 3:1 PS matching these referred to the values in the expanded matched datasets used for the outcome analysis. The values displayed were all means of the values taken from the datasets used in each run. These example simulations confirmed that the simulation coding worked as expected and had a mechanism to continue the simulation run, even if a single simulation failed. The performance measures of interest the mean, SD, bias and MSE of the log(HR) of the treatment effect and mean Model SE were collected and were displayed in tabular and graphical format for later simulations.

IPTW for ATT and PS stratification gave the treatment effect estimate with the lowest bias. 3:1 PS matching had the largest bias of the treatment effect estimate, which was positive so over-estimated the treatment effect, whereas the bias in the treatment effect estimates using the other PS methods was negative, which under-estimated the treatment effect. The SD of the treatment effect estimate varied from 0.2110 (PS stratification) to 0.2589 (3:1 PS matching). These performance measures showed that there was a difference in performance of the PS methods used, this is discussed further in Section 5.2. The simulations method was next enhanced by introducing measurement error, varying the effect size of the covariate with measurement error and varying the outcome prevalence.

When using 3:1 matching, a feature of the matching process was seen. The resampled datasets were built using *plasmode* simulations which were obtained using bootstrap sampling with replacement. In the simulated datasets the generated treatment was created using binomial distribution with probability of success defined by the amended PS, so it was possible that different occurrences of the same original case would have different generated treatments. They would have similar PS values so would be obvious choices for matching when using 3:1 matching and there was no restriction on this happening. Around 20% of Warfarin cases were matched to a Rivaroxaban case based on the same case in the original dataset. In this study this is referred to as ‘self-matching’.

Table 6: Preliminary results from the different PS methods with no added measurement error, using 1% prevalence and 100 datasets (N=100).

PS method	% M error	Mean*	SD*	Bias*	MSE*	MSE % change	Model SE mean	Num events	num WA	Num RI	N valid
IPTW for ATE	0	0.3361	0.2443	-0.0317	0.0607	0.0	0.2576	222.4	18344.8	2914.2	100
IPTW for ATT	0	0.3653	0.2165	-0.0022	0.0469	0.0	0.2037	180.4	18344.8	2914.2	100
3:1 PS match	0	0.4173	0.2589	0.0504	0.0695	0.0	0.2420	100.2	8748.1	2916.1	100
PS Strat-ification	0	0.3646	0.2110	-0.0029	0.0445	0.0	0.2005	220.7	18344.8	2914.2	100

*displayed as the log(HR)

4.4 Added measurement error

Guidance for the initial starting point for introducing measurement error into the variable *previous stroke* is informed by earlier work using the data on the REWARD study (Burnell, 2015) and Herrett et al. (2013) who compared recording of Myocardial Infarction (MI) from different data sources. The work from REWARD compared the recording of stroke events using an extract of primary care data from THIN only, with THIN and HES (secondary care) data. Although the REWARD data were primarily an extract from THIN (primary care data) 37% of the practices had HES-linked data, although these data were only available for the first three months of the study. Using THIN only, 343 strokes were recorded, but using both THIN and HES, 516 strokes were recorded. THIN only recorded 66.5% of strokes obtained when using THIN and HES. The HES and THIN combined result may not be a gold standard as both datasets are likely to be subject to different measurement errors, but the combined values are more likely to be more accurate than just using a single data source.

Herrett et al. (2013) used four data sources to compare the recording of MI and all-cause mortality in patients who had suffered acute MI. The data sources used were primary care (Clinical Practice Research Datalink - CPRD), secondary care (HES), disease registry (Myocardial Ischaemia National Audit Project - MINAP) and mortality register (Office for National Statistics - ONS). The number of subsequent Mis recorded differed across data sources. The number of events given by only one data source was underestimated by 25% to 30% compared to using three of the data sources. Looking at the incidence of MI, the number of events using only the CPRD was 25% lower than using three of the data sources.

The current study's author acknowledges that there can also be over-recording of events like stroke or MI, such as where an incorrect coding is used. However, it is more likely that the recording of an event is missed, leading to under-recording. The assumption of Herrett et al. (2013) is that the discrepancy between the different data sources is due to missing recordings

in each data source. Herrett et al. (2013) specifically investigates how parts of the healthcare system do not capture events which happen in other parts of the system. Both MI and stroke are events which are treated in the hospital setting and reported back to primary care. It is therefore a reasonable assumption that there will be under-recording of these events in primary care data.

Based on these two pieces of work, it would therefore appear that primary care data under-records stroke events by 25% to 35%. These are estimations as the data sources used for comparison may also be subject to recording error, not the true value (Herrett et al., 2013). However, it is possible, that there is over-recording in the hospital setting; for example all patients admitted with a suspected stroke are recorded as having had a stroke. To provide the complete picture, the current study looks at measurement error in both directions, under-recording and over-recording. The chosen range of the measurement error was expanded to -50% to +50%, which included the provisional estimate of 25% to 35%.

Measurement error was introduced into the previous stroke variable, to represent under- or over-recording, using the algorithm described (Table 7). The 'temporary variable' was calculated as the binomial(1,pr), where pr is the modulus of (p/100) and p is the chosen percentage measurement error. When positive measurement error was applied (over-recording), no change was made when a case had a previous stroke. When a case did not have a previous stroke the modified previous stroke became positive if the temporary variable was equal to 1. If the temporary variable was equal to 0, the modified previous stroke remained negative. When applying negative measurement error (under-recording), if a case did not have previous stroke no change was made. When a case did have a previous stroke the modified previous stroke became negative if the temporary variable was equal to 0. The modified previous stroke remained positive if the temporary variable was equal to 1.

Table 7: Algorithm for generating the measurement error to create the modified previous stroke variable.

Change made	Original Previous Stroke Variable	Temporary variable	Modified Previous Stroke Variable
Adding events	Yes	Any	Yes
	No	0	No
	No	1	Yes
Removing events	Yes	0	No
	Yes	1	Yes
	No	Any	No

The simulations were then expanded to use added measurement error between -50% to +50% in previous stoke, a covariate in the PS model. The results of the full simulations are given in Section 5.3.

4.5 Change of the effect size in the PS model

In addition to running simulations to investigate the effect of under- or over-recording in a covariate in the treatment allocation model by introducing measurement error, the effect size of that variable in the treatment allocation model was also included as a parameter to vary in the simulations. This section introduces the method to change the effect size of the variable with measurement error to assess its impact on the treatment effect estimate. This meant changing the coefficient of previous stroke in the PS model. Suggested values were given to achieve low, medium and high effect size. These values were then applied to simulations reported in Section 5.4.

The Odds Ratio (OR) of interest is

$$OR = \frac{odds\ of\ receiving\ Rivaroxaban\ if\ had\ previous\ stroke}{odds\ of\ receiving\ Rivaroxaban\ if\ no\ previous\ stroke}$$

Previous stroke is coded 0 for No, 1 for Yes

If β_1 is the coefficient for previous stroke in the PS model, then $OR = \exp(\beta_1)$ Cohen's *d* is the standardised mean difference between two group means, the effect size underlying power calculations for the two-sample t-test (Cohen, 1988). Cohen's *d* = 0.2, 0.5, and 0.8 are often used to indicate a low, medium, and high effect size (Chen, Cohen & Chen, 2010). Chen et al. (2010) calculated the Odds Ratios (OR) equivalent to Cohen's *d*, for low, medium and large effect size and presented the OR for different disease rates in the non-exposed group.

The calculated values of ORs, which were relevant to this study, are given in Table 8 and informed the values of the coefficient used to represent the low, medium and large effect size based on Cohen's *d*. The coefficient of previous stroke in the PS model was generated $= \ln(OR)$. This value was supplied as a parameter to the simulations and used in the PS model.

This change of effect size related to the PS modelling which was used to correct for treatment allocation bias. The 'outcome' in this case was the generated treatment, which was created using the participants' PS value. The 'untreated' group consisted of those with no previous stroke, the 'outcome prevalence' was those with a generated treatment of Rivaroxaban (the NOAC). In the study data the outcome prevalence took values between 13% and 14.2%, so the values quoted for the 10% prevalence in Table 8 were used to change the effect size of previous

stroke in the PS model. Rounding these parameters for use in the simulation runs reported in Section 5.4, the coefficient of previous stroke in the PS model took the values of 0.5 for low effect size, 1.0 for medium effect size and 1.5 for large effect size. When the PS model was fitted to the original data, the effect size of previous stroke was 0.123. This was 'very low' compared with Cohen's classification. Simulations using this very low (or original) effect size are also presented in Section 5.4.

Table 8: Effect sizes for prevalence of Rivaroxaban is generated treatment of 1% and 10%.

	Low Effect Cohen's d=0.2			Medium Effect Cohen's d=0.5			Large Effect Cohen's d=0.8		
Preva-lence**	OR	Coeffi-cient	xorig*	OR	Coeffi-cient	xorig*	OR	Coeffi-cient	xorig*
0.01 (1%)	1.6814	0.519627	4.2	3.4739	1.24528	10.1	6.7128	1.90402	15.5
0.1 (10%)	1.4615	0.379463	3.1	2.4972	0.91517	7.4	4.1387	1.42038	11.6

* the multiple of the original coefficient, 0.1229108

** prevalence of Rivaroxaban (the NOAC) is generated treatment

4.6 Sparseness of outcome data

The simulation method was now expanded to investigate the effect of the sparseness of the outcome data. This section looks at the method needed to generate datasets with different outcome prevalences.

The study's data can be regarded as sparse due to its rare outcomes, despite there being 82,795 patient records in the original data set. After selecting patients who were NOAC/OAC-naïve, and prescribed Rivaroxaban or Warfarin during the time Rivaroxaban was available there were 21,259 cases with 232 outcomes (hence an outcome prevalence of 1.1%). This dataset was used for PS stratification, IPTW for ATT and IPTW for ATE. Applying 3:1 PS matching, the matched dataset had 11,644 cases with only 98 outcomes. Indeed if 1:1PS matching had been applied there would only have been around 45 outcomes.

The model used to generate the survival times used in the simulations was a parametric survival model fitted to the original data. A Weibull distribution has parameters of γ, the shape parameter, and λ, the scale parameter, and was used for the baseline hazard because of the flexibility it offered. The modelling to the original data suggested the Weibull parameters of λ=0.000299 and γ=0.480355 (Section 3.5) and these values created the generated datasets with an outcome prevalence of around 1%. For this study, λ was varied and γ was kept constant to generate simulated datasets of different prevalences. This kept the shape of the baseline hazard function, γ, constant and as it was <1 it remained a monotone decreasing function.

To investigate the effect on the treatment effect estimate from variation of the sparseness of the outcomes, simulations were run with parameters to generate difference outcome prevalences. The values chosen were approximately 1% prevalence, which is similar to the original dataset and so is known to exist in real-world data, approximately 0.5% prevalence to investigate the effect of a lower prevalence, and approximately 10% prevalence to investigate the effect of data which does not suffer from sparseness of outcomes (Table 9). The results of the full simulations varying the outcome prevalence are given in Section 5.5.

Table 9: Baseline hazard changes for selected values – fixed γ and varying λ.

Prevalence	N	% m error	*Mean	*SD	Mean Number Events	Number WA	Number RI	N Valid	λ	γ
0.4%	100	0	0.4839	0.4688	50.5	8733.0	2911.0	100	0.00015	0.480355
0.9%	100	0	0.4016	0.2851	101.7	8733.0	2911.0	100	0.000299	0.480355
8.8%	100	0	0.3679	0.0868	1020.3	8733.0	2911.0	100	0.00325	0.480355

*log(HR) of estimated treatment effect

4.7 Sample size calculations

The initial simulation runs reported previously in this chapter had been made using 100 simulated datasets and were used to demonstrate the performance of the simulation method and functionality. The sample size of 100 datasets was arbitrary. This section explores the number of simulated datasets which should be used, based on the precision sample size calculation. The number of simulated datasets is regarded as the 'sample size'.

The sample size was determined by calculating CI widths of the mean treatment effect estimate from some additional simulations using 1,000 datasets, determining an acceptable CI width, and calculating the number of simulations required to give the acceptable CI width. Full details of these calculations are given in Appendix C.

It is acknowledged here that the selection of an acceptable CI is subjective. From visual inspection of plots of the mean of the treatment effect estimate, a CI of 0.04 was thought to be too high. A CI of 0.0367 is 10% of the true value of the treatment effect, 0.3674, using the dataset with the original characteristics. Combining this information and the visual inspection, an acceptable CI width of 0.035 was decided upon. The number of simulations required to generate this CI width are given in Table 10. For a CI width of 0.035, the required sample size for each prevalence was: for 10% prevalence, the lowest was 54 (PS stratification) and the highest was 106 (3:1 PS matching); for 1% prevalence, lowest was 560 (PS stratification) and the highest was 1124 (3:1 PS matching); for 0.5% prevalence, lowest was 1219 (PS stratification) and the highest

was 3081 (3:1 PS matching). Generally, the sample sizes were rounded up to the next 100 for the simulations runs presented in Chapter 5.

Table 10: Calculated Sample Size for CI width=0.035.

PS_Method	Prev-alence	Mean	SD of mean	N Calculated	N Rounded up
IPTW_ATE	0.5%	0.3165	0.4001	2008	2100
IPTW_ATE	1%	0.3483	0.2792	978	1000
IPTW_ATE	10%	0.3626	0.0872	95	100
IPTW_ATT	0.5%	0.3492	0.3174	1264	1300
IPTW_ATT	1%	0.3560	0.2142	576	600
IPTW_ATT	10%	0.3639	0.0664	55	100
3:1_match	0.5%	0.4684	0.4956	3081	3100
3:1_match	1%	0.4256	0.2994	1124	1200
3:1_match	10%	0.3716	0.0918	106	200
PS_strat	0.5%	0.3477	0.3117	1219	1300
PS_strat	1%	0.3575	0.2113	560	600
PS_strat	10%	0.3643	0.0656	54	100

4.8 Plan for simulation runs

To assess the effect of covariate measurement error and the sparseness of outcome data on the treatment effect, the parameters given in Table 11 were used. They were run for the PS methods assessed in this study and the results are presented and discussed in Chapter 5.

Table 11: Parameters and their values used in the simulation runs.

Parameter	Values
Measurement error in previous stroke in the PS model	-50%, -30%, -10%, 0%, +10%, +30%, +50%
Effect size of variable with measurement error	0.123 (original), 0.5 (low), 1.0 (medium) and 1.5 (high)
Outcome prevalence of future stroke	0.5%, 1%, 10%
Sample size, N, the number of simulated datasets	Specific to each PS method for each outcome prevalence

The chosen range of the measurement error was -50% to +50%, which included 25% to 35% given by the literature (Herrett et al., 2013). The effect size of a variable with measurement error could take the original, low, medium or large value. The outcome prevalence of future stroke could take the value 0.5%, 1%, 10%. 1% is close to the prevalence of the original data. The

number of simulated datasets used, N, was calculated for each PS method at each outcome prevalence (Section 4.7).

The variable for previous stroke contributes to the outcome model as part of the CHA2DS2-VASC score (Section 3.5). The CHA2DS2-VASC score is a validated score so the effect size of previous stroke is fixed within it. If the effect size of the CHA2DS2-VASC score were changed, it would also influence the effect of all the other variables used within it. Also, whether a patient has a future stroke (the outcome of interest) will be dependent on the true value for previous stroke (among other variables), that is if they actually had a stroke previously, not the error-prone variable, which is the one recorded in their EHR. Therefore, no measurement error was introduced in the outcome model; the scope of this study was to investigate the effect of measurement error in variables in the PS model (the treatment allocation model) only.

An extract of the Stata coding used to run these simulations is given in Appendix D.

4.9 Summary

The simulations method, presented in Section 4.2.2, established the simulations for use in this study. The parameters which could be varied were:

- measurement error applied to the variable for previous stroke (Section 4.4)
- the effect size of this variable in the PS model (treatment allocation model) (Section 4.5)
- the outcome prevalence (future stroke) (Section 4.6).

Preliminary runs using the simulations methods were applied to different PS methods, 3:1 PS matching, IPTW for ATT, IPTW for ATE and PS stratification. The simulations method allowed variation of sufficient parameters and correctly reported the performance measures of the treatment effect estimate to allow for comparison of the performance of the different PS methods used in this study. The results of these simulations are presented in Chapter 5 to compare the performance of the different PS methods under these conditions.

The work in this chapter developed the simulations method which allowed several parameters to vary: the effect of covariate measurement error; the effect size of this covariate in the PS model; and sparseness of data on the treatment effect estimate. The amount of measurement error in previous stroke, a variable in the PS model took the range -50% to +50% (Section 4.4). The effect size of this variable in the PS model could also be varied so that it had a low, medium or large effect (Section 4.5). The outcome prevalence could be varied by changing the number of outcome events (future stroke) in the generated datasets. The prevalences used were 0.5%,

1% and 10% (Section 4.6). The sample size for the simulations, the number of generated datasets, was informed by the work in Section 4.7. The results of the preliminary simulations were presented in tabular and graphical format. In Chapter 5 the simulations varying these parameters were run using the full sample size for each PS method. The results are presented and discussed.

Chapter 5 SIMULATIONS RESULTS

5.1 Introduction

In this chapter, simulation experiments were run using the method developed in Chapter 4. The simulated datasets were generated from the study data which compares the performance of Rivaroxaban (the novel treatment) with Warfarin (the control treatment) for patients with AF in the prevention of future stroke.

Section 5.2 compares the PS methods with no introduced measurement error. In Section 5.3, measurement error was introduced into a variable in the treatment allocation model to investigate the impact of both under-recording and over-recording of an event in a patient's EHR on the treatment effect estimate. In Section 5.4 the effect size in the treatment allocation model of the variable with measurement error is varied as well as the amount and direction of measurement error. In the original study's dataset the variable with assumed measurement error had a very low effect size in treatment allocation. In Section 5.5, the outcome prevalence was varied as well as the introduced measurement error to demonstrate the impact of sparse outcome data, which can be present in extracts from EHRs when there is under-recording or over-recording of an event which affects treatment allocation. Section 5.6 combines all these and varies the measurement error, its effect size in the PS model and the outcome prevalence. Simulations were run in all these scenarios using all four PS methods so that a comparison between them could be made. Recommendations for the PS method to use in the estimation of the ATE and ATT are made in Sections 5.7 and 5.8, respectively.

The method developed in Chapter 4 had demonstrated that the simulations generated the mean, SD, bias, absolute MSE, percentage change MSE and model SE of the treatment effect estimate. These were all used as performance measures of the treatment effect estimate and are reported in this chapter, in tabular and graphical format. The results display the estimate of the treatment effect (of Rivaroxaban over Warfarin) presented as the log(HR).

In this chapter the simulation results are demonstrated with the results using IPTW for ATE in graphical format. The results in tabular format and the results using the other PS methods are displayed in Appendix E (varying measurement error and outcome prevalence in tabular form and plotted for each PS method), Appendix F (varying measurement error and effect size in tabular form and plotted grouped by prevalence and PS method) and Appendix G (results from Appendix F plotting all PS methods together, grouped by effect size and prevalence). For integrity of results presentation, some plots given in the main body of the thesis are also repeated in the Appendices. The heat plots, generated using the Stata user written command -

heatplot- (Jann, 2019), (Figure 5 to Figure 10) present the results from all the simulations together. A summary of the findings is given in Table 13.

5.2 Results using original data characteristics

The four PS methods used in this study, IPTW for ATE, IPTW for ATT, 3:1 PS matching and PS stratification, were applied, using simulations, to the original study dataset to compare their performance with no added measurement error. The results are shown in Table 12. All values relate to the log of the HR. IPTW for ATE, IPTW for ATT and PS stratification used the full dataset for analysis, whereas 3:1 PS matching used the dataset containing matched cases only. The number of Warfarin cases (Num WA) was lower for 3:1 PS matching and the 3:1 PS matching simulated datasets also had a lower number of outcomes (Num Events). The treatment effect estimate using IPTW for ATE, IPTW for ATT and PS stratification all had negative bias and 3:1 PS matching had positive bias. PS stratification had the least biased estimate and 3:1 PS matching had the most biased. PS stratification had the highest precision (lowest SD) and lowest MSE of the treatment effect estimate and 3:1 PS matching had the lowest precision and highest MSE. PS stratification had the lowest mean of the Model SE and IPTW for ATE had the highest. Overall, 3:1 PS matching appeared to be performing the least well, giving a treatment effect estimate with the highest bias, lowest precision (from the highest SD) and the highest MSE. 3:1 PS matching was however retained in this chapter to investigate its performance in the presence of measurement error and sparse outcomes and for comparison in the estimation of the ATT.

Table 12: Simulation results comparing the PS methods with no added measurement error.

PS Method	Out-come Preva-lence	N	% m error	Mean*	SD*	Bias*	MSE*	Model SE mean	Num Future Stroke	Num WA	Num RI
IPTW for ATE	1%	1000	0	0.3494	0.2654	-0.0181	0.0707	0.2587	222.0	18350.2	2908.8
IPTW for ATT	1%	1000	0	0.3565	0.2083	-0.0110	0.0435	0.2035	181.5	18350.2	2908.8
3:1 PS Matching	1%	1200	0	0.4102	0.2857	0.0428	0.0834	0.2435	100.1	8726.2	2908.7
PS strati-fication	1%	1000	0	0.3575	0.2044	-0.0099	0.0418	0.2006	219.3	18350.2	2908.8

*displayed as the log(HR). *WA*: Warfarin. *RI*: Rivaroxaban.

5.3 Results with added measurement error

The impact of incorrect recording of a variable which affects the treatment allocation (previous stroke) was investigated. Under-recording of previous stroke, thought to be the more likely scenario (Section 4.4), was demonstrated by introducing negative measurement error in the range of -50% to 0%. Over-recording of previous stroke was demonstrated using positive measurement error in the range 0% to +50%. The original dataset characteristics were retained

and the simulations used the number of datasets given in Section 4.7. The results from the simulations are shown in Figure 3.

Over the measurement error range, 3:1 PS matching was the only PS method which produced a treatment effect estimate with positive bias. This bias had a magnitude of approximately three times that when using the other PS methods. 3:1 PS matching showed a rise in both the mean and bias of the treatment effect estimate at measurement values of +10% and + 20%. The other PS methods showed little variation in the mean and bias over the measurement error range.

The SD and MSE of the treatment effect estimate showed the same pattern for all four PS methods, a decrease in their values from higher negative measurement error toward no measurement error and then a slightly steeper decrease from no measurement error to higher positive values of measurement error. Over the measurement error range, 3:1 PS matching still had the highest values for the SD and MSE and PS stratification and IPTW for ATT had the lowest values (which were similar to each other). The model SE followed a similar pattern to the SD and MSE over the measurement error range, except that IPTW for ATE had the highest values. The percentage MSE change increased from higher negative measurement error towards no measurement error, then increased more steeply from no measurement error to higher positive measurement error. Generally, 3:1 PS matching had the lowest values.

Over the whole measurement error range, 3:1 PS matching had the treatment effect estimate with the highest bias and also had the highest SD and MSE, so had the treatment effect estimate with the lowest precision. The other three PS methods gave the treatment effect estimate with a similar bias, but PS stratification and IPTW for ATT both had the highest precision.

Neither under-recording (negative measurement error) nor over-recording (positive measurement error) of a variable affecting treatment allocation gave a biased estimate of the treatment effect using these data characteristics. Previous stroke is a positive contributor with a very low effect size in the treatment allocation model (PS model). However, 3:1 PS matching continued to give the treatment effect estimate with the largest bias, which was positive, whereas the other PS methods gave a smaller negative bias. Higher under-recording of a variable affecting treatment allocation gave a treatment effect estimate with lower precision and higher MSE. Higher over-recording of a variable affecting treatment allocation gave a treatment effect estimate with lower precision and higher MSE. These results are produced when the variable with measurement error is a positive contributor to the treatment allocation (PS model). PS stratification and IPTW for ATT gave the treatment effect estimates with the highest precision and lowest MSE.

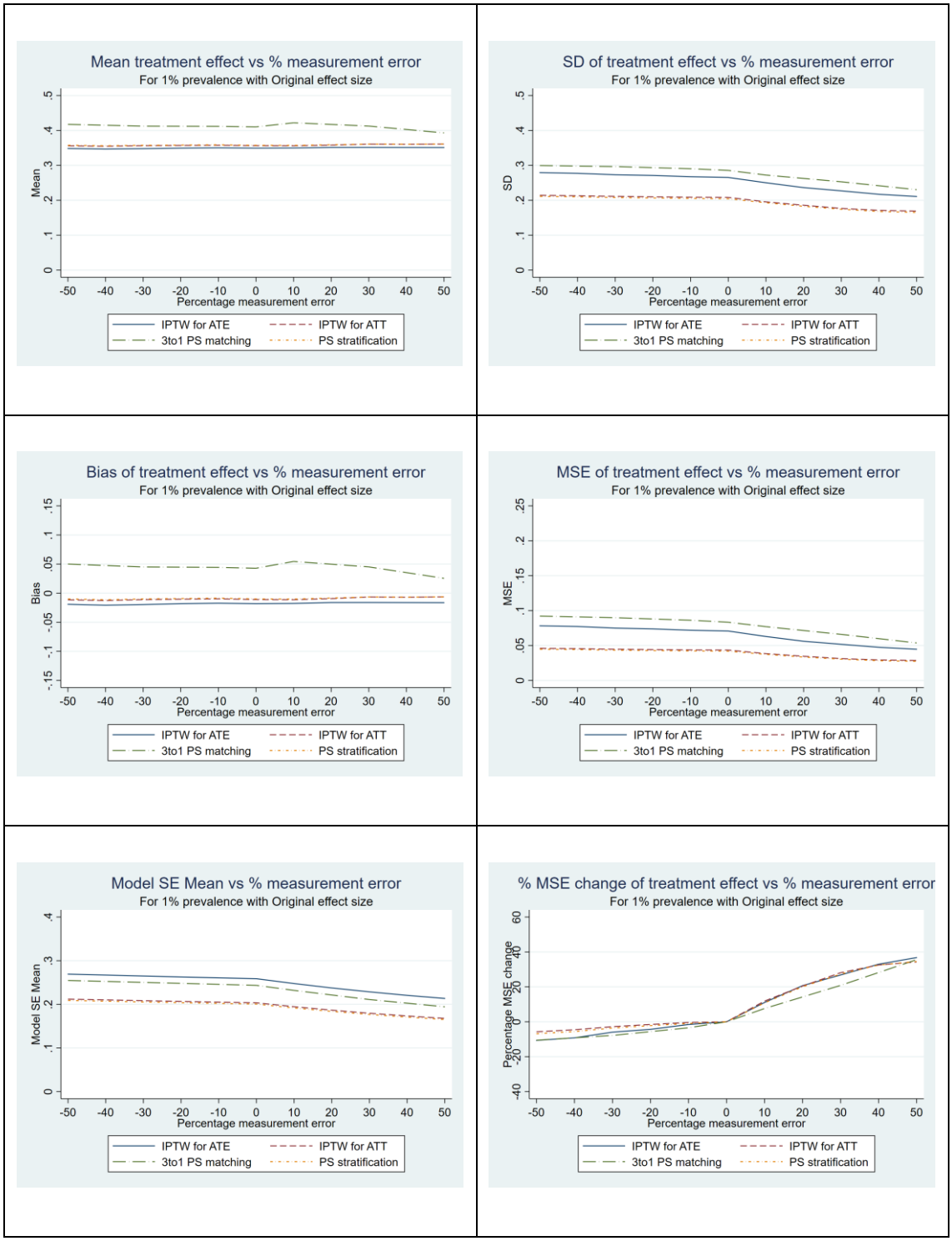


Figure 3: Study PS methods, 1% prevalence, original effect size - the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect displayed as log(HR).

5.4 Results varying measurement error and effect size

This section presents the results for simulations where the effect size on treatment allocation of the variable with measurement error (previous stroke) was varied by changing the coefficient in the PS model to investigate the effect on the treatment effect estimation. The under-recording

and over-recording remained the same used in Section 5.3 by using measurement error for previous stroke over the range, [-50%, +50%]. This part of the simulation experiment investigated change to the treatment effect estimate with different effect sizes, to make the work applicable to other datasets where the variable with measurement error has a higher impact. The values used for the coefficient of previous stroke in the PS model were 0.5 for the small effect size, 1.0 for a medium effect size and 1.5 for a high effect size (Section 4.5). These were compared with the simulations run in Section 5.3, using the original value of the effect size of the variable with measurement error in the PS model, 0.1229 (Section 3.3), which was regarded as very low in using this categorisation.

These simulations were run for all PS methods investigated in this study and use the same characteristics as the original data. The results from each PS method using the different effect sizes were plotted together for ease of comparison. Figure 4 displays the different effect sizes for IPTW for ATE, the plots for the other PS methods are given in Figure F-2 for IPTW for ATE, Figure F-5 for IPTW for ATT, Figure F-8 for 3:1 PS matching, and Figure F-11 for PS stratification. The heat plots (Figure 5 to Figure 10) present all the results together, the columns marked 1% prevalence on the plots relates to this section and the different effect sizes are marked on the y-axis.

Generally, for each PS method, the different effect sizes followed the same pattern for each performance measure of the treatment effect estimate. The bias of the treatment effect estimate reduced as the impact in the treatment allocation model of the variable with measurement error (effect size) increased. The bias was positive for all effect sizes using 3:1 PS matching. It was mainly negative for the other three PS methods, becoming positive for IPTW for ATT and PS stratification with high effect size and high positive measurement error. For all PS methods, the high impact in the treatment allocation model of the variable with measurement error (effect size) had the highest precision and lowest MSE. 3:1 PS matching and IPTW for ATE had the biggest differences between the SD and MSE for the original (very small) and the high effect size. For all PS methods, the Model SE was lower for the high effect size, again 3:1 PS matching and IPTW for ATE had the biggest differences between high and the original effect sizes. For all PS methods, the percentage MSE change had a higher magnitude for the high effect size than for the other effect sizes. These results all relate to the variable with measurement error being a positive contributor in the treatment allocation model (PS model).

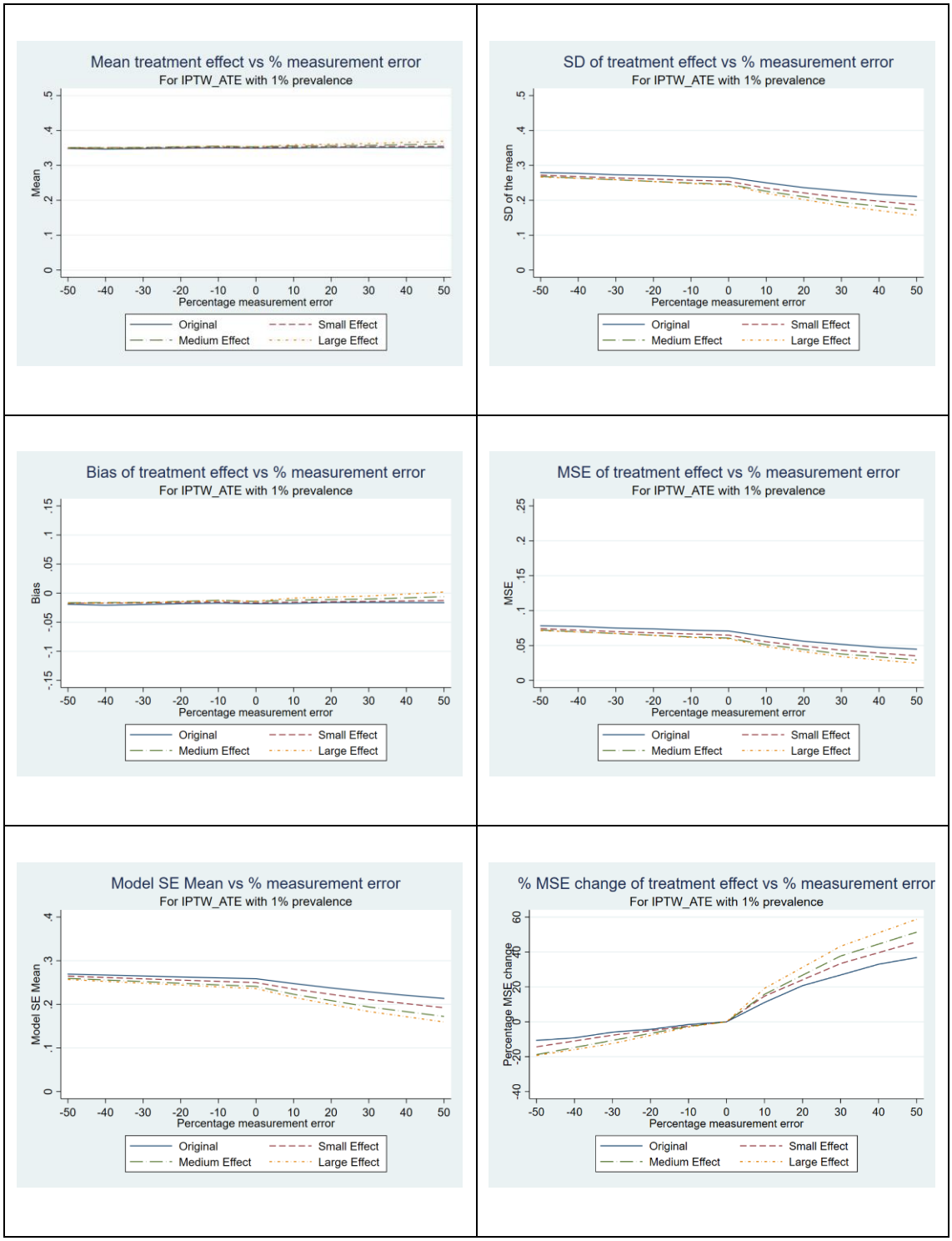


Figure 4: IPTW for ATE, 1% prevalence, displaying the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean for different effect sizes.

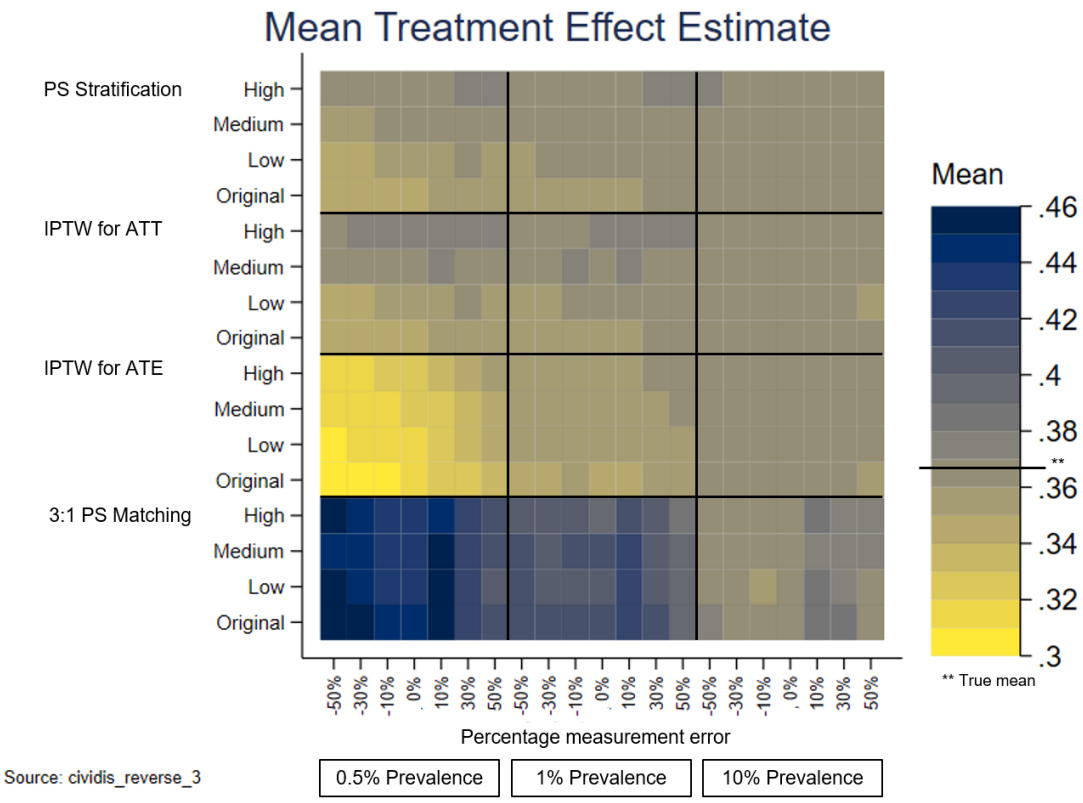


Figure 5: Heat plot for Mean treatment effect estimate. The x-axis shows the outcome prevalence and the introduced measurement error. The y-axis shows the PS method and the 'effect size' used.

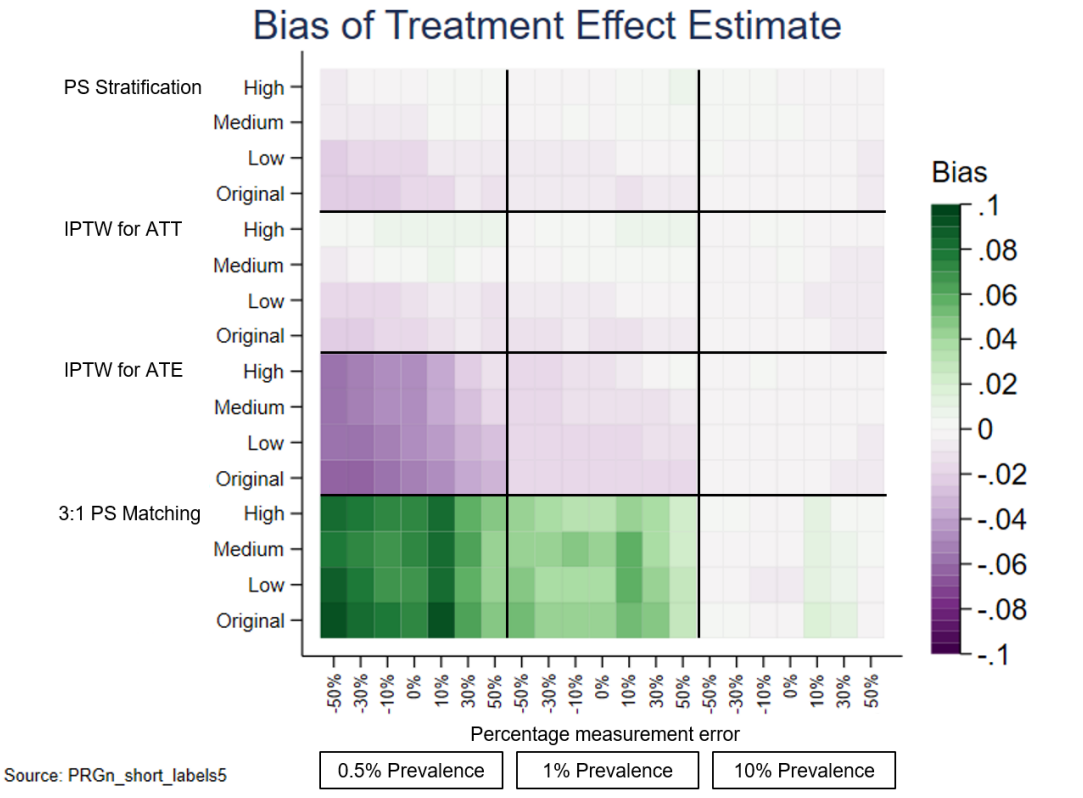


Figure 6: Heat plot for the Bias of the treatment effect estimate. The x-axis shows the outcome prevalence and the introduced measurement error. The y-axis shows the PS method and the 'effect size' used.

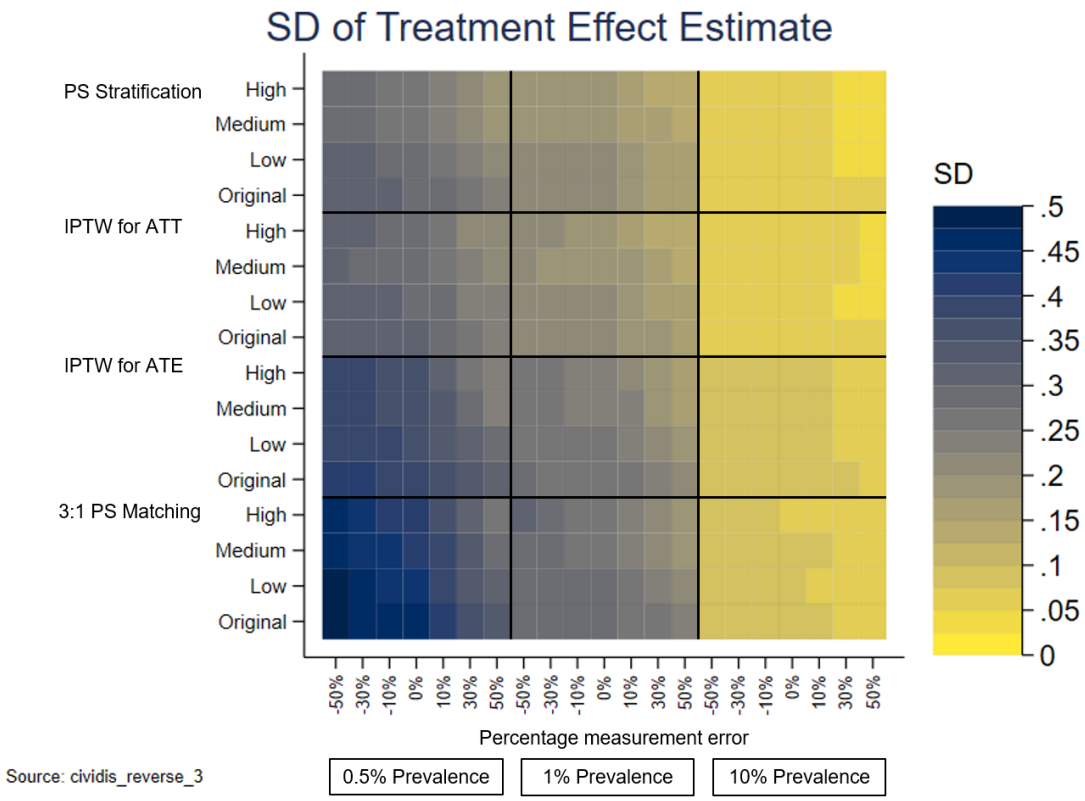


Figure 7: Heat plot of the SD of the treatment effect estimate. The x-axis shows the outcome prevalence and the introduced measurement error. The y-axis shows the PS method and the 'effect size' used.

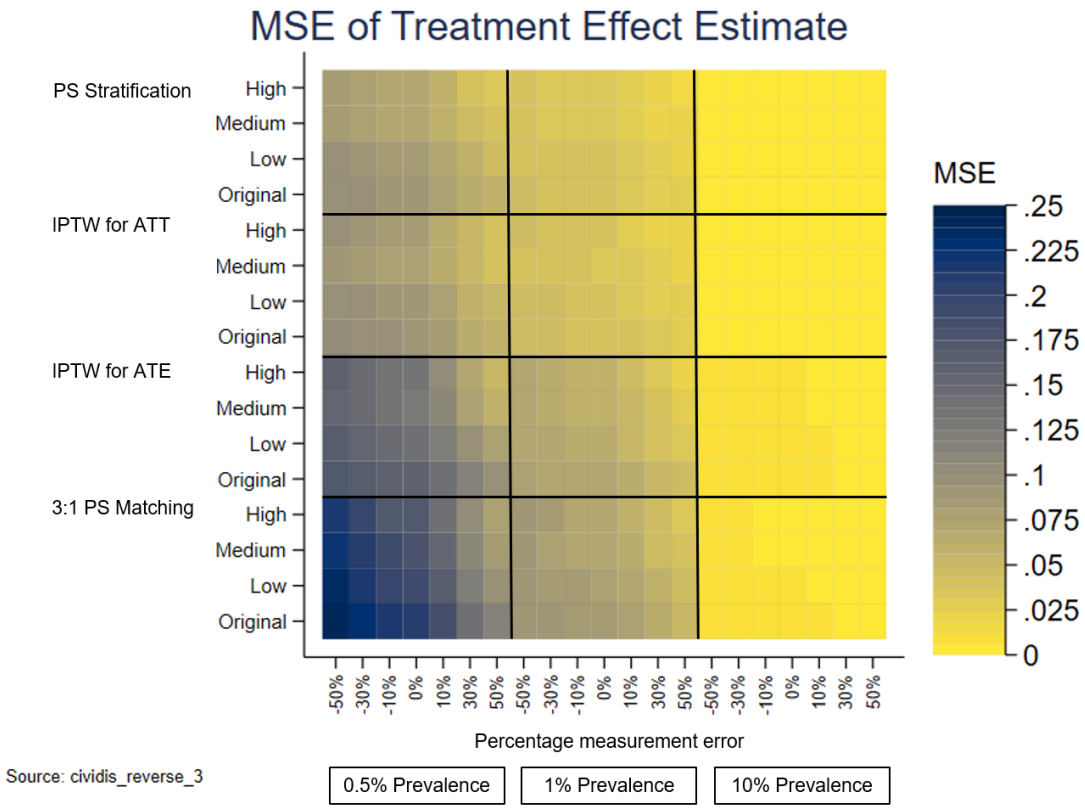


Figure 8: Heat plot of the MSE of the treatment effect estimate. The x-axis shows the outcome prevalence and the introduced measurement error. The y-axis shows the PS method and the 'effect size' used.

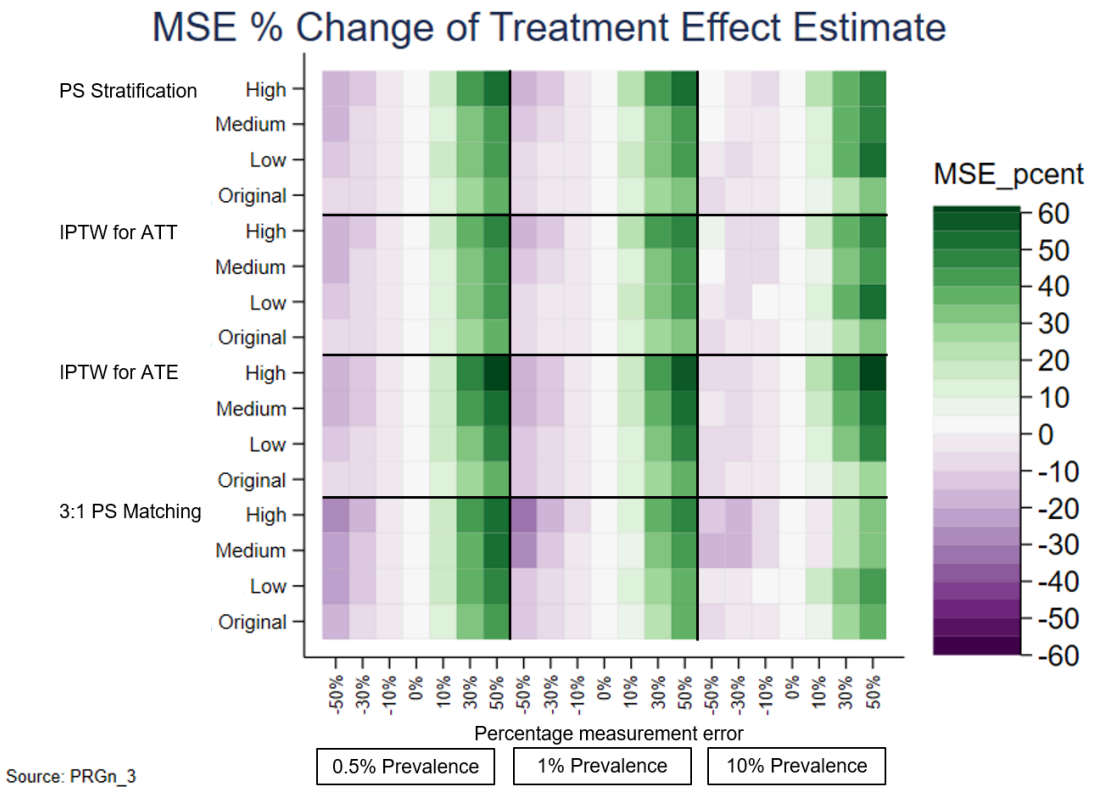


Figure 9: Heat plot of the MSE percentage change of the treatment effect estimate. The x-axis shows the outcome prevalence and the introduced measurement error. The y-axis shows the PS method and the 'effect size' used.

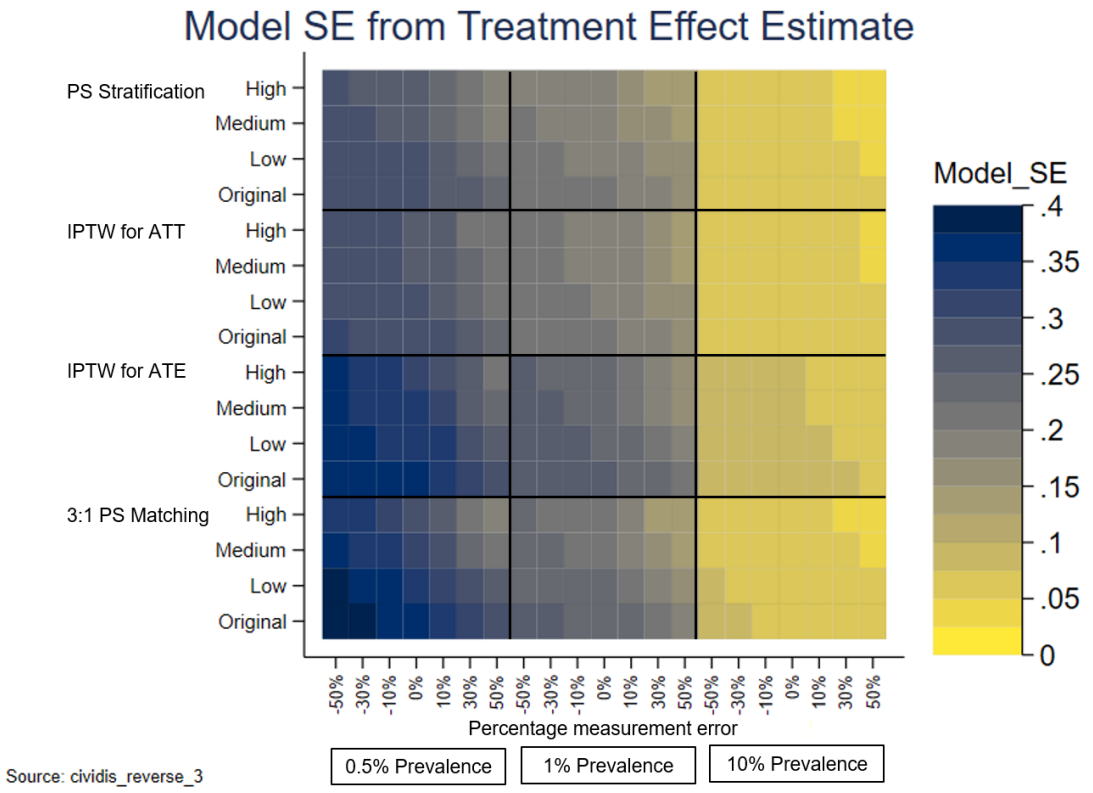


Figure 10: Heat plot of the Model SE of the treatment effect estimate. The x-axis shows the outcome prevalence and the introduced measurement error. The y-axis shows the PS method and the 'effect size' used.

5.5 Results varying measurement error and sparseness of outcome data

Simulations were run to investigate under-recording and over-recording of a variable which affects the treatment allocation (the PS model) and varying the sparseness of the outcome. The under-recording and over-recording was varied by introducing measurement error over the range -50% to +50% in previous stroke (Section 5.3). In the original dataset 1% of cases had a future stroke during the study period, hence an outcome prevalence of 1%. The outcome prevalence was varied by also generating datasets with outcome prevalence to 0.5% and 10% (Section 4.6). The original effect size in the treatment allocation model of the variable with measurement error was used. These results are displayed for IPTW for ATE (Figure 11). The results for the other PS methods are also displayed: IPTW for ATT (Figure E-6), 3:1 PS matching (Figure E-7) and PS stratification (Figure E-8). The heat plots (Figure 5 to Figure 10) present all the results together, the results for this section are shown in the rows marked 'Original' for each PS method. The different prevalences are given in the marked blocks.

The simulations for all PS methods showed some common features. Firstly, the highest prevalence runs, with 10% prevalence, were the closest to the true mean and so had the lowest bias. As the prevalence was reduced, the mean of the treatment effect estimate was further from the true mean and the bias increased. This was an example of 'sparse data bias' (Greenland et al., 2016) which gives an inflated treatment effect estimate (Section 2.5). Secondly, the performance measures from the simulations with different prevalences follow the same pattern over the measurement error range as the simulations with the original prevalence (1%) (Section 5.3). Thirdly, for all performance measures there was less variation over the measurement error range in the higher prevalence (10%) runs. The exceptions were the percentage MSE change for all PS methods and the mean and bias for 3:1 PS matching. For 3:1 PS matching, the 10% prevalence simulations showed the mean and the bias increasing at +10% and +20% measurement error, then decreasing for higher positive measurement error. Fourthly, the SD and MSE were the lowest for the highest prevalence run. Both these performance measures increased as the prevalence of the run was decreased and they followed a similar pattern for all the PS methods used. The last two points are due to the instability of the outcome modelling with fewer outcome events (lower prevalence). This reflects the fact that power (hence the SE) in time-to-event data is calculated using the number of outcomes (Section 4.6). There was no consistent pattern for the percentage MSE change. Fifth, the Model SE decreased gradually as the negative measurement error approached zero, then decreased at a higher rate as the positive measurement error increased. The Model SE was higher for the lower prevalence runs over the measurement error range. This was reflected in these results, as generally the higher

prevalence simulations gave treatment effect estimates which were less biased and had higher precision, they had lower SD and MSE.

IPTW for ATE, IPTW for ATT and PS stratification had negative bias for all prevalences. IPTW for ATT had the smallest difference between the bias at different prevalences and IPTW for ATE had the highest. 3:1 PS matching had positive bias for all the simulations except for the 10% prevalence runs with close to zero measurement error, which had low negative bias. Overall 3:1 PS matching had the largest difference between the bias using the different prevalences. There was much less difference for both the SD and the MSE for all the PS methods using the higher prevalence (10%) than for the lower prevalences (0.5% and 1%). This was also true for the Model SE mean. There was no general pattern for the percentage MSE change. Overall, 3:1 PS matching showed the biggest difference in all performance measures of the treatment effect estimate between the different prevalence simulations and IPTW for ATT and PS stratification showed the least difference.

A small set of simulations was run with a 5% outcome prevalence as it is close to the mid-point between 1% and 10% to investigate at what prevalence variability of the treatment effect estimate was introduced (Appendix G-1). The simulations were run for IPTW for ATE, which was selected because they were representative of IPTW for ATT and PS stratification. The simulation results for 5% prevalence followed a similar pattern to the 10% prevalence results more closely than those from the 1% prevalence simulations. The 5% prevalence simulations showed little variation of all performance measures over the measurement error range and little difference between the runs using the different effect sizes. The exception was a small variation in the SD and MSE for the different effect sizes at positive measurement error. However, this difference was less than that seen in the 1% prevalence runs. All PS methods used in this study showed the same pattern for different prevalences, more variability in the over the measurement error range and higher bias prevalence runs, SD and MSE for the lower prevalence. It was therefore assumed that 5% prevalence runs for the other PS methods would behave in a similar way, that is closer to the 10% prevalence runs than the 1% prevalence runs. In summary, measurement error and the effect size of the variable with measurement error showed a minimal change to the mean, SD, bias and MSE for the 5% and 10% prevalence runs. The highest variability of these performance measures due to sparse data were seen below 5% prevalence.

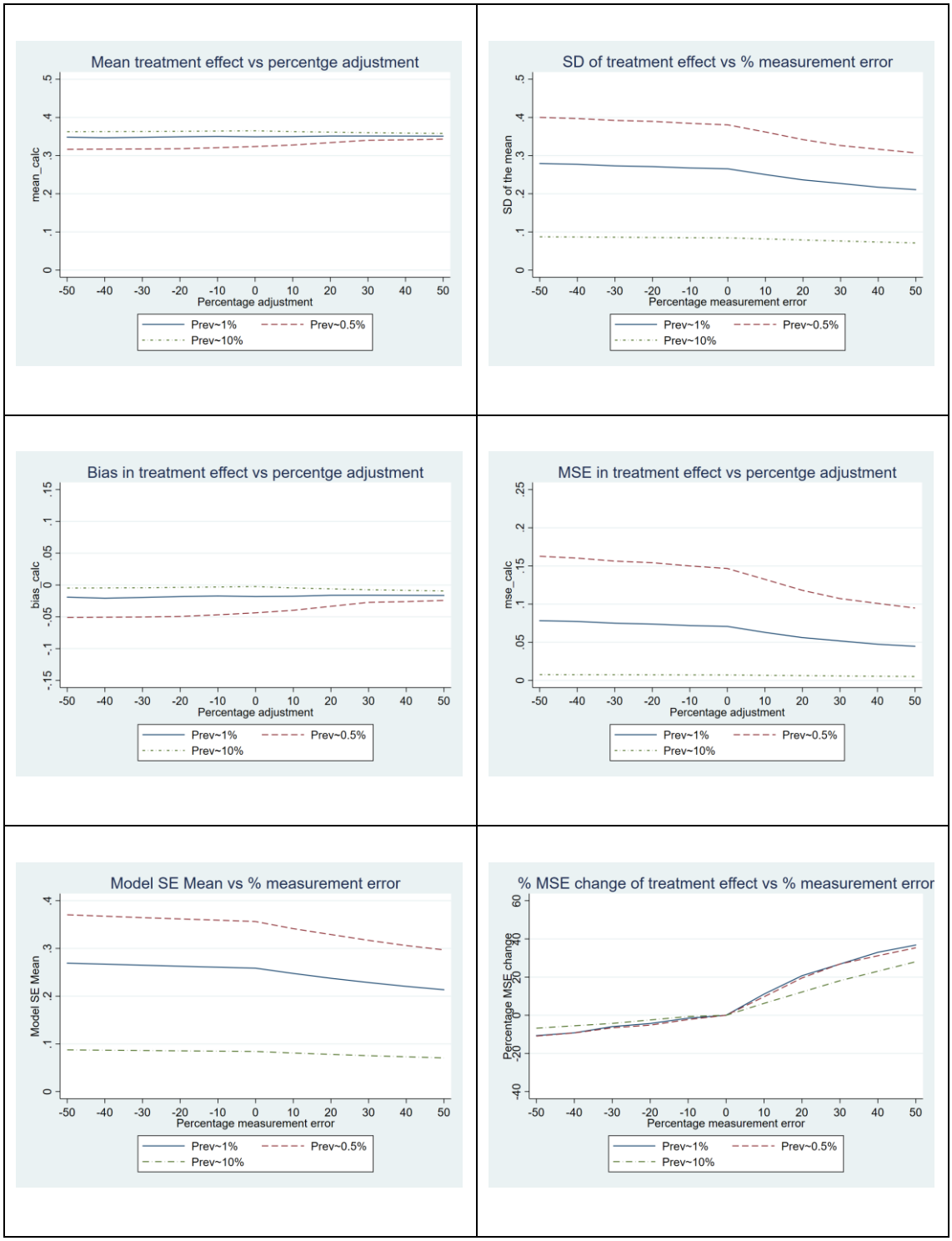


Figure 11: Using IPTW to generate ATE, the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean.

5.6 Results varying measurement error, effect size and sparseness of outcome data

5.6.1 Overview

Simulations were run varying the under-recording and over-recording of a variable which influences treatment allocation (previous stroke), the impact (effect size) this variable has on

treatment allocation (in the PS model) and the sparseness of the outcomes (future stroke). As in the previous section (Section 5.3), the under- or over-recording of previous stroke were represented by introducing measurement error into previous stroke between -50% and +50%. The effect size took the original values (very low), low, medium and high (Section 5.4). The outcome prevalence of the simulated datasets was set to 0.5%, 1% and 10% (Section 5.5). Full details of the simulation parameters are given in Section 4.8. The results for the different effect sizes plotted together for a given prevalence for IPTW for ATE, are shown in Figure 12 (0.5% prevalence), Figure 13 (1% prevalence) and Figure 14 (10% prevalence). Similar graphs for the other PS methods are given in Appendix F: for IPTW for ATE (Figure F-1 to Figure F-3), IPTW for ATT (Figure F-4 to Figure F-6), 3:1 PS matching (Figure F-7 to Figure F-9) and PS stratification (Figure F-10 to Figure F-12). The heat plots, Figure 5 to Figure 10, present all the results from all the simulations together for comparison. All performance measures relate to the log(HR) of the treatment effect estimate.

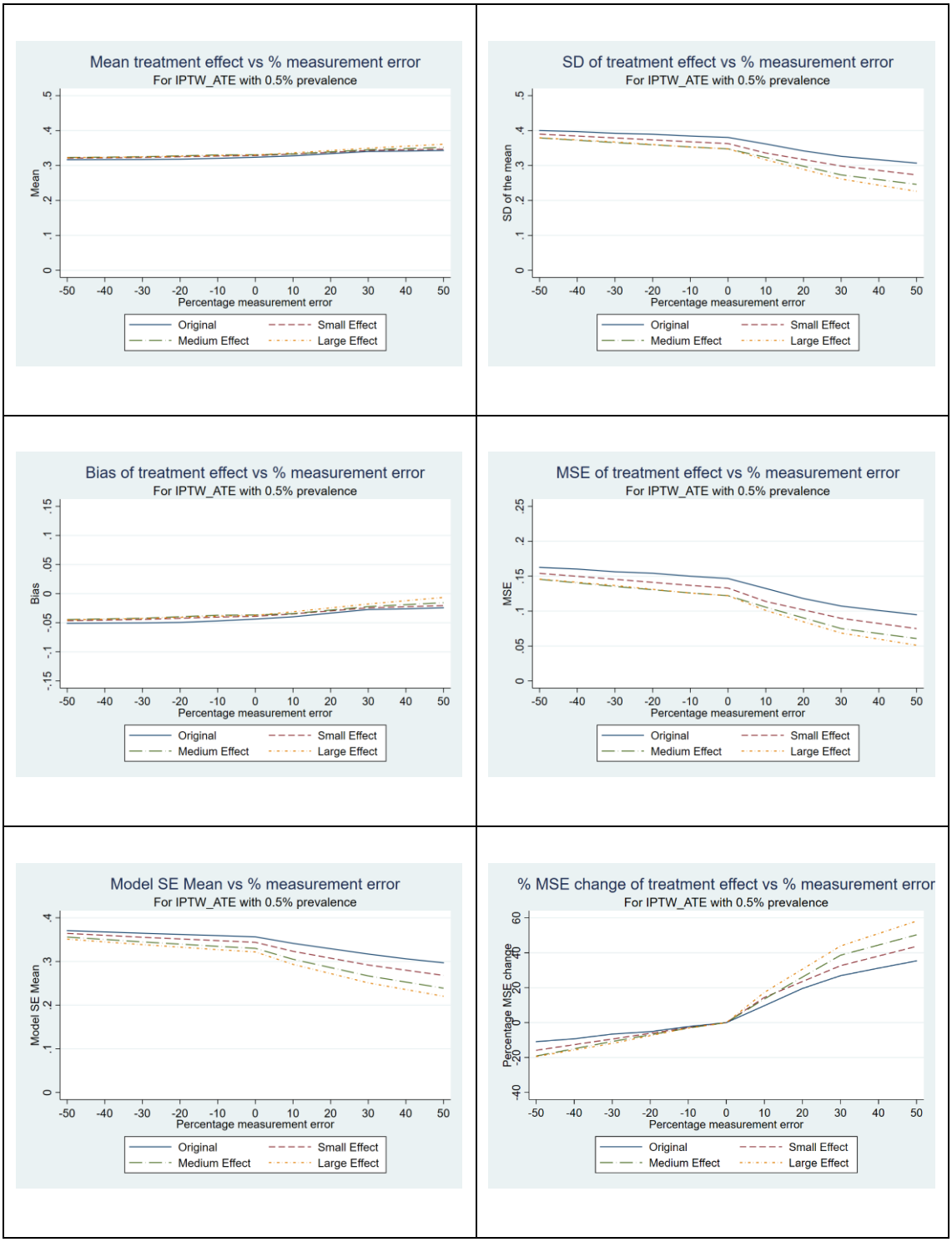


Figure 12: IPTW for ATE, 0.5% prevalence, displaying the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean for different effect sizes.

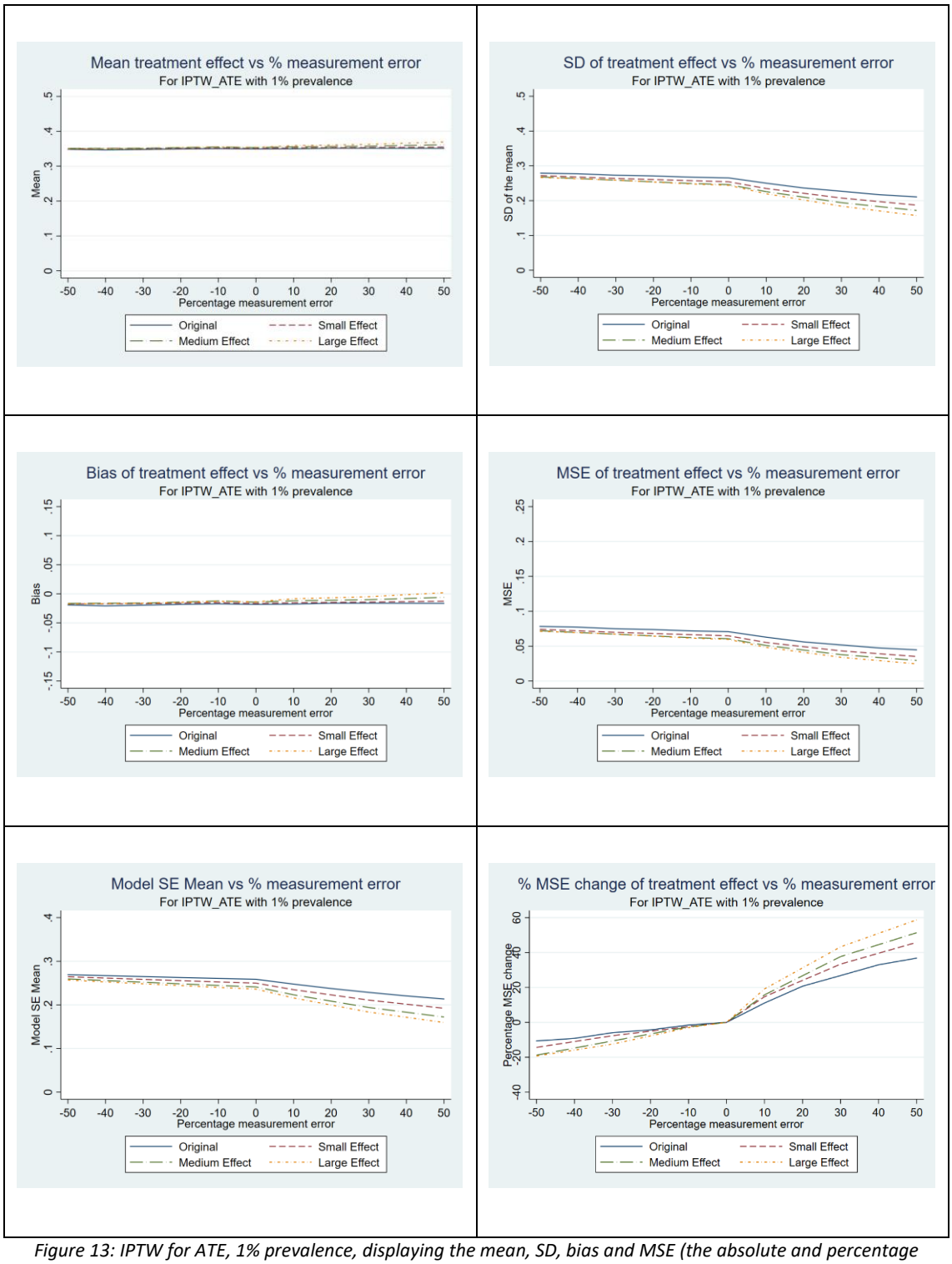


Figure 13: IPTW for ATE, 1% prevalence, displaying the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean for different effect sizes.

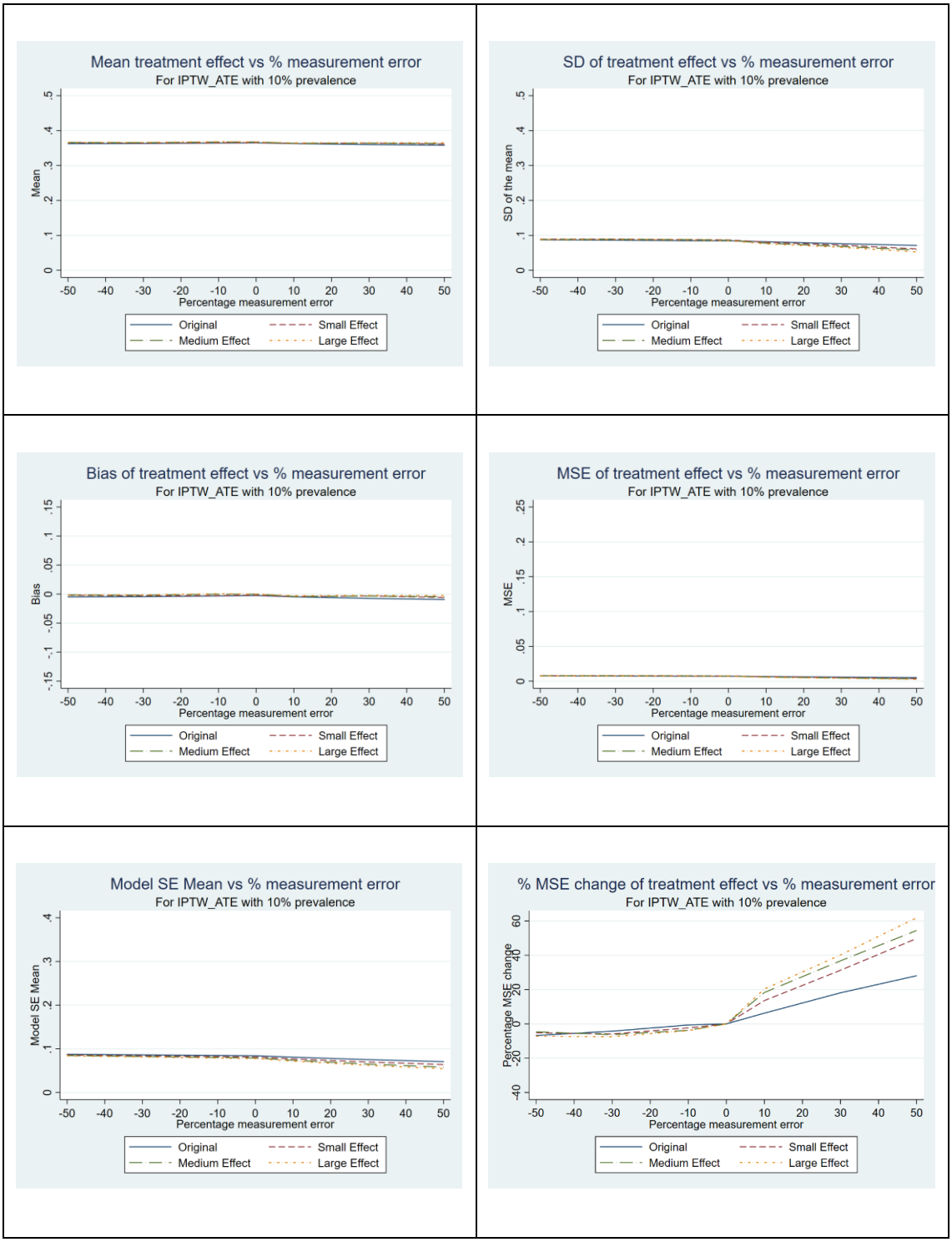


Figure 14: IPTW for ATE, 10% prevalence, displaying the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean for different effect sizes.

5.6.2 All PS methods

The general pattern of each performance measure (the mean, SD, bias, MSE absolute, MSE percentage change and model SE) over the percentage measurement error range did not change as the effect size was modified for each PS method used.

5.6.3 Comparison of PS methods - all outcome prevalence runs

For all PS methods analysed using all prevalences, some common features were seen. The following patterns in the results were seen which had previously been seen when measurement error and effect size were varied (Section 5.4): the different effect sizes followed the same pattern for each performance measure of the treatment effect estimate; the bias of the treatment effect estimate reduced (moved towards zero) as the effect size increased; the high effect size had the highest precision (lowest SD) and lowest MSE, hence; there was less of a pattern in the difference in the MSE percentage change for the different effect sizes; the Model SE was lower for the high effect size. These results also reflected the findings when measurement error and outcome prevalence were varied (Section 5.5) that is, the differences in the performance measures (mean, bias, SD, MSE and model SE) due to the change in effect size were greater for the lower prevalence runs.

5.6.4 Comparison of PS methods - higher outcome prevalence runs

As the higher prevalence runs, with 10% prevalence, showed less difference between the effect sizes for the performance measures used than the lower prevalence runs, using 1% and 0.5% prevalences, these are discussed in separate sub-sections. The graphs for the different combinations of prevalence and effect size are given in Appendix F, for IPTW for ATE (Figure F-3), IPTW for ATT (Figure F-6), 3:1 PS matching (Figure F-9), and PS stratification (Figure F-12). Appendix G Figure G-2 to Figure G-13, have plots with results from all the PS methods for a given effect size and prevalence plotted together. The heat plots (Figure 5 to Figure 10) present the performance measures of treatment effect. Figure 15 (taken from Figure G-3) is shown as an example and plots the treatment effect estimates for all the study PS methods for 10% outcome prevalence and the small effect size.

Generally, as the prevalence increased the difference in each performance measure of the treatment effect estimate (mean, SD, bias and MSE) due to the different effect sizes decreased. The difference in these performance measures for different effect sizes at 10% prevalence was minimal. The maximum differences in the mean between the high effect size (Figure G-5), and original run (Figure G-2), (using high effect – original effect) were 0.007 (1.9% of the true mean) for IPTW for ATE, 0.0047 (1.3% of the true mean) for IPTW for ATT, -0.0105 (-2.9% of the true

mean) for 3:1 PS matching and 0.0059 (1.6% of the true mean) for PS stratification. The highest differences were generally seen at higher positive measurement error. 3:1 PS matching was the only PS method where the bias was positive. However, for all PS methods these differences were small when the data had 10% prevalence.

At 10% prevalence, all values for the bias, for all effect sizes over the measurement error range were small. The maximum bias, that is the value furthest from the null, was negative for IPTW for ATE, IPTW for ATT and PS stratification and positive for 3:1 PS matching. 3:1 PS matching had the highest absolute maximum bias and PS stratification had the lowest. IPTW for ATE and IPTW for ATT had predominately negative bias with some positive values seen for the higher effect sizes. PS stratification had negative bias at lower effect size and positive bias at higher effect size with negative measurement error. 3:1 PS matching had positive bias with positive measurement error and negative bias with negative measurement error. In all methods the treatment effect estimate became less biased (the bias became closer to zero) as the effect size increased.

At 10% prevalence and for all PS methods, the SD was less than 0.1 for all effect sizes and over the measurement error range. This was much lower than for the lower prevalence runs, for the 0.5% prevalence runs where the SD took values between 0.2 and 0.5 using the different effect sizes and over the measurement error range. At 10% prevalence, the MSE was lower than its value in the equivalent runs using lower prevalences for all effect sizes and over the measurement error range.

There was little variation over the measurement error range for all the performance measures of the treatment effect estimate at 10% outcome prevalence. The exception was 3:1 matching where the mean and bias showed a small increase between +10% and +30% measurement error.

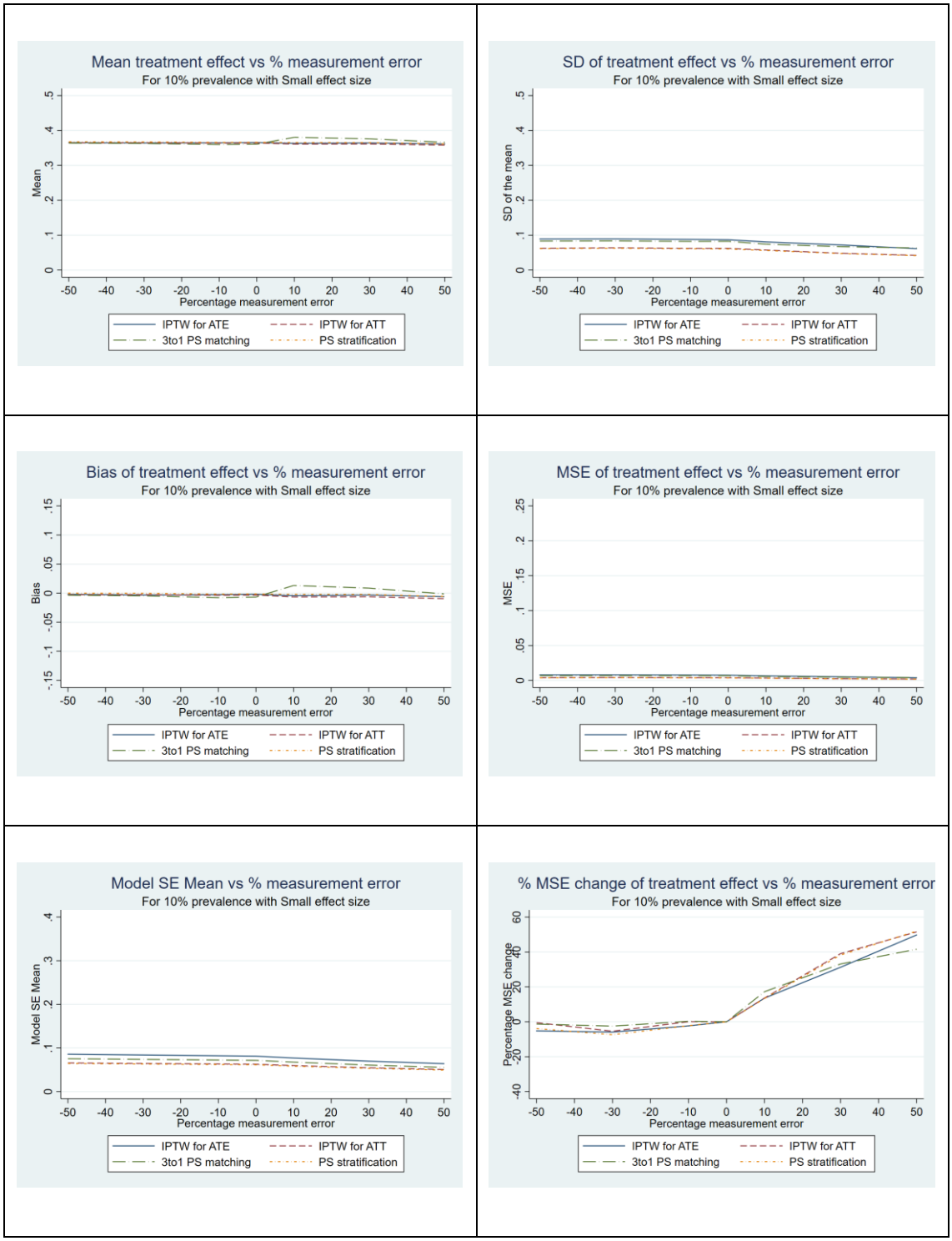


Figure 15: The study PS methods, 10% prevalence, small effect size – the mean, SD, bias and MSE of the estimated treatment effect are displayed as log(HR).

5.6.5 Comparison of PS methods - lower outcome prevalence runs

For all PS methods, for the runs with lower prevalence (0.5% and 1%), the bias, SD and MSE increased and so did the variation over the measurement error range compared to the 10% prevalence runs. However, 3:1 PS matching did show more variation over the measurement

error range than the other PS methods, particularly for the lower prevalence run of 0.5%. IPTW for ATE had predominately negative bias overall effect sizes over the measurement error range. IPTW for ATT had mostly negative bias for the lower effect sizes and mostly positive bias for the higher effect sizes. 3:1 PS matching had positive bias for all effect sizes over the measurement error range. PS stratification had mostly negative bias, but with the higher effect size and positive measurement error the bias was positive. At 1% prevalence IPTW for ATE, IPTW for ATT and PS stratification all had similar values of bias. IPTW for ATE showed the biggest difference between the bias from the original run and the high effect size run, 3:1 PS matching had the smallest difference.

In both the 0.5% and 1% prevalence runs, the SD and MSE behaved in a similar way as the effect size changed. As the effect size increased both the SD and MSE took lower values, with this difference being greater with increasing positive measurement error. The SD and MSE showed a gentle decline in values from -50% measurement error to no measurement error, then a steeper decline in values from no measurement error to +50%. The change in the values was more pronounced for the 0.5% prevalence runs (Figure G-10 to Figure G-13). 3:1 PS matching and IPTW for ATE showed the biggest fall over the measurement error range for both the SD and MSE (Figure G-13).

5.6.6 Summary of results

When the variable with measurement error that contributed to treatment allocation had a higher impact on the treatment allocation, the treatment effect estimate was less biased than when this variable had a lower impact on treatment allocation. This was seen in all PS methods used. These results may not be as expected, but can be explained by looking at the DGM. In the treatment allocation model fitted to the original data, previous stroke, the covariate with measurement error, was a positive contributor. In the original PS model its effect size was very small. When the effect size of previous stroke was increased in the simulations, those cases with previous stroke would have had a higher PS value. A higher PS value increased the probability of the generated treatment being Rivaroxaban. This in turn will increase the chance of those cases having a future stroke, so there were more outcome events in the simulated data, making the outcome modelling more stable. This would generate results with a lower SE, hence a lower SD and seemed to have generated lower bias and lower MSE (which a combination of SD and bias).

There was no clear pattern of the bias always being positive or negative. Only generalisations can therefore be made, such as for IPTW for ATE, IPTW for ATT and PS stratification generally had negative bias for lower effect sizes and positive bias at higher effect sizes. For 3:1 PS

matching, the bias was mostly positive with negative values seen in the 10% prevalence runs for negative measurement error. The variation in the bias due to the change in effect size among the PS methods used was small.

Sections 5.6.4 and 5.6.5 explored the change in the SD and MSE of the treatment effect estimate as the effect size of on the treatment allocation of the variable with measurement error was changed. For the lower prevalence runs, 0.5% and 1%, both the SD and MSE displayed the same patterns with a gentle decline in values from higher under-recording (-50% measurement error) to no measurement error, then a steeper decline in values from no measurement error to higher over-recording (+50% measurement error). The change in the values was more pronounced as the outcome prevalence reduced. When using the original, very low, effect size (of the variable affecting treatment allocation with measurement error), 3:1 PS matching and IPTW for ATE showed the biggest reduction over the measurement error range for both the SD and MSE. In the simulations run in this study, as the effect size on the treatment allocation of the variable with measurement error increased, the treatment effect estimate was less biased. It had a higher precision and a lower MSE, so the variability was lower. This was seen for all PS methods and all outcome prevalences. Both higher effect size and higher positive measurement error in the data reduced the bias, SD and MSE. This could be due to the DGM, above.

The results displayed for 3:1 PS matching were different to the other PS methods used in this study. For 3:1 PS matching, the bias was mostly positive for the lower prevalence runs with negative values seen in the 10% prevalence runs for negative measurement error. For the other PS methods, the bias was generally lower with a mixture of negative and positive values. The SD and MSE for 3:1 PS matching were higher than for the other PS methods used. There was more variation of all performance measures over the measurement error range for 3:1 PS matching. PS matching only uses the cases in the analysis dataset for which a match was found. In this study all Rivaroxaban (the novel treatment) cases were used but only the Warfarin (control) cases which were matched to a Rivaroxaban case and all other Warfarin cases were dropped. This meant the dataset used for the outcome analysis was smaller than that for the other PS methods, where all of the cases were used. This smaller dataset size, and hence fewer outcomes, could account for the higher values of bias, SD and MSE. This is supported by Franklin et al. (2017) for sparse data when a large number of unmatched cases were dropped the variance increased. Franklin et al. (2017) recommended full-matching. However, the reason for the bias being more positive was less clear.

5.7 Recommendations for estimations of ATE

IPTW for ATE and PS Stratification are PS methods which lead to estimates of the ATE. Both methods used the full dataset, with no trimming applied. For IPTW for ATE, weights were calculated and applied to the Cox regression in the outcome analysis. PS stratification offers more choice, both in the number of strata used and the outcome methods. Comparisons using different numbers of strata showed very little difference in the results in this study (Section 3.4.3) and 10 strata were used for these simulations. In the outcome analysis, Cox regression stratified on the PS strata was used to account for the nature of the data (Section 3.5).

When comparing IPTW for ATE and PS stratification the results were similar in all scenarios. When there was under-recording or over-recording of previous stroke, a variable in the treatment allocation model, and the original data characteristics retained (Figure 3) there was little variation in the bias of the treatment effect estimate over the measurement error range for both PS methods. Both IPTW for ATE and PS stratification showed a gentle decline in both SD and MSE as the under-recording reduced. The SD and MSE decreased more rapidly with increasing over-recording. For all values of under-recording and over-recording, PS stratification generated a treatment effect estimate with lower bias than IPTW for ATE. PS stratification had higher precision and lower values for MSE and Model SE than IPTW for ATE over the measurement error range. As the effect size in the treatment allocation model of the variable with measurement error (under- or over-recording) was varied between the original effect size and the high effect size, there was little difference in the amount the bias varied for the different PS methods. IPTW for ATE showed a greater variation in the SD, MSE and Model SE than PS stratification did (Figure 16 to Figure 18). For all effect sizes in the treatment allocation model, PS stratification had higher precision and lower values for SD, MSE and Model SE than IPTW for ATE. As the outcome prevalence was varied, IPTW for ATE showed a greater variation in the bias, SD, MSE and Model SE than PS stratification (Figure 16 to Figure 18). Again, for all effect sizes PS stratification had lower values for the bias, SD, MSE and Model SE than IPTW for ATE.

In all scenarios, the recommendation was to use PS stratification for estimations of the ATE, although the difference in the performance of PS stratification and IPTW for ATE was small. PS stratification gave a treatment effect estimate with lower bias, higher precision and lower MSE than IPTW for ATE. It should be remembered that characteristics of the data may guide the use of one PS method over another, regardless of their performance in the presence of measurement error (Section 5.10). The study dataset had good common support (it had a good overlap of PS values between the two treatment groups) so no trimming of the dataset was needed before the outcome analysis was performed. Although PS stratification had the best

performance in these simulation experiments, the balancing of the data by stratification did seem to be inconclusive (Section 3.4.3) whereas data from IPTW for ATE IPTW balanced well (Section 3.4.2)**.** This means that PS stratification may not have removed the same amount of systematic differences between the two treatment groups, but still has given a treatment effect estimate with comparable bias and precision, to IPTW for ATE. This does not reflect the recommendation of Caliendo and Kopeinig (2008) and Garrido et al. (2014) who recommend applying several PS conditioning methods and using the one which brings the best balance between the treatment groups.

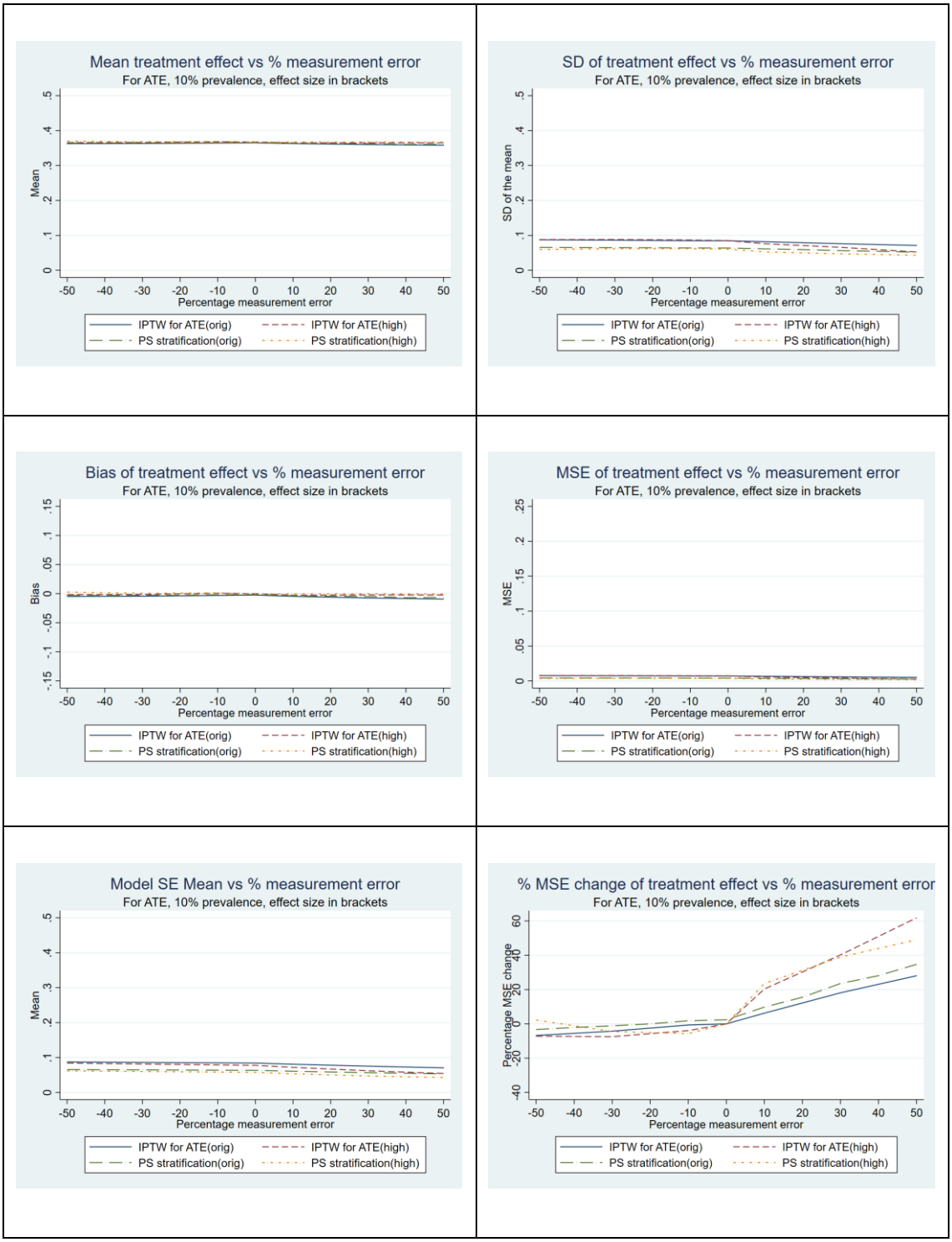


Figure 16: Study PS methods for ATE, 10% prevalence, displaying Original and High Effect Sizes – the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect are displayed as log(HR).

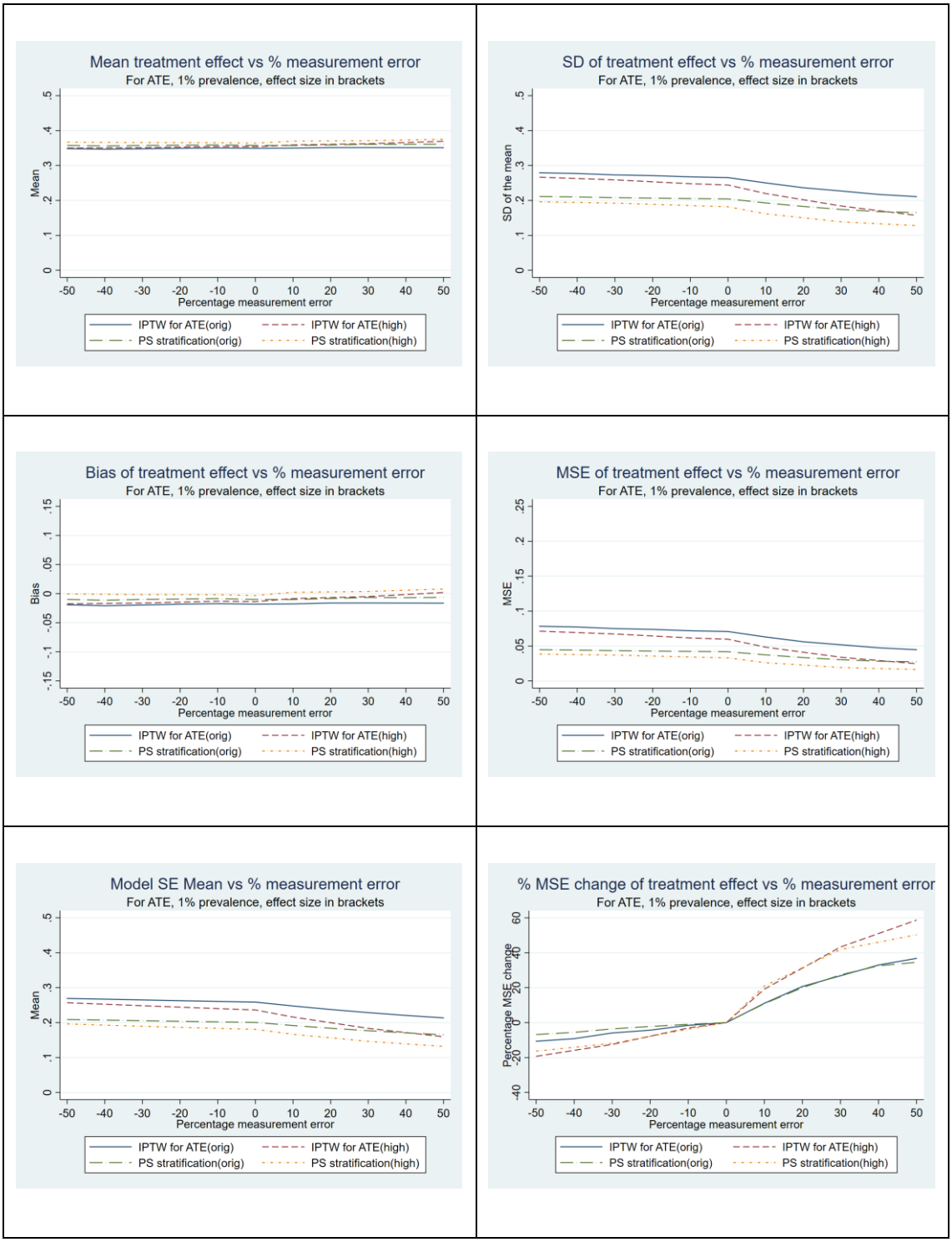


Figure 17: Study PS methods for ATE, 1% prevalence, displaying Original and High Effect Sizes – the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect are displayed as log(HR).

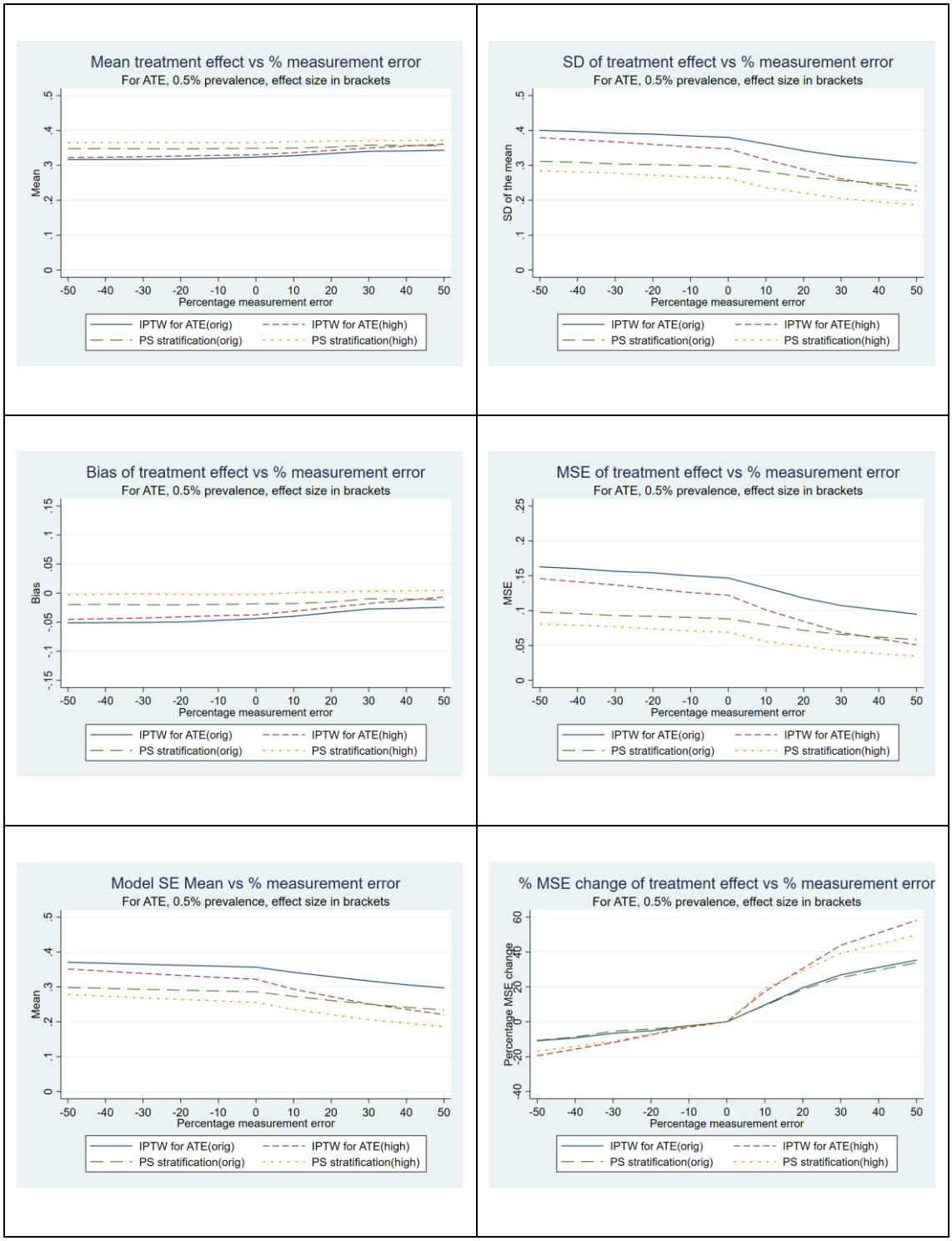


Figure 18: Study PS methods for ATE, 0.5% prevalence, displaying Original and High Effect Sizes – the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect are displayed as log(HR).

5.8 Recommendations for estimations of ATT

The PS methods used in this study which estimated ATT were 3:1 PS matching and IPTW for ATT. For IPTW for ATT, weights were calculated and applied to the Cox regression in the outcome analysis. For 3:1 PS matching, each Rivaroxaban (the NOAC) case was matched to 3 Warfarin

(control) cases on their PS values and the data stratified by these matched groups for the outcome analysis. IPTW for ATT used the full dataset with no trimming performed. 3:1 PS matching used the matched dataset, where non-matched cases were dropped, for the outcome analysis. Hence, 3:1 PS matching used a smaller dataset.

When there was under- or over-recording of previous stroke, a variable in the treatment allocation model, and the original data characteristics retained (Figure 3), 3:1 PS matching showed more variation in the bias over the measurement error range than IPTW for ATT did. The bias in the treatment effect estimate when using 3:1 PS matching was positive, whereas the bias using IPTW for ATT was negative. The absolute value of the bias using 3:1 PS matching was approximately three times higher than when using IPTW for ATT. The SD, MSE and model SE followed the same pattern for both the PS methods, and their values were lower for IPTW for ATT across the measurement error range. When the effect size in the treatment allocation of the variable with measurement error (under- or over-recording) was varied, it had less impact on the bias when using 3:1 PS matching (Figure F-7 to Figure F-9) than when using IPTW for ATT (Figure F-4 to Figure F-6). The SD, MSE and Model SE all had more variation between different effect sizes (in the treatment allocation model) when using 3:1 PS matching than when using IPTW for ATT. For all values of measurement error in the variable which was a contributor to the treatment allocation and its effect size in that model, the treatment effect estimate when using IPTW for ATT was less biased, had a lower precision and lower MSE and Model SE than when using 3:1 PS matching. Varying the sparseness of the outcome (future stroke) between 0.5% and 10% prevalence (Figure 19 to Figure 21) changed the bias, SD, MSE and Model SE by larger amounts when using 3:1 PS matching that it did for IPTW for ATT. Again, in all cases IPTW for ATT performed better than 3:1 PS matching.

Across all the different scenarios, IPTW for ATT was recommended for the estimation of ATT. It gave a treatment effect estimate which was less biased, had a higher precision and had a lower MSE. Situations which give low numbers of outcomes will disadvantage PS matching. Sparse outcome data will be problematic as the outcome analysis is performed on a subset of the sparse main dataset. Even using 3:1 matching, rather 1:1 matching did not give sufficient outcomes in the analysis dataset. Time-to-event data exacerbates this situation as it tends to have sparse outcomes, as cases which are censored will not have an outcome event (even if they have one after the study end date).

As the outcome prevalence decreased, the difference in the treatment effect performance measures between the PS methods increased. In this situation the recommendation for IPTW

for ATT would become even stronger. However, for 3:1 PS matching, as the positive measurement error increased, the treatment effect became less biased, had higher precision and MSE decreased. Introducing positive measurement error also increases the number of outcome events, which would mean a higher percentage increase in outcome events as 3:1 PS matching used the matched dataset which had a smaller number of cases and hence a smaller number of outcomes. For 3:1 PS matching, the high effect size showed more of an improvement (bias closer to zero and lower SD and MSE) over the original effect size, than IPTW for ATT. These may have been accounted for as the DGM used in this study generated higher levels of outcomes for higher positive measurement error and for higher effect size, see above. In data where there was under-recording of a variable in the PS model, 3:1 matching would also be disadvantaged.

It should be remembered that characteristics of the data may guide the use of one PS method over another (Section 5.10), regardless of their performance in the presence of measurement error. This dataset did not require trimming (Section 3.3), but if it had, the dataset used by IPTW for ATT may have been reduced in size and IPTW for ATT may not have performed so well.

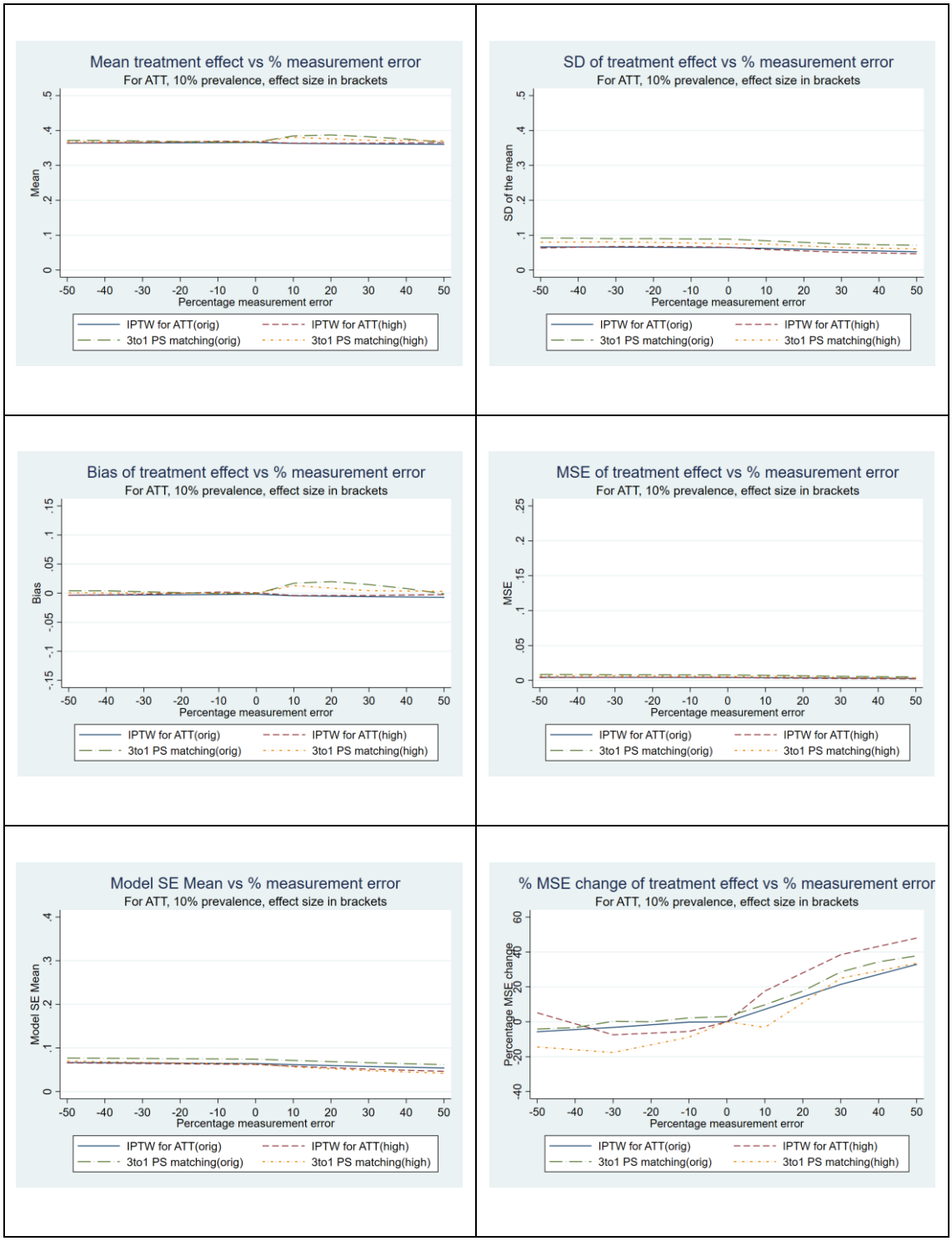


Figure 19: Study PS methods for ATT, 10% prevalence, displaying Original and High Effect Sizes – the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect are displayed as log(HR).

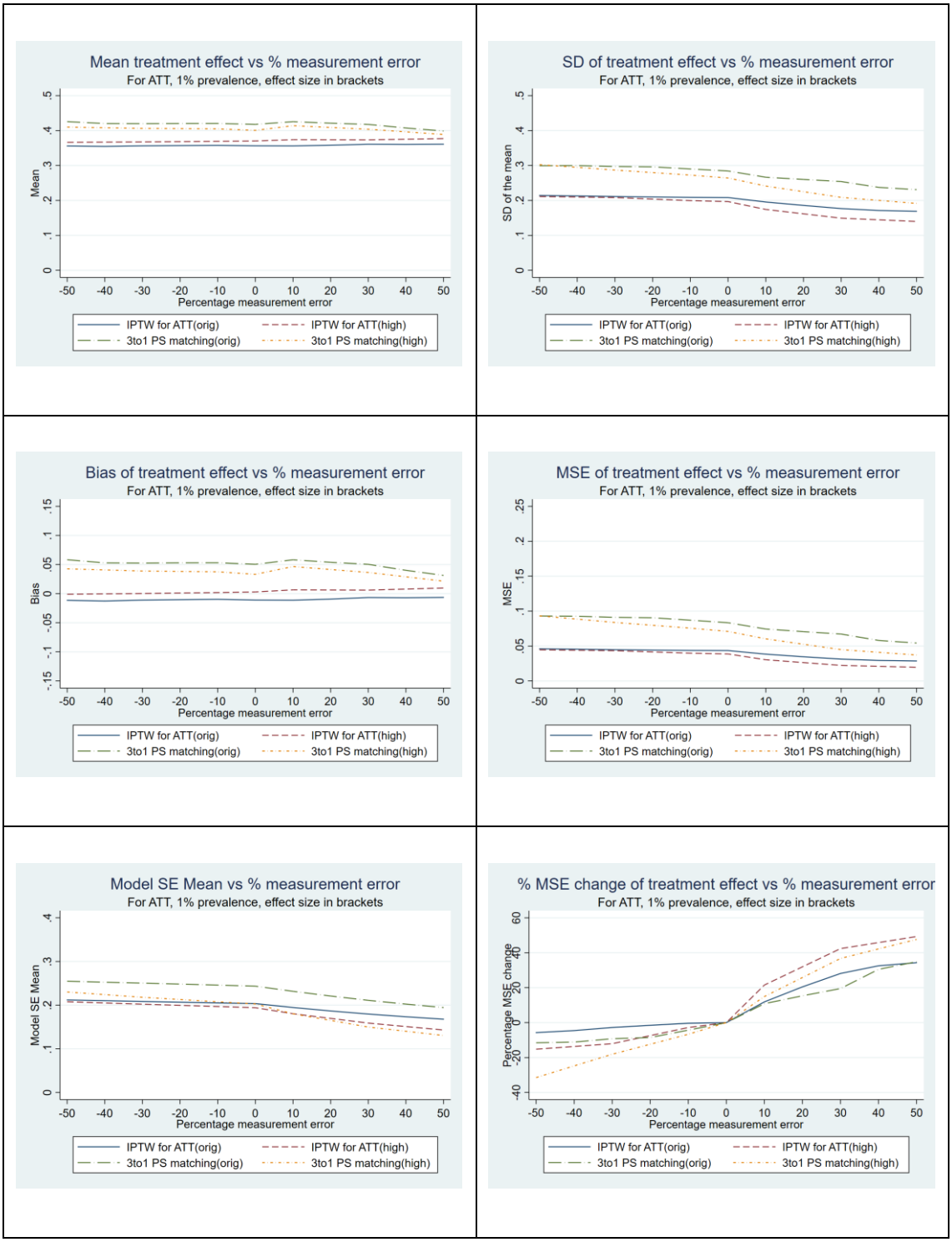


Figure 20: Study PS methods for ATT, 1% prevalence, displaying Original and High Effect Sizes – the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect are displayed as log(HR).

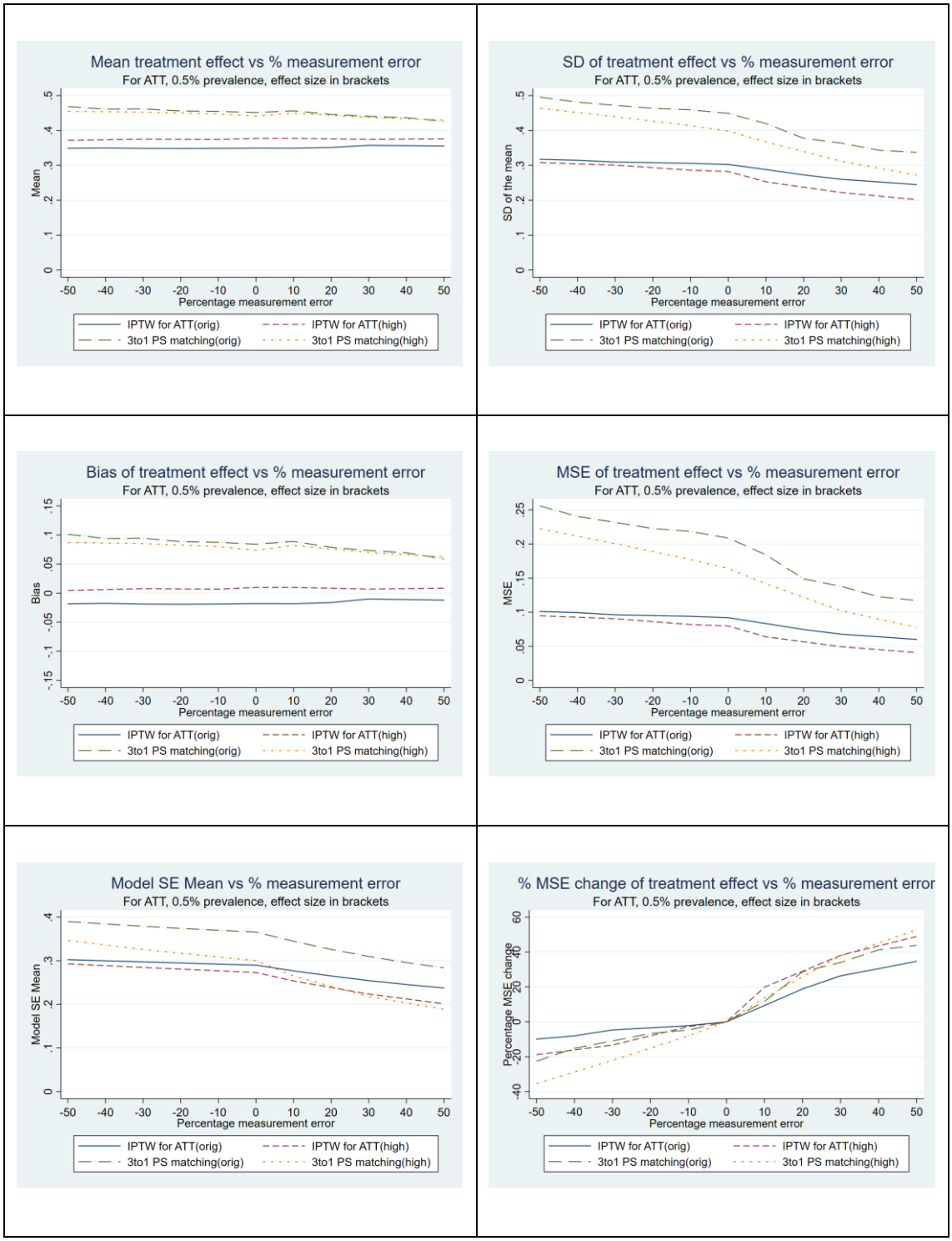


Figure 21: Study PS methods for ATT, 0.5% prevalence, displaying Original and High Effect Sizes – the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect are displayed as log(HR).

5.9 Summary of findings table

Table 13: Summary of findings.

Topic	Mea-sure-ment Error	Effect Size	Prev-alence	Summary
No measure-ment error	No	Original	Original	Comparing all four PS methods using simulations based on the original characteristics of the data – no measurement error, the original effect size of previous stroke in the treatment allocation model and an outcome prevalence of 1%. 3:1 PS matching appeared to perform the least well of the PS methods. It had larger bias and the bias was positive as opposed to negative for the other PS. 3:1 PS matching was retained for use in the later simulations to assess its performance with varying measurement error, effect size and outcome prevalence.
Negative measure-ment error	-ve	Original	Original	Under-recording of a variable in the treatment allocation model was implemented as negative measurement error of previous stroke [-50%, 0%]. Previous stroke had a very low effect size in the treatment allocation model. All PS methods used in this study showed there was little change in the bias of the treatment effect estimate and there was slightly lower precision and a small increase in the MSE, for increasing the magnitude of negative measurement error.
Positive measure-ment error	+ve	Original	Original	Over-recording of a variable in the treatment allocation model was implemented as positive measurement error of previous stroke [0%, +50%]. Previous stroke had a very low effect size in the treatment allocation model. All PS methods used in this study showed there was little change in the bias of the treatment effect estimate. As the size of the over-recording increased, there was higher precision in this estimate.
Effect Size	+ve & -ve	Varied	Original	The impact that the variable with under- or over-recording has on determining the treatment allocation (the effect size) was varied using values of Low, Medium and High for comparison with the Original (very low) effect size. There was still little variation in the mean, and bias, over the measurement error range for all the effect sizes. When the variable with measurement error had

				greater impact on the treatment allocation model, (the PS model), the treatment effect estimate had lower bias and higher precision. It does seem counterintuitive, could be due to the DGM used. This is discussed further in Section 5.6.6.
Prevalence	+ve & -ve	Original	Varied	To investigate the impact of sparse outcome data, the outcome prevalence was varied by generating data with different numbers of the primary outcome of future stroke. The lower prevalence (<5%) data gave treatment effect estimates with a higher bias and lower precision and using the higher prevalence data, with lower bias and higher precision. These results were to be expected, as higher EPV in the outcome model generates more stable models. At lower prevalences, there was more variation in the performance measures of the treatment effect estimate over the measurement error range of a variable in the treatment allocation model.
Effect size and Prevalence	+ve & -ve	Varied	Varied	When the variable with measurement error has greater impact on treatment decision-making, and hence in the PS model (effect size), the treatment effect estimate has lower bias and higher precision, this may be due to the DGM used. The differences in the performance measures of the treatment effect due to different effect sizes are greater when the data has lower outcome prevalence. The treatment effect estimates with the highest bias, lowest precision and highest MSE were obtained with low prevalence outcome data and when the variable with measurement error had a low (or very low) impact in the PS model.
PS Methods for ATE	+ve & -ve	Varied	Varied	PS stratification and IPTW for ATE were the PS methods used which estimate the ATE. Both methods gave treatment effect estimates which followed the patterns described above when measurement error was introduced in a variable in the treatment allocation model, when the effect size of this variable was changed in the treatment allocation model and for different outcome prevalences. There was only a little difference in the bias from both methods, but the bias was slightly closer to 0 for PS stratification. PS stratification had a higher precision and lower MSE, than those for IPTW for ATE over the measurement error range.

				Based on this study's data, the recommendation was to use PS stratification for estimating the ATE. However, the performance of the two PS methods was similar. There were some reservations about the balance produced by PS stratification.
PS Methods for ATT	+ve & -ve	Varied	Varied	3:1 PS Matching and IPTW for ATT were the PS methods used which estimate the ATT. Both methods gave treatment effect estimates which followed the patterns described above when measurement error was introduced in a variable in the treatment allocation model, when the effect size of this variable was changed in the treatment allocation model and for different outcome prevalences. Based on this study's data, the recommendation was to use IPTW for ATT for estimating the ATT, which showed superior performance over 3:1 PS matching. In all scenarios IPTW for ATT had lower bias and higher precision.

5.10 Summary

The first set of simulations were run to compare the PS methods used in this study. They used data with the characteristics of the original study dataset with no introduced measurement error (Section 5.2). This showed that 3:1 PS matching gave the treatment effect estimate with the highest bias and that this bias was positive whereas all the other PS methods gave estimates with negative bias. The precision of the estimate using 3:1 PS matching was lower than that of the other methods. Although 3:1 matching performed poorly, it was retained in the study to assess its performance in the presence of measurement error and sparse outcome data.

One of the main objectives of this study was to investigate the effect of measurement error in the treatment effect estimate. In Section 5.3 measurement error was introduced into the variable for previous stroke, a positive contributor to the treatment allocation model (the PS model). The measurement error included under- or over-recording and the other characteristics of the original data set were retained. When using IPTW for ATE, IPTW for ATT and PS stratification, there was only a small amount of variation in the bias of the treatment effect estimate over the measurement error range. 3:1 PS matching did show more variation of the treatment effect estimate over the measurement error range and still generated a more biased treatment effect estimate than the other PS methods.

For under-recording of previous stroke (negative measurement error), for all the PS methods, the treatment effect estimate had higher precision, lower MSE and model SE as the measurement error moved towards zero. In the cases which misrecorded previous stroke as negative, their PS values was lower and hence their simulated treatment more likely to be the control, Warfarin. There would therefore be fewer cases on the active treatment, Rivaroxaban, in the simulated dataset. The generated outcome data model uses the generated treatment value as a covariate (with a positive coefficient), so if there were fewer generated treatments of Rivaroxaban, there would also be fewer outcome events. The DGM was generating fewer outcome events and as discussed below, lower outcome prevalence gives less stability of the treatment effect estimate. Larger values of under-recording of the variable prone to measurement error, thought to be more likely in primary care data (Section 4.4) produce a treatment effect estimate with a low bias and a lower precision.

For over-recording of previous stroke (positive measurement error), for all the PS methods the precision of the treatment effect estimate increased and the MSE and model SE decreased at a steeper rate as the added measurement error increased. Higher values of over-recording of the variable prone to measurement error, thought to be less likely in primary care data, also do not have a large effect on the bias of the treatment effect, but the treatment effect estimate had a higher precision. This is likely to be caused by a higher number of cases recording that they have had a previous stroke, so their PS values are higher. This means in the simulations there will be more cases with the generated treatment of Rivaroxaban, which in turn means there will be a higher number of cases with an outcome event, so making the outcome modelling more stable (see Section 5.6.6 for the effect of the DGM).

To summarise, when introducing measurement error (under- or over-recording) and using the original data characteristics, there was not a large impact on the bias of the treatment effect estimate. The variable with introduced measurement error (previous stroke) was a weak prognostic variable in treatment allocation (it had a very low coefficient in the PS model), so this could have been expected. The precision of the treatment effect estimate increased as the positive measurement error increased and conversely the precision of the treatment effect estimate decreased as the introduced negative measurement error increased. These results related to the number of outcomes in the generated datasets under these conditions and follow the principle that a higher number of outcome events provides more stable modelling.

So with the assumption that there is generally negative measurement error in primary care data of previous stroke (Section 4.4) in other words it is under-recorded, and the variable with the

measurement error is a weak, positive prognostic variable of treatment allocation, then the treatment effect estimate will only have small bias measurement error although the precision of this estimate will be lower.

The variable with measurement error, previous stroke, had a very small impact in the treatment allocation model, the PS model, when it was fitted to the original study dataset. To investigate the generalisability of the effect of measurement error, the effect size was varied, that is changing the impact of the variable with measurement error in the PS model. Values to represent high, medium, and low were chosen and simulations run using the characteristics of the original data (Section 5.5). Varying the effect size did not change the 'pattern' of any of the performance measures of the treatment effect estimate (mean, bias, SD, MSE and Model SE) over the measurement error range. As the effect size increased, the treatment effect estimate was less biased, had a higher precision and lower MSE and model SE, across the measurement error range. These results may not be as expected, but can be explained by looking at the DGM (Section 5.6.6). When the variable with measurement error had a high effect size in the PS model, for those with a previous stroke, their PS value will be higher than if the effect size were low. A higher PS value increases the probability of the generated treatment being Rivaroxaban. This in turn will generate more outcome events in the simulations, making the outcome modelling more stable. At higher introduced positive measurement error, more outcome events are generated (Section 5.2), so there may be a cumulative effect, hence there being a larger difference between the original and high effect sizes at high positive measurement error and high negative measurement error. Whether this is a genuine effect or is caused by the DGM used in this study is discussed in Section 6.10. The effect size in the treatment allocation model (PS model) of the variable with measurement error (under-recording and over-recording) appeared to have more impact on the on the bias and precision of the treatment effect estimate, than the amount of measurement error in this variable. The implication for future work could be to be aware that when a variable with measurement error in the PS model has a low effect size, the treatment effect estimate will be more biased, have a lower precision, and a higher MSE. Applying these results, if previous stroke (subject to measurement error) had a high impact on the decision to prescribe Rivaroxaban or Warfarin, then the treatment effect estimate would have a lower bias and higher precision, than if previous stroke had less impact on the prescribing or Rivaroxaban or Warfarin. The difference in the treatment effect estimates when using the different impact sizes would be less for under-recording of previous stroke (negative measurement error) than for over-recording of previous stroke (positive measurement error).

Sparse data bias is widely reported in the literature (Greenland et al., 2016). Simulations were run varying the outcome prevalence as well as the measurement error of previous stroke, in the treatment allocation model (Section 5.5). In addition to the 1% outcome prevalence, which is similar to the original study data, outcome prevalences of 0.5% and 10% were also generated. Section 5.6 reports the results from simulations where the effect size of the variable with measurement error in the treatment allocation model (the PS model) was also varied. The effect of changing the outcome prevalence was considerable. Overall, as the outcome prevalence was lowered, the treatment effect estimate had higher bias, lower precision and higher MSE and Model SE. The results were as expected as higher EPV in the outcome model produced more stable models. Lower outcome prevalence 'amplify' any variation in bias, SD and MSE across the measurement error range and between different effect sizes. This was seen for all the PS methods used. When using EHR to estimate the real-world treatment effect, the outcome prevalence should be considered, even when using large datasets. Using PS methods to remove systematic differences between the treatment groups before running the outcome analysis of time-to-event data, the effect of measurement error in a variable in the treatment allocation model and the effect size of that variable in the treatment allocation model have a greater impact on the treatment effect estimate when the outcome prevalence is low, <5%. These findings strongly suggest that more consideration should be given to covariate measurement error in the presence of low outcome prevalence. The use or development of methods to adjust for measurement error in this type of data should consider a range of outcome prevalences and be targeted towards lower outcome prevalence data.

The combination of these parameters which would give the treatment effect estimate with the highest bias and lowest precision are: high under-recording of the variable with measurement error; being a low contributor in the treatment allocation model; low outcome prevalence. The study dataset met these criteria: suspected under-recording of previous stroke of approximately 35%; low effect size of previous stroke in the PS model of 0.1229; and prevalence of future stroke, the primary outcome of 1%. Therefore, it could be expected that the estimate of the treatment effect using the original data would be biased and of a low precision.

The characteristics of the analysis dataset used for the different PS methods varied. In the study dataset, there were 21,259 cases of which 2,911 cases were prescribed Rivaroxaban, the novel treatment. When using PS matching, only matched cases are used for the analysis (S. Guo & Fraser, 2015, p. 132). In the study's matched dataset, all Rivaroxaban cases were used and 3:1 PS matching was used to maximise the number of Warfarin (control) cases. This meant that the analysis dataset had 11,644 cases, but not all the Warfarin cases were unique. In contrast, there

was no trimming of cases with very high or very low PS scores (Sturmer, Rothman, Avorn & Glynn, 2010) meaning that the full dataset was used for the analysis following IPTW. Other datasets may require trimming. Data with different characteristics may give different results for the different PS methods and hence recommend the use of different PS methods for estimating the ATE and ATT to those given below.

Although four PS methods were compared in this study, IPTW for ATE and PS stratification estimated the ATE and IPTW for ATT and 3:1 PS matching estimated the ATT. For the ATE, PS stratification performed better than IPTW for ATE. The difference in the performance measures of the treatment effect, (mean, SD, bias and MSE) and the Model SE over the measurement error range, between the PS methods was seen at the lower prevalence runs (Figure 17 and Figure 18) although the differences were small. For the 10% prevalence runs (Figure 16) the performance of these PS methods was similar. For the ATT, IPTW for ATT performed better than 3:1 PS matching. Again, the difference in the performance of these PS methods increased for the lower prevalence runs (Figure 20 and Figure 21) and the difference between them at the 10% prevalence runs were smaller (Figure 19). The recommendations in terms of the performance measures of treatment effect estimate, mean bias, SD and MSE, were to use PS stratification for ATE and IPTW for ATT for estimating the ATT, meaning these methods gave a less biased treatment effect estimate with a higher precision. However, all four PS methods showed little variation in the bias of the treatment effect estimate over the measurement error range. The precision of this estimate increased as the under-recording reduced (negative measurement error moved towards zero) and the precision increased at a higher rate as the over-recording (positive measurement) increased. For the ATE, PS stratification did perform better than IPTW for ATE, however there were problems achieving balance using the PS in PS stratification (Section 3.4.3). Perhaps this indicates that the PS methods which achieves the best balance is not necessarily the best one to use. Although Franklin et al. (2017) used different study parameters, their data recommended PS stratification (using 10 strata) over IPTW for ATE, but would have recommended 1:1 PS matching over IPTW for ATT (they did not use 3:1 PS matching). It is difficult to make a definitive recommendation for which PS methods to use. For example, if the prevalence of the data is higher, ≥ 5%, the recommendation is less clear. More generally, characteristics of the data may guide the use of one PS method over another, regardless of their performance in the presence of measurement error. In the study dataset, no trimming was required before implementing IPTW, so IPTW used the full dataset for the analysis, but 3:1 PS matching dropped cases. This may not be the case for other data. Franklin et al. (2017) compared over 35 different implementations of PS methods when using sparse data. They did

not recommend 1:1 PS matching, stratification or IPTW when the data had sparse outcomes, non-overlap of the PS distributions and when there were extreme weights. Apart from the sparse outcomes, these features do not appear in the study dataset, so the study recommendations differ to those of (Franklin et al., 2017).

In summary, measurement error of a single variable with a very low prognostic effect in the treatment allocation model under-recording by 50% to over-recording to 50%, had only a small effect on the bias and precision of treatment effect estimate. If the effect size in the treatment allocation model of this variable was increased (from low to high), the bias decreased only by <2% and the precision increased. This was seen in this implementation of the simulations, which may be due to the DGM used (Section 5.6.6), where the PS value was used to create a generated treatment which in turn was used to generate the outcome data. When the outcome prevalence was <5% the differences in bias and precision of the treatment effect estimate, when using different amounts of measurement error and effect sizes, were much higher that for the higher prevalence runs, ≥5%. Hence when the outcome prevalence is lower, <5%, measurement error of this type could be a possible source of error when estimating the treatment effect when using routinely collected data.

When estimating the ATT, the study's recommendation was to use IPTW for ATT rather than 3:1 PS matching. When estimating the ATE using IPTW for ATE and PS Stratification, there was no clear recommendation. It should be noted that characteristics of the dataset can affect the performance of the different PS methods. When there is poor common support (overlap of the PS distributions), PS matching performs better than IPTW. So this study's recommendations relate to a dataset which generated fewer matched pairs than the original dataset, hence 3:1 PS matching was at a disadvantage, and there was no trimming of cases with extreme PS values, which benefitted the IPTW methods.

Chapter 6 DISCUSSION

6.1 Background

Electronic Health Records can be used to conduct observational studies to give a real-world treatment effect estimate. The real-world treatment effect estimate is often different to that generated in a RCT, but is a better reflection of the treatment's performance in the general population. PS methods are widely used to adjust for treatment allocation bias in observational studies. Different PS methods perform better according to the data characteristics. The aim was to select the PS methods which gave the best performing treatment effect estimate, that is those with the lowest bias and highest precision, or the lowest MSE. A comparison of four PS methods (PS matching, IPTW for ATE, IPTW for ATT and PS Stratification) was made in this study. The study's outcome data are in the form of time-to-event data, which is not widely reported on in the literature when a comparison of PS methods is made.

The impact of under- or over-recording of a variable which influenced the treatment allocation was investigated by introducing negative and positive measurement error into the covariate for previous stroke, a variable in the treatment allocation model (the PS model). The impact of the misrecorded binary variable was further examined by changing its effect size in the treatment allocation model, that is changing the coefficient of previous stroke in the treatment allocation model. The changes due to sparse outcome data were demonstrated by generating the simulated datasets with different outcome prevalences.

6.2 No introduced measurement error

Comparing the four PS methods using the study dataset characteristics (with no introduced under-recording or over-recording, the original effect size of this variable in the treatment allocation model and the original prevalence), 3:1 PS matching appeared to perform the least well. 3:1 PS matching had bias considerably higher than the other three PS methods. The bias for 3:1 PS matching was positive but was negative for the other PS methods. The MSE was highest for 3:1 PS matching and IPTW for ATE. All other performance measures were similar for all the PS methods (Section 5.2). The poor performance of 3:1 matching is discussed in Section 6.8.

In practice, several PS methods may be used and compared. Caliendo and Kopeinig (2008) and Garrido et al. (2014) recommend applying several different PS methods and selecting the one which gives the best balance. It is acknowledged that the characteristics of the data will influence the performance of the different PS methods and hence any recommendations.

Although the aims of this study were to compare PS methods in the presence of covariate measurement error and sparse outcome data, comparing the performance of the selected PS methods before these were applied was a valid starting point. Much of the literature which compares PS methods does so in scenarios with no measurement error and with outcome data which are not 'sparse'.

In this study the performance of the treatment effect estimate in the outcome analysis was compared between simulations using the different PS methods. Austin (2009b) compared the amount by which the different PS methods balance the data between the treatment groups. In the current study, balance checks were performed to ensure that sufficient treatment allocation bias had been removed, there was no direct comparison of the amount of balance given by each PS method. There was no consideration of PS model misspecification. If the PS model was incorrectly specified, IPTW may be more sensitive to its effect (Rubin, 2004), although weighting is a 'doubly-robust' property so is more stable to model misspecification (Lunceford & Davidian, 2004).

The study's results were, that if applying PS methods to adjust for treatment allocation bias when conducting an observational study and the outcome is of the format of time-to-event, the suggestion for data thought to have no measurement error and an outcome prevalence of approximately 1% would be not to use 3:1 PS matching. This is different to the recommendation from the literature. When using data with a binary outcome, PS matching and IPTW perform equally well, removing more systematic differences between treatment groups than PS Stratification and Covariate Adjustment (Austin, 2009b, 2011a, 2011b). When the common support is not good, matching performs better (Busso et al., 2014).

The study dataset characteristics may have favoured IPTW as there was good common support of the PS (Section 5.10). The PS method to use would depend on whether the outcome of interest is the ATE or the ATT. Although the literature contains studies which compare the performance of PS methods when the data have a binary outcome, generally there is little reporting of comparison of PS methods using time-to-event data. Although Austin (2013) recommends PS matching and PS weighting (IPTW) for use with time-to-event data.

6.3 Introduced covariate measurement error

Although EHR can be used to estimate the real-world treatment effect of a novel product, the data may be subject to measurement error, particularly as they are collected for clinical purposes and not for research. Measurement error is commonly not taken into account in applied medical research. The focus of this study is in covariate measurement error

demonstrated by introducing under-recording or over-recording into 'previous stroke', a binary variable in the treatment allocation model (the PS model). Previous stroke will be more likely to be under-recorded than over-recorded (Herrett et al., 2013), but the simulations were run to investigate the effect of under-recording or over-recording of a single covariate. It is acknowledged that in EHR the measurement error is likely to be a combination of both under-recording and over-recording. Also, there is likely to be measurement error in several covariates and the primary outcome variable (future stroke). These were not considered in this study.

The simulations from this study showed that higher under-recording of a variable in the treatment allocation model (negative measurement error) gave lower precision (and higher MSE). With higher over-recording of a variable in the treatment allocation model (positive measurement error) there was higher precision (and lower MSE) (Section 5.3). The bias remained fairly constant over the measurement error range, which may be due to the variable with measurement error having a small impact in the treatment allocation model.

Studies which report on the effect of varying measurement error and compare PS methods are rare. However, De Gil et al. (2015), Conover et al. (2021) and Hong et al. (2019) ran simulations which included covariate measurement error. These studies could be used to draw some comparisons, but had differences from the current study: they varied different parameters to the current study; they used 'reliability' of a covariate whereas the current study used only positive measurement error or negative measurement error; the scenarios used were often with parameters varied which were not considered in the current study; if they compared PS methods, they included different PS methods to those in the current study, and they used different performance measures of the treatment effect.

The current study's results are different to the findings in De Gil et al. (2015), where covariate measurement error affected bias but not the RMSE. Conover et al. (2021) reported that in scenarios with only false positive misclassifications (over-recording) produced higher bias than scenarios with only false negative misclassifications (under-recording). This was how the measurement error was implemented in the current study. In the current study there was little variation in the bias and any change was in the opposite direction, with slightly lower bias for over-recording (positive measurement error). Hong et al. (2019) showed that the bias and MSE reduced as the reliability of mismeasured confounders approached one, in terms of the current study this was as measurement error approached zero. These results are different to the current study, as positive measurement error behaved in a different manner than negative measurement error. This could be due to the small changes to the number of outcomes that the

measurement error produced, which is noticeable when the outcome prevalence is low, close to 1%.

In the current study, although there were differences in the performance measures for the PS methods considered, introduced under-recording or over-recording did not change the relative performance of the four PS methods considered (Figure G-2 to Figure G-13). There was a small difference with higher over-recording, which increased the precision for 3:1 PS matching and IPTW for ATE. These were the PS methods with the lowest precision. This may be an artefact of the DGM. Higher over-recording generates data with more outcomes (Section 5.6.6), so any differences between the performance of the PS methods are reduced. De Gil et al. (2015) compared PS methods but did not report that any PS method performed better in the presence of measurement error. Conover et al. (2021) reported 1:1 PS matching had lower bias and higher precision than IPTW using the untrimmed dataset (this was seen for a strong contra indication of a rare exposure and a strong indication of a common exposure), but following trimming generally IPTW performed better than 1:1 PS matching. This agrees with the literature when measurement error is not considered, but was different to the results from the current study. However the current study's dataset had good common support which did not require trimming and so would have favoured IPTW over PS matching. Conover et al. (2021) report that 'modest' amounts of measurement error, in around ≤5% of observations can introduce bias.

The results of this study show that when using EHR to estimate the real-world treatment effect and adjusting for treatment allocation bias with PS methods, there is little change to the bias of the treatment effect estimate when a variable which is used to predict treatment allocation is either under- or over-recorded. This finding may be specific to this study. When using PS methods to correct for treatment allocation bias, the bias of the treatment effect estimate appears to be robust to measurement error/misclassification of a variable with very low impact in the treatment allocation model. When there is more under-recording of this variable, there is lower precision in this estimate. When there is more over-recording of this variable, there is higher precision in this estimate. Under-recording or over-recording in a covariate which affects the treatment allocation does not affect the relative performance of the PS methods used in this study. The recommendations for the PS methods to use are the same as when no covariate measurement error was introduced (Section 6.2).

6.4 Effect size and measurement error

In this study, the effect size of the covariate with under-recording or over-recording (previous stroke) in the treatment allocation model fitted to the data was 'very low', using the

categorisation in Chen et al. (2010). The coefficient of previous stroke in the PS model was 0.12 (Section 4.8), so varying the measurement error only in this variable was unlikely to generate large changes in the treatment effect estimate. To make this work more widely applicable and to investigate the impact when the covariate with under- or over-recording has a higher effect size in the treatment allocation model, the effect size was varied in the simulations so that in addition to the original (very low) value it also took the values equivalent to Low, Medium and High. This meant that the covariate with under- or over-recording had more impact in the treatment allocation model.

In the study dataset, when previous stroke had a higher impact on the treatment allocation, that is a higher effect size in the treatment allocation model (PS Model), the treatment effect estimate had a lower bias and higher precision than if previous stroke had a lower impact on the treatment allocation. 14.1% of cases recorded a previous stroke in the full dataset (Section 4.5). The effect size of the variable with measurement error in the PS model appeared to have more impact on the treatment effect estimate, and hence its bias and precision, than the amount of under- or over-recording in this variable (Section 5.4). It should be noted that the 'true value' of previous stroke (as opposed to the recorded value with the measurement error) also contributed to the outcome. Its true value was included in the CHA2DS2-VASc score in the outcome model and the true value of previous stroke (not the value recorded with measurement error) would have influenced whether or not the participant had a future stroke.

When the variable with under- or over-recording had a higher effect size in the treatment allocation model it produced a treatment effect estimate with a lower bias and higher precision which may seem counterintuitive. The most likely explanation is that this effect is an artefact of the DGM used in the simulations method (Section 5.6.6) and may be masking the actual effect. This is discussed further in Section 6.10. The changes in covariate measurement error and its effect size did not change the relative performance of the different PS methods. The recommendation for the PS methods to use remains unchanged and the same as Section 6.2.

Varying the effect size in the treatment allocation model of the variable with under- or over-recording was not widely reported in the studies which compared PS methods. However, De Gil et al. (2015) did vary both covariate measurement error (covariate reliability) and effect size (strength of relationship between covariates and treatment assignment). Both of these, as individual parameters and their interaction, were among parameters which affected CI coverage. Conover et al. (2021) considered misclassification in a strong indicator and a strong

contra-indicator of treatment assignment, so they changed the sign and not the value of the effect size.

This study adds to the literature where there is a lack of the reporting comparisons of PS methods with covariate under- or over-recording and varied effect size in the treatment allocation model. This may lead to recommendations for the DGM to use in simulation studies which estimate real-world treatment effect estimates from EHR.

6.5 Sparseness of outcome data and measurement error

In the study dataset 14.1% of cases had had a previous stroke (Section 4.5), 13.7% of cases were prescribed Rivaroxaban (the NOAC) (Section B-5) and 1.1% of cases had a future stroke (the primary outcome) (Section 4.6). The form of sparse data considered was rare outcome events, which could lead to a low number of events per variable (EPV). Cox modelling (among other methods) can be subject to bias if there are small numbers in any of the treatment-outcome combinations (Greenland et al., 2016). Sparse data bias produces treatment effect estimates away from the null, so inflated treatment effect estimates are produced. Sparse data bias can occur in large datasets with a low number of outcomes and the study dataset is an example of this. PS conditioning is one of the methods to improve the handling of sparse data (Greenland et al., 2016), PS methods combine the information from several variables into one, making the EPV lower in the outcome model. So, comparison of different PS methods in the presence of sparse outcomes could further inform the best PS method(s) to use. In this study IPTW for ATT and PS Stratification generated the treatment effect with the lowest bias and highest precision at low outcome prevalence, 0.5% (Figure G-10).

This study also varied under- or over-recording of a variable in the treatment allocation model in addition to varying the outcome prevalence. The lower prevalence (<5%) data gave treatment effect estimates with a higher bias and lower precision, whereas using the higher prevalence data gave a treatment effect estimate with lower bias and higher precision. At lower prevalences, there was more variation in the performance measures of the treatment effect estimate over the measurement error range of a variable in the treatment allocation model (Section 5.5). This study's recommendations for estimating the ATE and ATT were compared with the findings from Franklin et al. (2017) and Hajage et al. (2016). These studies varied different parameters in their simulations to the current study. However, the 'averaged' results in Franklin et al. (2017) allowed a comparison of the performance of the PS methods in the current study.

The recommendation from the current study for the estimation of the ATE was to use PS stratification (using 10 strata) over IPTW for ATE. PS stratification had slightly lower bias and slightly higher precision (and lower MSE) than IPTW for ATE over the measurement error range and for all prevalences. The differences were small but increased with lower outcome prevalence. This agreed with Franklin et al. (2017)'s averaged results, where PS stratification (using 10 strata) performed slightly better than IPTW for ATE in terms of absolute bias and absolute MSE. Additionally, Franklin et al. (2017) reported that PS stratification (using 10 strata) had slightly lower bias than IPTW for ATE at 5% prevalence, but no difference was seen at 2% and 1% prevalence. Hajage et al. (2016) only studied one PS method to estimate the ATE so did not make a recommendation.

The recommendation from the current study for the estimation of the ATT was to use IPTW for ATT over 3:1 PS matching. In all cases (over the measurement error range and all prevalences) IPTW for ATT had lower bias and higher precision (so lower MSE). The differences in the performance measures increased as the outcome prevalence decreased. This recommendation was stronger as these differences were larger than those seen in the ATE comparison. Franklin et al. (2017)'s averaged results reported that PS 1:1 matching performed slightly better than IPTW for ATT in terms of absolute bias and absolute MSE. These findings are different to the current study findings, but tend to agree with other findings in the literature. The dataset in the current study had good common support, but rare outcomes, so this may account for the better performance of IPTW for ATT over 3:1 matching. The recommendation from Hajage et al. (2016) was IPTW for ATT over PS matching for estimates of the ATT, for rare exposure/outcomes. This did agree with the results of the current study. Focusing on studies which use time-to-event data (without considering the outcome prevalence), Austin (2013) and Austin (2014b) did not find performance issues with PS matching when there was low exposure, which disagrees with Hajage et al. (2016). Hajage et al. (2016) suggested that Austin (2013) had used a higher sample size in their simulations. This is discussed further in Section 6.8.

This type of sparse data, a large dataset with rare outcomes, is not uncommon (Chao, 1994; Franklin et al., 2017; Paul & Deng, 2000). An example of rare outcomes in observational studies is comparing treatment effect or Serious Adverse Events in drug safety studies (Ross et al., 2015). Even though the outcome may be rare, it can be serious, such as Das et al. (2016) who looked at the challenges of the Neonatal Research Network in trial design under these circumstances.

Data with a higher outcome prevalence, ≥5%, should not be of great concern in EHR (with large data sets) as the treatment effect estimate is likely to have lower bias and higher precision.

However, for lower outcome prevalence (<5%) the treatment effect estimates are likely to have a higher bias and lower precision. At lower prevalences, <5%, there was more variation in the performance measures of the treatment effect estimate over the measurement error range of a variable in the treatment allocation model and its effect size in the treatment allocation model. The recommendations from the current study were to use PS Stratification for estimating the ATE and to use IPTW for ATT to estimate the ATT. These are the same as the recommendations where the outcome prevalence was not varied (Section 5.3).

There is little reporting in the literature of comparison of PS methods with time-to-event outcomes and sparse outcome, although Franklin et al. (2017) did highlight this for future work. There has been no reporting of the comparison of PS methods in the presence of covariate measurement error and sparse outcome. This may provide more evidence to support IPTW over PS matching for estimates of ATT in sparseness of exposure/outcomes.

6.6 Effect size, sparseness of outcome data and measurement error

This study estimates the real-world treatment effect estimate of a novel treatment when the data is in the format of time-to-event, using PS methods to adjust for treatment allocation bias. The findings in this section combine the effects of two other real-world problems in EHR – measurement error and sparse outcome data. Here three parameters were varied: under- or over-recording in a covariate in the treatment allocation model; the effect size of this variable in the treatment allocation model; and the outcome prevalence. A comparison of the PS methods used in this study in these simulations was made.

Simulations which compared the performance of PS methods outcomes with under- or over-recording of a variable in the treatment allocation model, varying its effect size in the treatment allocation model and varying the outcome prevalence have not been reported in the literature before.

To summarise the findings, when the variable with measurement error has greater impact on treatment decision-making, and hence in the PS model (effect size), the treatment effect estimate has lower bias and higher precision, for the study dataset. The differences in the performance measures of the treatment effect due to different effect sizes are greater when the data has lower outcome prevalence. The treatment effect estimates with the highest bias and lowest precision (highest SD and MSE) were obtained with low prevalence outcome data and when the variable with measurement error had a low (or very low) impact in the PS model (Section 5.6).

Recommendations for the PS methods remain the same as those from the data using the original characteristics (Section 6.3): for ATE the recommendation is to use PS stratification; and for ATT the recommendation is to use 'IPTW for ATT'.

6.7 Recommendations for PS methods

This section brings together the recommendations for the PS methods to use in the different scenarios (Sections 6.2 to 6.6). There are four general categories of PS methods: PS Matching, PS Weighting, which include Inverse Probability of Treatment Weighting (IPTW), PS Stratification and Covariate Adjustment on the PS, although there are now many different variations of these basic categories. This study compared the performance of 3:1 PS matching, IPTW for ATT, IPTW for ATE and PS stratification. The matched or weighted nature of the data should be taken into account in the outcome analysis following PS conditioning and there are different options available to do this.

For all scenarios considered in this study the recommendations for the PS methods to use for ATE and ATT are the same. For the ATE, this study's recommendation was to use PS stratification. Its bias was similar to IPTW for ATE but its precision was higher (with a lower MSE). However, the performance of the two PS methods was similar (Section 5.7). This recommendation was not as definitive as that for ATT and there were some reservations about the balance produced by PS stratification. If applied to a dataset with different characteristics, the advice may be different. In the literature, of the PS methods which estimate the ATE, IPTW for ATE is recommended as it provides the best balance (Austin, 2009b). The literature provided no comparison of PS methods when covariate measurement error was introduced. When the outcome data were sparse, this study's recommendation to use IPTW for ATE differed from Franklin et al. (2017), whose averaged results over all the simulations, reported PS stratification (using 10 strata) performed slightly better than IPTW for ATE in terms of absolute bias and absolute MSE. This is supported by the findings of the current study.

For the ATT the recommendation was to use IPTW for ATT, which had a smaller bias and higher precision (lower MSE) than 3:1 PS matching (Section 5.8). If applied to a dataset with different characteristics, the advice may have been different, as the literature advises that PS matching performs better than IPTW when there is not good common support in the data. The study data did have good common support so would have favoured IPTW. The recommendation from this study is different to the literature which reports PS matching as performing better than IPTW for ATT. Many of these studies relate to data with binary outcomes, but Austin (2014b) and Austin (2013) use data with time-to-event outcomes (Section 6.9) and they too recommend PS

matching over IPTW for ATT (Section 6.8). When sparseness of data was considered, the study's recommendation to use IPTW for ATT differed to the averaged results over the simulations in Franklin et al. (2017) but agreed with Hajage et al. (2016). Franklin et al. (2017) had binary outcomes, but varied exposure prevalence and outcome prevalence and they reported 1:1 PS matching performed slightly better than IPTW for ATT in terms of absolute bias and absolute MSE. Hajage et al. (2016) recommended IPTW for ATT over PS matching for estimates of the ATT, for rare exposure/outcomes. This is discussed further in Section 6.8.

These findings are based on under- or over-recording in a variable which is a positive contributor to the treatment allocation, that is, if a participant had had a previous stroke they were more likely to be prescribed Rivaroxaban (the NOAC) than Warfarin (the control). With the study dataset there was limited access to validation data which indicated that there was more likely to be under-recording for the variable for previous stroke. For other datasets there may be no validation data, so no indication of the direction or magnitude of the measurement error. In such cases both under- and over-recording should be investigated and a judgment made about the magnitude of the measurement error.

The characteristics of the dataset used can influence the optimal choice PS method. The study dataset had good common support which favoured IPTW over 3:1 matching. A dataset with different characteristics may have produced different recommendations. Caliendo and Kopeinig (2008) and Garrido et al. (2014) recommend applying several different PS methods and selecting the one which gives the best balance. This would allow the influence of the characteristics of the dataset to be accounted for.

The study dataset data contained patients with AF, and the comparison was of the performance of Rivaroxaban, the novel treatment, over Warfarin, the previous standard treatment, in the prevention of a future stroke. There were contraindications which prevented the prescribing of Rivaroxaban to certain patients. The ATT was a more appropriate measure as an estimate of the treatment effect on those for whom the treatment was appropriate. However, in this study, the dataset was also used to demonstrate the performance of PS methods which estimated the ATE when there was covariate measurement error and low outcome prevalence.

6.8 Poor performance of 3:1 PS matching

This study's recommendation for estimating the ATT is to use IPTW for ATT rather than 3:1 PS matching (Section 6.7). This applied to all scenarios considered. 3:1 PS matching had a more biased estimate of the treatment effect estimate and also had lower precision (and higher MSE). This is different to the literature which reports that PS matching and IPTW generally perform

equally well (Austin, 2009b, 2011a, 2011b). Mostly this referred to binary outcome data, but (Austin, 2013) recommended PS matching and PS weighting (IPTW) for estimating the marginal HR in time-to-event data. Generally, the circumstances will dictate which method to use (Austin, 2014b).

Franklin et al. (2017) and Hajage et al. (2016) ran simulations to compare the performance of PS methods in sparse data conditions and both have methods similar to this study. Franklin et al. (2017)'s simulations had sparse, binary outcomes and Hajage et al. (2016) used data with sparse exposure and time-to-event outcomes. The study's finding for recommending IPTW for ATT over 3:1 PS matching for estimating the ATT disagreed with the general literature, including Franklin et al. (2017). The exception is Hajage et al. (2016) whose recommendation was the same as the current study.

Possible reasons for the poor performance of 3:1 matching were first, the study dataset's characteristics. From the literature, PS matching performs well (in comparison to IPTW for ATT) when there is poor common support. The study dataset had good common support, so would favour IPTW for ATT to some extent. Second, the implementation of 3:1 PS matching was found to be more difficult than the other PS methods. Following PS matching, the outcome analysis had to account for the matched nature of the data. The 'strata' option was used to group the matched cases together (one Rivaroxaban and three Warfarin cases) so that they were regarded as having the same baseline hazard. Other options were explored but were not suitable to run in the simulations. Generally, implementing Cox regression for many matched pairs/groups was more problematic than applying weights, generated by IPTW, to Cox regression.

There are existing Stata user-written programs which were recommended by the literature and other researchers in the field. However, these offered limited functionality. Some programs could perform a full analysis using PS matching but only for a binary outcome, not for time-to-event data like the study's data. The Stata user written program -psmatch2- was chosen to perform the PS matching because it recorded the Warfarin (control treatment) IDs that the Rivaroxaban (novel treatment) cases were matched to. This was needed in the outcome analysis when matched groups were formed. However -psmatch2- did not offer a many:1 matching with 'no replacement'. A 'no replacement' option could have been useful for a sensitivity analysis in view of the poor performance of 3:1 PS matching. 3:1 PS matching had been chosen over 1:1 matching to boost the number of outcomes in a sparse dataset. As all matching was done using replacement, there could have been an over reliance on certain Warfarin cases, which were matched many times to different Rivaroxaban cases. There were 18,348 Warfarin (control) cases

compared to 2,911 Rivaroxaban cases in the original dataset (Section B-5) which meant that there were sufficient Warfarin cases to have matched the Rivaroxaban cases at a ratio of 3:1 had 'no replacement' been used. A 'no replacement' option may have changed the cases matched and hence the performance of the treatment effect estimate, although this in turn would have changed the underlying distribution of the matched sample.

When using 3:1 PS matching, 'self-matching' was seen (Section 4.3). The simulated datasets were generated using plasmode simulations by making random draws, using replacement, from the original dataset. All simulated datasets were likely to contain multiple copies of some of the original cases. Introducing measurement error generated an amended value for previous stroke, giving an amended PS value which then gave a generated treatment. It was possible for multiple copies of the same original case to generate cases which have generated treatment of both Rivaroxaban and Warfarin. These Warfarin cases became an obvious match for Rivaroxaban cases generated. This was known as 'self-matching'. It is not clear to what extent self-matching affected the treatment effect estimate. The number of self-matched cases were identified and could have been excluded from the outcome analysis dataset. This would have reduced the size of the analysis dataset further, which would be likely to increase the bias and reduce the precision. To have excluded matching to the same original case and to match to the nearest different case would have meant amending/re-writing part of -psmatch2-. No studies were found in the literature which reported the topic of self-matching, but it may be common in bootstrapped samples. Franklin et al. (2017) found that in full matching in areas of poor overlap a single treated case could be matched to as many as 500 controls and suggested limiting this number, however, this could give higher variance if the number of outcome events reduces.

Although implementing 3:1 matching with no replacement or excluding self-matched cases may not have reduced the bias generated in the treatment effect estimate, they would have allowed sensitivity analysis to be performed. This may have informed different implementations of 3:1 matching which would have generated a less biased treatment effect estimate. Future work to investigate the poor performance of 3:1 PS matching could include further development of a many:1 matching program in Stata to allow for a 'no replacement' option and to include an option to exclude self-matching. Further exploration of the options to use when implementing Cox regression to account for the matched nature of the data could be made.

6.9 Use of time-to-event data

This study recommended PS stratification over IPTW for ATE for estimates of the ATE (Section 5.7) and IPTW for ATT over 3:1 PS matching for estimates of the ATT (Section 5.8). The literature

which compared the performance of different PS methods mainly covered data with a binary outcome. There was little research published on situations where the outcome is of the form time-to-event. For estimates of the ATE, this study's results do not agree with the finding from Austin (2014b), Austin (2013), but when there is sparseness of outcome data this study's results agree with Franklin et al. (2017). For estimates of the ATT, the study's results agree with and Hajage et al. (2016), but not with that from Austin (2014b), Austin (2013) and Franklin et al. (2017) so further work may consolidate the recommendations. Franklin et al. (2017) also highlighted data with time-to-event outcomes as an area for future work. Guidance of the implementation of PS methods, particularly PS matching, for time-to-event data should be considered for future work.

6.10 Changing effect size and the Data Generating Mechanism

To further explore the effect of measurement error, the effect size of the variable with under- or over-recording in the treatment allocation model was also varied. The covariate with measurement error, previous stroke, had a coefficient in the PS fitted to the original data, which was classified as 'very low' based on the categorisation in Chen et al. (2010). The simulations using this value showed little variation in the bias of the treatment effect estimate (Section 5.3). To investigate the impact of a higher effect size, the effect size of the covariate with measurement error was increased in the simulations. When the variable with measurement error had a higher effect size in the PS model, the treatment effect estimate had lower bias and higher precision. This seemed to be counterintuitive, the measurement error in the PS model had a higher impact in the treatment allocation model and gave a treatment effect estimate with lower bias and higher precision. This can be explained by looking at the DGM, as described in Section 5.6.6. In the treatment allocation model fitted to the original data, previous stroke, the covariate with measurement error, was a positive contributor and its effect size was very small. When the effect size of previous stroke was increased in the simulations, those cases with previous stroke would have had a higher PS value. A higher PS value increased the probability of the generated treatment being Rivaroxaban. This in turn will increase the chance of those cases having a future stroke, so there were more outcome events in the simulated data, making the outcome modelling more stable. This would generate results with a lower SE, hence a lower SD and seemed to have generated lower bias and lower MSE (which is a combination of SD and bias).

There would be datasets extracted from EHR where the covariate with under- or over-recording was a negative contributor in the treatment allocation model and/or receiving the novel treatment would have a negative coefficient in the outcome model. For example, consider the

case where the covariate with measurement error was a negative contributor to the treatment allocation and receiving the novel treatment was still a positive contributor in the outcome model. With higher under-recording of the variable with measurement error, there would be more cases with a higher PS, hence more cases with a generated novel treatment. This in turn would generate more cases with an outcome event, leading to a treatment effect with a lower bias and higher precision. This is in contrast to the study models, where higher under-recording of the variable with measurement error led to a treatment effect estimate with a higher bias and lower precision.

It is still not clear if this change (a reduced bias and increased precision of the treatment effect estimate) seen when the effect size of the covariate with measurement error in the treatment allocation is increased is a genuine effect, just an artefact of the DGM, or if the DGM is masking the underlying effect.

Future work could be to extend the simulations so that the coefficient of the covariate with measurement error is a negative contributor to the treatment allocation model and/or the treatment is also a negative contributor to the outcome. Varying these negative values would investigate the DGM used.

6.11 Summary

The real-world treatment effect can be estimated from EHR in the form of an observational study, so the treatment allocation was not randomised and PS methods were used to adjust for treatment allocation bias. The aims of this study were to investigate the effect of under- or over-recording of a dichotomous covariate in the treatment allocation model and sparse data when estimating the real-world treatment effect in this way, and to compare the performance of different PS methods in these scenarios. Simulations were run based on a dataset which compared the performance of Rivaroxaban (the novel treatment) with Warfarin (the control treatment) on the prevention of future stroke for patients with AF. The outcomes were in the format of time-to-event. Parameters were varied for measurement error in a covariate in the treatment allocation model (previous stroke), the effect size of this covariate in the PS model and the outcome prevalence (future stroke). Using simulations, the performance measures of the treatment effect estimate generated when using four PS methods were collected to allow a comparison of the PS methods to be made.

Introducing measurement error replicating under-recording or over-recording of a single dichotomous covariate in the treatment allocation model, covered a range of 50% under-recording to 50% over-recording. Changes in the magnitude of the measurement error had little

effect on the bias of the treatment effect estimate. When there was under-recording of that variable, as the size of the under-recording increased, there was lower precision in this estimate. When there was over-recording of this variable, as the size of the over-recording increased, there was higher precision in this estimate. Under- or over-recording in a single covariate in the treatment allocation model did not have a large impact on the treatment effect estimate when its effect size (in the treatment allocation model) was very small.

As the effect size of the covariate with measurement error in the treatment allocation model was very small, to make the work more useful, its effect size was varied. When the effect size was high, the treatment effect estimate had lower bias and higher precision. In simulations based on this dataset, the variable with measurement error was a positive contributor in the PS model, meaning they would be more likely to be allocated Rivaroxaban (novel treatment) and the novel treatment was also a positive contributor in the outcome model, giving a slightly increased risk of future stroke. Both these together increased the probability of that patient having an outcome event. Simulations generated with higher outcome prevalence produced a treatment effect with a lower bias and higher precision. It is not clear if this result is due to the impact of changing the effect size or is an artefact of the DGM which is generating more outcomes, hence the outcome modelling is more stable. A different simulation experiment may help to resolve this. Generally there is little in the literature about comparing PS methods in the presence of covariate measurement error and the impact of changing its effect size in the PS model.

Although the original study dataset may have appeared large, with 21,259 cases, the outcome prevalence was 1%. Varying this prevalence showed that for the lower outcome prevalence (≤1%) the treatment effect estimates are likely to have a higher bias and lower precision. At this lower prevalence there was more variation in the performance measures of the treatment effect estimate over the measurement error range of a variable in the PS model and its effect size in the PS model. Data with a higher outcome prevalence, >1%, had a treatment effect estimate with lower bias and higher precision and variation due to measurement error and its effect size in the PS model was much less. These findings were as expected.

The recommendation for the PS methods to use remained the same in all simulation scenarios (no introduced measurement error, introduced measurement error, introduced measurement error and varied effect size, introduced measurement error and varied outcome prevalence and introduced measurement error, varied effect size and varied outcome prevalence). For estimations of the ATE, PS stratification performed slightly better than IPTW for ATE. The

difference in performance was small and often the literature recommends IPTW for ATE. For estimation of the ATT, IPTW for ATT performed considerably better than 3:1 PS matching. Often the literature recommends PS matching for estimating the ATT, but Hajage et al. (2016) also found IPTW for ATT performed better than PS matching using time-to-event data and sparseness in the data. In their study that was sparseness of exposure and in the current study it was sparseness in the outcome.

This study has shown that PS methods recommended in the literature may not perform well for individual datasets. Although Caliendo and Kopeinig (2008) and Garrido et al. (2014) recommend applying several PS methods and selecting the one which produces the best balance for the outcome analysis, this study showed that PS methods which give relatively poor balance can produce a treatment effect estimate with lower bias and higher precision. So, it is unclear if selecting a PS method based on how well it balances the treatment groups is valid. Studies which use EHR to conduct observational studies should consider the impact of sparse outcome data, not just on the bias and precision of the treatment effect estimate, but also on the effect that covariate measurement error and its impact on the treatment allocation will have. For low outcome prevalence data, the effect of measurement error and its impact on the treatment allocation on the treatment effect estimate will be greater. For data with outcomes in the form of time-to-event, not all outcome events will be recorded as some will be censored, making the data more sparse, which compounds these problems.

When the data are of the form time-to-event, guidance on the implementation of PS methods, particularly PS matching, and comparison of PS methods should be considered for future work. Further exploration of the options to account for the matched nature following PS matching of the data when Cox regression is performed could be undertaken.

The simulations could be extended to include different scenarios: by defining measurement error in terms of reliability, so a combination of under- and over-recording (positive and negative measurement error) would be included; by using the current measurement error definition but amending the number of over-recorded cases to match the number of under-recorded cases (there were more cases with no previous stroke than those with previous stroke); changing the variable with measurement error to be a negative contributor in the treatment allocation model and/or changing the treatment to be a negative contributor in the outcome model; introducing measurement error into a continuous variable in the treatment allocation model. A possible enhancement could be to include both variables which affect the treatment allocation and

variables which affect the outcome in the PS model. This would then generate an estimate of the marginal treatment effect.

This study successfully ran simulations using four PS methods and varying under- or over-recording in a variable in the treatment allocation model, varying this variable's effect size in the treatment allocation model and varying the outcome prevalence. The performance measures of the treatment effect estimate were collected and displayed in tabular and graphical format.

This study used only four PS methods (each with one outcome option) which is a limitation, as there are many variations of these basic categories of PS methods. Also, it is currently not clear if the changes to the treatment effect estimate are due to changes in the effect size or the DGM used. The parameters varied in the simulations only took a limited range of values (see above for suggested extensions of the simulations). The results using additional PS methods or expanded parameter ranges could be compared with those of this study.

Another limitation of this work is that in the estimation of treatment effect variance no account was taken of the uncertainty in estimating the PS. This means the CIs of the treatment effect produced were too wide. For PS stratification, using the analogous marginal variance method, which accounts for the uncertainty from estimating the PS model from the data, can reduce the variance by up to 12% depending on the data characteristics, compared to the commonly used variance estimation. For IPTW the variance reduction can be 18% using the analogous marginal variance method compared to the commonly used variance estimation. Such changes are particularly noticeable for larger samples, n>1000 (Williamson et al., 2012b).

This study's findings contribute to the body of knowledge, particularly when using PS methods and varying covariate measurement error, the covariate's effect size in the treatment allocation model, the outcome prevalence, and combinations of these. These simulations were applied to time-to-event data, which are generally not widely reported on. This study did make recommendations for the PS methods to use, but the recommendations have been guided by the characteristics of the study dataset, which may mean that they are limited to datasets with similar characteristics.

REFERENCES

Ali, M. S., Groenwold, R. H. H., Pestman, W. R., Belitser, S. V., Roes, K. C. B., Hoes, A. W., . . . Klungel, O. H. (2014). Propensity score balance measures in pharmacoepidemiology: a simulation study. *Pharmacoepidemiology and drug safety, 23*(8), 802. doi:10.1002/pds.3574

Altman, D. G. (1991). *Practical statistics for medical research* London: Chapman and Hall/CRC.

Austin, P. C. (2008a). Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score. *Pharmacoepidemiology and Drug Safety, 17*(12), 1218-1225. doi:10.1002/pds.1674

Austin, P. C. (2008b). Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. *Pharmacoepidemiology and Drug Safety, 17*(12), 1202-1217. doi:10.1002/pds.1673

Austin, P. C. (2009a). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine, 28*(25), 3083-3107. doi:10.1002/sim.3697

Austin, P. C. (2009b). The Relative Ability of Different Propensity Score Methods to Balance Measured Covariates Between Treated and Untreated Subjects in Observational Studies. *Medical Decision Making, 29*(6), 661-677. doi:10.1177/0272989x09341755

Austin, P. C. (2011a). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research, 46*(3), 399-424. doi:10.1080/00273171.2011.568786

Austin, P. C. (2011b). A Tutorial and Case Study in Propensity Score Analysis: An Application to Estimating the Effect of In-Hospital Smoking Cessation Counseling on Mortality. *Multivariate Behavioral Research, 46*(1), 119-151. doi:10.1080/00273171.2011.540480

Austin, P. C. (2013). The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in Medicine, 32*(16), 2837-2849. doi:10.1002/sim.5705

Austin, P. C. (2014a). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine, 33*(6), 1057-1069. doi:10.1002/sim.6004

Austin, P. C. (2014b). The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Statistics in Medicine, 33*(7), 1242-1258. doi:10.1002/sim.5984

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine, 26*(4), 734-753. doi:10.1002/sim.2580

Austin, P. C., & Mamdani, M. M. (2006). A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine, 25*(12), 2084-2106. doi:10.1002/sim.2328

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine, 34*(28), 3661-3679. doi:10.1002/sim.6607

Banerjee, A., Benedetto, V., Gichuru, P., Burnell, J., Antoniou, S., Schilling, R. J., . . . Sutton, C. J. (2020). Adherence and persistence to direct oral anticoagulants in atrial fibrillation: a population-based study. *Heart, 106*(2), 119-+. doi:10.1136/heartjnl-2019-315307

Black, D. A., Berger, M. C., & Scott, F. A. (2000). Bounding parameter estimates with nonclassical measurement error. *Journal of the American Statistical Association, 95*(451), 739-748. doi:10.2307/2669454

Blackwell, M., Honaker, J., & King, G. (2017). A Unified Approach to Measurement Error and Missing Data: Overview and Applications. *Sociological Methods & Research, 46*(3), 303-341. doi:10.1177/0049124115585360

Braun, D., Gorfine, M., Parmigiani, G., Arvold, N. D., Dominici, F., & Zigler, C. (2017). Propensity scores with misclassified treatment assignment: a likelihood-based adjustment. *Biostatistics, 18*(4), 695-710. doi:10.1093/biostatistics/kxx014

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*(12), 1149-1156. doi:10.1093/aje/kwj149

Burnell, J. (2015). *Sensitivity Analysis for HES vs HES and THIN*. University of Central Lancashire.

Busso, M., DiNardo, J., & McCrary, J. (2014). NEW EVIDENCE ON THE FINITE SAMPLE PROPERTIES OF PROPENSITY SCORE REWEIGHTING AND MATCHING ESTIMATORS. *Review of Economics and Statistics, 96*(5), 885-897. doi:10.1162/REST_a_00431

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31-72. doi:10.1111/j.1467-6419.2007.00527.x

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective* (2nd ed.): Chapman and Hall/CRC.

Carroll, R. J., & Stefanski, L. A. (1990). APPROXIMATE QUASI-LIKELIHOOD ESTIMATION IN MODELS WITH SURROGATE PREDICTORS. *Journal of the American Statistical Association, 85*(411), 652-663. doi:10.2307/2290000

Cham, H. N., & West, S. G. (2016). Propensity Score Analysis With Missing Data. *Psychological Methods, 21*(3), 427-445. doi:10.1037/met0000076

Chang, Y. C., Perng, C. H., & Shiau, C. Y. (2000). On estimating stratified PH model with single covariate from sparse data with application to brain metastases study. *Biometrical Journal, 42*(5), 569. doi:10.1002/1521-4036(200009)42:5<569::Aid-bimj569>3.0.Co;2-#

Chao, A. (1994). Population-Size Estimation for Sparse Data - Reply. *Biometrics, 50*(1), 303.

Chen, H. N., Cohen, P., & Chen, S. (2010). How Big is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies. *Communications in Statistics-Simulation and Computation, 39*(4), 860-864. doi:10.1080/03610911003650383

Choi, B. C., & Pak, A. W. (2005). Peer reviewed: a catalog of biases in questionnaires. *Preventing chronic disease, 2*(1).

Choi, L., Carroll, R. J., Beck, C., Mosley, J. D., Roden, D. M., Denny, J. C., & Van Driest, S. L. (2018). Evaluating statistical approaches to leverage large clinical datasets for uncovering therapeutic and adverse medication effects. *Bioinformatics, 34*(17), 2988-2996. doi:10.1093/bioinformatics/bty306

Chu, R., Walter, S. D., Guyatt, G., Devereaux, P. J., Walsh, M., Thorlund, K., & Thabane, L. (2012). Assessment and Implication of Prognostic Imbalance in Randomized Controlled Trials with a Binary Outcome - A Simulation Study. *Plos One, 7*(5). doi:10.1371/journal.pone.0036677

Cleves, M., Gould, W. W., & Marchenko, Y. V. (2016). *An introduction to survival analysis using Stata* (Revised 3rd ed.): Stata press.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*(2), 295–313. doi:https://doi.org/10.2307/2528036

Cochran, W. G., & Chambers, S. P. (1965). The Planning of Observational Studies of Human Populations. *Journal of the Royal Statistical Society: Series A, 128*, 234-255. Retrieved from http://dx.doi.org/10.2307/2344179

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.): Hillsdale, N.J: L. Erlbaum Associates.

Conover, M. M., Rothman, K. J., Sturmer, T., Ellis, A. R., Poole, C., & Funk, M. J. (2021). Propensity score trimming mitigates bias due to covariate measurement error in inverse probability of treatment weighted analyses: A plasmode simulation. *Statistics in Medicine, 40*(9), 2101-2112. doi:10.1002/sim.8887

Cook, J. R., & Stefanski, L. A. (1994). SIMULATION-EXTRAPOLATION ESTIMATION IN PARAMETRIC MEASUREMENT ERROR MODELS. *Journal of the American Statistical Association, 89*(428), 1314-1328. doi:10.2307/2290994

Das, A., Tyson, J., Pedroza, C., Schmidt, B., Gantz, M., Wallace, D., . . . Higgins, R. D. (2016). Methodological issues in the design and analyses of neonatal research studies: Experience of the NICHD Neonatal Research Network. *Seminars in perinatology, 40*(6), 374. doi:10.1053/j.semperi.2016.05.005

De Gil, P. R., Bellara, A. P., Lanehart, R. E., Lee, R. S., Kim, E. S., & Kromrey, J. D. (2015). How Do Propensity Score Methods Measure Up in the Presence of Measurement Error? A Monte Carlo Study. *Multivariate Behavioral Research, 50*(5), 520-532. doi:10.1080/00273171.2015.1022643

DiPrete, T. A., & Gangl, M. (2004). Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology, 2004, Vol 34, 34*, 271-310. doi:10.1111/j.0081-1750.2004.00154.x

Dong, H., & Millimet, D. L. (2020). Propensity Score Weighting with Mismeasured Covariates: An Application to Two Financial Literacy Interventions. *Journal of Risk and Financial Management, 13*(11). doi:10.3390/jrfm13110290

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*: Chapman and Hall.

Fabiani, M., Bella, A., Rota, M. C., Clagnan, E., Gallo, T., D'Amato, M., . . . Rizzo, C. (2015). A/H1N1 pandemic influenza vaccination: A retrospective evaluation of adverse maternal, fetal and neonatal outcomes in a cohort of pregnant women in Italy. *Vaccine, 33*(19), 2240. doi:10.1016/j.vaccine.2015.03.041

Firth, D. (1995). BIAS REDUCTION OF MAXIMUM-LIKELIHOOD-ESTIMATES (VOL 80, PG 27, 1993). *Biometrika, 82*(3), 667-667. Retrieved from <Go to ISI>://WOS:A1995RY02600021

Franklin, J. M., Eddings, W., Austin, P. C., Stuart, E. A., & Schneeweiss, S. (2017). Comparing the performance of propensity score methods in healthcare database studies with rare outcomes. *Statistics in medicine, 36*(12), 1946. doi:10.1002/sim.7250

Franklin, J. M., Schneeweiss, S., Polinski, J. M., & Rassen, J. A. (2014). Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Computational Statistics & Data Analysis, 72*, 219-226. doi:10.1016/j.csda.2013.10.018

Garrido, M. M., Kelley, A. S., Paris, J., Roza, K., Meier, D. E., Morrison, R. S., & Aldridge, M. D. (2014). Methods for Constructing and Assessing Propensity Scores. *Health Services Research, 49*(5), 1701-1720. doi:10.1111/1475-6773.12182

Gastwirth, J. L., Krieger, A. M., & Rosenbaum, P. R. (1998). Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika, 85*(4), 907-920. doi:10.1093/biomet/85.4.907

Gayat, E., Resche-Rigon, M., Mary, J. Y., & Porcher, R. (2012). Propensity score applied to survival data analysis through proportional hazards models: a Monte Carlo study. *Pharmaceutical Statistics, 11*(3), 222-229. doi:10.1002/pst.537

Gravel, C. A., & Platt, R. (2018). Weighted estimation for confounded binary outcomes subject to misclassification. *Statistics in Medicine, 37*(3), 425-436. doi:10.1002/sim.7522

Greenland, S., Mansournia, M. A., & Altman, D. G. (2016). Sparse data bias: a problem hiding in plain sight. *Bmj-British Medical Journal, 353*, i1981. doi:10.1136/bmj.i1981

Greenland, S., Schwartzbaum, J. A., & Finkle, W. D. (2000). Problems due to small samples and sparse data in conditional logistic regression analysis. *American Journal of Epidemiology, 151*(5), 531. Retrieved from https://watermark.silverchair.com/151-5-531.pdf?token=AQECAHi208BE49Ooan9kkhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAAmw

wggJoBgkqhkiG9w0BBwagggJZMIICVQIBADCCAk4GCSqGSIb3DQEHATAeBglghkgBZQMEAS4wEQQMrIImWWQMZgLUfoQpAgEQgIICH30hol-tGnDTCpBK5EWD31qbGWuejz9tBD2jLQEN4k63xlxHOPYrF9pluv5WjObnnLaIkwpcFlan2zuuA5o0gfWZWhp-15W0p_eHiCCTWGrdtJqD4Xx2gxdwFV34LWb-lGUb7gomFk9fOJs0ZNqro9zNBQ7f49TQAqj5LcuDhs28sEkaw05-a_09WWHKF8PYc8UXZXmubAuq99pm1o07QR5lNOzFZVx7YnPC_SdiS7M5_1j1uuhQHluNot83gXDBq4R7Sq9po5TxlqTI6TH8BpkaHcxhKqmRxgnJCHnQPZfuBouaArmjfHqpO0s8G34L9p8UU86hayqI9sphKyrnlyb689zac0Vl3D_HuC8unsiEthUxMtQ7efZwZJtf3TAPGIjm7mlQJN16RFBJb1loMxbfJQteJ56ydCWDatr3eyx4dgk5EQKj4GsY9jEows2TvIp5LfEd45eF2Exh-nLdjOV8IrJ_YPgYuu17b-uwW_7SGQEVAQ42Ir6f2_PyHGKC9J5dAtD-gQ9_KLkWtQYjWF61sGhZm236o0enKHElJXURzj54s-zu79pRXaRjYTpIa4g-ElcbARSKFlVzaw_zLbqLih1WSYRGJWe5lgjJKXnupVza6LFafYI9R9nq3y2jb0v5cwyGvF_RJTYK1obfctfL1YQPP6XdQIDbLoiposDF3MRJW_WeP_MJb4ncjp_hNafRYFBAYCn1A3NamqUKBg

Guo, S., & Fraser, M. W. (2015). *Propensity score analysis: Statistical methods and applications* (2nd ed.): SAGE publications.

Guo, Y., Little, R. J., & McConnell, D. S. (2012). On Using Summary Statistics From an External Calibration Sample to Correct for Covariate Measurement Error. *Epidemiology, 23*(1), 165-174. doi:10.1097/EDE.0b013e31823a4386

Hajage, D., Tubach, F., Steg, P. G., Bhatt, D. L., & De Rycke, Y. (2016). On the use of propensity scores in case of rare exposure. *Bmc Medical Research Methodology, 16*. doi:10.1186/s12874-016-0135-1

Herrett, E., Shah, A. D., Boggon, R., Denaxas, S., Smeeth, L., van Staa, T., . . . Hemingway, H. (2013). Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *Bmj-British Medical Journal, 346*. doi:10.1136/bmj.f2350

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica, 71*(4), 1161-1189. doi:10.1111/1468-0262.00442

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*(3), 199-236. doi:10.1093/pan/mpl013

Hong, H., Aaby, D. A., Siddique, J., & Stuart, E. A. (2019). Propensity Score-Based Estimators With Multiple Error-Prone Covariates. *American Journal of Epidemiology, 188*(1), 222-230. doi:10.1093/aje/kwy210

Hong, H., Rudolph, K. E., & Stuart, E. A. (2017). Bayesian Approach for Addressing Differential Covariate Measurement Error in Propensity Score Methods. *Psychometrika, 82*(4), 1078-1096. doi:10.1007/s11336-016-9533-x

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society Series a-Statistics in Society, 171*, 481-502. doi:10.1111/j.1467-985X.2007.00527.x

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics, 86*(1), 4-29. doi:10.1162/003465304323023651

Imbens, G. W., & Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature, 47*(1), 5-86. doi:10.1257/jel.47.1.5

Jann, B. (2019). heatplot: Stata module to create heat plots and hexagon plots. Retrieved from http://ideas.repec.org/c/boc/bocode/s458598.html

Joffe, M. M., Ten Have, T. R., Feldman, H. I., & Kimmel, S. E. (2004). Model selection, confounder control, and marginal structural models: Review and new applications. *American Statistician, 58*(4), 272-279. doi:10.1198/000313004x5824

Kaplan, D. (1999). An extension of the propensity score adjustment method for the analysis of group differences in MIMIC models. *Multivariate Behavioral Research, 34*(4), 467-492. doi:10.1207/s15327906mbr3404_4

Keogh, R. H., Shaw, P. A., Gustafson, P., Carroll, R. J., Deffner, V., Dodd, K. W., . . . Freedman, L. S. (2020). STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1-Basic theory and simple methods of adjustment. *Statistics in Medicine, 39*(16), 2197-2231. doi:10.1002/sim.8532

Kuss, O. (2002). Global goodness-of-fit tests in logistic regression with sparse data. *Statistics in medicine, 21*(24), 3789. doi:10.1002/sim.1421

Lee, B. K. (2010). Propensity Score Weighting and Doubly Robust Adjustment in Sparse Data Situations. *American Journal of Epidemiology, 171*, S142.

Leuven, E., & Sianesi, B. (2003). PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing (Version 4.0.12 30jan2016). Retrieved from http://ideas.repec.org/c/boc/bocode/s432001.html

Li, M. X. (2013). Using the Propensity Score Method to Estimate Causal Effects: A Review and Practical Guide. *Organizational Research Methods, 16*(2), 188-226. doi:10.1177/1094428112447816

Lip, G. Y. H., Nieuwlaat, R., Pisters, R., Lane, D. A., & Crijns, H. (2010). Refining Clinical Risk Stratification for Predicting Stroke and Thromboembolism in Atrial Fibrillation Using a Novel Risk Factor-Based Approach The Euro Heart Survey on Atrial Fibrillation. *Chest, 137*(2), 263-272. doi:10.1378/chest.09-1584

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores - An introduction and experimental test. *Evaluation Review, 29*(6), 530-558. doi:10.1177/0193841x05275596

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine, 23*(19), 2937-2960. doi:10.1002/sim.1903

Lunt, M., & Linden, A. (n.d.). propwt: Generating Weights for Propensity Analysis. Retrieved from http://personalpages.manchester.ac.uk/staff/mark.lunt

Millimet, D. L. (2011). The Elephant in the Corner: A Cautionary Tale about Measurement Error in Treatment Effects Models. *Advances in Econometrics, 27*, 1-39. doi:https://doi.org/10.1108/S0731-9053(2011)000027A004

Morgan, S. L., & Todd, J. J. (2008). A DIAGNOSTIC ROUTINE FOR THE DETECTION OF CONSEQUENTIAL HETEROGENEITY OF CAUSAL EFFECTS. In Y. Xie (Ed.), *Sociological Methodology, Vol 38* (Vol. 38, pp. 231-281).

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine, 38*(11), 2074-2102. doi:10.1002/sim.8086

Nguyen, T. Q., & Stuart, E. A. (2020). Propensity Score Analysis With Latent Covariates: Measurement Error Bias Correction Using the Covariate's Posterior Mean, aka the Inclusive Factor Score. *Journal of Educational and Behavioral Statistics, 45*(5), 598-636. doi:10.3102/1076998620911920

Patorno, E., Glynn, R. J., Hernandez-Diaz, S., Liu, J., & Schneeweiss, S. (2014). Studies with Many Covariates and Few Outcomes Selecting Covariates and Implementing Propensity-Score-Based Confounding Adjustments. *Epidemiology, 25*(2), 268-278. doi:10.1097/ede.0000000000000069

Paul, S. R., & Deng, D. L. (2000). Goodness of fit of generalized linear models to sparse data. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 62*, 323. doi:10.1111/1467-9868.00234

Peikes, D. N., Moreno, L., & Orzol, S. M. (2008). Propensity score matching: A note of caution for evaluators of social programs. *American Statistician, 62*(3), 222-231. doi:10.1198/000313008x332016

Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology, 53*(5), 793-808. doi:10.1080/10635150490522304

Pruzek, R. M. (2011). Introduction to the Special Issue on Propensity Score Methods in Behavioral Research. *Multivariate Behavioral Research, 46*(3), 389-398. doi:10.1080/00273171.2011.576618

Raykov, T. (2012). Propensity Score Analysis With Fallible Covariates: A Note on a Latent Variable Modeling Approach. *Educational and Psychological Measurement, 72*(5), 715-733. doi:10.1177/0013164412440999

Rosenbaum, P. R. (1987). MODEL-BASED DIRECT ADJUSTMENT. *Journal of the American Statistical Association, 82*(398), 387-394. doi:10.2307/2289440

Rosenbaum, P. R., & Rubin, D. B. (1983). THE CENTRAL ROLE OF THE PROPENSITY SCORE IN OBSERVATIONAL STUDIES FOR CAUSAL EFFECTS. *Biometrika, 70*(1), 41-55. doi:10.1093/biomet/70.1.41

Rosenbaum, P. R., & Rubin, D. B. (1984). REDUCING BIAS IN OBSERVATIONAL STUDIES USING SUBCLASSIFICATION ON THE PROPENSITY SCORE. *Journal of the American Statistical Association, 79*(387), 516-524. doi:10.2307/2288398

Ross, M. E., Kreider, A. R., Huang, Y.-S., Matone, M., Rubin, D. M., & Localio, A. R. (2015). Propensity Score Methods for Analyzing Observational Data Like Randomized Experiments: Challenges and Solutions for Rare Outcomes and Exposures. *American Journal of Epidemiology, 181*(12), 989. doi:10.1093/aje/kwu469

Rubin, D. B. (2004). On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and Drug Safety, 13*(12), 855-857. doi:10.1002/pds.968

Rudolph, K. E., & Stuart, E. A. (2018). Using Sensitivity Analyses for Unobserved Confounding to Address Covariate Measurement Error in Propensity Score Methods. *American Journal of Epidemiology, 187*(3), 604-613. doi:10.1093/aje/kwx248

Sackett, D. L. (1979). Bias in analytic research. In *The case-control study consensus and controversy* (pp. 51-63): Elsevier.

Schafer, J. L., & Kang, J. (2008). Average Causal Effects From Nonrandomized Studies: A Practical Guide and Simulated Example. *Psychological Methods, 13*(4), 279-313. doi:10.1037/a0014268

Sengewald, M. A., Steiner, P. M., & Pohl, S. (2019). When does measurement error in covariates impact causal effect estimates? Analytic derivations of different scenarios and an empirical illustration. *British Journal of Mathematical & Statistical Psychology, 72*(2), 244-270. doi:10.1111/bmsp.12146

Shu, D., & Yi, G. Y. (2019a). Causal inference with measurement error in outcomes: Bias analysis and estimation methods. *Statistical Methods in Medical Research, 28*(7), 2049-2068. doi:10.1177/0962280217743777

Shu, D., & Yi, G. Y. (2019b). Inverse-probability-of-treatment weighted estimation of causal parameters in the presence of error-contaminated and time-dependent confounders. *Biometrical Journal, 61*(6), 1507-1525. doi:10.1002/bimj.201600228

Shu, D., & Yi, G. Y. (2019c). Weighted causal inference methods with mismeasured covariates and misclassified outcomes. *Statistics in Medicine, 38*(10), 1835-1854. doi:10.1002/sim.8073

Sibbald, B., & Roland, M. (1998). Understanding controlled trials - Why are randomised controlled trials important? *British Medical Journal, 316*(7126), 201-201. Retrieved from <Go to ISI>://WOS:000071616400030

Siino, M., Fasola, S., & Muggeo, V. M. R. (2018). Inferential tools in penalized logistic regression for small and sparse data: A comparative study. *Statistical methods in medical research, 27*(5), 1365. doi:10.1177/0962280216661213

Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the Importance of Reliable Covariate Measurement in Selection Bias Adjustments Using Propensity Scores. *Journal of Educational and Behavioral Statistics, 36*(2), 213-236. doi:10.3102/1076998610375835

Sturmer, T., Rothman, K. J., Avorn, J., & Glynn, R. J. (2010). Treatment Effects in the Presence of Unmeasured Confounding: Dealing With Observations in the Tails of the Propensity Score Distribution-A Simulation Study. *American Journal of Epidemiology, 172*(7), 843-854. doi:10.1093/aje/kwq198

Sturmer, T., Schneeweiss, S., Avorn, J., & Glynn, R. J. (2005). Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American Journal of Epidemiology, 162*(3), 279-289. doi:10.1093/aje/kwi192

Subbiah, M., & Srinivasan, M. R. (2008). Classification of 2 x 2 sparse data sets with zero cells. *Statistics & Probability Letters, 78*(18), 3212. doi:10.1016/j.spl.2008.06.023

Sullivan, S. G., & Greenland, S. (2013). Bayesian regression in SAS software. *International journal of epidemiology, 42*(1), 308. doi:10.1093/ije/dys213

Thoemmes, F. J., & Kim, E. S. (2011). A Systematic Review of Propensity Score Methods in the Social Sciences. *Multivariate Behavioral Research, 46*(1), 90-118. doi:10.1080/00273171.2011.540475

Thomopoulos, N. T. (2013). *Essentials of Monte Carlo simulation : statistical methods for building simulation models*: New York: Springer.

Tumlinson, S. E., Sass, D. A., & Cano, S. M. (2014). The Search for Causal Inferences: Using Propensity Scores Post Hoc to Reduce Estimation Error With Nonexperimental Research. *Journal of Pediatric Psychology, 39*(2), 246-257. doi:10.1093/jpepsy/jst143

VanderWeele, T. J., & Arah, O. A. (2011). Bias Formulas for Sensitivity Analysis of Unmeasured Confounding for General Outcomes, Treatments, and Confounders. *Epidemiology, 22*(1), 42-52. doi:10.1097/EDE.0b013e3181f74493

Vaughan, L. K., Divers, J., Padilla, M. A., Redden, D. T., Tiwari, H. K., Pomp, D., & Allison, D. B. (2009). The use of plasmodes as a supplement to simulations: A simple example evaluating individual admixture estimation methodologies. *Computational Statistics & Data Analysis, 53*(5), 1755-1766. doi:10.1016/j.csda.2008.02.032

Wallace, M. (2020). Analysis in an imperfect world. *Significance, 17*(1), 14-19. doi:https://doi.org/10.1111/j.1740-9713.2020.01353.x

Webb-Vargas, Y., Rudolph, K. E., Lenis, D., Murakami, P., & Stuart, E. A. (2017). An imputation-based solution to using mismeasured covariates in propensity score analysis. *Statistical Methods in Medical Research, 26*(4), 1824-1837. doi:10.1177/0962280215588771

Whittaker, T. A. (2020). The Comparison of Latent Variable Propensity Score Models to Traditional Propensity Score Models under Conditions of Covariate Unreliability. *Multivariate Behavioral Research, 55*(4), 625-646. doi:10.1080/00273171.2019.1663136

Williamson, E., Morley, R., Lucas, A., & Carpenter, J. (2012a). Propensity scores: From naive enthusiasm to intuitive understanding. *Statistical Methods in Medical Research, 21*(3), 273-293. doi:10.1177/0962280210394483

Williamson, E., Morley, R., Lucas, A., & Carpenter, J. (2012b). Variance estimation for stratified propensity score estimators. *Statistics in Medicine, 31*(15), 1617-1632. doi:10.1002/sim.4504

Yoshida, K., Hernandez-Diaz, S., Solomon, D. H., Jackson, J. W., Gagne, J. J., Glynn, R. J., & Franklin, J. M. (2017). Matching Weights to Simultaneously Compare Three Treatment Groups Comparison to Three-way Matching. *Epidemiology, 28*(3), 387. doi:10.1097/ede.0000000000000627

APPENDIX A – LITERATURE SEARCH

Background

The searches were run as a scoping review to investigate the literature for Propensity Score methods combined with measurement error or combined with sparse data.

All searches used the Web of Science.

Searches for Propensity Score including measurement error

Original search conducted on 16/11/16

Table A-1: Original search - 16/11/16.

General search			
Name	Search	Matches	Details
Sa1	Propensity score* (in Title)	2596	
Targeted searches			
Name	Search	Matches	Details
Sa2	Propensity score* (in Title) AND “statistics probability” category	190	
Sa3	Propensity score* (in Title) AND missing OR incomplete (in Title)	15	
Sa4	Propensity score* (in Title) AND measurement error (in Title)	1	
Sa5	Propensity score* (in Title) AND selection bias (in Title)	29	
Sa6	Propensity score* (in Title) AND selection bias (in Topic) AND “statistics probability” category	31	
Sa7	Propensity score* (in Title) AND "multiple imputation" OR "Multiple Imputation" OR "Multiple imputation" (in Topic)	22	
Sa8	Propensity score* (in Title) AND logistic regression (in Title)	14	
Sa9	Propensity score* (in Title) AND meta analysis (in Title)	13	
Sa10	Propensity score* (in Title) AND cost effective* (in Title)	13	
Sa11	Propensity score* (in Title) AND methodological (in Title)	2	
Sa12	Propensity score* (in Title) AND observational (in Title) AND “statistics probability” category	19 (103 w 1st 2 searches)	
Sa13	Propensity score* (in Title) AND “CPRD” (in Topic)	3	
Sa14	Propensity score* (in Title) AND “THIN” (in Topic)	6	

Re-run of original searches conducted in April 2018, looking for papers published since 2016.

Table A-2: Re-run of searches – April 2018.

General search			
Name	Search	Matches	Details
Sa1	Propensity score* (in Title)	1,441	
Targeted searches			
Name	Search	Matches	Details
Sa2	Propensity score* (in Title) AND "statistics probability" category	35	included 4 which presented measurement error methods for use with PS methods*

*(Webb-Vargas, Rudolph, Lenis, Murakami & Stuart, 2017), (Braun et al., 2017), (Hong, Rudolph & Stuart, 2017), (Rudolph & Stuart, 2018).

Re-run of searches conducted 23/03/21, searched for years 2018 to 2021.

Table A-3: Re-run of searches – 23/03/21.

Search	Matches	Details
"propensity scor*" in title (original Sa1)	4070	No further action
"propensity scor*" in title in Stats and Prob category (original Sa2)	115	Screened
"propensity scor*" AND "measurement error" in topic	34*	Screened
"propensity scor*" AND "misclassification" in topic	31*	Screened

*after de-duplication had a total of 57

Propensity Scores and Sparse data

Original search conducted on 18/04/19

Table A-4: Sparse data search – 18/04/19.

Search	Matches	Details
"sparse data" in Title & Stats & probability category	45	
"sparse data" & "propensity score" in Topic	5	
"rare outcome" & "propensity score" in Topic	8	
"rare outcomes" & "propensity score" in Topic	11	
"sparse data bias" in Topic	11	

Screening produced 30 papers of relevance to Sparse Data.

Table A-5: Relevant papers from sparse data search - 18/04/19.

Topic	Number
Cox PH 1 theoretical 1 study with rare outcomes 1 rare exposure but methods are similar	3
General Sparse Data method/background	10
Key papers Franklin (2017) Greenland (2016)	2
Method – Bayesian	1
Method - data augmentation	2
Method - penalisation	3
Method – stratification	1
Method - Bayesian/data augmentation	1
PS methods	5
Tutorials for related methods	2
Grand Total	30

Searches re-run on 22/03/21, searched for years 2019 and 2021.

Table A-6: Re-run of sparse data search – 22/03/21.

Search	Number	Details
“Sparse data” in title in Stats and Prob category	5	
“Sparse data” in title	116	
“sparse data” AND “propensity scor*” in topic	3	Saved to EndNote
“rare outcome” AND “propensity scor*” in topic	0	
“rare outcomes” AND “propensity scor*” in topic	2	Saved to EndNote
“spare data bias” in topic	12	Saved to EndNote

16 papers saved to EndNote for screening.

APPENDIX B – ADDITIONAL INFORMATION FOR METHODS

Contents

B-1 Introduction

B-1.1 Overview

This Appendix provides supplementary information about the selection of the NOAC to use in the study dataset, the modelling of the PS, the balance checking following PS conditioning, the outcome modelling and generating the baseline hazard function.

B-1.2 Additional information about the study dataset

This study used data supplied to the Performance-Based Innovation Rewards project (REWARD). The aims of REWARD were to increase access to pharmaceutical products, particularly in low- and middle-income countries, by financially rewarding pharmaceutical companies for the performance of their products. Performance-based Reimbursement (PBR) tools were assessed to evaluate the real-world effectiveness of new products in both high and low- and middle-income countries. The assessment of real-world effectiveness, of the group of Novel Oral Anticoagulants (NOAC) compared to the existing or control treatment, Warfarin, referred to as an Oral Anti-Coagulant (OAC), on stroke incidence amongst patients with Atrial Fibrillation (AF) in the UK using an extract from The Health Improvement Network (THIN), was an example of a PBR tool.

The data for REWARD had been supplied in episode format with a new episode starting when the patient's AF status changed. This was different to the start date for the NOAC/OAC prescribing. The episode with the start time closest to the patient's first prescription of an NOAC/OAC was taken as their baseline data. This meant that for some patients the baseline data were measured after the first prescription date. Data which were taken several months after the first prescription was considered more current than that several years before, but this is acknowledged as a source of measurement error.

B-2 The choice of model assessment criteria

The Stata estimation command *estat ic* provided three ways to assess the model fit: Likelihood Ratio Test (LRT); Akaike Information Criteria (AIC); Bayesian Information Criteria (BIC). The LRT was not used as when additional parameters are added to the model the LRT will always improve. The AIC represents the amount of data lost when using model_x to represent model_y. The AIC is designed to select the model which is the best approximation to the truth (Posada &

Buckley, 2004). The BIC finds the model which is the best approximation to the true model to fit the data. Generally, the BIC tends to select a simpler model than the AIC (Posada & Buckley, 2004). For these reasons the BIC was used. The BIC is defined as $BIC = -2 * ln(likelihood) + ln(n) * k$ where *n* is the sample size and *k* is the number of parameters.

The criteria used for assessing the fit of a model was:

- The standard errors for the variables should not be large to ensure the model converged.
- Models using the highest number of observations were favoured.
- The model with the lowest BIC.

B-3 Modelling the Propensity Score for the RI-WA dataset

B-3.1 Selection of the PS modelling method

Both logistic regression and probit regression apply a function to transform outcomes from a linear model so that they fall in the range of [0, 1]. Probit models can be used when there are non-constant error variances (heteroskedastic probit models), but this was not the case in the study data so either probit or logistic regression could have been used. Logistic regression was selected over probit regression because logistic regression is widely used in applied medical research and it can display the log odds which is useful in understanding the influence of a variable on the treatment allocation.

Logistic regression gives the logged odds, *L*, of the probability of receiving the treatment, *P*, the PS. For the ith patient

$$L_i = \ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \underline{\beta}^T X_i$$

where β_0 is the intercept, $\underline{\beta}$ is the vector of the model coefficients for X_i the ith patient's baseline covariates, i=1, …, n, n is the number of patients.

Hence P_i , the PS for the ith patient

$$P_i = 1/(1 + \exp\left(-1 * (\beta_0 + \underline{\beta}^T X_i)\right))$$

and P_i will have values in the range [0, 1].

B-3.2 Variable selection for the PS model

The variables which are known to clinically influence the decision to prescribe a NOAC or Warfarin were stroke (Hankey et al., 2012; Toso, 2014), alcohol misuse (Baczek, Chen, Kluger &

Coleman, 2012), chronic kidney disease (Boriani et al., 2016), liver disease (Lai et al., 2016), CHA2DS2-VASc score (Giralt-Steinhauer et al., 2013; Lee, Monz, Clemens, Brueckmann & Lip, 2012), HAS-BLED score (O'Caoimh et al., 2017), ischaemic heart disease used to indicate previous myocardial infarction (Bhatia & Lip, 2004), and age (Wolff, Shantsila, Lip & Lane, 2015).

Other non-clinically relevant variables which appeared to affect prescribing were included in the PS model. The time, in days, from the first AF diagnosis to the first NOAC/OAC prescription date was included as it is likely to be a surrogate for other factors affecting prescribing. A variable to represent the date of the first NOAC/OAC prescription, was also included. This appeared to strongly affect prescribing, as time progressed during the study a higher proportion of patients were prescribed NOACs compared to Warfarin.

B-3.3 Generating the PS model

All clinically relevant variables were kept in the PS model, regardless of their statistical significance during the model selection process. Other non-clinically relevant variables were kept in the model if their p-value <=0.05, showing them to be statistically significant.

Table B-1 describes the variables considered for the model. The functional form of the variable is its relationship with the dependent variable. For a continuous variable this included linear, squared or expressed as a categorical variable. The functional forms for the continuous variables, age and date of first prescription, were assessed using the criteria in Section B-2. The functional forms of the variable age that were considered, were age86_gen (below 86 age=86, above 86 age=age), age+age^2+age^3, age+age^2, age(continuous), no_age. These were each tested in a model containing the clinically relevant variables. All these forms were considered further except no age.

For the RI-WA dataset the first prescription date was expressed as the difference, in days, between prescription date and the NICE licence date for Rivaroxaban. It was represented by the variable licence_to_noac. Different functional forms were considered (Table B-2). The forms which showed the best fit, using BIC, were date_by_qtr_adj3 (any dates in or before 2012q4 were set to missing), date+date^2 and date(continuous).

Table B-1: Variables used in PS model and outcome model.

Variable name	Variable description	Variable type	Considered for PS model	Considered for Outcome model
af_to_noac_gen	First NOAC/OAC prescription was ≤ 28 days of first AF diagnosis?	Binary	Y	
age_65_gen	Age ≥ 65?	Binary		Y
age_75_gen	Age ≥75?	Binary		Y
age86_gen	=86 if age≤86, else =age	Binary	Y	
alcohol_misuse_gen	Alcohol misuse?	Binary	Y	
chads2_vasc_calculated	CHA2DS2-VASc score calculated from the data	Quanti-tative	Y	
ckd_gen	Chronic kidney disease?	Binary	Y	
congestive_card_fail_gen	Congestive cardiac failure?	Binary		Y
diabetes_gen	Diabetes?	Binary		Y
first_noac_date	Date of first NOAC/OAC prescription	Cont-inuous	Y	
hasbled_gen	HAS-BLED score calculated from the data	Quanti-tative	Y	
hypercholesterol_gen	Hypercholesterolemia?	Binary		Y
hypertension_gen	Hypertension?	Binary		Y
ihd_gen	Ischemic heart disease?	Binary	Y	
licence_to_noac	Date of first NOAC/OAC, used in RI-WA dataset. It is the RI licence date to date of first prescription, in days.	Cont-inuous	Y	
licence_to_noac30	licence_to_noac/30 RI licence date to date of first prescription, in months	Cont-inuous	Y	
liver_disease_gen	Liver disease?	Binary	Y	
number_of_prescriptions	Number of medications currently prescribed to this patient	Quanti-tative		Y
on_cvd_antiplatelet	On antiplatelet?	Binary		Y
on_cvd_bp_lowering	On blood pressure lowering medication?	Binary		Y
on_cvd_statin	On statins?	Binary		Y
Sex	Sex of patient	Binary		Y
smoke_now_gen	Smokes now?	Binary		Y
smoke_prev_gen	Smoked previously?	Binary		Y
stroke_or_tia_gen	Previous stroke or TIA?	Binary	Y	
Townsend	Townsend deprivation score. (Quintiles)	*		Y

Treatment	Name of NOAC/OAC treatment	Categ-orical		Y
vascular_disease_gen	Vascular disease?	Binary		Y

*Townsend was first used as a categorical variable then as a factor

Table B-2: Functional forms of the variable date of first prescription assessed, represented by licence_to_noac.

Form of Date	Description
date_by_qtr_adj3	If date <=2012q4 set date = missing
date+date^2	date+date^2
date(cont)	date(continuous)
date_by_qtr_adj1	If date=2012q2 set date = 2012q3
date_by_qtr_adj2	If <=2012q4 set date = 2012q4
date_by_qtr	date_by_qtr
date_by_year	date_by_year
no_date	no_date

The selected functional forms of the two variables were added in combinations of clinically relevant variables and the resulting models assessed. The assessment of these models is shown in Table B-3. Model selection algorithm was then applied as follows:

- Ensure there are no high SEs – all models passed.
- The lowest was BIC preferred – suggested models 2, 14, 10 or 6, but these use fewer observations. Observations were dropped for some of the early time periods where there were fewer Rivaroxaban patients.
- Consider lowest BIC using full observations – these are models 4, 12, 16, 8 and 3. Model 4 (age86_gen + date+date^2) (Table B-3) is chosen due to parsimony.

Table B-3: Assessment of combination of different functional forms of age and date for PS model in RI-WA dataset, ordered by BIC.

Model	Age	Prescription date	N	LL0	LL1	df	AIC	BIC
2	age86_gen	date_by_qtr_adj3	18551	-7998.38	-7308.94	20	14657.89	14814.45
14	age+age^2+age^3	date_by_qtr_adj3	18551	-7998.38	-7307.23	22	14658.46	14830.68
10	age+age^2	date_by_qtr_adj3	18551	-7998.38	-7312.31	21	14666.63	14831.02
6	age(cont)	date_by_qtr_adj3	18551	-7998.38	-7319.43	20	14678.85	14835.42
4	**age86_gen**	**date+date^2**	**21259**	**-8489.82**	**-7507.96**	**12**	**15039.91**	**15135.48**
12	age+age^2	date+date^2	21259	-8489.82	-7510.46	13	15046.92	15150.46
16	age+age^2+age^3	date+date^2	21259	-8489.82	-7506.59	14	15041.19	15152.69
3	age86_gen	date(cont)	21259	-8489.82	-7523.93	11	15069.86	15157.47
8	age(cont)	date+date^2	21259	-8489.82	-7519.13	12	15062.26	15157.83
11	age+age^2	date(cont)	21259	-8489.82	-7526.91	12	15077.81	15173.38
15	age+age^2+age^3	date(cont)	21259	-8489.82	-7522.84	13	15071.69	15175.23
7	age(cont)	date(cont)	21259	-8489.82	-7535.62	11	15093.24	15180.85
1	age86_gen	no_date	21259	-8489.82	-8458.9	10	16937.8	17017.45
13	age+age^2+age^3	no_date	21259	-8489.82	-8458.03	12	16940.06	17035.64
9	age+age^2	no_date	21259	-8489.82	-8464.23	11	16950.45	17038.06
5	age(cont)	no_date	21259	-8489.82	-8471.54	10	16963.09	17042.73
0	no_age	no_date	21259	-8489.82	-8481.12	9	16980.24	17051.92

The 'best' treatment allocation model was identified. For the RI-WA dataset the model (Table B-4) using age (which increased when the patient was over 86 years) and date of first prescription used the difference, in days, between prescription date and the NICE licence date for Rivaroxaban plus its squared format, in addition to the clinically relevant variables selected.

Table B-4: 'Best' PS model for the Rivaroxaban-Warfarin dataset.

Covariate	Coefficient	SE of coefficient	z	P>|z|	[95% CI]
Previous stroke	0.108	0.065	1.67	0.095	(-0.019, 0.235)
Alcohol misuse	0.117	0.136	0.86	0.39	(-0.149, 0.383)
Chronic kidney disease	0.005	0.069	0.07	0.944	(-0.130, 0.139)
Liver disease	0.036	0.439	0.08	0.935	(-0.825, 0.897)
CHA2DS2-VASc score (calculated)	0.020	0.023	0.84	0.404	(-0.026, 0.065)
HAS-BLED score (calculated)	-0.009	0.040	-0.23	0.818	(-0.087, 0.069)
Ischemic heart disease	-0.091	0.052	-1.76	0.078	(-0.193, 0.010)
First NOAC/OAC prescription was ≤ 28 days of first AF diagnosis?	-0.194	0.042	-4.59	<0.001	(-0.277, -0.111)
=86 if age≤86, else =age	0.075	0.013	5.84	<0.001	(0.050, 0.100)
licence_to_noac*	0.005	0.000	13.56	<0.001	(0.004, 0.006)
licence_to_noac^2	-1.62E-06	2.92E-07	-5.54	<0.001	(-2.19E-06, -1.04E-06)
Constant term	-10.669	1.110	-9.61	<0.001	(-12.846, -8.493)

*licence_to_noac30 is the Rivaroxaban licence date to date of first NOAC/OAC prescription, in days

B-3.4 PS Model refinement

In the Rivaroxaban-Warfarin dataset, the PS model was assessed to determine if any of the variables could be dropped to simplify its use in the simulations phase. Table B-5 shows the assessment of the variables, marked (1), have a similar effect size with coefficients close to either 0.1 or -0.1 and were kept in the model. For alcohol_misuse_gen the standard error (of the coefficient estimate) is higher, but this could be due to alcohol misuse having a lower prevalence. The variables, marked (2), have low p-values and low standard errors and were kept in the model. These variables could have been surrogates for other, unmeasured variables. The variables, marked (3), were kept in the model. Despite the fact they had low coefficients and high p-values they were contraindications to prescribing NOACs, so are clinically relevant. The variables, marked (4), were discarded. They all had low coefficients and high p-values and are not clinically relevant when prescribing NOACs or Warfarin. The constant, marked (5) was included. The refined PS model used in this study is given in Table B-6.

Table B-5: Assessment of variables to retain in the treatment allocation model for the RI-WA dataset.

Covariate	Coefficient	SE of coefficient	z	P>|z|	[95% CI]	Keep?
Previous stroke	0.108	0.065	1.67	0.095	(-0.019, 0.235)	1 - keep
Alcohol misuse	0.117	0.136	0.86	0.390	(-0.149, 0.383)	1 - keep
Ischemic heart disease	-0.091	0.052	-1.76	0.078	(-0.193, 0.01)	1 – keep
First NOAC/OAC prescription was ≤ 28 days of first AF diagnosis?	-0.194	0.042	-4.59	<0.001	(-0.277, -0.111)	2 – keep
=86 if age≤86, else =age	0.075	0.013	5.84	<0.001	(0.05, 0.1)	2 – keep
licence_to_noac*	0.005	0.000	13.56	<0.001	(0.004, 0.006)	2 – keep
licence_to_noac^2	-1.620E-06	2.920E-07	-5.54	<0.001	(-2.19E-06, -1.04E-06)	2 – keep
Chronic kidney disease	0.005	0.069	0.07	0.944	(-0.13, 0.139)	3 – keep
Liver disease	0.036	0.439	0.08	0.935	(-0.825, 0.897)	3 – keep
CHA2DS2-VASc score (calculated)	0.020	0.023	0.84	0.404	(-0.026, 0.065)	4- discard
HAS-BLED score (calculated)	-0.009	0.040	-0.23	0.818	(-0.087, 0.069)	4- discard
Constant term	-10.669	1.110	-9.61	<0.001	(-12.846, -8.493)	5 – keep

*licence_to_noac30 is the Rivaroxaban licence date to date of first NOAC/OAC prescription, in days

The variables for the CHA2DS2-VASc score (Lip et al., 2010) and HAS-BLED score (Pisters et al., 2010) were removed from the model. The variable to measure the date of first prescription was adjusted to be in units of 30 day and renamed licence_to_noac30 (

Table B-6). Re-fitting the PS model after removing the two variables changed the coefficients. This could be due to the variable chads2_vasc_calculated being correlated to other variables such as stroke_or_tia_gen (previous stroke or TIA).

Table B-6: The refined treatment allocation model for the RI-WA dataset.

Covariate	Coefficient	SE of coefficient	z	P>|z|	[95% CI]
Previous stroke	0.123	0.061	2.03	0.042	(0.004, 0.242)
Alcohol misuse	0.098	0.128	0.76	0.446	(-0.153, 0.348)
Chronic kidney disease	0.008	0.051	0.16	0.871	(-0.093, 0.109)
Liver disease	0.033	0.437	0.07	0.941	(-0.825, 0.890)
Ischemic heart disease	-0.082	0.051	-1.61	0.108	(-0.181, 0.018)
First NOAC/OAC prescription was ≤ 28 days of first AF diagnosis?	-0.192	0.042	-4.56	<0.001	(-0.275, -0.110)
=86 if age≤86, else =age	0.077	0.013	6.13	<0.001	(0.053, 0.102)
licence_to_noac30 *	0.153	0.011	13.56	<0.001	(0.131, 0.175)
$(licence\_to\_noac30)^2$	-0.001	<0.001	-5.54	<0.001	(-0.002, -0.001)
Constant term	-10.830	1.096	-9.88	<0.001	(-12.979, -8.682)

*licence_to_noac30 is the Rivaroxaban licence date to date of first NOAC/OAC prescription, in months

B-3.5 PS balance checking

The literature advises that a check for common support, or overlap, should be carried out once the PS model has been defined and hence the PS value calculated. This ensures the two treatment groups have sufficient participants with similar PS values to make the PS conditioning meaningful. This can be checked visually using a density plot of the PS for each treatment group. Figure B-1 shows that there was good common support in both datasets. A simple match on the PS showed that all NOAC cases were matched to a Warfarin case in both datasets. In both datasets the PS model was sufficiently well defined to continue the analysis.

Balance checking verifies that the estimated PS, which was generated here, was sufficiently close to the true PS. Cases with the same true PS will have the same covariate distribution. If the distribution of the covariates is similar for the matched cases with the same estimated PS, then the estimated PS is sufficiently well defined (Ho, Imai, King & Stuart, 2007) (Section 2.3.5 main text). Although Garrido et al. (2014) suggests that the balance checks be carried out before PS conditioning , but not all authors agree and only perform balance checking after the PS method has been applied. In this study, covariate balance checking was not undertaken at this stage, it was done after PS conditioning.

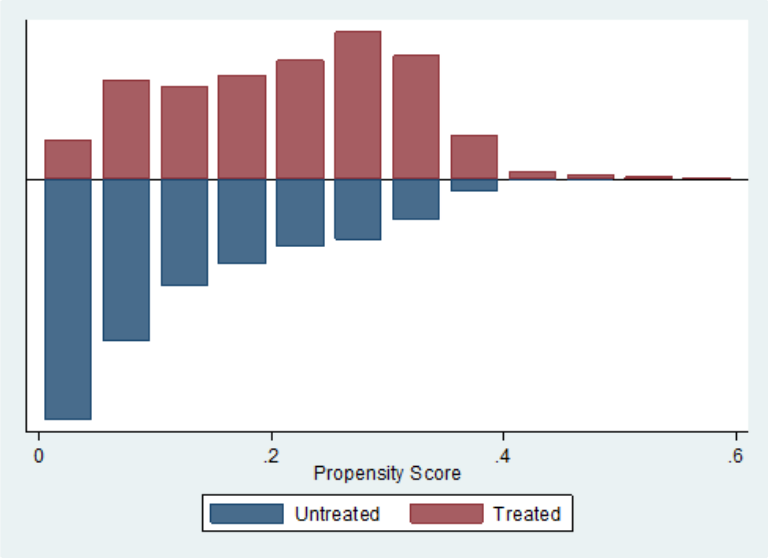


Figure B-1: Histogram of Propensity Score, using Stata's -psgraph-, for Rivaroxaban (Treated) and Warfarin (Untreated) for the RI-WA dataset.

B-4 Apixaban vs Warfarin Dataset

B-4.1 PS modelling for the AP-WA dataset

The PS model was fitted to the Apixaban-Warfarin dataset, following the same method described for the Rivaroxaban-Warfarin dataset. This helped inform the decision of which dataset to select as the study dataset. The functional forms for the continuous variables, age and date of first prescription, were assessed using the criteria (Section B-0). The functional forms of age that were considered, were age86_gen (below 86 age=86, above 86 age=age), age+age^2, age+age^2+age^3, age(continuous) and no age. These were each tested in a model containing the clinically relevant variables. The forms which showed the best fit, using BIC, were age86_gen and age(continuous). The functional forms of the date of first prescription (to be referred to as *date*) that were considered were, date (continuous), date by year, date by quarter, date by adjusted quarter (date before July 2013 set to quarter214 (July to Sept 2013)), date further adjusted quarter (date before July 2013 (qtr214) =missing). These adjustments were made due to the low number of RI patients at the beginning of the study. Only linear terms were considered due to the form of this variable. The forms which showed the best fit, using BIC, were date (continuous), date by quarter, date further adjusted quarter (date before July 2013 (qtr214) set to missing). The latter lost observations so was dropped. Date (continuous) and date by quarter were considered the best forms.

The selected forms of the age and date of first prescription were added in all combinations to the clinically relevant variables and the resulting models assessed. The model selection algorithm was then applied as follows:

- Ensure no high SEs – all models passed.
- Lowest BIC preferred – suggests models 4, 8 16 or 12, but these use fewer observations. Observations were dropped for some of the early time periods where there were fewer Apixaban patients.
- Consider lowest BIC using full observations – these are models 3 (age86_gen & date_continuous) and 7 (age_continuous & date_continuous). Model 3 (Table B-7) was chosen as the "best" model due to parsimony.

Table B-7: Assessment of combination of different functional forms of age and date for PS model in AP-WA dataset, ordered by BIC.

M	age	pr_date	N	LL0	LL1	df	AIC	BIC
M4	age86_gen	date_by_qtr(fthr_adj)	9665	-3814.49	-3533.37	16	7098.737	7213.557
M8	age(cont)	date_by_qtr(fthr_adj)	9665	-3814.49	-3539.25	16	7110.508	7225.328
M16	age+age^2+age^3	date_by_qtr(fthr_adj)	9665	-3814.49	-3531.18	18	7098.368	7227.541
M12	age+age^2	date_by_qtr(fthr_adj)	9665	-3814.49	-3536.31	17	7106.623	7228.62
M3	**age86_gen**	**date(cont)**	**13897**	**-4403**	**-3765.55**	**11**	**7553.108**	**7636.042**
M7	**age(cont)**	**date(cont)**	**13897**	**-4403**	**-3771.11**	**11**	**7564.224**	**7647.158**
M15	age+age^2+age^3	date(cont)	13897	-4403	-3763.59	13	7553.169	7651.182
M11	age+age^2	date(cont)	13897	-4403	-3768.61	12	7561.221	7651.695
M2	age86_gen	date_by_year	13897	-4403	-3856.23	12	7736.453	7826.926
M6	age(cont)	date_by_year	13897	-4403	-3860.69	12	7745.376	7835.849
M10	age+age^2	date_by_year	13897	-4403	-3858.71	13	7743.42	7841.433
M14	age+age^2+age^3	date_by_year	13897	-4403	-3854.3	14	7736.591	7842.143
M1	age86_gen	no_date	13897	-4403	-4384.07	10	8788.145	8863.54
M0	no_age	no_date	13897	-4403	-4391.38	9	8800.757	8868.612
M5	age(cont)	no_date	13897	-4403	-4389.33	10	8798.656	8874.05
M13	age+age^2+age^3	no_date	13897	-4403	-4382.22	12	8788.442	8878.916
M9	age+age^2	no_date	13897	-4403	-4387.6	11	8797.191	8880.125

The best model for the AP-WA dataset (Table B-8) used variables to represent age (which increased when the patient was over 86 years) and the date of first NOAC/OAC prescription as a continuous variable in addition to the clinically relevant variables.

Table B-8: 'Best' PS model for the Apixaban-Warfarin dataset.

Covariate	Coefficient	SE of Coefficient	z	P>|z|	95% CI
Previous stroke	0.266	0.092	2.88	0.004	(0.085, 0.446)
Alcohol misuse	0.543	0.170	3.18	0.001	(0.209, 0.877)
Chronic kidney disease	-0.096	0.101	-0.95	0.341	(-0.293, 0.101)
Liver disease	0.514	0.507	1.01	0.31	(-0.479, 1.508)
CHA2DS2-VASc score (calculated)	0.035	0.034	1.02	0.309	(-0.032, 0.102)
HAS-BLED score (calculated)	-0.035	0.057	-0.62	0.537	(-0.148, 0.077)
Ischemic heart disease	-0.076	0.075	-1.01	0.313	(-0.223, 0.071)
First NOAC/OAC prescription was ≤ 28 days of first AF diagnosis?	-0.168	0.061	-2.74	0.006	(-0.288, -0.048)
=86 if age≤86, else =age	0.070	0.019	3.71	<0.001	(0.033, 0.107)
Date of first NOAC/OAC prescription	0.005	0.000	30.98	<0.001	(0.005, 0.005)
Constant term	-107.017	3.609	-29.66	<0.001	(-114.089, -99.944)

B-4.2 PS balance checking – AP-WA

Figure B-2 shows that there was good common support in the AP-WA dataset. A simple match on the PS showed that all Apixaban cases were matched to a Warfarin case.

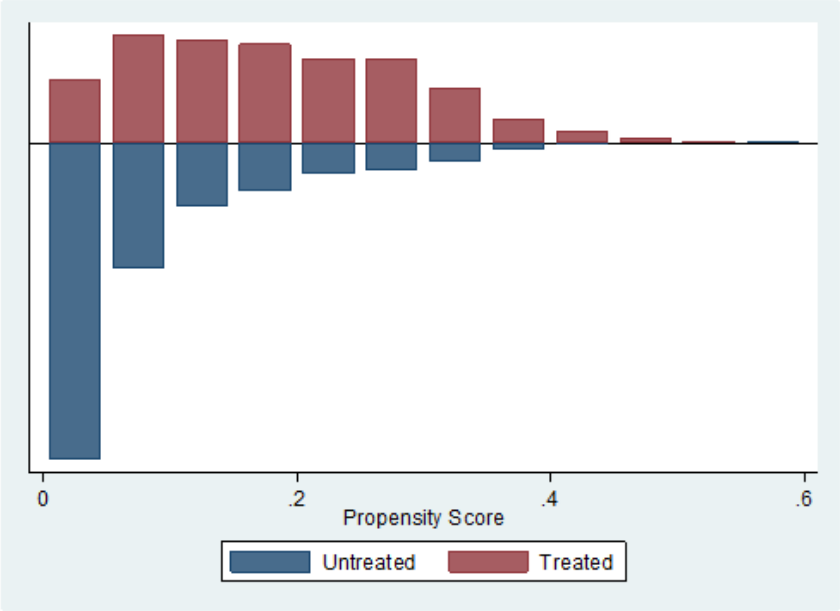


Figure B-2: Histogram of Propensity Score, using Stata's -psgraph-, for Apixaban (Treated) and Warfarin (Untreated) for AP-WA dataset.

B-5 Selection of the study dataset

The two datasets, Apixaban-Warfarin and Rivaroxaban-Warfarin, were compared to decide which one to continue working with during the study. The first NOAC/OAC prescription date after the National Institute for Health and Care Excellence (NICE) approval date for Rivaroxaban was May 2012 and for Apixaban was February 2013. Although common support was shown in both datasets, the NOAC patients in the Rivaroxaban-Warfarin dataset had more patients with a higher PS so more contrast to the Warfarin patients, compared to the Apixaban-Warfarin dataset (Table B-9). There were more patients overall in the Rivaroxaban-Warfarin dataset as the Rivaroxaban NICE licence was earlier than the Apixaban NICE licence date and also a larger percentage of the patients are prescribed Rivaroxaban (13.7%) compared to Apixaban (9.6%) (Table B-9). This meant there were likely to be more outcome events in the Rivaroxaban-Warfarin dataset which would make the outcome modelling more stable. Considering that both datasets were performing reasonably well in terms of PS matching, the precondition for model stability was regarded as an important characteristic and therefore the Rivaroxaban-Warfarin dataset was chosen to take forward.

Table B-9: The number of patients on each treatment in the AP-WA and RI-WA datasets.

Treatment	Frequency	Percent
Warfarin	12,559	90.37
Apixaban	1,338	9.63
Total	13,897	100

Treatment	Frequency	Percent
Warfarin	18,348	86.31
Rivaroxaban	2,911	13.69
Total	21,259	100

B-6 PS conditioning methods

Details of the implementation of the PS conditioning methods are given in the main text. Supplementary information and the results of the balance checks are given in this section.

B-6.1 Stata matching programs

The early simulations gave the opportunity to explore alternative matching functions in Stata. The function -kmatch- (Jann, 2017) could perform PS matching and calculated the ATT and ATE only for binary outcomes. It did not record details of which Warfarin cases were matched to each Rivaroxaban case in the 3to1 matching performed in this study. This was needed to form the matched groups which had the same baseline hazard in the outcome analysis. -nnmatch- (Herr, Drukker, Imbens & Abadie, n.d.) could also perform matching, but only calculated the ATE or ATT for binary outcomes. -psmatch2- (Leuven & Sianesi, 2003) offered additional functionality so was used for 3to1 PS matching in this study.

B-6.2 Additional PS matching types

Different PS matching methods were investigated for this study (Table B-10) in order to find the most appropriate one to take forward. All use greedy matching, where the best match was made by each treated case from the available untreated cases. All except A5 and A6 used matching with no replacement, when a match was made with an untreated case that case was no longer available for matching to subsequent treated cases. A5 and A6 used matching with replacement, where after a match was made the untreated case was still available for subsequent cases, meaning that untreated cases could be matched to more than one treated case. All of the methods used nearest neighbour matching, the treated cases matched the available untreated cases with the closest PS. There was no restriction on the difference of the PS values of the matched pairs. A2, A3 and A4 matched on the logit(PS) and imposed a caliper, meaning that a treated case could only match an untreated case if the difference in their logit(PS) was within the value of the caliper. A2 used the standard caliper of 2 x the SD of the logit(PS), 0.227264. A3 used a caliper of 0.1 and A4 used a caliper of 0.01. All cases, except A5 and A6, used 1 to 1 matching. A5 matched 2 untreated cases to each treated case and A6 matched 3 untreated cases to each treated case. A7 used the common support option, so dropped treated cases whose PS was outside the range of the PS for the untreated group. A8 used common support by dropping 10% of the treatment observations at which the PS density of the control observations was the lowest. A1, 1:1 nearest neighbour (default settings) no-replacement, was considered for use, but A6, 3:1 nearest neighbour (default settings & with replacement), was used for PS matching in this study.

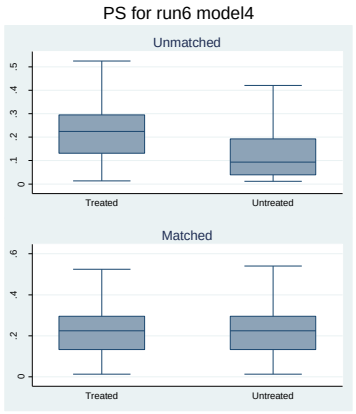
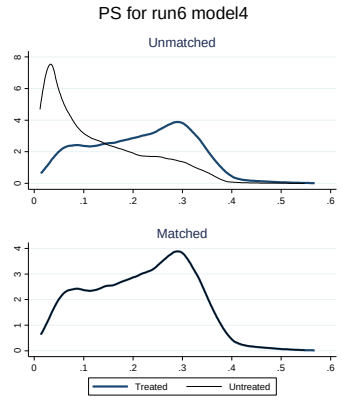
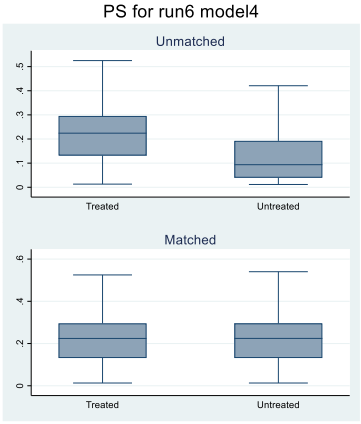
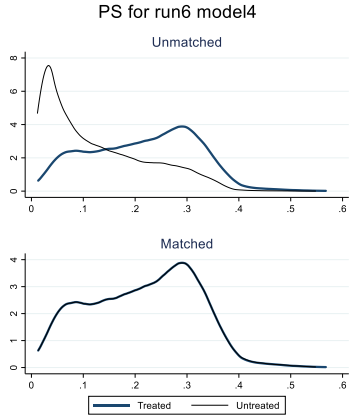
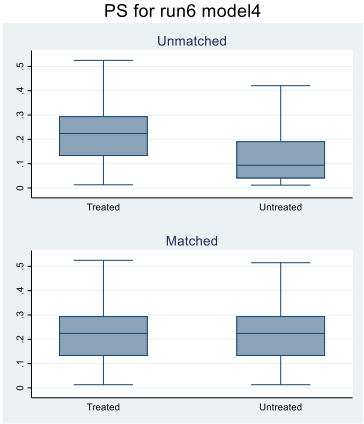
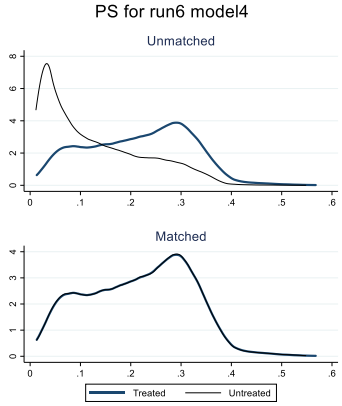
Table B-10: Propensity Score matching methods applied to the RI-WA dataset.

Ref	Description
A1	1:1 nearest neighbour (default settings) no-replacement
A2	1:1 nearest neighbour, using caliper 0.227264, no-replacement
A3	1:1 nearest neighbour, using caliper 0.1, no-replacement
A4	1:1 nearest neighbour, using caliper 0.01, no-replacement
A5	2:1 nearest neighbour (default settings & with replacement)
A6	3:1 nearest neighbour (default settings & with replacement)
A7	1:1 nearest neighbour (default settings) no-replacement common support
A8	1:1 nearest neighbour (default settings) no-replacement trimming 10% from treatment group

B-6.3 PS matching - balance checks

PS balance results from some of the PS matching methods given in Table B-10. A3 and A4, 1:1 nearest neighbour using caliper 0.1 and 0.01, respectively were not considered as A2, using a caliper of 0.227264, showed very little difference to A1, 1:1 nearest neighbour.

Table B-11: PS box plot and PS density before and after PS matching.

PS Match Method	Box plot of PS (*pstest ps_calc1, box both*)	Distribution of PS (*pstest ps_calc1, density both*)
A1 1:1 nearest neighbour, no-replacement		
A2 1:1 nearest neighbour, using caliper, no-replacement		
A5 2:1 nearest neighbour, with replacement		

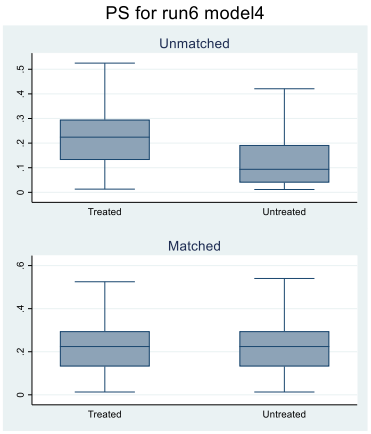
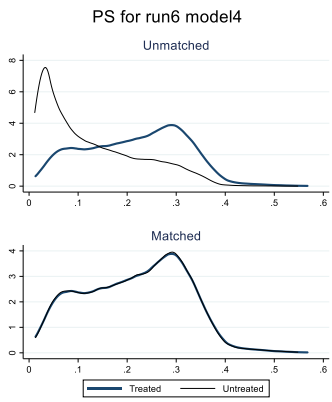
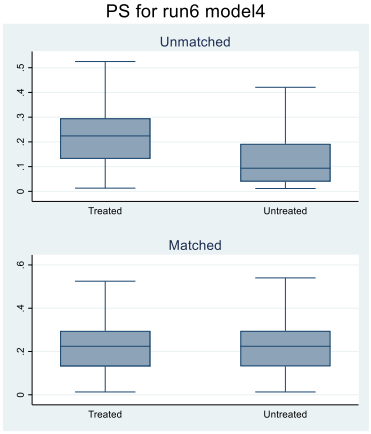
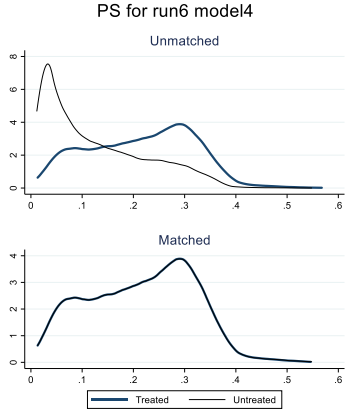
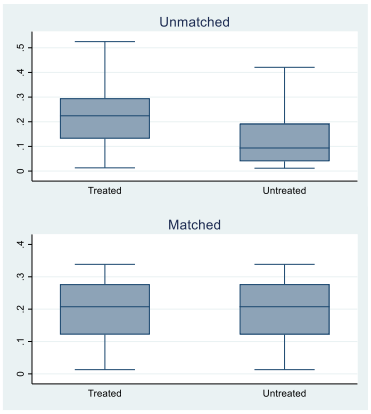
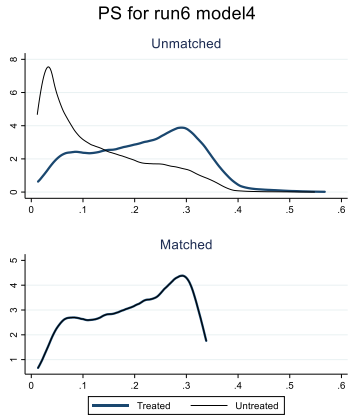
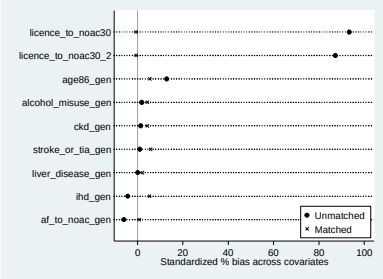
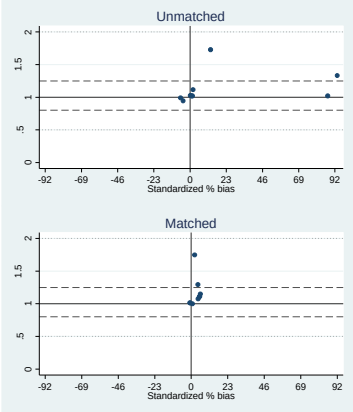
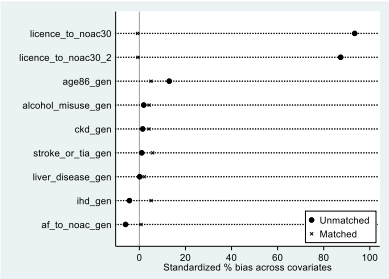
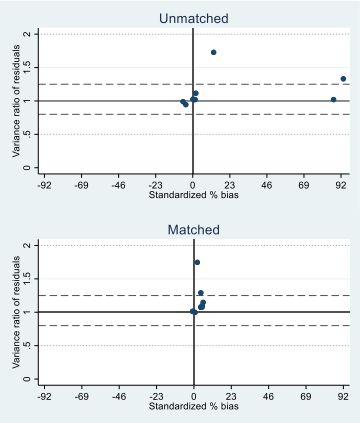
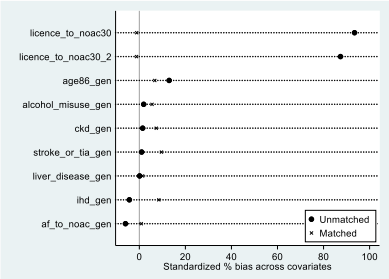
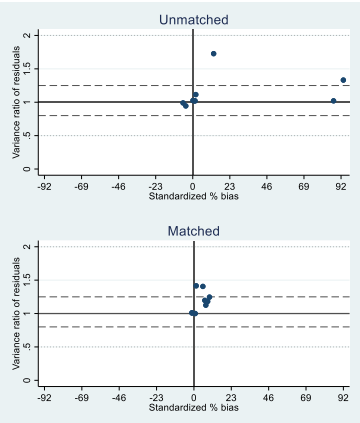
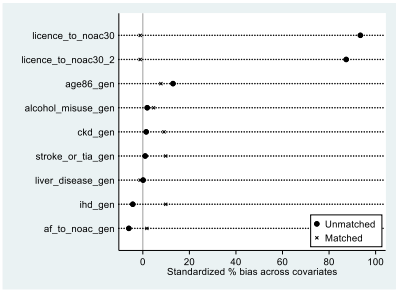
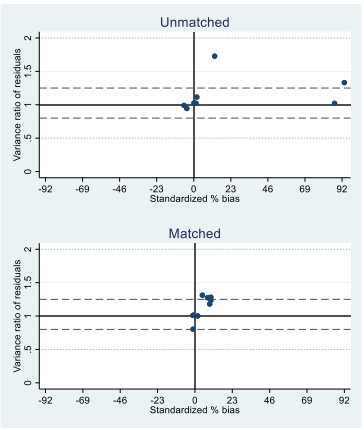
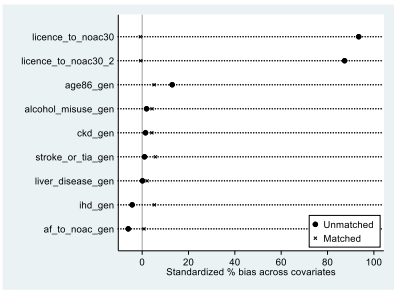
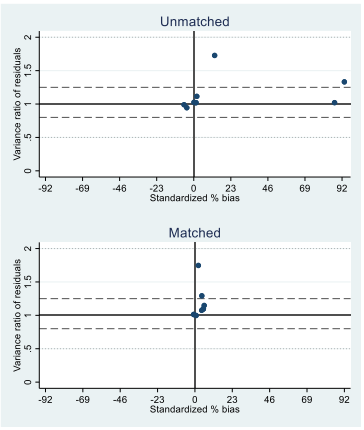
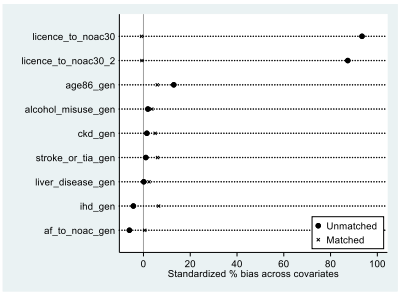
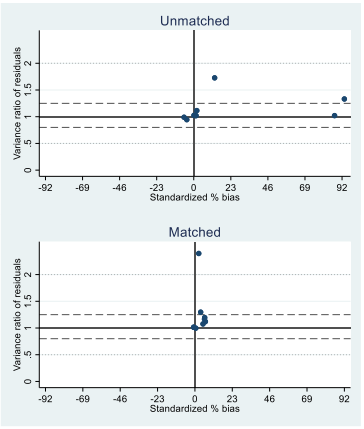
A6 3:1 nearest neighbour, with replacement (PS matching method used in this study)		
A7 1:1 nearest neighbour, no-replacement, common support imposed		
A8 1:1 nearest neighbour, no-replacement, trimming 10% from treatment group		

Table B-12: Variable standardised Differences and Variable ratio of residuals vs standardised %bias, before and after PS matching.

PS Match Method	Variable Standardised Differences (Using, graph)	Variable Ratio of Residual vs Standardised %bias (Using, scatter)
A1 1:1 nearest neighbour, no-replacement		
A2 1:1 nearest neighbour, using caliper, no-replacement		
A5 2:1 nearest neighbour, with replacement		

A6 3:1 nearest neighbour, with replacement (PS matching method used in this study)		
A7 1:1 nearest neighbour, no-replacement, common support imposed		
A8 1:1 nearest neighbour, no-replacement, trimming 10% from treatment group		

B-6.4 Balance checking for IPTW

The results of the balance checking for IPTW for ATT and IPTW for ATE are shown in this section.

Table B-13: Standardised mean differences for the original data and that using the IPTW weights for ATT and ATE.

	Original			ATT weights			ATE weights		
Covariate	Mean treated	Mean un-treated	Stdz'd diff*	Mean treated	Mean un-treated	Stdz'd diff*	Mean treated	Mean un-treated	Stdz'd diff*
Previous stroke	0.14	0.14	0.01	0.14	0.14	-0.003	0.15	0.14	0.021
Alcohol misuse	0.03	0.03	0.019	0.03	0.03	0.005	0.02	0.03	-0.022
Chronic kidney disease	0.22	0.22	0.014	0.22	0.23	-0.004	0.23	0.22	0.025
Liver disease	0	0	0.001	0	0	-0.002	0	0	0.037
Ischemic heart disease	0.22	0.24	-0.043	0.22	0.22	-0.001	0.24	0.24	-0.003
af_to_noac_gen†	0.44	0.47	-0.06	0.44	0.44	0.005	0.45	0.46	-0.019
=86 if age≤86, else =age	0.58	0.37	0.13	0.58	0.58	0.001	0.39	0.4	-0.005
licence_to_noac30††	24.32	15.79	0.934	24.32	24.3	0.002	17.5	16.96	0.059
(licence_to_noac30)^2	660.82	346.86	0.874	660.82	659.47	0.004	400.33	389.68	0.03

*Standardised Difference

†First NOAC/OAC prescription was ≤ 28 days of first AF diagnosis?

††The Rivaroxaban licence date to date of first NOAC/OAC prescription, in months

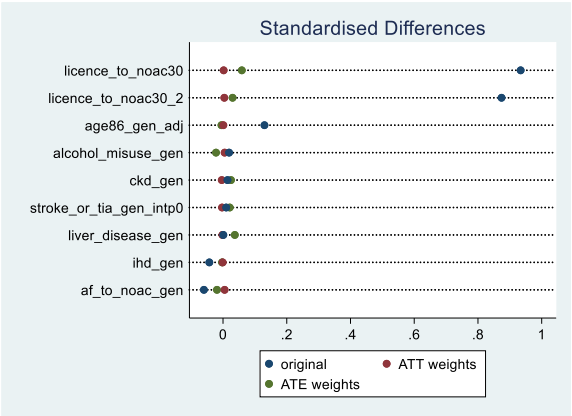
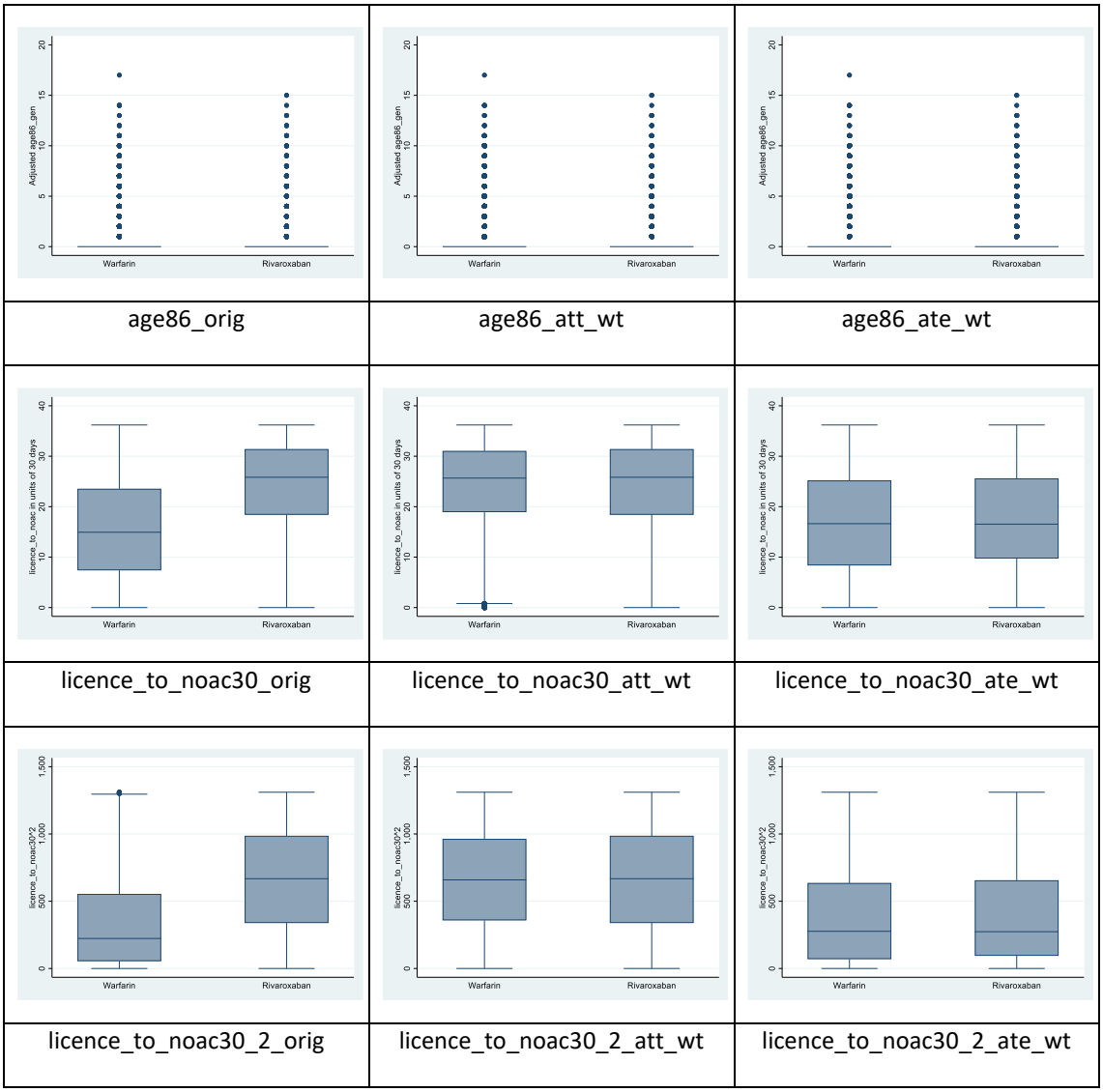


Figure B-3: Dot plot of standardised mean differences for the original data and that using the IPTW weights for ATT and ATE.



age86_orig	age86_att_wt	age86_ate_wt
licence_to_noac30_orig	licence_to_noac30_att_wt	licence_to_noac30_ate_wt
licence_to_noac30_2_orig	licence_to_noac30_2_att_wt	licence_to_noac30_2_ate_wt

Figure B-4: Plots of continuous variables in the PS model from the original data and with IPTW weights applied for ATT and ATE.

B-6.5 Balance checking for PS Stratification

The results of the balance checking following PS stratification are given in this section.

Table B-14: Standardised mean differences for the original data and that stratified on the PS with 5, 10 and 50 strata.

Number of Strata	Covariate	Mean in treated	Mean in untreated	Standardised diff.
Original Data	Previous stroke	0.14	0.14	0.010
	Alcohol misuse	0.03	0.03	0.019
	Chronic kidney disease	0.22	0.22	0.014
	Liver disease	0	0	0.001
	Ischemic heart disease	0.22	0.24	-0.043
	af_to_noac_gen†	0.44	0.47	-0.060
	=86 if age≤86, else =age	0.58	0.37	0.130
	licence_to_noac30††	24.32	15.79	0.934
	$(licence\_to\_noac30)^2$	660.82	346.86	0.874
Number of Strata	Covariate	Mean in treated	Mean in untreated	Standardised diff.
5 Strata	Previous stroke	0.14	0.14	0.009
	Alcohol misuse	0.03	0.03	0.004
	Chronic kidney disease	0.22	0.22	0.008
	Liver disease	0	0	0.003
	Ischemic heart disease	0.22	0.22	-0.006
	af_to_noac_gen†	0.44	0.45	-0.019
	=86 if age≤86, else =age	0.58	0.51	0.043
	licence_to_noac30††	24.32	23.90	0.046
	$(licence\_to\_noac30)^2$	660.82	641.67	0.053
Number of Strata	Covariate	Mean in treated	Mean in untreated	Standardised diff.
10 Strata	Previous stroke	0.14	0.14	0.004
	Alcohol misuse	0.03	0.03	-0.001
	Chronic kidney disease	0.22	0.22	0.004
	Liver disease	0	0	0
	Ischemic heart disease	0.22	0.22	-0.002
	af_to_noac_gen†	0.44	0.44	-0.006
	=86 if age≤86, else =age	0.58	0.53	0.029
	licence_to_noac30††	24.32	24.24	0.010
	$(licence\_to\_noac30)^2$	660.82	656.18	0.013
Number of Strata	Covariate	Mean in treated	Mean in untreated	Standardised diff.
50 Strata	Previous stroke	0.14	0.14	-0.001
	Alcohol misuse	0.03	0.03	-0.004
	Chronic kidney disease	0.22	0.22	0.002
	Liver disease	0	0	-0.002
	Ischemic heart disease	0.22	0.22	0.001
	af_to_noac_gen†	0.44	0.44	0

	=86 if age≤86, else =age	0.58	0.56	0.010
	licence_to_noac30††	24.32	24.30	0.002
	(licence_to_noac30)^2	660.82	659.25	0.004

†First NOAC/OAC prescription was ≤ 28 days of first AF diagnosis?
††The Rivaroxaban licence date to date of first NOAC/OAC prescription, in months

B-7 Outcome modelling

B-7.1 Outcome model – background and theory

The REWARD data were extracted to compare the effect of NOACs compared with Warfarin in the prevention of future stroke, the primary outcome. The outcome analysis was performed on time-to-event data, that is time to first stroke following the first NOAC/OAC prescription, using survival analysis methods.

The Cox model is a semi-parametric approach used to analyse survival data. The hazard function – for the ith patient at time t

$$h(t|X_i) = h_0(t)exp\left(X_i^T \underline{\beta_X}\right) \quad (1)$$

where $\underline{\beta_X}$ is the vector of coefficients estimated in the outcome model, and X_i , the baseline variables for the ith patient, X_i^T its transpose.

The Cox model requires the Proportional Hazards (PH) assumption to be valid, that is the ratio of the hazards for any two individuals is constant over time. The baseline hazard, $h_0(t)$, is not estimated and its form is not important. Comparing the ith patient to the mth patient

$$h(t|X_i)/h(t|X_m) = exp(X_i^T \underline{\beta_X})/exp(X_m^T \underline{\beta_X})$$

Cox regression was used for the analysis and to estimate the treatment effect. Different implementations of Cox regression were used to account for the PS conditioning used to address the systematic differences between the treatment groups.

B-7.2 Nature of the data

For PS matching, account had to be taken for the matched nature of the data when using running the survival analysis, here Cox regression (Austin, 2011). The options considered were frailty, marginal survival models and stratified survival models. Frailty, a measure of a participant's predisposition to stroke in addition to the values described in the regression model (O'Quigley & Stare, 2002), is a latent random effect used multiplicity in the hazard function and directly influences the outcome (Cleves, Gould & Marchenko, 2016). Frailty was not suitable as the PS is not an inherent trait, it can depend on the algorithm used. The marginal survival models method

is known by various names, Huber (Huber, 1967) and White (White, 1980), sandwich estimator of variance and robust estimator of variance. It gives an alternative method for estimating standard errors, although the point estimate will be the same as the standard method, but it was not selected to use. The method chosen was Cox regression stratified by matched pair or group. The matched nature of the data is accounted for by each group (or strata) taking a different baseline hazard (Cleves et al., 2016, p. 115) . Cox regression with stratification ran successfully in Stata.

$$h(t|x_j) = h_{01}(t)exp(x_j\beta_x) \text{ , if j is in group 1.}$$

$$h(t|x_j) = h_{02}(t)exp(x_j\beta_x) \text{ , if j is in group 2.}$$

When using IPTW, for both the ATE and ATT, variance estimates had to take account of the weighted nature of the data and robust variance estimation is commonly used. The weights generated by IPTW were used directly as options in the Cox regression. (The weights were used as Stata's *pweights* as arguments in the -stcox- command).

When using PS stratification, there are three methods to estimate the HR in the outcome analysis (Austin, 2013, 2014). First, the treatment effect is estimated using Cox PH within each stratum and these are pooled or averaged to give the ATE (pooled). Second, Cox PH is run with an indicator variable for treatment and strata as a categorical variable (adjusted). Third, regress on an indicator variable for treatment and stratify on the strata (stratified). This allows the baseline hazard to vary across strata. The *stratified* method was chosen for use in this study It was used with PS matching and allowed for a closer comparison of PS stratification and PS matching.

B-7.3 Outcome model

The outcome model was fitted to the analysis dataset used following PS matching. The CHA2DS2-VASc (Lip, Nieuwlaat, Pisters, Lane & Crijns, 2010) is a score which indicates the risk of stroke for patients with AF. As the study outcome is time to first stroke after first NOAC/OAC prescription, the variables for the outcome model were selected from the variables which contribute to the CHA2DS2-VASc score

Table B-15) unless they had been fully accounted for in the PS (the treatment allocation model). Expert clinical advice also suggested additional variables to be included in the outcome model, such as hypercholesterolemia, prescribed antiplatelets, statins or blood pressure lowering medication, the number of different medications being prescribed, smoking history (current and past) and the Townsend deprivation quintile (Table B-16).

Table B-15: CHA2DS2-VASc risk factors and the models in which they were accounted for.

Risk Factor	Value in CV Score	In PS model?	In Outcome model?
C ongestive heart failure/LV dysfunction	1	Y	Y
H ypertension	1		Y
A ge ≥ 75 y	2	*	Y
D iabetes mellitus	1		Y
S troke/TIA/TE	2	y	
V ascular disease (prior myocardial infarction, peripheral artery disease, or aortic plaque)	1		Y
A ge 65-74 y	1	*	Y
S ex C ategory (i.e. female gender)	1		Y

*age only represented by age86_gen, no further categorisation used.

Table B-16: Variables considered in the outcome model.

Variable	Reason for choice of variable
Treatment	Variable of interest
Sex	In CHA2DS2-VASc
Congestive cardiac failure	In CHA2DS2-VASc
Hypertension	In CHA2DS2-VASc
Diabetes	In CHA2DS2-VASc
Vascular disease	In CHA2DS2-VASc
Hypercholesterolemia	Expert Advice
On antiplatelet	Expert Advice
On statins	Expert Advice
On blood pressure lowering medication	Expert Advice
Townsend deprivation quintile	Expert Advice
Number of medications currently prescribed	Expert Advice
Smokes now	Expert Advice*
Smoked previously	Expert Advice*
Age ≥ 65	In CHA2DS2-VASc**
Age ≥75	In CHA2DS2-VASc**

*these variables were considered together

**these variables were considered together. The PS used a variable age86_gen these additional variables were used in the outcome modelling as they were used in the CHA2DS2-VASc.

There were two PS matching methods under consideration, 1:1 matching and 3:1 matching (Section B-6.2). These each generated using their own dataset and the outcome modelling was applied to each of these datasets. The selection of the best outcome model used the following options. All variables are those listed in Table B-16:

- All variables included with Townsend's deprivation quintile as a continuous variable.
- All variables included with Townsend's deprivation quintile as a factor variable.
- Treatment only.

- Treatment and each variable individually (univariate).
- Selection of models using combinations of the 'best' univariate variables and treatment.

The criteria used for the outcome model selection were:

- No high standard errors (SE). High SEs indicate the model was not converging well.
- The highest number of cases used. Some models drop cases when there was missing data in the variables used.
- The lowest BIC.
- A reasonable AIC and Log-Likelihood.

B-7.4 Outcome modelling options

The model assessment began by running the 'univariate' models, that is just one variable plus treatment in the model (Table B-17). Additional variables, age_65_gen and age_75_gen, were added to the model to match more closely the CHA2DS2-VASc scoring system. Some pairs of parameters were considered together. smoke_now_gen was retained and smoke_prev_gen dropped and both age_65_gen and age_75_gen were kept in the outcome model.

Table B-17: Results from the 'univariate' models, 1 variable plus treatment, sorted by p-value.

Variable	Coef.	SE	z	P>|z|	95% CI	Chads-vasc variable?
On blood pressure lowering medication	-1.099	0.292	-3.77	<0.001	(-1.671, -0.527)	
On statins	-0.837	0.280	-2.99	0.003	(-1.386, -0.288)	
On antiplatelet	-0.748	0.299	-2.50	0.012	(-1.334, -0.162)	
Hypercholesterolemia	-0.609	0.286	-2.13	0.033	(-1.170, -0.048)	
Hypertension	-0.473	0.264	-1.79	0.073	(-0.990, 0.045)	Y
Sex	0.407	0.252	1.61	0.107	(-0.087, 0.901)	Y
Age ≥75	0.725	0.517	1.40	0.161	(-0.288, 1.738)	Y
Diabetes	-0.442	0.352	-1.26	0.209	(-1.132, 0.248)	Y
Townsend deprivation quintile	-0.118	0.102	-1.16	0.246	(-0.317, 0.081)	
Smoked previously	0.269	0.267	1.01	0.315	(-0.255, 0.793)	
Vascular disease	0.279	0.379	0.74	0.462	(-0.464, 1.022)	Y
Number medications currently prescribed	-0.013	0.019	-0.70	0.485	(-0.050, 0.024)	
Congestive cardiac failure	-0.199	0.809	-0.25	0.805	(-1.786, 1.387)	Y
Age ≥ 65	-0.061	0.570	-0.11	0.914	(-1.178, 1.055)	Y
Smokes now	-0.049	0.615	-0.08	0.937	(-1.254, 1.157)	

Table B-17 identified four variables which may be regarded as significant with p-values <0.05, being prescribed blood pressure lowering medication, statins or antiplatelets and

hypercholesterolemia. Models were investigated which included these four variables, these four variables and the CHA2DS2-VASc score, and these four variables and the variables used in the CHA2DS2-VASc score. The models all included treatment but non-clinical and non-significant variables were excluded. The model assessment criteria is presented, the models are sorted by the number of outcomes (high is preferred) then by BIC (low is preferred).

Missing data, mostly in Townsend and smoking status, meant that the univariate models for townsend_only, smoking_only, and full_townsend_factors used fewer failures, so these models were disregarded. The model which used the best four variables and the CHA2DS2-VASc score was selected as the model to use. It was a good compromise; it did not lose cases due to missing data and so had the maximum number of outcome events, its BIC, the model selection criteria, was higher than other models, but it included more information as it included the CHA2DS2-VASc score. The CHA2DS2-VASc information was contained in a single variable so the degrees of freedom were lower than the model which used all the variables which contribute to the CHA2DS2-VASc. The chosen model is shown in full in Table B-18.

All the considered models showed treatment had a positive coefficient, so being prescribed the NOAC increases the risk of future stroke compared to prescribing Warfarin. Only the full model (with Townsend deprivation quintile as continuous) show this as significant. The coefficient varies from, 0.33 in some of the univariate models to 0.56 in the full model (Townsend deprivation quintile as continuous) and the standard errors show little variation, 0.23 to 0.29. These values are the log(hazard ratio).

Table B-18: The outcome model selected for use. The model includes treatment, the 4 most significant univariate variables and the CHA2DS2-VASc score.

Covariate	HR	SE of HR	95% CI of HR	Coeffi-cient*	SE of Coeffi-cient	95% CI of Coefficient	p-value
Treatment	1.534	0.383	(0.940, 2.504)	0.428	0.250	(-0.062, 0.918)	0.087
Prescribed blood pressure lowering meds	0.339	0.110	(0.180, 0.639)	-1.081	0.323	(-1.714, -0.448)	0.001
Prescribed statins	0.677	0.245	(0.333, 1.378)	-0.390	0.362	(-1.100, 0.321)	0.282
Prescribed antiplatelets	0.646	0.225	(0.326, 1.279)	-0.437	0.349	(-1.121, 0.246)	0.210
Hypercholesterolemia	0.729	0.269	(0.354, 1.502)	-0.316	0.369	(-1.039, 0.407)	0.391
CHA2DS2-VASc score	1.360	0.165	(1.073, 1.725)	0.308	0.121	(0.070, 0.545)	0.011

*Coefficient is the log(hazard ratio)

B-7.5 Baseline hazard function

The outcome of interest was time to first future stroke following the first NOAC/OAC prescription. When the outcome model was generated using the Cox model the baseline hazard $h_0(t)$ was not calculated. However, in the simulations, developed in Chapter 4, it was needed

to generate the simulated survival time. This can be done using a parametric survival model. Empirical investigations had suggested a Weibull model would be an appropriate baseline hazard function. A Weibull distribution offered flexibility, its parameters are γ, the shape parameter, and λ, the scale parameter. By varying the shape parameter, γ, the distribution of the function changes, for $\gamma = 1$ this distribution is an exponential so, the hazard is constant.

From *equation* (1) the hazard function – for the i[th] patient at time t

$$h(t|X_i) = h_0(t)exp(X_i^T \underline{\beta_X})$$

where $exp(X_i^t \underline{\beta_X})$ can be calculated from $\underline{\beta_X}$, the vector of coefficients estimated in the outcome model, and X_i , the baseline variables for the i[th] patient.

Using the Weibull model for the baseline hazard

$$h_0(t) = pt^{p-1} \exp(a)$$

where p and exp(a) are shape and scale parameters, respectively, and they were obtained empirically and comparing with the original dataset.

Hence

$$h(t|X_i) = pt^{p-1} exp(\beta_0)exp(X_i^T \underline{\beta_X})$$

The cumulative hazard is defined as

$$H(t|X) = \int_0^t h(u|X)du$$

And the survival function

$$S(t|X) = \exp[-H(t|X)]$$

Hence using the Weibull model for the baseline hazard

$$H(t|X_i) = \exp\left(\beta_0 + X_i{}^T \beta_X\right) t^p$$

$$S(t|X_i) = \exp\left[-\exp\left(\beta_0 + X_i{}^T \beta_X\right) t^p\right] \quad (2)$$

The survival times were generated using *equation* (2), which was censored at the end of the patient's baseline episode, and an indicator variable set with the event being future stroke within the time of observation.

B-7.6 Baseline hazard - method and results

The function of the baseline hazard was assessed empirically. The smoothed baseline hazard function was plotted from the data (Figure B-5). It showed that the values for the baseline hazard were small and that there was a slight increase around 350 days in the analysis time. To match this function a Mixture-Weibull function, implemented by a Stata user written function, was considered. However the function took too long to run and would have made running the simulations infeasible. A standard Weibull function gave a suitable match to the study data. Cox regression using a Weibull baseline hazard function was fitted to the data and gave baseline hazard parameters of $\lambda = 0.00029933$ and $\gamma = 0.480355$. Figure B-6 plots these values.

In the simulated datasets, for each participant their survival time was simulated using the chosen baseline hazard and other variables. The censoring variable was generated by comparing the survival time to the end of the episode (the observed time) for each participant. If the survival time was less than the episode length the patient was recorded as having a future stroke. If the survival time was longer than the episode length the patient was recorded as not having a future stroke.

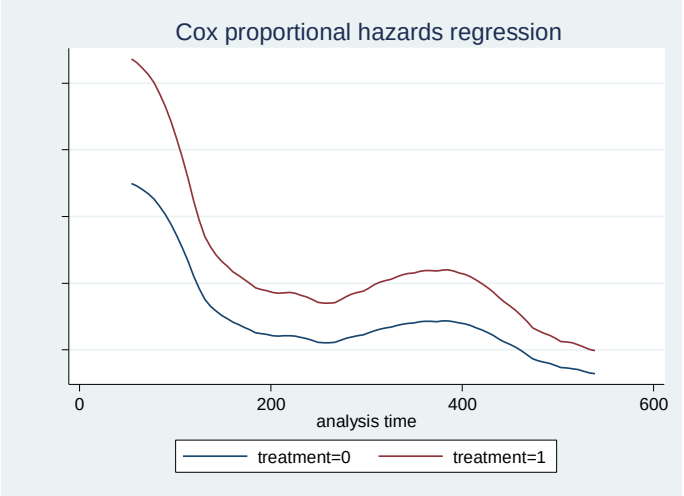


Figure B-5: Smoothed baseline hazard function – Analysis Time is in days.

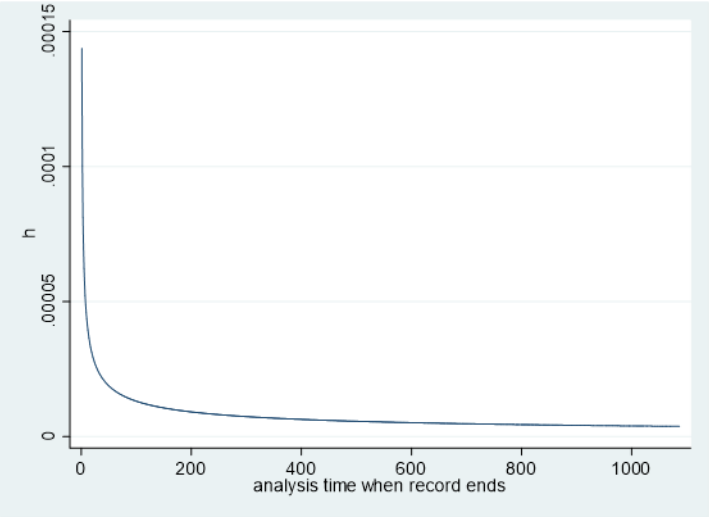


Figure B-6: Baseline hazard function generated using λ = 0.00029933 and γ = 0.480355.

References for Appendix B

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research, 46*(3), 399-424. doi:10.1080/00273171.2011.568786

Austin, P. C. (2013). The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in Medicine, 32*(16), 2837-2849. doi:10.1002/sim.5705

Austin, P. C. (2014). The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Statistics in Medicine, 33*(7), 1242-1258. doi:10.1002/sim.5984

Baczek, V. L., Chen, W. T., Kluger, J., & Coleman, C. I. (2012). Predictors of warfarin use in atrial fibrillation in the United States: a systematic review and meta-analysis. *Bmc Family Practice, 13*. doi:10.1186/1471-2296-13-5

Bhatia, G. S., & Lip, G. Y. (2004). Atrial fibrillation post-myocardial infarction: frequency, consequences, and management. *Current heart failure reports, 1*(4), 149-155.

Boriani, G., Laroche, C., Diemberger, I., Popescu, M. I., Rasmussen, L. H., Petrescu, L., . . . Lip, G. Y. H. (2016). Glomerular filtration rate in patients with atrial fibrillation and 1-year outcomes. *Scientific Reports, 6*. doi:10.1038/srep30271

Cleves, M., Gould, W. W., & Marchenko, Y. V. (2016). *An introduction to survival analysis using Stata* (Revised 3rd ed.): Stata press.

Garrido, M. M., Kelley, A. S., Paris, J., Roza, K., Meier, D. E., Morrison, R. S., & Aldridge, M. D. (2014). Methods for Constructing and Assessing Propensity Scores. *Health Services Research, 49*(5), 1701-1720. doi:10.1111/1475-6773.12182

Giralt-Steinhauer, E., Cuadrado-Godia, E., Ois, A., Jiménez-Conde, J., Rodríguez-Campello, A., Soriano, C., & Roquer, J. (2013). Comparison between CHADS 2 and CHA 2 DS 2-VAS c score in a stroke cohort with atrial fibrillation. *European journal of neurology, 20*(4), 623-628.

Hankey, G. J., Patel, M. R., Stevens, S. R., Becker, R. C., Breithardt, G., Carolei, A., . . . Comm, R. A. S. (2012). Rivaroxaban compared with warfarin in patients with atrial fibrillation and previous stroke or transient ischaemic attack: a subgroup analysis of ROCKET AF. *Lancet Neurology, 11*(4), 315-322. doi:10.1016/s1474-4422(12)70042-x

Herr, J. L., Drukker, D., Imbens, G. W., & Abadie, A. (n.d.). nnmatch: Nearest neighbor matching estimation for average treatment effects. Retrieved from http://www.stata-journal.com/software/sj4-3

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*(3), 199-236. doi:10.1093/pan/mpl013

Huber, P. J. (1967). *UNDER NONSTANDARD CONDITIONS.* Paper presented at the Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Weather modification.

Jann, B. (2017). 'KMATCH': module module for multivariate-distance and propensity-score matching, including entropy balancing, inverse probability weighting, (coarsened) exact matching, and regression adjustment. Retrieved from http://fmwww.bc.edu/RePEc/bocode/k

Lai, H. C., Chien, W. C., Chung, C. H., Lee, W. L., Wu, T. J., Wang, K. Y., . . . Liu, T. J. (2016). Atrial fibrillation, liver disease, antithrombotics and risk of cerebrovascular events: A population-based cohort study. *International Journal of Cardiology, 223*, 829-837. doi:10.1016/j.ijcard.2016.08.297

Lee, S., Monz, B. U., Clemens, A., Brueckmann, M., & Lip, G. Y. H. (2012). Representativeness of the dabigatran, apixaban and rivaroxaban clinical trial populations to real-world atrial fibrillation patients in the United Kingdom: a cross-sectional analysis using the General Practice Research Database. *Bmj Open, 2*(6). doi:10.1136/bmjopen-2012-001768

Leuven, E., & Sianesi, B. (2003). PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing (Version 4.0.12 30jan2016). Retrieved from http://ideas.repec.org/c/boc/bocode/s432001.html

Lip, G. Y. H., Nieuwlaat, R., Pisters, R., Lane, D. A., & Crijns, H. (2010). Refining Clinical Risk Stratification for Predicting Stroke and Thromboembolism in Atrial Fibrillation Using a Novel Risk Factor-Based Approach The Euro Heart Survey on Atrial Fibrillation. *Chest, 137*(2), 263-272. doi:10.1378/chest.09-1584

O'Caoimh, R., Igras, E., Ramesh, A., Power, B., O'Connor, K., & Liston, R. (2017). ASSESSING THE APPROPRIATENESS OF ORAL ANTICOAGULATION FOR ATRIAL FIBRILLATION IN ADVANCED FRAILTY: USE OF STROKE AND BLEEDING RISK-PREDICTION MODELS. *Journal of Frailty & Aging, 6*(1), 46-52. Retrieved from <Go to ISI>://WOS:000449832100009

O'Quigley, J., & Stare, J. (2002). Proportional hazards models with frailties and random effects. *Statistics in Medicine, 21*(21), 3219-3233. doi:10.1002/sim.1259

Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology, 53*(5), 793-808. doi:10.1080/10635150490522304

Toso, V. (2014). Recommendations for the use of new oral anticoagulants (NOACs) after TIA or stroke caused by atrial fibrillation (AF), after a consensus conference among Italian neurologists (the Venice group). *Neurological Sciences, 35*(5), 723-727. doi:10.1007/s10072-013-1590-7

White, H. (1980). A heteroskedasticity-consistent covariance-matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48*(4), 817-838. doi:10.2307/1912934

Wolff, A., Shantsila, E., Lip, G. Y. H., & Lane, D. A. (2015). Impact of advanced age on management and prognosis in atrial fibrillation: insights from a population-based study in general practice. *Age and Ageing, 44*(5), 874-878. doi:10.1093/ageing/afv071

APPENDIX C – ADDITIONAL INFORMATION FOR SIMULATIONS

C-1 Sample size calculations for the simulations

The initial simulation runs reported in Section 4.3 had been made using 100 simulated datasets and were used to demonstrate the performance of the simulation method and functionality. The sample size of 100 datasets was arbitrary. This section explores the number of simulated datasets which should be used, based on the precision sample size calculation. The number of simulated datasets is regarded as the 'sample size'.

The sample size was determined by calculating CI widths of the mean treatment effect estimate from some additional simulations using 1,000 datasets, determining an acceptable CI, calculating the number of simulations required to give the acceptable CI and then using this number of simulated datasets (the sample size) in the study simulations.

Precision sample size calculation

The precision sample size calculation uses:

$SE = \frac{SD}{\sqrt{N}}$ (Altman, 1991, p. 154) where SE is the Standard Error, SD is the Standard Deviation and N the sample size, here the number of simulated datasets.

The population variance was not known and had to be estimated from the data, using the simulations with N=1000. In cases where the population variance has to be approximated from the sample, the t-distribution, with $n - 1$ degrees of freedom is used to calculate the Confidence Intervals (CI). However, as n was large, the Normal distribution is a good approximation to the t-distribution. Also, the Central Limits Theorem states that regardless of the distribution of the variable in the population, the distribution of the sample means will be nearly Normal providing the samples are large enough (Altman, 1991, p. 154). Altman (1991, p. 181) suggests that n>=100 is large enough for these conditions to hold. Therefore, the normal distribution was used and the formula to calculate the 95% CIs was $CIs = \hat{\theta} \pm 1.96 * SE$ where $\hat{\theta}$ is the mean of θ_i

So N, the number of simulations (or sample size), is chosen to give the required CI width. The SD from earlier simulation runs was used.

Calculation of the CI widths from the existing simulations

The results from the simulations with no change of effect size of the variable with measurement error, using N=1000 generated datasets were used to calculate their CI widths. Only the runs using -50% measurement error and +50% measurement error were used. Generally, the -50% measurement error run had the highest SD and the +50% run had the lowest SD, these runs gave

the extreme values of the SD in each simulation and hence the extreme values of the CI width. All measurement errors related to the log(HR) of the treatment effect estimate when comparing RI against WA.

Table C-1: 95% CI widths from N=1000 simulations for the different PS methods used with prevalences of 0.5%, 1% and 10%.

Prevalence	**0.5%**	**0.5%**	**1%**	**1%**	**10%**	**10%**
% M error	**-50%**	**50%**	**-50%**	**50%**	**-50%**	**50%**
IPTW for ATE	0.0496	0.0381	0.0346	0.0261	0.0108	0.0088
IPTW for ATT	0.0393	0.0303	0.0266	0.0209	0.0082	0.0065
3to1 Matching	0.0614	0.0418	0.0371	0.0286	0.0114	0.0088
PS Stratification	0.0386	0.0299	0.0262	0.0205	0.0081	0.0064

The CI widths from the 10% prevalence runs were much smaller than the 0.5% and 1% prevalence runs in the same simulation set, but there was little difference between the 0.5% and 1% prevalence runs in the same simulations set. There was not a large difference between the -50% and +50% measurement error runs in the same simulation. This is a reflection that the power in time-to-event data is calculated using the number of outcomes (Altman, 1991, p. 393; Guo & Fraser, 2015, p. 352), as opposed to the number of participants in data with binary or a continuous outcome. In this study the higher prevalence simulations, by definition, have a higher number of outcomes (hence a smaller CI). The DGM used generated slightly more outcomes in the data for +50% measurement error than for -50% measurement error (hence a slightly smaller CI). The PS methods which had the lowest SD, and hence lowest SE, IPTW for ATT and PS Stratification, had the lowest CI widths.

Determining the required CI width

These interpretations were based on the information available before the full simulations were run, that is example simulation runs using 1000 simulated dataset and varying the outcome prevalence. IPTW for ATE was chosen as it neither had the largest nor the lowest difference between the mean treatment effect estimates from the different prevalences. A visual inspection showed the CI widths for the 10% prevalence run were <0.01, Table C-1, which did not include the mean treatment effect estimate from any of the other prevalences used, meaning the other plots of the treatment effect estimates were easily separated with this value of the CI. For the 1% prevalence runs, the CI widths were 0.03, which did not include any other prevalence runs. The CI width for the 0.5% prevalence run, at +50% was 0.04 and this did include the other two prevalence runs. To summarise, a CI of 0.01 was deemed good, a CI between 0.02 and 0.03 was thought to be acceptable, and a CI of 0.04 and above was thought to be too high.

It is acknowledged that the interpretation of these values is subjective and that the treatment effect estimate means generated in the full simulations (from varying other parameters) were not known.

An alternative way of assessing an acceptable CI was to consider 10% of the true value of the treatment effect, 0.3674, giving 0.0367 as an acceptable CI, when applied to the 1% prevalence simulations. The original data had a prevalence of approximately 1%. Combining this information and the visual inspection, an acceptable CI width of 0.035 was decided upon.

Calculating the number of simulations (sample size) from given CI

By fixing the CI width at 0.02, 0.03, 0.035 and 0.04 the sample size, the number of simulations required N, were calculated to achieve these, Table 10. Only the -50% measurement error runs are displayed, which have the highest SD. Table C-1 showed that there was little difference between the CI widths for the simulations with different amounts of measurement error for the same prevalence. The results using the SD from the N=1000 runs are displayed.

Table C-2: Calculated Sample Size for fixed CI widths.

PS_Method	Prev-alence	Mean	SD of mean	N for CI width = 0.02	N for CI width = 0.03	N for CI width = 0.035	N for CI width = 0.04
IPTW_ATE	0.5%	0.3165	0.4001	6150	2733	2008	1537
IPTW_ATE	1%	0.3483	0.2792	2995	1331	978	749
IPTW_ATE	10%	0.3626	0.0872	292	130	95	73
IPTW_ATT	0.5%	0.3492	0.3174	3870	1720	1264	968
IPTW_ATT	1%	0.3560	0.2142	1763	783	576	441
IPTW_ATT	10%	0.3639	0.0664	169	75	55	42
3to1_match	0.5%	0.4684	0.4956	9436	4194	3081	2359
3to1_match	1%	0.4256	0.2994	3444	1530	1124	861
3to1_match	10%	0.3716	0.0918	324	144	106	81
PS_strat	0.5%	0.3477	0.3117	3732	1659	1219	933
PS_strat	1%	0.3575	0.2113	1715	762	560	429
PS_strat	10%	0.3643	0.0656	165	73	54	41

For a CI width of 0.035, the required sample size for each prevalence was: for 10% prevalence, the lowest was 54 (PS stratification) and the highest was 106 (3to1 PS matching); for 1% prevalence, lowest was 560 (PS stratification) and the highest was 1124 (3to1 PS matching); for 0.5% prevalence – lowest was 1219 (PS stratification) and the highest was 3081 (3to1 PS

matching). Generally, the sample sizes were rounded up to the next 100 for the simulations runs presented in Chapter 5.

References for Appendix C

Altman, D. G. (1991). *Practical statistics for medical research* London: Chapman and Hall/CRC.

Guo, S., & Fraser, M. W. (2015). *Propensity score analysis: Statistical methods and applications* (2nd ed.): SAGE publications.

APPENDIX D – STATA CODING

This Appendix includes samples of the Stata .do file code used for running the simulations. It includes ONLY the code for salient parts of the process. There are no program headers, no arguments passed to the code, no verification of the arguments and no error checking.

Due to the formatting required for the thesis, some of the commands in the coding wrap over more than one line.

This code will not run in this format, it is for example purposes only.

```
*Samples of the Stata coding for running the simulations
*It includes ONLY the salient parts of the process
*NOTE - this coding will not run in this format.
*There are no program headers, there are no arguments passed to the code,
*there is no verification of the arguments and no error checking.
*This coding is for example purposes only
*
*Jane Burnell - September 2021
**************************************

*an example call
log_local_ext EXT APP_PS_strat_MED_R1A_M50 300 m 50 stratification inf_chng
treatment strata1 PS4_86this stratified stcox_best4 Interim_Log seed_append
0.367433 new_id temp 0 local streg W 0.00015 0.480355 10 1.0
*the arguments from log_local_ext()
args level main_tag num action pcent ps_method inflnce treat strata PS_score
outcome_call output_model main_log seed_type true_mean sorting location
zero_val mlog_loc out_gen_model out_dist lambda1 gamma1 v1 v2

*log_local_ext calls
emain()
*does argument validation
*creates the folder to save results for this simulation run
*creates summary_log file - stores results from all simulated datasets
*creates seed_log - where Stata's RNG is stored after each dataset is analysed

*loop to run simulations using `num' datasets
forvalues j = 1(1)`num'{
       estep2()
       *generates a bootstrapped sample from the original dataset
       *creates the new variable with added measurement error -
       *"stroke_or_tia_gen_int`action'`pcent'"
       if (`pc_val'==0) {
              gen stroke_or_tia_gen_int`action'`pcent' = stroke_or_tia_gen
```

```
        *add dummy call for no measurement error
        gen temp_rbinomial = rbinomial(1,0.5)
}
else {
        gen temp_rbinomial = rbinomial(1,`pc_val')
        capture drop stroke_or_tia_gen_int`action'`pcent'
        gen stroke_or_tia_gen_int`action'`pcent' = stroke_or_tia_gen
        if ("`action'" == "p" | "`action'" == "P") {
                *add extra strokes
                replace stroke_or_tia_gen_int`action'`pcent' = 1 if
                (stroke_or_tia_gen == 0 & temp_rbinomial==1)
        }
        else if ("`action'" == "m" | "`action'" == "M") {
                *remove some strokes
                replace stroke_or_tia_gen_int`action'`pcent' = 0 if
                (stroke_or_tia_gen == 1 & temp_rbinomial==1)
        }
}
*END estep2()
```

estep3a()

```
*generates the PS values using
        *1)the PS model fitted to the original data
        *2)previous stroke variable with added measurement error
        *3)any changes to its "effect size"
        capture drop temp_`PS_score'_`action'`pcent'

if ("`inflnce'" == "inf_std") {
        *the original model - but using stroke_adjusted
        * with original coeff for prev_stroke
        gen temp_`PS_score'_`action'`pcent' =
        0.1229108*stroke_or_tia_gen_int`action'`pcent' + ///
        0.0975508*alcohol_misuse_gen + ///
        0.0083667*ckd_gen + 0.0326479*liver_disease_gen -
        0.0816038*ihd_gen ///
        -0.1923699*af_to_noac_gen + 0.077417*age86_gen_adj +
        0.1529995*licence_to_noac30 ///
        -0.0014545*c.licence_to_noac30_2 -4.172514
}
else if ("`inflnce'" == "inf_chng") {
        *the original model - but using stroke_adjusted
        *with changed coeff for previous_stroke
        gen temp_`PS_score'_`action'`pcent' =
        `v2'*stroke_or_tia_gen_int`action'`pcent' + ///
        0.0975508*alcohol_misuse_gen + ///
```

```
        0.0083667*ckd_gen + 0.0326479*liver_disease_gen -
        0.0816038*ihd_gen ///
        -0.1923699*af_to_noac_gen + 0.077417*age86_gen_adj +
        0.1529995*licence_to_noac30 ///
        -0.0014545*c.licence_to_noac30_2 -4.172514
}
*generate the PS value
capture drop `PS_score'_`action'`pcent'
gen `PS_score'_`action'`pcent' = 1/(1 + exp(-
1*(temp_`PS_score'_`action'`pcent')))
label var `PS_score'_`action'`pcent' "Propensity Score for LR Model4
with age86_adj for this dataset `action'`pcent'- Manual calc"

*cv_score calculated with previous stroke original and with measurement
*error
*END estep3a()
```

estep10()

```
*generates the generated treatment - based on the patient's PS value
capture drop treat_gen
gen treat_gen = rbinomial(1,`PS_score'_`action'`pcent')
label var treat_gen "Generated Treatment with `action'`pcent'"

*generates the outcome variables
        *these are - 1)time to next stroke and 2)did the event happen
        *before the enddate
        *they are based on the generated treatment and Weibull parameters
        *for the outcome prevalence
survsim stime_gen, distribution(exponential)lambdas(`lambda1') ///
covariates(treat_gen 0.2775938 on_cvd_bp_lowering -0.3989188
on_cvd_statin -0.2432114 ///
on_cvd_antiplatelet 0.3181043 hypercholesterol_gen -0.0687236 ///
chads2_vasc_calculated_int`action'`pcent' 0.1658724)
*END estep10()

*PS conditioning - PS method determines which branch to follow
if ("`ps_method'" == "3to1match") {
        *for 3to1 PS matching
        erun_simulations_gen()
        *creates 3to1 matched dataset
        psmatch2 `treat', pscore(`PS_score') neighbor(3)
        set more off
        keep if !missing(_weight)
        *and saves it to the working folder
        *END erun_simulations_gen()
```

```
        eexpand_dataset()
        *expands the matched dataset - so there is a record for each time
        *a case is used
              **CODING IS VERBOSE SO NOT INCLUDED HERE
        *END eexpand_dataset()

        int_out_stcox_best4()
        *runs the outcome analysis for PS matching with specified options
        stset s_enddate_gen, failure(future_stroke_gen)
        origin(first_noac_date)
        stcox `treat' on_cvd_bp_lowering on_cvd_statin
        on_cvd_antiplatelet hypercholesterol_gen
        chads2_vasc_calculated_orig, strata(`strata') nohr
        *writes results to summary_log
        *END int_out_stcox_best4()
}
else if ("`ps_method'" == "weight") {
        *for IPTW for ATE & IPTW for ATT
        erun_simulations_iptw()
        *generates PS weights
        propwt `treat' `PS_score' , `this_w'
        *END erun_simulations_iptw()

        int_out_stcox_best4_iptw()
        *runs the outcome analysis for IPTW with specified options
        stset s_enddate_gen [pweight=`this_w'_wt],
        failure(future_stroke_gen) origin(first_noac_date)
        stcox `treat' on_cvd_bp_lowering on_cvd_statin
        on_cvd_antiplatelet hypercholesterol_gen
        chads2_vasc_calculated_orig, nohr vce(robust)
        *writes results to summary_log
        *END int_out_stcox_best4_iptw()
}
else if ("`ps_method'" == "stratification") {
        *for PS stratification
        erun_simulations_strat()
        *generates strata using PS
        capture drop strat`nstrat'
        xtile strat`nstrat'=`PS_score'_`action'`pcent', n(`nstrat')
        *END erun_simulations_strat()

        int_out_stcox_best4_strat()
        *runs the outcome analysis for PS stratification with specified
        *options
```

```
            stset s_enddate_gen, failure(future_stroke_gen)
            origin(first_noac_date)
            stcox `treat' on_cvd_bp_lowering on_cvd_statin
            on_cvd_antiplatelet hypercholesterol_gen
            chads2_vasc_calculated_orig, strata(`nstrat') nohr
            *writes results to summary_log
            *END int_out_stcox_best4_strat()
      }
      *deletes the working dataset as the results have been stored in
      *summary_log
}
      *all simulated datasets have been analysed and results collected
      *closes the summary_log & seed_log files

      erun_int_calc()
      *generates the treatment effect estimate performance measures
      *from the results from all the datasets in summary_log
            *get mean
            quietly su v3
            local v3mean=r(mean)
            *also capture the number on non-missing observations
            local valid_n=r(N)

            *v4 is now the Model SE, so change temp names
            capture drop v3a_temp
            gen v3a_temp=(v3 - `v3mean')^2

            *get SD - is the SD but still called se_calc
            quietly su v3a_temp
            local se_calc=sqrt(r(sum)/(r(N)-1))

            *get bias as used in Austin(2013)
            capture drop v3b_temp
            gen v3b_temp=(v3 - `true_mean')
            quietly su v3b_temp
            local bias_calc=r(sum)/(r(N))

            *MSE - mean error squared in Austin(2013)
            capture drop v3c_temp
            gen v3c_temp=v3b_temp^2
            quietly su v3c_temp
            local mse_calc=r(sum)/(r(N))

            *Model SE
            capture drop v4a_temp
```

```
            gen v4a_temp=v4^2
            capture drop sum_v4a_temp
            egen sum_v4a_temp=sum(v4a_temp)
            local tot_v4a = sum_v4a_temp[1]
            local model_se = sqrt((1/`se_n')*`tot_v4a')

      *writes the performance measures of the simulation run to the main
      *logfile
      *END erun_int_calc()

*END emain()
```

APPENDIX E – TABLES AND GRAPHS VARYING OUTCOME PREVALENCE

E-1 Example simulations using N=100

In the headers in all tables in this Appendix, *Prevalence* is Outcome Prevalence, *Num Events* is the Number of Future Strokes, *Num WA* is the Number of participants prescribed Warfarin and *Num RI* is the Number of participants prescribed Rivaroxaban.

Table E-1: Simulation runs using IPTW for ATE, N=100.

Preva-lence	N	% Adj	Mean*	SE*	Bias*	MSE*	MSE % change	Model SE mean	Num events	num WA	Num RI	N valid
0.5%	100	-50	0.2972	0.3979	-0.0710	0.1633	-9.4	0.3652	104.6	18363.8	2895.2	100
0.5%	100	-40	0.2915	0.3995	-0.0767	0.1654	-10.8	0.3627	106.0	18359.9	2899.1	100
0.5%	100	-30	0.2963	0.3940	-0.0719	0.1603	-7.4	0.3602	107.7	18356.1	2902.9	100
0.5%	100	-20	0.2906	0.3927	-0.0776	0.1602	-7.3	0.3579	109.3	18352.3	2906.7	100
0.5%	100	-10	0.2942	0.3839	-0.0740	0.1528	-2.4	0.3560	110.9	18348.4	2910.6	100
0.5%	100	0	0.2962	0.3797	-0.0720	0.1493	0.0	0.3544	112.7	18344.8	2914.2	100
0.5%	100	10	0.2895	0.3752	-0.0787	0.1469	1.6	0.3407	121.1	18319.7	2939.3	100
0.5%	100	20	0.3049	0.3474	-0.0632	0.1247	16.5	0.3289	130.5	18295.1	2963.9	100
0.5%	100	30	0.3158	0.3318	-0.0522	0.1128	24.4	0.3185	140.2	18270.5	2988.5	100
0.5%	100	40	0.3084	0.3241	-0.0597	0.1086	27.3	0.3077	149.8	18245.1	3013.9	100
0.5%	100	50	0.3059	0.3151	-0.0621	0.1031	31.0	0.2974	159.1	18220.2	3038.8	100
1%	100	-50	0.3423	0.2657	-0.0253	0.0712	-17.3	0.2659	207.3	18363.8	2895.2	100
1%	100	-40	0.3380	0.2648	-0.0297	0.0710	-17.0	0.2643	210.4	18359.9	2899.1	100
1%	100	-30	0.3406	0.2572	-0.0271	0.0669	-10.1	0.2629	213.3	18356.1	2902.9	100
1%	100	-20	0.3379	0.2547	-0.0299	0.0658	-8.3	0.2614	216.5	18352.3	2906.7	100
1%	100	-10	0.3373	0.2461	-0.0304	0.0615	-1.3	0.2597	219.5	18348.4	2910.6	100
1%	100	0	0.3361	0.2443	-0.0317	0.0607	-0.0	0.2576	222.4	18344.8	2914.2	100
1%	100	10	0.3353	0.2278	-0.0324	0.0529	12.8	0.2465	240.6	18319.7	2939.3	100
1%	100	20	0.3321	0.2099	-0.0357	0.0453	25.4	0.2374	258.5	18295.1	2963.9	100
1%	100	30	0.3315	0.2068	-0.0363	0.0441	27.4	0.2289	276.6	18270.5	2988.5	100
1%	100	40	0.3295	0.1900	-0.0383	0.0376	38.1	0.2206	295.1	18245.1	3013.9	100
1%	100	50	0.3288	0.1940	-0.0390	0.0392	35.5	0.2131	313.0	18220.2	3038.8	100
10%	100	-50	0.3654	0.0876	-0.0021	0.0077	-0.1	0.0872	2063.4	18363.8	2895.2	100
10%	100	-40	0.3641	0.0880	-0.0034	0.0078	-1.2	0.0866	2089.7	18359.9	2899.1	100
10%	100	-30	0.3631	0.0886	-0.0044	0.0079	-2.6	0.0860	2116.7	18356.1	2902.9	100
10%	100	-20	0.3639	0.0897	-0.0036	0.0081	-5.0	0.0852	2144.8	18352.3	2906.7	100
10%	100	-10	0.3638	0.0883	-0.0037	0.0078	-1.9	0.0847	2173.0	18348.4	2910.6	100
10%	100	0	0.3645	0.0875	-0.0029	0.0077	-0.0	0.0840	2199.8	18344.8	2914.2	100
10%	100	10	0.3604	0.0851	-0.0071	0.0073	5.0	0.0808	2363.1	18319.7	2939.3	100
10%	100	20	0.3600	0.0819	-0.0075	0.0068	11.9	0.0777	2526.6	18295.1	2963.9	100
10%	100	30	0.3603	0.0765	-0.0072	0.0059	22.9	0.0751	2686.9	18270.5	2988.5	100
10%	100	40	0.3578	0.0763	-0.0097	0.0059	22.8	0.0728	2851.4	18245.1	3013.9	100
10%	100	50	0.3558	0.0722	-0.0117	0.0053	30.2	0.0706	3013.4	18220.2	3038.8	100

*displayed as the log(HR)

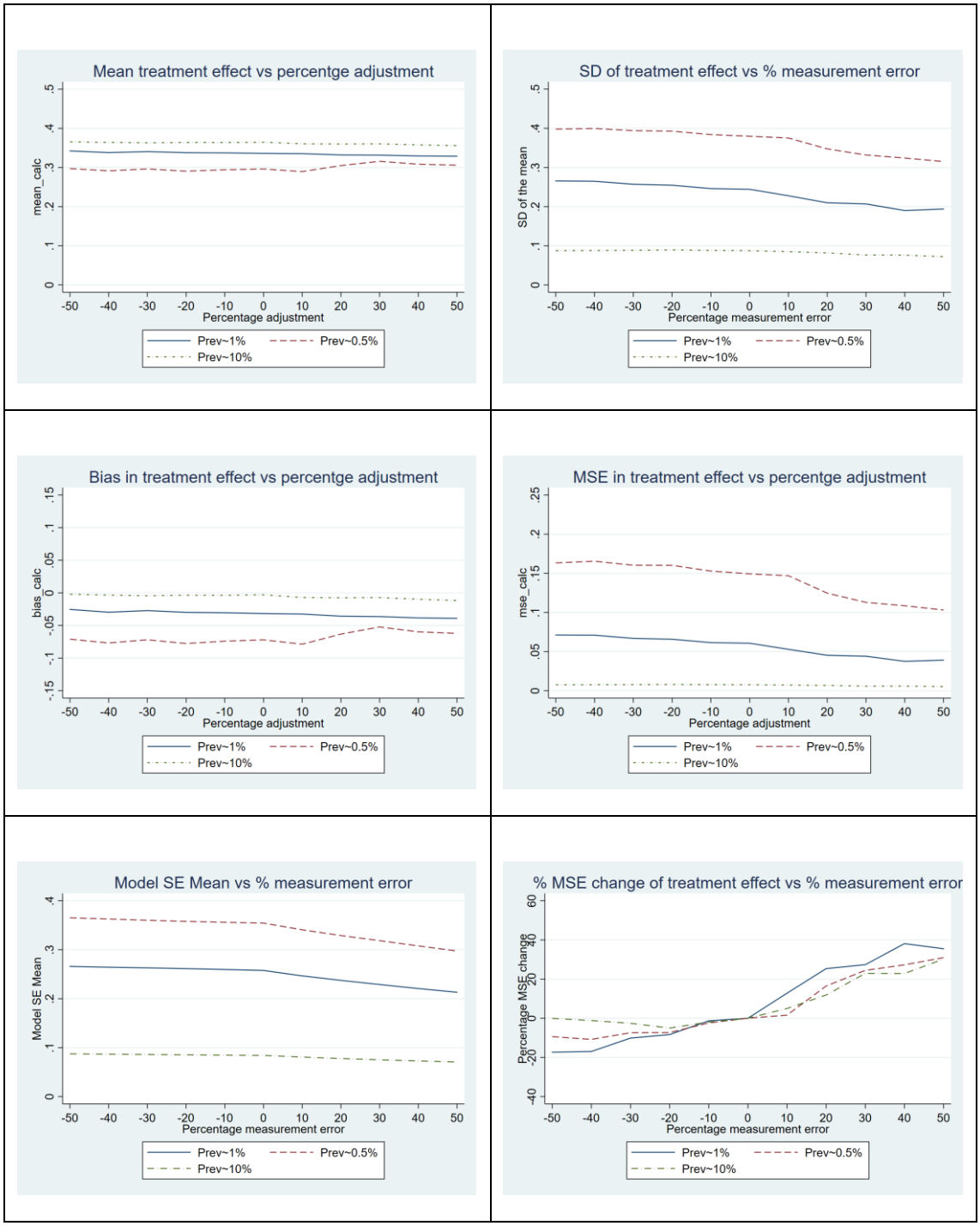


Figure E-1: Using IPTW for ATE, N=100, the mean, SE, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean.

Table E-2: Simulation runs using IPTW for ATT, N=100.

Preva-lence	N	% Adj	Mean*	SE*	Bias*	MSE*	MSE % change	Model SE mean	Num events	num WA	Num RI	N valid
0.5%	100	-50	0.3579	0.3191	-0.0097	0.1019	-15.1	0.3015	84.4	18363.8	2895.2	100
0.5%	100	-40	0.3552	0.3152	-0.0124	0.0995	-12.4	0.2993	85.7	18359.9	2899.1	100
0.5%	100	-30	0.3559	0.3071	-0.0116	0.0944	-6.7	0.2964	87.1	18356.1	2902.9	100
0.5%	100	-20	0.3537	0.3022	-0.0139	0.0915	-3.4	0.2941	88.6	18352.3	2906.7	100
0.5%	100	-10	0.3495	0.2980	-0.0182	0.0891	-0.6	0.2919	89.9	18348.4	2910.6	100
0.5%	100	0	0.3489	0.2970	-0.0187	0.0885	-0.0	0.2894	91.5	18344.8	2914.2	100
0.5%	100	10	0.3433	0.2969	-0.0244	0.0887	-0.2	0.2779	99.1	18319.7	2939.3	100
0.5%	100	20	0.3511	0.2730	-0.0165	0.0748	15.5	0.2660	106.7	18295.1	2963.9	100
0.5%	100	30	0.3529	0.2682	-0.0147	0.0722	18.5	0.2560	114.3	18270.5	2988.5	100
0.5%	100	40	0.3436	0.2654	-0.0241	0.0710	19.8	0.2473	122.6	18245.1	3013.9	100
0.5%	100	50	0.3442	0.2617	-0.0234	0.0690	22.1	0.2390	130.2	18220.2	3038.8	100
1%	100	-50	0.3743	0.2238	0.0069	0.0502	-7.0	0.2111	167.8	18363.8	2895.2	100
1%	100	-40	0.3689	0.2240	0.0015	0.0502	-7.1	0.2098	170.6	18359.9	2899.1	100
1%	100	-30	0.3702	0.2252	0.0028	0.0507	-8.2	0.2082	172.8	18356.1	2902.9	100
1%	100	-20	0.3662	0.2241	-0.0012	0.0502	-7.2	0.2068	175.4	18352.3	2906.7	100
1%	100	-10	0.3640	0.2171	-0.0035	0.0471	-0.6	0.2052	178.0	18348.4	2910.6	100
1%	100	0	0.3653	0.2165	-0.0022	0.0469	0.0	0.2037	180.4	18344.8	2914.2	100
1%	100	10	0.3638	0.2072	-0.0037	0.0430	8.3	0.1948	196.3	18319.7	2939.3	100
1%	100	20	0.3634	0.1881	-0.0040	0.0354	24.5	0.1871	211.2	18295.1	2963.9	100
1%	100	30	0.3650	0.1857	-0.0025	0.0345	26.4	0.1801	226.1	18270.5	2988.5	100
1%	100	40	0.3613	0.1795	-0.0062	0.0323	31.2	0.1740	241.8	18245.1	3013.9	100
1%	100	50	0.3617	0.1731	-0.0058	0.0300	36.0	0.1683	256.8	18220.2	3038.8	100
10%	100	-50	0.3669	0.0588	-0.0005	0.0035	3.9	0.0666	1694.0	18363.8	2895.2	100
10%	100	-40	0.3651	0.0589	-0.0023	0.0035	3.5	0.0661	1718.3	18359.9	2899.1	100
10%	100	-30	0.3644	0.0605	-0.0030	0.0037	-1.8	0.0657	1742.1	18356.1	2902.9	100
10%	100	-20	0.3638	0.0599	-0.0037	0.0036	0.2	0.0652	1767.6	18352.3	2906.7	100
10%	100	-10	0.3631	0.0590	-0.0043	0.0035	2.7	0.0648	1792.2	18348.4	2910.6	100
10%	100	0	0.3632	0.0599	-0.0042	0.0036	-0.0	0.0643	1816.1	18344.8	2914.2	100
10%	100	10	0.3603	0.0585	-0.0072	0.0035	3.7	0.0618	1964.6	18319.7	2939.3	100
10%	100	20	0.3592	0.0552	-0.0083	0.0031	13.7	0.0596	2110.1	18295.1	2963.9	100
10%	100	30	0.3605	0.0517	-0.0070	0.0027	24.4	0.0576	2250.1	18270.5	2988.5	100
10%	100	40	0.3572	0.0532	-0.0104	0.0029	18.6	0.0558	2392.0	18245.1	3013.9	100
10%	100	50	0.3561	0.0502	-0.0114	0.0026	26.6	0.0541	2529.7	18220.2	3038.8	100

*displayed as the log(HR)

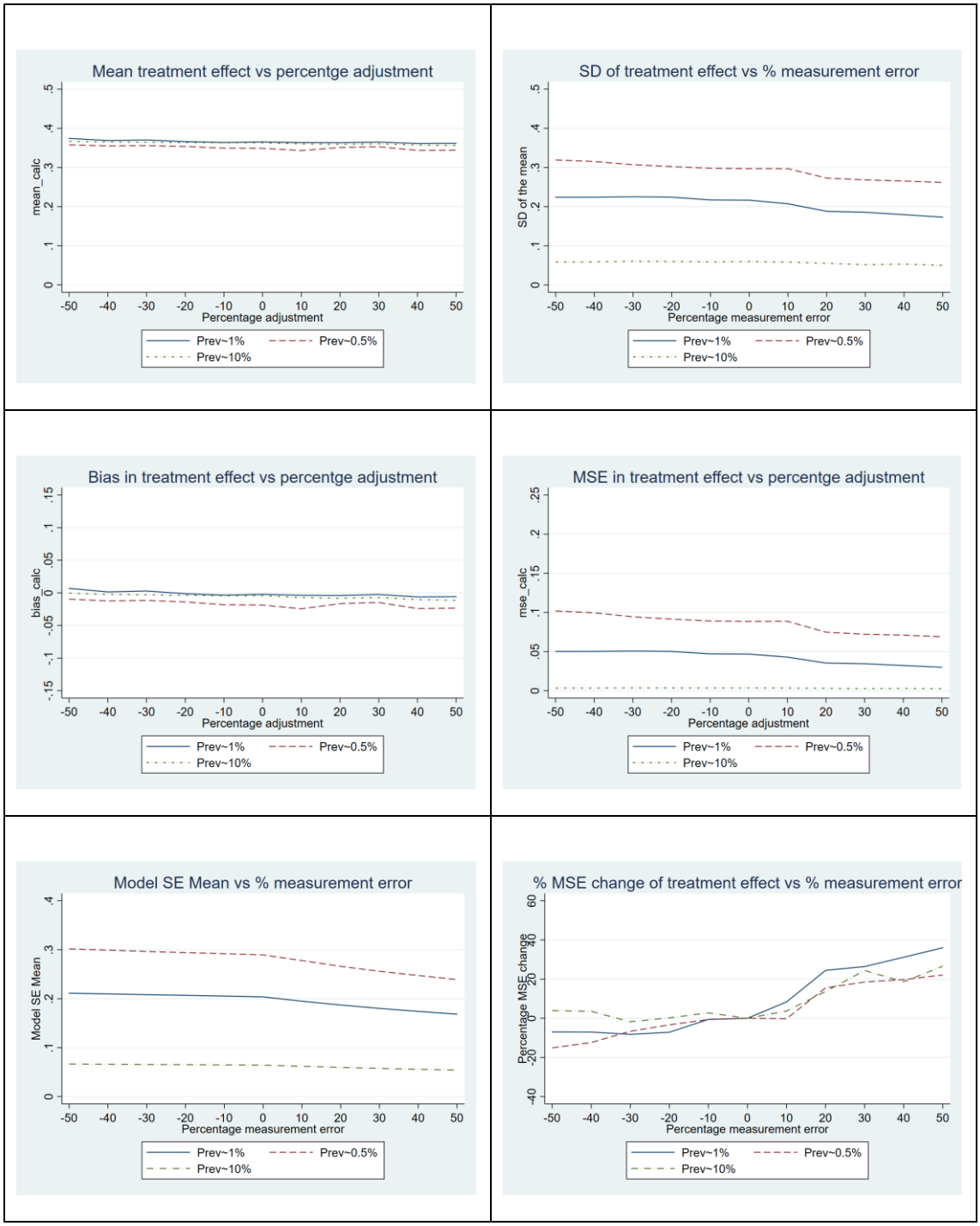


Figure E-2: Using IPTW for ATT, N=100, the mean, SE, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean.

Table E-3: Simulation runs using 3to1 PS matching, N=100.

Preva-lence	N	% Adj	Mean*	SE*	Bias*	MSE*	MSE % change	Model SE mean	Num events	num WA	Num RI	N valid
0.5%	100	-50	0.4714	0.4972	0.1050	0.2581	-59.2	0.3943	46.1	8689.6	2896.6	100
0.5%	100	-40	0.4468	0.4486	0.0802	0.2076	-28.0	0.3861	47.2	8701.1	2900.4	100
0.5%	100	-30	0.4611	0.4467	0.0946	0.2084	-28.5	0.3816	48.0	8713.2	2904.4	100
0.5%	100	-20	0.4315	0.4209	0.0647	0.1813	-11.8	0.3723	48.8	8724.5	2908.2	100
0.5%	100	-10	0.4337	0.3933	0.0669	0.1591	1.8	0.3668	49.8	8735.7	2911.9	100
0.5%	100	0	0.4215	0.3989	0.0546	0.1621	-0.0	0.3633	50.8	8748.1	2916.1	100
0.5%	100	10	0.4281	0.3846	0.0613	0.1516	6.5	0.3420	55.8	8822.9	2941.0	100
0.5%	100	20	0.4132	0.3746	0.0462	0.1425	12.1	0.3231	61.6	8897.5	2965.8	100
0.5%	100	30	0.4128	0.3530	0.0458	0.1267	21.8	0.3072	66.9	8971.9	2990.6	100
0.5%	100	40	0.3899	0.3329	0.0227	0.1114	31.3	0.2937	73.0	9048.1	3016.0	100
0.5%	100	50	0.4048	0.3189	0.0377	0.1031	36.4	0.2816	78.0	9123.9	3041.3	100
1%	100	-50	0.4206	0.2970	0.0537	0.0911	-30.9	0.2540	92.3	8689.6	2896.6	100
1%	100	-40	0.4155	0.2841	0.0486	0.0830	-19.4	0.2512	93.8	8701.1	2900.4	100
1%	100	-30	0.4283	0.2883	0.0615	0.0868	-24.9	0.2494	95.2	8713.2	2904.4	100
1%	100	-20	0.4229	0.2794	0.0560	0.0812	-16.7	0.2465	96.8	8724.5	2908.2	100
1%	100	-10	0.4266	0.2708	0.0598	0.0769	-10.5	0.2441	98.6	8735.7	2911.9	100
1%	100	0	0.4173	0.2589	0.0504	0.0695	0.0	0.2420	100.2	8748.1	2916.1	100
1%	100	10	0.4057	0.2591	0.0387	0.0686	1.4	0.2300	110.0	8822.9	2941.0	100
1%	100	20	0.4147	0.2606	0.0478	0.0702	-0.9	0.2189	120.7	8897.5	2965.8	100
1%	100	30	0.4063	0.2485	0.0393	0.0633	9.0	0.2085	132.0	8971.9	2990.6	100
1%	100	40	0.3898	0.2331	0.0226	0.0548	21.1	0.1998	142.8	9048.1	3016.0	100
1%	100	50	0.3945	0.2281	0.0273	0.0527	24.2	0.1924	153.4	9123.9	3041.3	100
10%	100	-50	0.3782	0.0926	0.0108	0.0087	-17.7	0.0767	939.1	8689.6	2896.6	100
10%	100	-40	0.3772	0.0904	0.0099	0.0083	-12.0	0.0763	952.1	8701.1	2900.4	100
10%	100	-30	0.3778	0.0885	0.0105	0.0079	-7.5	0.0758	966.4	8713.2	2904.4	100
10%	100	-20	0.3736	0.0888	0.0062	0.0079	-7.3	0.0753	981.8	8724.5	2908.2	100
10%	100	-10	0.3720	0.0872	0.0046	0.0076	-3.1	0.0748	997.7	8735.7	2911.9	100
10%	100	0	0.3710	0.0859	0.0036	0.0074	0.0	0.0743	1011.9	8748.1	2916.1	100
10%	100	10	0.3907	0.0821	0.0235	0.0073	1.4	0.0714	1098.8	8822.9	2941.0	100
10%	100	20	0.3857	0.0767	0.0184	0.0062	15.7	0.0685	1196.2	8897.5	2965.8	100
10%	100	30	0.3830	0.0759	0.0157	0.0060	18.7	0.0661	1290.8	8971.9	2990.6	100
10%	100	40	0.3766	0.0749	0.0092	0.0057	22.9	0.0639	1389.8	9048.1	3016.0	100
10%	100	50	0.3687	0.0679	0.0013	0.0046	37.5	0.0617	1491.8	9123.9	3041.3	100

*displayed as the log(HR)

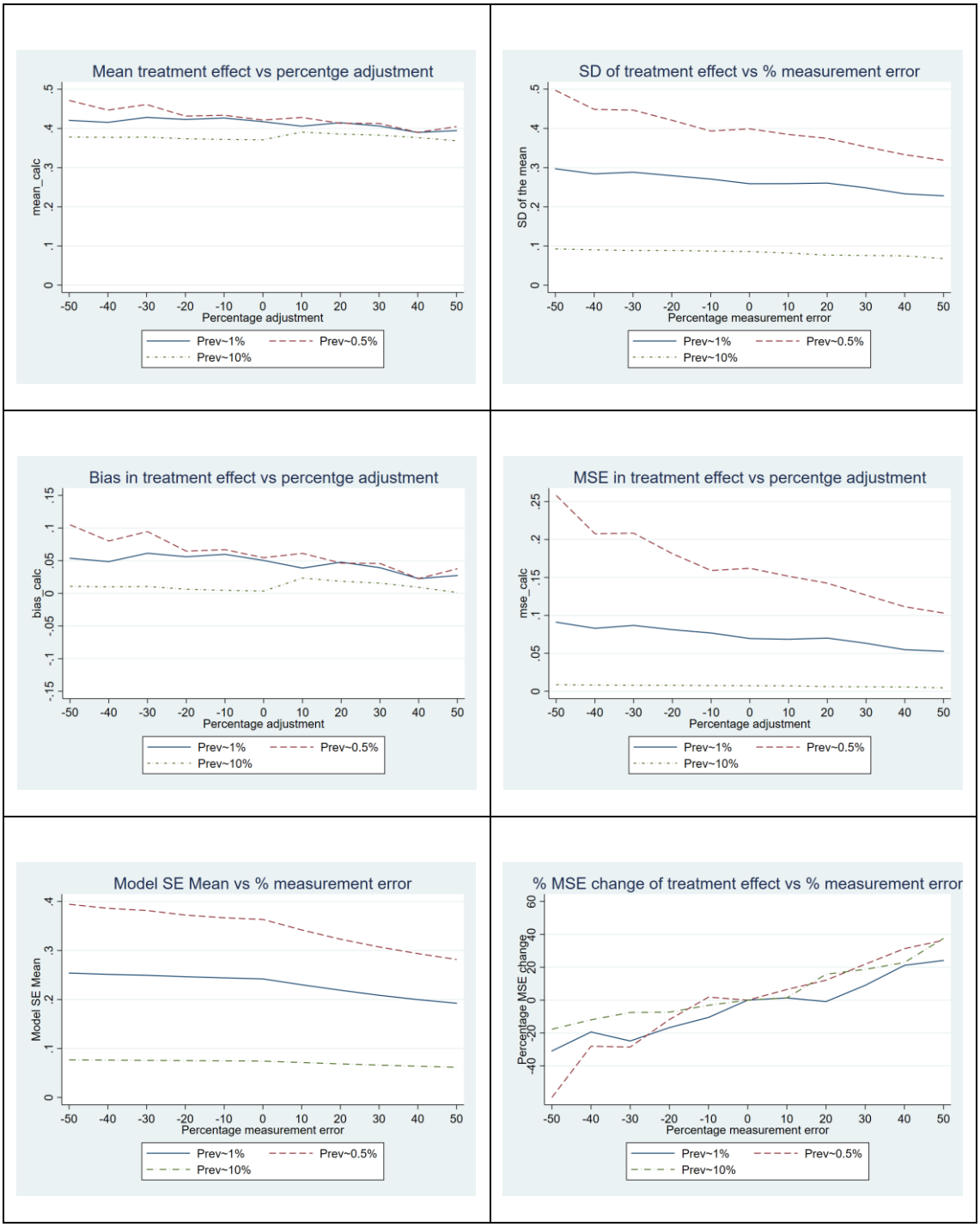


Figure E-3: Using 3to1 PS matching, N=100, the mean, SE, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean.

Table E-4: Simulation runs using PS stratification, with 10 strata, N=100.

Preva-lence	N	% Adj	Mean*	SE*	Bias*	MSE*	MSE % change	Model SE mean	Num events	num WA	Num RI	N valid
0.5%	100	-50	0.3521	0.3098	-0.0155	0.0962	-13.8	0.2971	103.8	18363.8	2895.2	100
0.5%	100	-40	0.3483	0.3062	-0.0193	0.0941	-11.3	0.2950	105.3	18359.9	2899.1	100
0.5%	100	-30	0.3497	0.2984	-0.0179	0.0894	-5.7	0.2922	106.8	18356.1	2902.9	100
0.5%	100	-20	0.3478	0.2946	-0.0198	0.0872	-3.1	0.2897	108.5	18352.3	2906.7	100
0.5%	100	-10	0.3459	0.2888	-0.0218	0.0839	0.8	0.2875	110.0	18348.4	2910.6	100
0.5%	100	0	0.3453	0.2899	-0.0223	0.0845	0.0	0.2849	111.7	18344.8	2914.2	100
0.5%	100	10	0.3405	0.2886	-0.0272	0.0840	0.6	0.2736	120.2	18319.7	2939.3	100
0.5%	100	20	0.3493	0.2660	-0.0183	0.0711	15.9	0.2617	129.4	18295.1	2963.9	100
0.5%	100	30	0.3525	0.2635	-0.0151	0.0697	17.6	0.2518	138.9	18270.5	2988.5	100
0.5%	100	40	0.3436	0.2610	-0.0241	0.0687	18.8	0.2432	148.5	18245.1	3013.9	100
0.5%	100	50	0.3430	0.2561	-0.0247	0.0662	21.7	0.2351	157.8	18220.2	3038.8	100
1%	100	-50	0.3742	0.2185	0.0068	0.0478	-7.4	0.2078	205.6	18363.8	2895.2	100
1%	100	-40	0.3682	0.2182	0.0008	0.0476	-7.0	0.2065	208.6	18359.9	2899.1	100
1%	100	-30	0.3699	0.2194	0.0025	0.0481	-8.2	0.2049	211.5	18356.1	2902.9	100
1%	100	-20	0.3658	0.2181	-0.0017	0.0475	-6.8	0.2036	214.7	18352.3	2906.7	100
1%	100	-10	0.3637	0.2110	-0.0038	0.0446	-0.1	0.2021	217.7	18348.4	2910.6	100
1%	100	0	0.3646	0.2110	-0.0029	0.0445	0.0	0.2005	220.7	18344.8	2914.2	100
1%	100	10	0.3621	0.2026	-0.0054	0.0411	7.7	0.1917	238.8	18319.7	2939.3	100
1%	100	20	0.3615	0.1840	-0.0060	0.0339	23.9	0.1840	256.8	18295.1	2963.9	100
1%	100	30	0.3625	0.1803	-0.0050	0.0325	26.9	0.1772	274.9	18270.5	2988.5	100
1%	100	40	0.3595	0.1715	-0.0080	0.0295	33.8	0.1711	293.4	18245.1	3013.9	100
1%	100	50	0.3600	0.1669	-0.0075	0.0279	37.3	0.1655	311.1	18220.2	3038.8	100
10%	100	-50	0.3677	0.0593	0.0003	0.0035	3.1	0.0655	2045.1	18363.8	2895.2	100
10%	100	-40	0.3661	0.0594	-0.0014	0.0035	2.7	0.0651	2071.3	18359.9	2899.1	100
10%	100	-30	0.3653	0.0610	-0.0022	0.0037	-2.8	0.0647	2098.3	18356.1	2902.9	100
10%	100	-20	0.3646	0.0603	-0.0029	0.0036	-0.4	0.0642	2126.0	18352.3	2906.7	100
10%	100	-10	0.3640	0.0597	-0.0034	0.0036	1.2	0.0638	2153.7	18348.4	2910.6	100
10%	100	0	0.3642	0.0601	-0.0033	0.0036	0.0	0.0634	2180.3	18344.8	2914.2	100
10%	100	10	0.3620	0.0589	-0.0055	0.0035	3.5	0.0609	2343.6	18319.7	2939.3	100
10%	100	20	0.3616	0.0557	-0.0059	0.0031	13.4	0.0586	2506.0	18295.1	2963.9	100
10%	100	30	0.3627	0.0524	-0.0048	0.0028	23.6	0.0566	2666.1	18270.5	2988.5	100
10%	100	40	0.3593	0.0538	-0.0082	0.0030	18.3	0.0549	2829.6	18245.1	3013.9	100
10%	100	50	0.3581	0.0505	-0.0094	0.0026	27.1	0.0532	2991.1	18220.2	3038.8	100

*displayed as the log(HR)

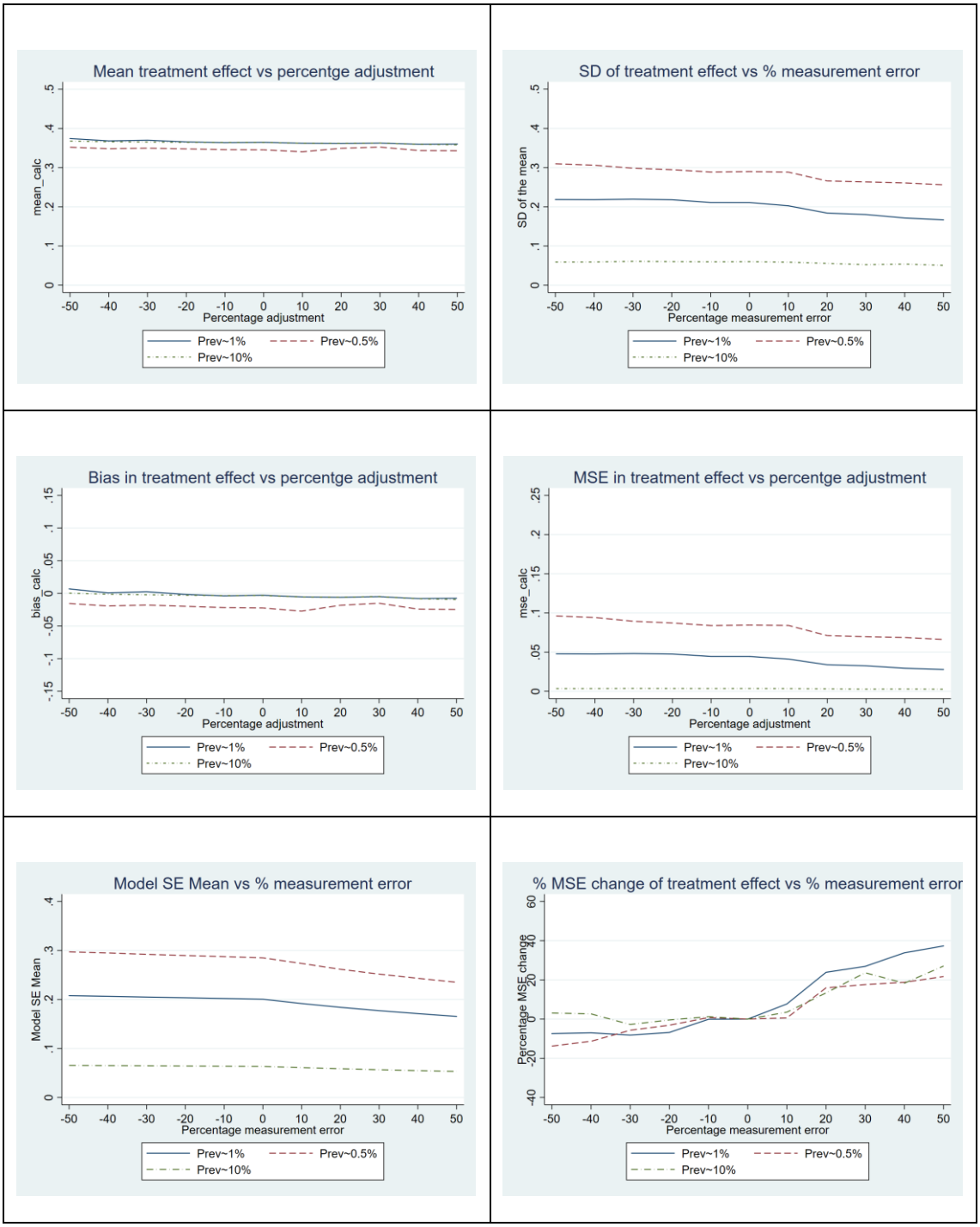


Figure E-4: Using PS stratification, using 10 strata, N=100, the mean, SE, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean.

E-2 Full simulations using original data characteristics

In the headers in all tables in this Appendix, *Prevalence* is Outcome Prevalence, *Num Events* is the Number of Future Strokes, *Num WA* is the Number of participants prescribed Warfarin and *Num RI* is the Number of participants prescribed Rivaroxaban.

Table E-5: Full simulation runs using IPTW to generate the ATE, using original effect size.

Preva-lence	N	% Adj	Mean*	SE*	Bias*	MSE*	MSE % change	Model SE mean	Num events	num WA	Num RI	N valid
0.5%	2100	-50	0.3038	0.4107	-0.0636	0.1727	-8.9	0.3728	103.5	18368.7	2890.3	2100
0.5%	2100	-30	0.3064	0.4044	-0.0611	0.1672	-5.5	0.3668	106.8	18360.9	2898.1	2100
0.5%	2100	-10	0.3100	0.3980	-0.0575	0.1616	-2.0	0.3610	110.0	18353.2	2905.8	2100
0.5%	2100	0	0.3130	0.3945	-0.0544	0.1585	-0.0	0.3581	111.6	18349.4	2909.6	2100
0.5%	2100	10	0.3208	0.3712	-0.0467	0.1399	11.7	0.3426	120.9	18324.5	2934.5	2100
0.5%	2100	30	0.3289	0.3342	-0.0385	0.1131	28.6	0.3173	139.5	18275.0	2984.0	2100
0.5%	2100	50	0.3365	0.3124	-0.0310	0.0985	37.9	0.2972	158.3	18225.0	3034.0	2100
1%	1000	-50	0.3483	0.2792	-0.0191	0.0783	-10.7	0.2691	206.4	18369.6	2889.4	1000
1%	1000	-30	0.3479	0.2731	-0.0196	0.0750	-5.9	0.2649	212.6	18361.7	2897.3	1000
1%	1000	-10	0.3502	0.2675	-0.0172	0.0719	-1.6	0.2607	218.8	18354.0	2905.0	1000
1%	1000	0	0.3494	0.2654	-0.0181	0.0708	-0.0	0.2587	222.0	18350.2	2908.8	1000
1%	1000	10	0.3498	0.2502	-0.0176	0.0629	11.1	0.2476	240.4	18325.5	2933.5	1000
1%	1000	30	0.3514	0.2270	-0.0161	0.0518	26.8	0.2286	276.8	18276.2	2982.8	1000
1%	1000	50	0.3511	0.2109	-0.0164	0.0447	36.8	0.2135	313.3	18226.3	3032.7	1000
10%	1000	-50	0.3626	0.0872	-0.0048	0.0076	-6.8	0.0875	2058.8	18369.6	2889.4	1000
10%	1000	-30	0.3631	0.0862	-0.0044	0.0075	-4.3	0.0861	2113.4	18361.7	2897.3	1000
10%	1000	-10	0.3643	0.0848	-0.0031	0.0072	-0.7	0.0848	2168.5	18354.0	2905.0	1000
10%	1000	0	0.3650	0.0845	-0.0025	0.0071	-0.0	0.0842	2195.5	18350.2	2908.8	1000
10%	1000	10	0.3628	0.0817	-0.0046	0.0067	6.2	0.0809	2357.8	18325.5	2933.5	1000
10%	1000	30	0.3600	0.0761	-0.0074	0.0059	18.1	0.0752	2681.4	18276.2	2982.8	1000
10%	1000	50	0.3581	0.0711	-0.0094	0.0051	28.1	0.0706	3005.7	18226.3	3032.7	1000

*displayed as the log(HR)

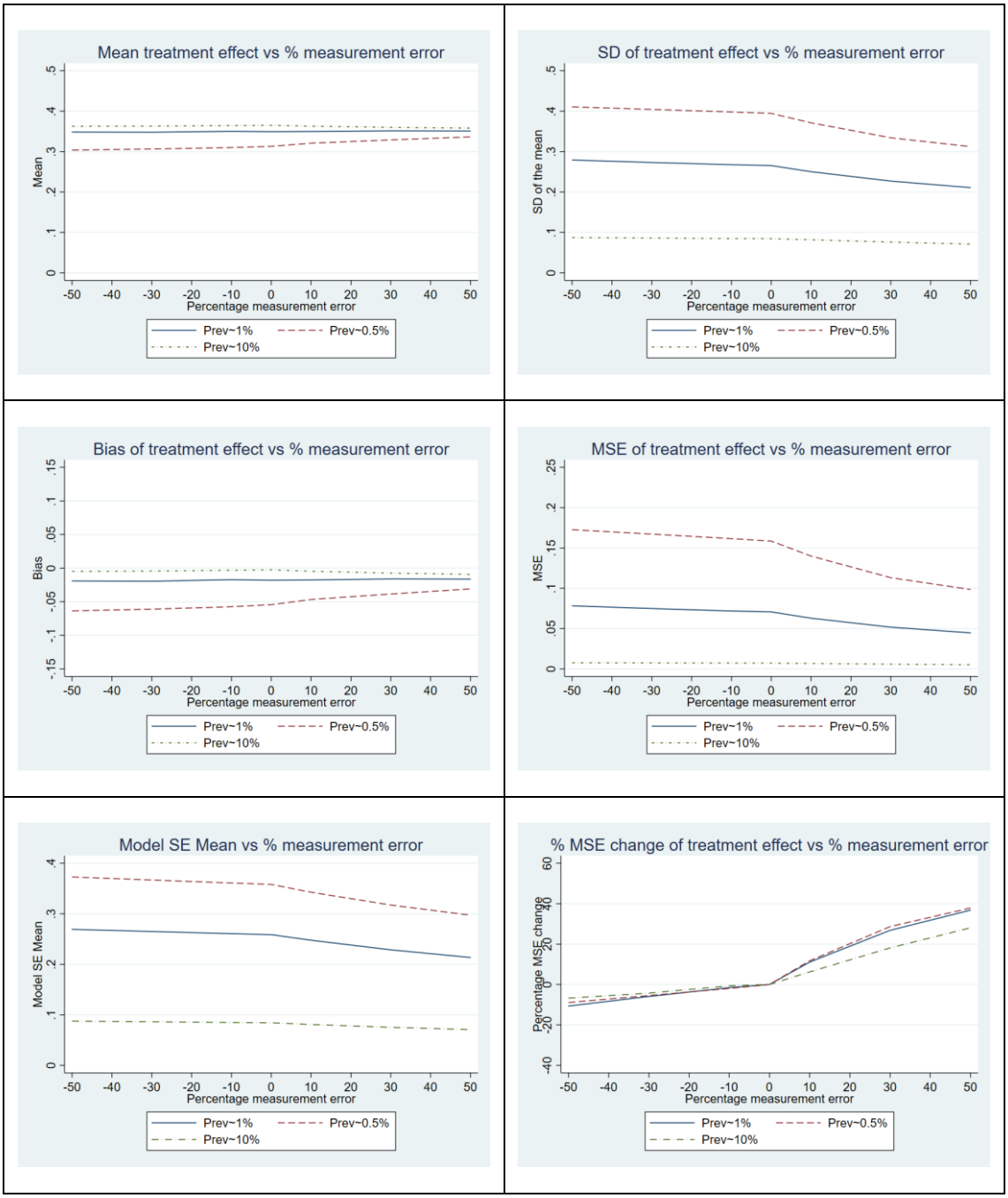


Figure E-5: Using IPTW to generate the ATE, original effect size – the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean.

Table E-6: Full simulation runs using IPTW to generate the ATT, using original effect size.

Preva-lence	N	% Adj	Mean*	SE*	Bias*	MSE*	MSE % change	Model SE mean	Num events	num WA	Num RI	N valid
0.5%	1300	-50	0.3467	0.3186	-0.0207	0.1018	-9.2	0.3034	84.5	18369.5	2889.5	1300
0.5%	1300	-30	0.3471	0.3130	-0.0203	0.0983	-5.4	0.2981	87.3	18361.7	2897.3	1300
0.5%	1300	-10	0.3485	0.3088	-0.0189	0.0956	-2.5	0.2930	90.1	18353.9	2905.1	1300
0.5%	1300	0	0.3497	0.3050	-0.0178	0.0933	-0.0	0.2904	91.4	18350.1	2908.9	1300
0.5%	1300	10	0.3531	0.2874	-0.0143	0.0828	11.3	0.2769	99.5	18325.4	2933.6	1300
0.5%	1300	30	0.3590	0.2611	-0.0084	0.0682	26.9	0.2548	115.3	18276.2	2982.8	1300
0.5%	1300	50	0.3573	0.2453	-0.0101	0.0602	35.4	0.2378	130.6	18226.3	3032.7	1300
1%	1000	-50	0.3560	0.2142	-0.0115	0.0460	-5.7	0.2119	168.2	18369.6	2889.4	1000
1%	1000	-30	0.3564	0.2112	-0.0111	0.0447	-2.8	0.2084	173.6	18361.7	2897.3	1000
1%	1000	-10	0.3577	0.2088	-0.0097	0.0437	-0.4	0.2051	178.8	18354.0	2905.0	1000
1%	1000	0	0.3565	0.2083	-0.0110	0.0435	-0.0	0.2035	181.5	18350.2	2908.8	1000
1%	1000	10	0.3562	0.1955	-0.0113	0.0384	11.9	0.1946	197.6	18325.5	2933.5	1000
1%	1000	30	0.3608	0.1768	-0.0066	0.0313	28.1	0.1796	228.4	18276.2	2982.8	1000
1%	1000	50	0.3610	0.1690	-0.0064	0.0286	34.3	0.1678	258.6	18226.3	3032.7	1000
10%	1000	-50	0.3639	0.0664	-0.0035	0.0044	-5.7	0.0666	1695.6	18369.6	2889.4	1000
10%	1000	-30	0.3644	0.0656	-0.0030	0.0043	-3.2	0.0656	1744.1	18361.7	2897.3	1000
10%	1000	-10	0.3651	0.0647	-0.0023	0.0042	-0.2	0.0647	1792.7	18354.0	2905.0	1000
10%	1000	0	0.3656	0.0646	-0.0018	0.0042	0.0	0.0643	1816.5	18350.2	2908.8	1000
10%	1000	10	0.3630	0.0622	-0.0045	0.0039	7.1	0.0618	1962.9	18325.5	2933.5	1000
10%	1000	30	0.3616	0.0570	-0.0058	0.0033	21.4	0.0576	2247.6	18276.2	2982.8	1000
10%	1000	50	0.3603	0.0525	-0.0072	0.0028	32.9	0.0541	2524.5	18226.3	3032.7	1000

*displayed as the log(HR)

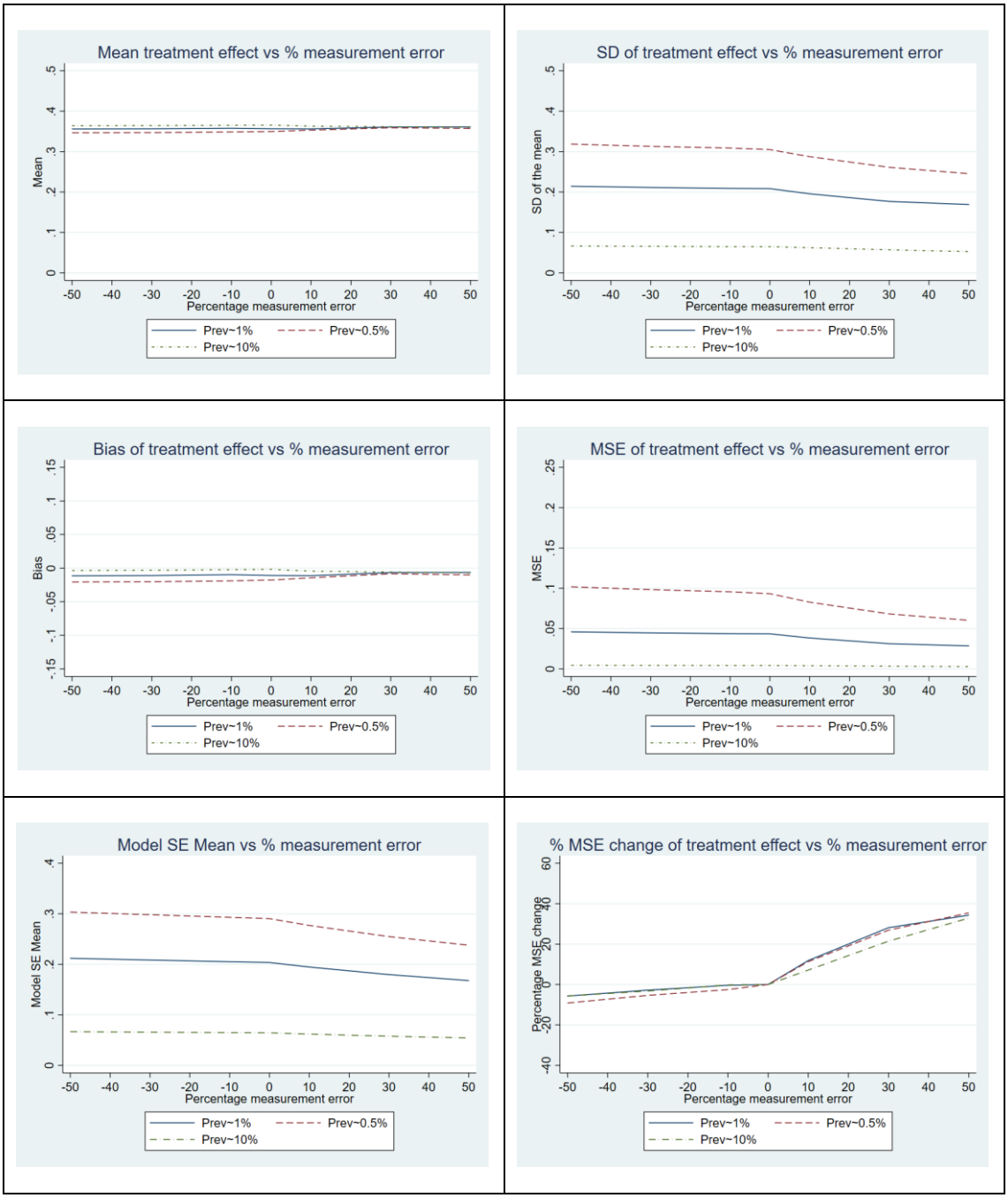


Figure E-6: Using IPTW to generate the ATT, original effect size – the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean.

Table E-7: Full simulation runs using 3to1 PS matching, using original effect size.

Preva-lence	N	% Adj	Mean*	SE*	Bias*	MSE*	% MSE change	Model SE mean	Num events	num WA	Num RI	N valid
0.5%	3100	-50	0.4578	0.4851	0.0904	0.2434	-16.9	0.3878	46.5	8675.9	2892.0	3100
0.5%	3100	-30	0.4501	0.4693	0.0827	0.2270	-9.0	0.3787	48.1	8698.7	2899.6	3100
0.5%	3100	-10	0.4429	0.4572	0.0754	0.2146	-3.1	0.3697	49.8	8721.7	2907.2	3100
0.5%	3100	0	0.4422	0.4503	0.0748	0.2083	-0.0	0.3655	50.6	8733.3	2911.1	3100
0.5%	3100	10	0.4578	0.4180	0.0903	0.1828	12.2	0.3453	55.4	8808.1	2936.0	3100
0.5%	3100	30	0.4278	0.3706	0.0604	0.1409	32.3	0.3097	66.2	8957.6	2985.9	3100
0.5%	3100	50	0.4126	0.3386	0.0452	0.1167	44.0	0.2837	77.1	9107.5	3035.8	3100
1%	1200	-50	0.4175	0.2996	0.0500	0.0922	-10.6	0.2547	92.6	8669.0	2889.7	1200
1%	1200	-30	0.4124	0.2965	0.0450	0.0899	-7.8	0.2503	95.5	8691.8	2897.3	1200
1%	1200	-10	0.4118	0.2904	0.0443	0.0862	-3.4	0.2458	98.6	8714.5	2904.8	1200
1%	1200	0	0.4102	0.2857	0.0428	0.0834	0.0	0.2435	100.1	8726.2	2908.7	1200
1%	1200	10	0.4219	0.2722	0.0545	0.0770	7.6	0.2319	109.5	8800.9	2933.6	1200
1%	1200	30	0.4126	0.2530	0.0451	0.0660	20.9	0.2109	130.8	8950.2	2983.4	1200
1%	1200	50	0.3930	0.2304	0.0255	0.0537	35.6	0.1943	152.4	9100.0	3033.3	1200
10%	1000	-50	0.3716	0.0918	0.0042	0.0085	-7.3	0.0768	940.9	8667.2	2889.1	1000
10%	1000	-30	0.3699	0.0899	0.0025	0.0081	-2.8	0.0759	968.3	8690.1	2896.7	1000
10%	1000	-10	0.3670	0.0891	-0.0004	0.0079	-0.8	0.0749	998.3	8712.7	2904.2	1000
10%	1000	0	0.3665	0.0888	-0.0010	0.0079	-0.0	0.0744	1013.4	8724.4	2908.1	1000
10%	1000	10	0.3845	0.0839	0.0171	0.0073	7.0	0.0714	1100.0	8799.2	2933.1	1000
10%	1000	30	0.3822	0.0747	0.0148	0.0058	26.4	0.0662	1290.0	8948.7	2982.9	1000
10%	1000	50	0.3663	0.0711	-0.0011	0.0050	35.9	0.0618	1490.3	9098.3	3032.8	1000

*displayed as the log(HR)

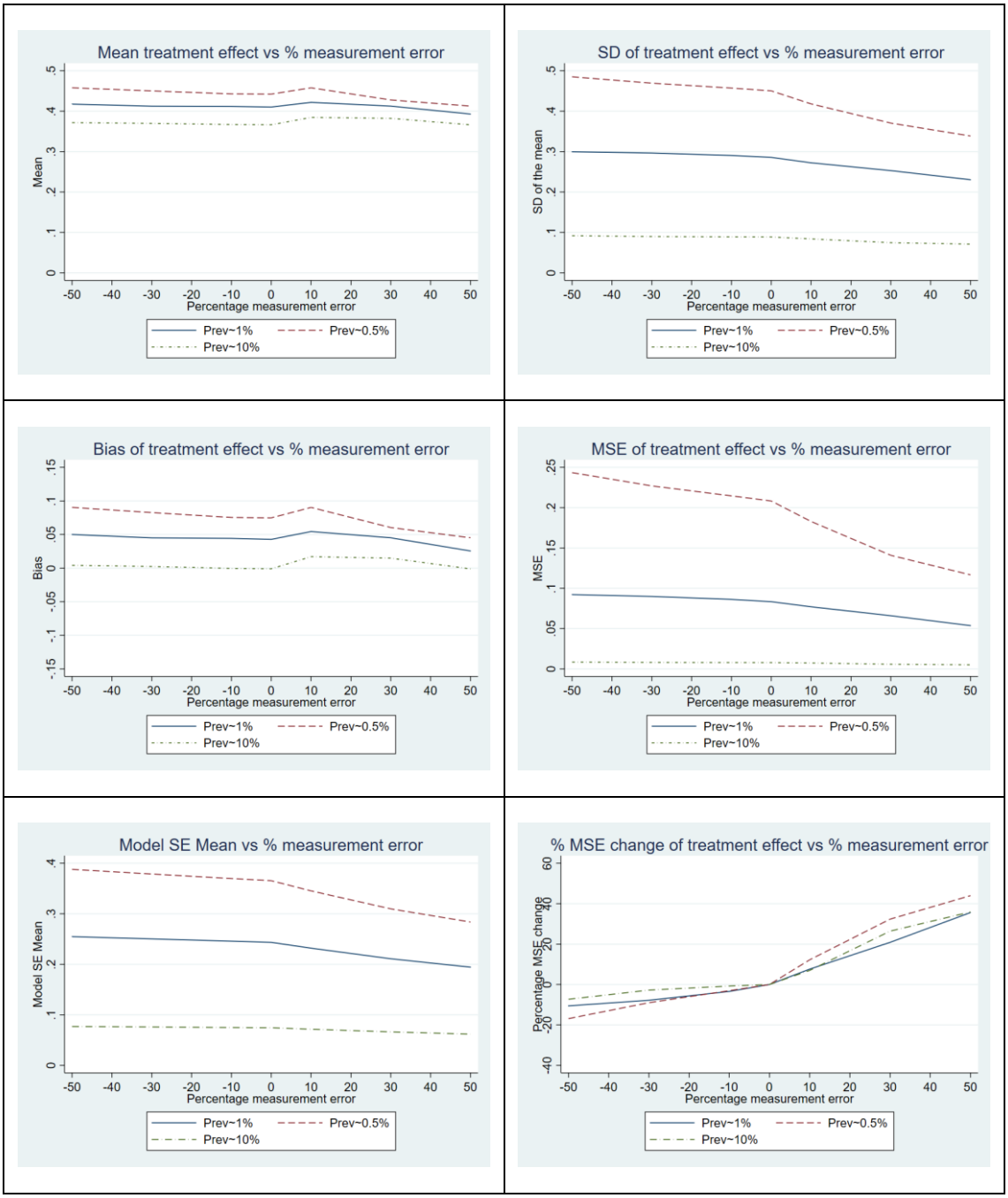


Figure E-7: Using 3to1 PS matching, original effect size – the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean.

Table E-8: Full simulation runs using PS Stratification, using original effect size.

Preva-lence	N	% Adj	Mean*	SE*	Bias*	MSE*	MSE % change	Model SE mean	Num events	num WA	Num RI	N valid
0.5%	1300	-50	0.3443	0.3129	-0.0232	0.0983	-9.7	0.2993	102.5	18369.5	2889.5	1300
0.5%	1300	-30	0.3447	0.3074	-0.0228	0.0949	-5.9	0.2941	105.7	18361.7	2897.3	1300
0.5%	1300	-10	0.3467	0.3028	-0.0207	0.0920	-2.7	0.2890	108.9	18353.9	2905.1	1300
0.5%	1300	0	0.3481	0.2989	-0.0193	0.0897	0.0	0.2864	110.4	18350.1	2908.9	1300
0.5%	1300	10	0.3522	0.2820	-0.0153	0.0797	11.1	0.2729	119.7	18325.4	2933.6	1300
0.5%	1300	30	0.3583	0.2570	-0.0091	0.0661	26.3	0.2509	138.2	18276.2	2982.8	1300
0.5%	1300	50	0.3569	0.2410	-0.0106	0.0581	35.2	0.2341	156.7	18226.3	3032.7	1300
1%	1000	-50	0.3575	0.2113	-0.0099	0.0447	-6.8	0.2089	203.9	18369.6	2889.4	1000
1%	1000	-30	0.3575	0.2080	-0.0099	0.0434	-3.6	0.2054	210.0	18361.7	2897.3	1000
1%	1000	-10	0.3589	0.2055	-0.0085	0.0423	-1.0	0.2021	216.2	18354.0	2905.0	1000
1%	1000	0	0.3575	0.2044	-0.0099	0.0419	-0.0	0.2006	219.3	18350.2	2908.8	1000
1%	1000	10	0.3573	0.1929	-0.0102	0.0373	10.9	0.1916	237.6	18325.5	2933.5	1000
1%	1000	30	0.3609	0.1743	-0.0065	0.0304	27.3	0.1768	273.9	18276.2	2982.8	1000
1%	1000	50	0.3610	0.1654	-0.0064	0.0274	34.5	0.1651	310.1	18226.3	3032.7	1000
10%	1000	-50	0.3643	0.0656	-0.0031	0.0043	-5.8	0.0656	2041.7	18369.6	2889.4	1000
10%	1000	-30	0.3646	0.0649	-0.0028	0.0042	-3.6	0.0647	2096.2	18361.7	2897.3	1000
10%	1000	-10	0.3654	0.0640	-0.0021	0.0041	-0.6	0.0638	2150.9	18354.0	2905.0	1000
10%	1000	0	0.3658	0.0638	-0.0016	0.0041	-0.0	0.0634	2177.8	18350.2	2908.8	1000
10%	1000	10	0.3637	0.0613	-0.0037	0.0038	7.5	0.0609	2339.4	18325.5	2933.5	1000
10%	1000	30	0.3625	0.0563	-0.0049	0.0032	21.7	0.0567	2662.2	18276.2	2982.8	1000
10%	1000	50	0.3611	0.0518	-0.0064	0.0027	33.1	0.0533	2985.0	18226.3	3032.7	1000

*displayed as the log(HR)

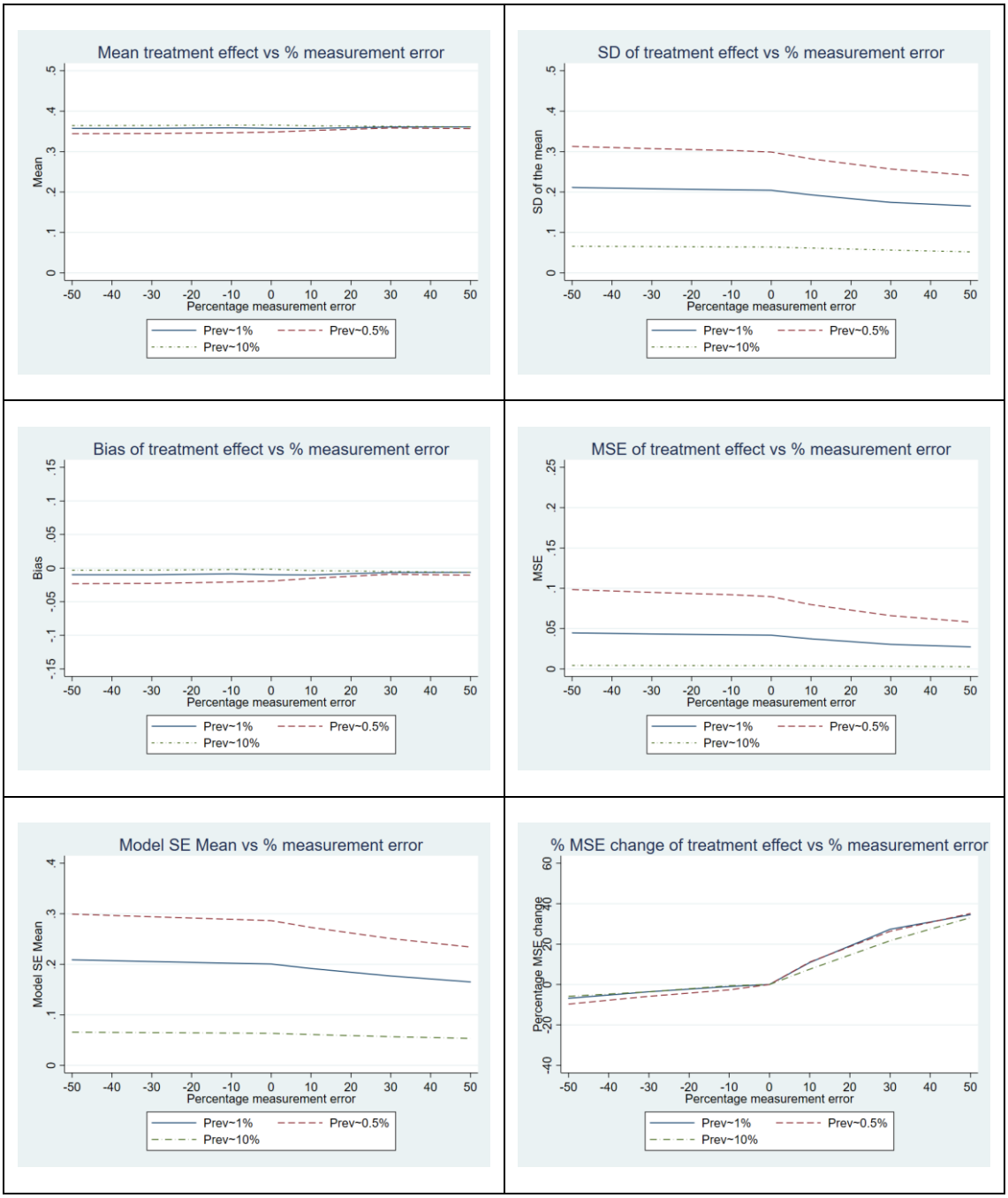


Figure E-8: Using PS Stratification, original effect size – the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean.

APPENDIX F – TABLES AND GRAPHS VARYING EFFECT SIZE

F-1 Tables for different prevalence and fixed effect size

In the headers in all tables in this Appendix, *Prevalence* is Outcome Prevalence, *Num Events* is the Number of Future Strokes, *Num WA* is the Number of participants prescribed Warfarin and *Num RI* is the Number of participants prescribed Rivaroxaban.

IPTW for ATE

Table F-1: Simulation runs using IPTW for ATE – Small effect.

Preva-lence	N	% m error	Mean*	SE*	Bias*	MSE*	MSE* %**	Model SE Mean	Num event	Num WA	Num RI
0.5%	2100	-50	0.3075	0.3998	-0.0600	0.1634	-14.3	0.3662	103.7	18300.5	2958.5
0.5%	2100	-30	0.3103	0.3908	-0.0572	0.1559	-9.1	0.3577	107.0	18265.3	2993.7
0.5%	2100	-10	0.3152	0.3803	-0.0523	0.1473	-3.1	0.3496	110.2	18230.5	3028.6
0.5%	2100	0	0.3180	0.3749	-0.0494	0.1429	-0.0	0.3456	111.9	18213.0	3046.0
0.5%	2100	10	0.3255	0.3451	-0.0420	0.1208	15.5	0.3252	121.4	18100.8	3158.2
0.5%	2100	30	0.3349	0.3059	-0.0325	0.0946	33.8	0.2934	140.6	17875.5	3383.5
0.5%	2100	50	0.3416	0.2757	-0.0258	0.0767	46.4	0.2689	160.1	17650.0	3609.0
1%	1000	-50	0.3503	0.2720	-0.0172	0.0742	-14.4	0.2645	206.7	18301.5	2957.5
1%	1000	-30	0.3505	0.2639	-0.0170	0.0698	-7.7	0.2585	212.9	18266.4	2992.6
1%	1000	-10	0.3522	0.2575	-0.0152	0.0665	-2.5	0.2526	219.3	18231.8	3027.2
1%	1000	0	0.3508	0.2543	-0.0166	0.0649	-0.0	0.2499	222.4	18214.5	3044.5
1%	1000	10	0.3520	0.2348	-0.0154	0.0553	14.7	0.2348	241.2	18102.5	3156.5
1%	1000	30	0.3533	0.2075	-0.0141	0.0432	33.4	0.2109	278.7	17877.4	3381.6
1%	1000	50	0.3548	0.1871	-0.0127	0.0351	45.8	0.1922	316.6	17651.8	3607.2
10%	200	-50	0.3654	0.0891	-0.0020	0.0079	-5.2	0.0858	2067.9	18298.3	2960.7
10%	200	-30	0.3643	0.0893	-0.0031	0.0080	-5.9	0.0839	2122.2	18262.9	2996.1
10%	200	-10	0.3649	0.0879	-0.0026	0.0077	-2.4	0.0821	2178.2	18228.2	3030.8
10%	200	0	0.3656	0.0869	-0.0019	0.0075	-0.0	0.0812	2205.8	18210.7	3048.3
10%	200	10	0.3631	0.0807	-0.0043	0.0065	13.6	0.0770	2373.1	18099.0	3160.0
10%	200	30	0.3644	0.0720	-0.0031	0.0052	31.2	0.0697	2706.5	17872.3	3386.7
10%	200	50	0.3615	0.0613	-0.0059	0.0038	49.8	0.0640	3042.1	17646.6	3612.4

*displayed as the log(HR) **percent change

Table F-2: Simulation runs using IPTW for ATE – Medium Effect.

Preva-lence	N	% m error	Mean*	SE*	Bias*	MSE*	MSE* %**	Model SE mean	Num event	Num WA	Num RI
0.5%	2100	-50	0.3103	0.3907	-0.0572	0.1558	-19.4	0.3581	103.8	18189.3	3069.7
0.5%	2100	-30	0.3128	0.3791	-0.0546	0.1466	-12.3	0.3472	107.2	18109.8	3149.2
0.5%	2100	-10	0.3191	0.3656	-0.0483	0.1359	-4.2	0.3367	110.5	18030.4	3228.6
0.5%	2100	0	0.3215	0.3584	-0.0460	0.1305	-0.0	0.3319	112.2	17990.7	3268.3
0.5%	2100	10	0.3295	0.3270	-0.0380	0.1083	17.0	0.3069	122.1	17736.8	3522.3
0.5%	2100	30	0.3400	0.2768	-0.0275	0.0773	40.7	0.2681	142.3	17228.1	4030.9
0.5%	2100	50	0.3484	0.2441	-0.0190	0.0599	54.1	0.2394	162.9	16720.1	4538.9
1%	1000	-50	0.3510	0.2682	-0.0165	0.0721	-18.8	0.2595	207.0	18190.5	3068.5
1%	1000	-30	0.3515	0.2588	-0.0159	0.0671	-10.7	0.2520	213.4	18111.0	3148.0
1%	1000	-10	0.3552	0.2494	-0.0123	0.0623	-2.6	0.2445	220.0	18031.4	3227.6
1%	1000	0	0.3532	0.2460	-0.0142	0.0607	0.0	0.2412	223.3	17991.7	3267.3
1%	1000	10	0.3555	0.2259	-0.0119	0.0511	15.8	0.2230	242.9	17737.4	3521.6
1%	1000	30	0.3574	0.1943	-0.0101	0.0378	37.7	0.1940	282.3	17228.2	4030.8
1%	1000	50	0.3614	0.1718	-0.0060	0.0295	51.4	0.1722	322.6	16719.9	4539.1
10%	200	-50	0.3661	0.0880	-0.0013	0.0077	-4.7	0.0847	2072.1	18187.6	3071.4
10%	200	-30	0.3655	0.0887	-0.0019	0.0078	-6.2	0.0824	2128.2	18108.5	3150.5
10%	200	-10	0.3673	0.0877	-0.0001	0.0076	-3.7	0.0801	2185.9	18028.8	3230.2
10%	200	0	0.3674	0.0861	-0.0000	0.0074	0.0	0.0790	2214.1	17988.4	3270.6
10%	200	10	0.3635	0.0777	-0.0039	0.0060	18.3	0.0738	2386.9	17736.2	3522.8
10%	200	30	0.3640	0.0684	-0.0034	0.0047	36.8	0.0650	2735.9	17225.9	4033.1
10%	200	50	0.3630	0.0579	-0.0044	0.0034	54.5	0.0580	3087.3	16714.1	4544.9

*displayed as the log(HR) **percent change

Table F-3: Simulation runs using IPTW for ATE – High Effect.

Preva-lence	N	% m error	Mean*	SE*	Bias*	MSE*	MSE* %**	Model SE mean	Num event	Num WA	Num RI
0.5%	2100	-50	0.3104	0.3932	-0.0570	0.1578	-19.8	0.3533	104.1	18057.7	3201.3
0.5%	2100	-30	0.3147	0.3805	-0.0528	0.1475	-12.0	0.3406	107.6	17925.6	3333.4
0.5%	2100	-10	0.3201	0.3667	-0.0473	0.1366	-3.8	0.3289	111.0	17793.5	3465.5
0.5%	2100	0	0.3222	0.3601	-0.0452	0.1317	-0.0	0.3234	112.8	17727.6	3531.4
0.5%	2100	10	0.3317	0.3233	-0.0357	0.1058	19.7	0.2951	123.1	17308.3	3950.7
0.5%	2100	30	0.3443	0.2658	-0.0231	0.0711	46.0	0.2521	144.3	16471.0	4788.0
0.5%	2100	50	0.3547	0.2268	-0.0127	0.0516	60.8	0.2208	166.2	15633.9	5625.1
1%	1000	-50	0.3500	0.2666	-0.0174	0.0713	-19.3	0.2569	207.5	18058.8	3200.2
1%	1000	-30	0.3513	0.2589	-0.0162	0.0672	-12.5	0.2484	214.2	17926.7	3332.3
1%	1000	-10	0.3545	0.2480	-0.0129	0.0616	-3.0	0.2400	221.0	17794.6	3464.4
1%	1000	0	0.3537	0.2442	-0.0137	0.0598	-0.0	0.2362	224.4	17728.5	3530.5
1%	1000	10	0.3590	0.2198	-0.0084	0.0483	19.2	0.2159	244.8	17309.5	3949.5
1%	1000	30	0.3625	0.1842	-0.0050	0.0339	43.3	0.1836	286.4	16471.8	4787.2
1%	1000	50	0.3694	0.1571	0.0020	0.0247	58.7	0.1595	329.1	15635.0	5624.0
10%	200	-50	0.3656	0.0880	-0.0018	0.0077	-7.2	0.0840	2077.6	18056.1	3202.9
10%	200	-30	0.3660	0.0881	-0.0014	0.0077	-7.5	0.0814	2135.9	17923.7	3335.3
10%	200	-10	0.3681	0.0867	0.0006	0.0075	-3.9	0.0789	2194.7	17791.8	3467.2
10%	200	0	0.3674	0.0850	-0.0000	0.0072	-0.0	0.0776	2224.4	17724.9	3534.1
10%	200	10	0.3647	0.0758	-0.0028	0.0057	20.3	0.0719	2405.1	17307.7	3951.3
10%	200	30	0.3652	0.0657	-0.0022	0.0043	40.2	0.0622	2769.6	16471.5	4787.5
10%	200	50	0.3651	0.0525	-0.0024	0.0027	61.8	0.0544	3141.0	15635.5	5623.5

*displayed as the log(HR) **percent change

IPTW for ATT

Table F-4: Simulation runs using IPTW ATT – Small Effect.

Preva -lence	N	% m error	Mean*	SE*	Bias*	MSE*	MSE* % **	Model SE mean	Num event	Num WA	Num RI
0.5%	1300	-50	0.3484	0.3151	-0.0190	0.0996	-12.0	0.2986	87.0	18301.1	2957.9
0.5%	1300	-30	0.3498	0.3087	-0.0176	0.0955	-7.5	0.2917	90.7	18266.0	2993.0
0.5%	1300	-10	0.3522	0.3015	-0.0152	0.0911	-2.5	0.2852	94.2	18231.3	3027.7
0.5%	1300	0	0.3536	0.2979	-0.0138	0.0889	-0.0	0.2820	95.9	18213.9	3045.1
0.5%	1300	10	0.3586	0.2756	-0.0088	0.0760	14.5	0.2656	105.9	18101.8	3157.2
0.5%	1300	30	0.3610	0.2491	-0.0064	0.0620	30.2	0.2403	123.8	17876.9	3382.1
0.5%	1300	50	0.3558	0.2281	-0.0116	0.0521	41.3	0.2213	139.8	17651.6	3607.4
1%	1000	-50	0.3578	0.2130	-0.0097	0.0454	-5.9	0.2087	173.0	18301.5	2957.5
1%	1000	-30	0.3599	0.2102	-0.0075	0.0442	-3.0	0.2042	180.0	18266.4	2992.6
1%	1000	-10	0.3610	0.2079	-0.0064	0.0432	-0.8	0.2000	186.8	18231.8	3027.2
1%	1000	0	0.3608	0.2071	-0.0066	0.0429	0.0	0.1981	190.2	18214.5	3044.5
1%	1000	10	0.3629	0.1895	-0.0045	0.0359	16.3	0.1870	209.9	18102.5	3156.5
1%	1000	30	0.3658	0.1707	-0.0016	0.0291	32.2	0.1696	244.8	17877.4	3381.6
1%	1000	50	0.3654	0.1584	-0.0020	0.0251	41.6	0.1563	276.3	17651.8	3607.2
10%	200	-50	0.3649	0.0624	-0.0026	0.0039	-0.5	0.0656	1744.4	18298.3	2960.7
10%	200	-30	0.3650	0.0639	-0.0025	0.0041	-5.4	0.0644	1807.3	18262.9	2996.1
10%	200	-10	0.3640	0.0621	-0.0034	0.0039	0.1	0.0633	1870.7	18228.2	3030.8
10%	200	0	0.3643	0.0622	-0.0032	0.0039	0.0	0.0627	1901.5	18210.7	3048.3
10%	200	10	0.3610	0.0575	-0.0064	0.0033	13.6	0.0596	2081.0	18099.0	3160.0
10%	200	30	0.3613	0.0482	-0.0061	0.0024	39.0	0.0545	2401.6	17872.3	3386.7
10%	200	50	0.3581	0.0423	-0.0094	0.0019	51.5	0.0506	2687.6	17646.6	3612.4

*displayed as the log(HR) **percent change

Table F-5: Simulation runs using IPTW for ATT – Medium Effect.

Preva-lence	N	% m error	Mean*	SE*	Bias*	MSE*	MSE* %**	Model SE mean	Num event	Num WA	Num RI
0.5%	1300	-50	0.3615	0.3007	-0.0060	0.0904	-18.5	0.2937	90.7	18190.3	3068.7
0.5%	1300	-30	0.3642	0.2883	-0.0032	0.0831	-8.9	0.2854	95.7	18110.8	3148.2
0.5%	1300	-10	0.3687	0.2780	0.0013	0.0772	-1.2	0.2777	100.3	18031.3	3227.7
0.5%	1300	0	0.3698	0.2763	0.0023	0.0763	0.0	0.2741	102.5	17991.7	3267.3
0.5%	1300	10	0.3728	0.2593	0.0053	0.0672	11.9	0.2559	114.8	17737.5	3521.5
0.5%	1300	30	0.3691	0.2306	0.0017	0.0531	30.4	0.2279	135.2	17228.8	4030.2
0.5%	1300	50	0.3652	0.2082	-0.0022	0.0433	43.2	0.2071	151.5	16720.9	4538.1
1%	1000	-50	0.3665	0.2033	-0.0009	0.0413	-10.7	0.2061	180.6	18190.5	3068.5
1%	1000	-30	0.3672	0.1982	-0.0002	0.0392	-5.2	0.2008	190.1	18111.0	3148.0
1%	1000	-10	0.3703	0.1939	0.0029	0.0376	-0.7	0.1957	199.2	18031.4	3227.6
1%	1000	0	0.3688	0.1933	0.0013	0.0373	-0.0	0.1934	203.7	17991.7	3267.3
1%	1000	10	0.3706	0.1800	0.0032	0.0324	13.2	0.1807	228.0	17737.4	3521.6
1%	1000	30	0.3692	0.1599	0.0017	0.0255	31.6	0.1612	267.6	17228.2	4030.8
1%	1000	50	0.3694	0.1460	0.0020	0.0213	42.9	0.1466	299.7	16719.9	4539.1
10%	200	-50	0.3657	0.0631	-0.0018	0.0040	0.4	0.0651	1814.5	18187.6	3071.4
10%	200	-30	0.3664	0.0657	-0.0010	0.0043	-7.9	0.0636	1899.6	18108.5	3150.5
10%	200	-10	0.3670	0.0652	-0.0005	0.0042	-6.0	0.0623	1982.7	18028.8	3230.2
10%	200	0	0.3680	0.0633	0.0005	0.0040	0.0	0.0616	2021.9	17988.4	3270.6
10%	200	10	0.3627	0.0599	-0.0048	0.0036	9.9	0.0580	2241.6	17736.2	3522.8
10%	200	30	0.3604	0.0520	-0.0070	0.0027	31.1	0.0520	2604.3	17225.9	4033.1
10%	200	50	0.3622	0.0478	-0.0052	0.0023	42.2	0.0476	2890.2	16714.1	4544.9

*displayed as the log(HR) **percent change

Table F-6: Simulation runs using IPTW for ATT – High Effect.

Preva-lence	N	% m error	Mean*	SE*	Bias*	MSE*	MSE* %**	Model SE mean	Num event	Num WA	Num RI
0.5%	1300	-50	0.3688	0.3123	0.0013	0.0974	-18.2	0.2939	95.3	18058.4	3200.6
0.5%	1300	-30	0.3720	0.3052	0.0045	0.0931	-12.9	0.2855	101.6	17926.4	3332.6
0.5%	1300	-10	0.3742	0.2920	0.0068	0.0852	-3.4	0.2776	107.4	17794.1	3464.9
0.5%	1300	0	0.3772	0.2871	0.0098	0.0825	-0.0	0.2736	110.1	17728.1	3530.9
0.5%	1300	10	0.3769	0.2579	0.0095	0.0666	19.3	0.2543	124.7	17309.1	3949.9
0.5%	1300	30	0.3751	0.2243	0.0076	0.0503	39.0	0.2240	146.7	16471.5	4787.5
0.5%	1300	50	0.3733	0.2040	0.0059	0.0416	49.5	0.2016	163.3	15635.0	5624.0
1%	1000	-50	0.3664	0.2112	-0.0010	0.0446	-15.2	0.2077	189.6	18058.8	3200.2
1%	1000	-30	0.3677	0.2083	0.0002	0.0434	-12.1	0.2022	202.0	17926.7	3332.3
1%	1000	-10	0.3693	0.1995	0.0018	0.0398	-2.8	0.1968	213.3	17794.6	3464.4
1%	1000	0	0.3703	0.1967	0.0029	0.0387	-0.0	0.1941	218.5	17728.5	3530.5
1%	1000	10	0.3739	0.1742	0.0065	0.0304	21.5	0.1803	247.0	17309.5	3949.5
1%	1000	30	0.3735	0.1492	0.0060	0.0223	42.4	0.1589	290.0	16471.8	4787.2
1%	1000	50	0.3771	0.1398	0.0097	0.0196	49.3	0.1431	322.4	15635.0	5624.0
10%	200	-50	0.3639	0.0634	-0.0036	0.0040	5.1	0.0658	1898.8	18056.1	3202.9
10%	200	-30	0.3656	0.0675	-0.0018	0.0045	-7.5	0.0643	2007.5	17923.7	3335.3
10%	200	-10	0.3693	0.0669	0.0019	0.0045	-5.5	0.0628	2107.2	17791.8	3467.2
10%	200	0	0.3682	0.0652	0.0007	0.0042	-0.0	0.0621	2156.9	17724.9	3534.1
10%	200	10	0.3635	0.0590	-0.0039	0.0035	17.6	0.0580	2414.2	17307.7	3951.3
10%	200	30	0.3635	0.0510	-0.0040	0.0026	38.4	0.0514	2803.7	16471.5	4787.5
10%	200	50	0.3650	0.0469	-0.0024	0.0022	48.0	0.0465	3090.9	15635.5	5623.5

*displayed as the log(HR) **percent change

3to1 PS Matching

Table F-7: Simulation runs using 3to1 PS matching – Small Effect.

Prevalence	N	% m error	Mean*	SE*	Bias*	MSE*	MSE* %**	Model SE mean	Num event	Num WA	Num RI
0.5%	3100	-50	0.4538	0.4752	0.0864	0.2332	-20.9	0.3787	48.8	8881.4	2960.5
0.5%	3100	-30	0.4442	0.4570	0.0768	0.2147	-11.3	0.3663	51.5	8986.1	2995.4
0.5%	3100	-10	0.4360	0.4404	0.0685	0.1986	-2.9	0.3543	54.2	9090.7	3030.2
0.5%	3100	0	0.4372	0.4337	0.0698	0.1929	0.0	0.3486	55.5	9143.2	3047.7
0.5%	3100	10	0.4518	0.3953	0.0844	0.1634	15.3	0.3223	63.3	9481.2	3160.4
0.5%	3100	30	0.4270	0.3448	0.0595	0.1224	36.5	0.2811	80.1	10157.3	3385.8
0.5%	3100	50	0.4077	0.3098	0.0402	0.0976	49.4	0.2516	97.2	10833.6	3611.2
1%	1200	-50	0.4150	0.2975	0.0475	0.0907	-12.2	0.2497	97.1	8874.8	2958.3
1%	1200	-30	0.4059	0.2908	0.0385	0.0860	-6.4	0.2430	102.3	8979.6	2993.2
1%	1200	-10	0.4056	0.2848	0.0382	0.0825	-2.1	0.2366	107.3	9083.8	3027.9
1%	1200	0	0.4055	0.2818	0.0381	0.0808	0.0	0.2335	109.7	9135.9	3045.3
1%	1200	10	0.4228	0.2639	0.0553	0.0726	10.1	0.2181	124.6	9473.3	3157.8
1%	1200	30	0.4087	0.2375	0.0413	0.0581	28.1	0.1924	158.0	10149.5	3383.2
1%	1200	50	0.3934	0.2198	0.0260	0.0489	39.4	0.1734	191.8	10824.5	3608.2
10%	200	-50	0.3640	0.0832	-0.0035	0.0069	-1.3	0.0753	987.9	8893.8	2964.6
10%	200	-30	0.3629	0.0836	-0.0045	0.0070	-2.5	0.0738	1033.7	8998.7	2999.6
10%	200	-10	0.3598	0.0823	-0.0076	0.0068	0.3	0.0723	1081.4	9103.1	3034.4
10%	200	0	0.3608	0.0825	-0.0067	0.0068	0.0	0.0716	1104.5	9155.9	3052.0
10%	200	10	0.3806	0.0741	0.0132	0.0056	17.2	0.0675	1247.2	9492.2	3164.1
10%	200	30	0.3761	0.0671	0.0087	0.0046	33.1	0.0607	1555.9	10170.5	3390.2
10%	200	50	0.3662	0.0632	-0.0012	0.0040	41.6	0.0554	1874.0	10849.3	3616.4

*displayed as the log(HR) **percent change

Table F-8: Simulation runs using 3to1 PS Matching – Medium Effect.

Preva-lence	N	% m error	Mean*	SE*	Bias*	MSE*	MSE* %**	Model SE mean	Num event	Num WA	Num RI
0.5%	3100	-50	0.4472	0.4647	0.0797	0.2223	-23.1	0.3641	52.9	9215.5	3071.8
0.5%	3100	-30	0.4421	0.4481	0.0746	0.2063	-14.3	0.3479	57.1	9453.3	3151.1
0.5%	3100	-10	0.4370	0.4288	0.0696	0.1886	-4.5	0.3321	61.4	9691.5	3230.5
0.5%	3100	0	0.4399	0.4187	0.0725	0.1805	-0.0	0.3250	63.5	9810.6	3270.2
0.5%	3100	10	0.4520	0.3785	0.0846	0.1504	16.7	0.2927	76.2	10573.7	3524.6
0.5%	3100	30	0.4286	0.3278	0.0612	0.1111	38.4	0.2467	103.0	12099.4	4033.1
0.5%	3100	50	0.4103	0.2857	0.0428	0.0834	53.8	0.2163	130.1	13622.5	4540.8
1%	1200	-50	0.4101	0.2936	0.0427	0.0879	-26.0	0.2410	105.4	9207.6	3069.2
1%	1200	-30	0.4089	0.2801	0.0414	0.0801	-14.7	0.2316	113.5	9445.7	3148.6
1%	1200	-10	0.4133	0.2657	0.0459	0.0726	-4.0	0.2228	121.4	9682.7	3227.6
1%	1200	0	0.4117	0.2606	0.0442	0.0698	0.0	0.2187	125.5	9801.1	3267.0
1%	1200	10	0.4224	0.2467	0.0550	0.0638	8.5	0.1989	150.7	10564.9	3521.6
1%	1200	30	0.4062	0.2172	0.0387	0.0486	30.3	0.1696	204.1	12093.0	4031.0
1%	1200	50	0.3902	0.1973	0.0228	0.0394	43.5	0.1497	258.4	13615.6	4538.5
10%	200	-50	0.3649	0.0834	-0.0025	0.0069	-19.9	0.0730	1062.1	9230.3	3076.8
10%	200	-30	0.3656	0.0824	-0.0018	0.0068	-17.1	0.0708	1137.6	9467.4	3155.8
10%	200	-10	0.3647	0.0792	-0.0027	0.0062	-8.2	0.0686	1215.3	9702.1	3234.0
10%	200	0	0.3650	0.0762	-0.0024	0.0058	0.0	0.0676	1253.1	9822.0	3274.0
10%	200	10	0.3794	0.0753	0.0120	0.0058	-0.2	0.0620	1493.4	10581.0	3527.0
10%	200	30	0.3743	0.0659	0.0068	0.0044	24.3	0.0539	1992.6	12107.3	4035.8
10%	200	50	0.3703	0.0617	0.0029	0.0038	34.2	0.0482	2495.1	13629.0	4543.0

*displayed as the log(HR) **percent change

Table F-9: Simulation runs using 3to1 PS matching – High Effect.

Preva-lence	N	% m error	Mean*	SE*	Bias*	MSE*	MSE* %**	Model SE mean	Num event	Num WA	Num RI
0.5%	3100	-50	0.4510	0.4566	0.0836	0.2154	-27.4	0.3461	58.1	9608.3	3202.8
0.5%	3100	-30	0.4465	0.4375	0.0791	0.1976	-16.8	0.3259	64.4	10003.7	3334.6
0.5%	3100	-10	0.4393	0.4118	0.0719	0.1747	-3.3	0.3077	70.8	10399.9	3466.6
0.5%	3100	0	0.4377	0.4052	0.0703	0.1691	0.0	0.2998	73.9	10597.1	3532.4
0.5%	3100	10	0.4481	0.3670	0.0806	0.1412	16.5	0.2643	92.8	11855.0	3951.7
0.5%	3100	30	0.4264	0.3108	0.0589	0.1001	40.8	0.2172	131.9	14368.0	4789.3
0.5%	3100	50	0.4161	0.2741	0.0486	0.0775	54.2	0.1882	170.7	16879.6	5626.5
1%	1200	-50	0.4076	0.3012	0.0402	0.0922	-30.1	0.2300	115.7	9598.9	3199.6
1%	1200	-30	0.4041	0.2869	0.0367	0.0836	-17.9	0.2180	128.0	9995.3	3331.8
1%	1200	-10	0.4021	0.2715	0.0346	0.0748	-5.5	0.2075	140.1	10389.5	3463.2
1%	1200	0	0.3993	0.2645	0.0319	0.0709	-0.0	0.2026	146.1	10586.8	3528.9
1%	1200	10	0.4119	0.2441	0.0445	0.0615	13.2	0.1801	183.4	11845.3	3948.4
1%	1200	30	0.4038	0.2078	0.0364	0.0445	37.3	0.1496	261.1	14361.5	4787.1
1%	1200	50	0.3891	0.1895	0.0217	0.0364	48.7	0.1303	339.1	16871.8	5623.9
10%	200	-50	0.3680	0.0798	0.0006	0.0063	-14.5		1159.9	9628.3	3209.4
10%	200	-30	0.3678	0.0809	0.0003	0.0065	-17.6	0.0671	1274.3	10021.3	3340.4
10%	200	-10	0.3673	0.0778	-0.0001	0.0060	-8.8	0.0644	1388.7	10412.5	3470.8
10%	200	0	0.3670	0.0746	-0.0004	0.0055	-0.0	0.0631	1446.1	10611.3	3537.1
10%	200	10	0.3805	0.0746	0.0130	0.0057	-3.2	0.0566	1800.5	11869.3	3956.4
10%	200	30	0.3717	0.0645	0.0043	0.0042	24.9	0.0479	2528.8	14375.8	4791.9
10%	200	50	0.3705	0.0607	0.0031	0.0037	33.5	0.0422	3259.1	16888.9	5629.6

*displayed as the log(HR) **percent change

PS Stratification

Table F-10: Simulation runs using PS Stratification – Small Effect.

Preva-lence	N	% m error	Mean*	SE*	Bias*	MSE*	MSE* %**	Model SE mean	Num event	Num WA	Num RI
0.5%	1300	-50	0.3470	0.3072	-0.0204	0.0947	-12.5	0.2936	102.8	18301.1	2957.9
0.5%	1300	-30	0.3481	0.3009	-0.0193	0.0909	-7.9	0.2866	106.1	18266.0	2993.0
0.5%	1300	-10	0.3506	0.2935	-0.0168	0.0864	-2.6	0.2799	109.3	18231.3	3027.7
0.5%	1300	0	0.3518	0.2899	-0.0157	0.0842	0.0	0.2767	111.0	18213.9	3045.1
0.5%	1300	10	0.3591	0.2671	-0.0084	0.0714	15.2	0.2599	120.5	18101.8	3157.2
0.5%	1300	30	0.3617	0.2409	-0.0057	0.0580	31.1	0.2346	139.6	17876.9	3382.1
0.5%	1300	50	0.3576	0.2218	-0.0098	0.0493	41.5	0.2159	158.6	17651.6	3607.4
1%	1000	-50	0.3595	0.2085	-0.0079	0.0435	-8.4	0.2051	204.4	18301.5	2957.5
1%	1000	-30	0.3604	0.2046	-0.0070	0.0419	-4.3	0.2005	210.7	18266.4	2992.6
1%	1000	-10	0.3609	0.2020	-0.0065	0.0408	-1.7	0.1962	217.1	18231.8	3027.2
1%	1000	0	0.3600	0.2003	-0.0074	0.0401	-0.0	0.1942	220.3	18214.5	3044.5
1%	1000	10	0.3643	0.1842	-0.0032	0.0339	15.6	0.1828	239.2	18102.5	3156.5
1%	1000	30	0.3658	0.1645	-0.0017	0.0270	32.6	0.1655	276.7	17877.4	3381.6
1%	1000	50	0.3657	0.1530	-0.0018	0.0234	41.7	0.1524	314.0	17651.8	3607.2
10%	200	-50	0.3675	0.0615	0.0001	0.0038	-3.9	0.0645	2051.6	18298.3	2960.7
10%	200	-30	0.3673	0.0626	-0.0001	0.0039	-7.4	0.0632	2106.6	18262.9	2996.1
10%	200	-10	0.3659	0.0610	-0.0016	0.0037	-2.2	0.0620	2162.6	18228.2	3030.8
10%	200	0	0.3661	0.0603	-0.0013	0.0036	-0.0	0.0615	2190.2	18210.7	3048.3
10%	200	10	0.3654	0.0562	-0.0020	0.0031	13.3	0.0583	2357.7	18099.0	3160.0
10%	200	30	0.3653	0.0474	-0.0021	0.0022	38.3	0.0532	2689.5	17872.3	3386.7
10%	200	50	0.3619	0.0415	-0.0055	0.0017	51.9	0.0493	3022.0	17646.6	3612.4

*displayed as the log(HR) **percent change

Table F-11: Simulation runs using PS Stratification – Medium Effect.

Preva-lence	N	% m error	Mean*	SE*	Bias*	MSE*	MSE* %**	Model SE mean	Num event	Num WA	Num RI
0.5%	1300	-50	0.3577	0.2899	-0.0097	0.0841	-18.0	0.2856	103.2	18190.3	3068.7
0.5%	1300	-30	0.3591	0.2791	-0.0083	0.0779	-9.3	0.2766	106.7	18110.8	3148.2
0.5%	1300	-10	0.3619	0.2696	-0.0056	0.0727	-2.0	0.2684	110.1	18031.3	3227.7
0.5%	1300	0	0.3621	0.2670	-0.0054	0.0713	0.0	0.2647	111.8	17991.7	3267.3
0.5%	1300	10	0.3686	0.2494	0.0012	0.0621	12.8	0.2454	121.8	17737.5	3521.5
0.5%	1300	30	0.3688	0.2184	0.0014	0.0477	33.1	0.2177	141.8	17228.8	4030.2
0.5%	1300	50	0.3645	0.1998	-0.0030	0.0399	44.0	0.1978	161.9	16720.9	4538.1
1%	1000	-50	0.3668	0.1966	-0.0007	0.0386	-13.8	0.2000	205.2	18190.5	3068.5
1%	1000	-30	0.3664	0.1911	-0.0010	0.0365	-7.5	0.1942	211.9	18111.0	3148.0
1%	1000	-10	0.3677	0.1856	0.0003	0.0344	-1.4	0.1888	218.6	18031.4	3227.6
1%	1000	0	0.3650	0.1843	-0.0024	0.0339	0.0	0.1863	222.0	17991.7	3267.3
1%	1000	10	0.3696	0.1716	0.0021	0.0294	13.3	0.1731	241.9	17737.4	3521.6
1%	1000	30	0.3681	0.1507	0.0007	0.0227	33.2	0.1539	281.3	17228.2	4030.8
1%	1000	50	0.3683	0.1385	0.0009	0.0192	43.5	0.1399	320.6	16719.9	4539.1
10%	200	-50	0.3696	0.0606	0.0022	0.0037	0.6	0.0631	2058.0	18187.6	3071.4
10%	200	-30	0.3683	0.0622	0.0009	0.0039	-4.8	0.0615	2115.7	18108.5	3150.5
10%	200	-10	0.3682	0.0622	0.0007	0.0038	-4.6	0.0600	2174.3	18028.8	3230.2
10%	200	0	0.3682	0.0608	0.0008	0.0037	-0.0	0.0593	2203.2	17988.4	3270.6
10%	200	10	0.3672	0.0561	-0.0002	0.0031	14.9	0.0555	2378.2	17736.2	3522.8
10%	200	30	0.3665	0.0477	-0.0009	0.0023	38.4	0.0497	2725.7	17225.9	4033.1
10%	200	50	0.3646	0.0438	-0.0028	0.0019	47.9	0.0454	3074.0	16714.1	4544.9

*displayed as the log(HR) **percent change

Table F-12: Simulation runs using PS Stratification – High Effect.

Preva-lence	N	% m error	Mean*	SE*	Bias*	MSE*	MSE* %**	Model SE mean	Num event	Num WA	Num RI
0.5%	1300	-50	0.3618	0.2878	-0.0056	0.0828	-16.9	0.2793	103.8	18058.4	3200.6
0.5%	1300	-30	0.3628	0.2803	-0.0047	0.0785	-10.9	0.2693	107.4	17926.4	3332.6
0.5%	1300	-10	0.3642	0.2698	-0.0032	0.0727	-2.7	0.2604	111.0	17794.1	3464.9
0.5%	1300	0	0.3651	0.2662	-0.0023	0.0708	0.0	0.2561	112.8	17728.1	3530.9
0.5%	1300	10	0.3693	0.2396	0.0018	0.0574	19.0	0.2356	123.5	17309.1	3949.9
0.5%	1300	30	0.3723	0.2061	0.0049	0.0425	40.0	0.2066	144.7	16471.5	4787.5
0.5%	1300	50	0.3715	0.1878	0.0041	0.0353	50.2	0.1863	165.9	15635.0	5624.0
1%	1000	-50	0.3668	0.1962	-0.0006	0.0385	-16.3	0.1962	206.2	18058.8	3200.2
1%	1000	-30	0.3657	0.1924	-0.0018	0.0370	-11.9	0.1895	213.2	17926.7	3332.3
1%	1000	-10	0.3655	0.1853	-0.0019	0.0343	-3.8	0.1835	220.3	17794.6	3464.4
1%	1000	0	0.3641	0.1819	-0.0033	0.0331	0.0	0.1808	223.9	17728.5	3530.5
1%	1000	10	0.3698	0.1617	0.0023	0.0261	21.0	0.1665	244.9	17309.5	3949.5
1%	1000	30	0.3712	0.1387	0.0037	0.0192	41.8	0.1463	286.7	16471.8	4787.2
1%	1000	50	0.3753	0.1280	0.0079	0.0164	50.3	0.1319	328.4	15635.0	5624.0
10%	200	-50	0.3702	0.0597	0.0027	0.0036	2.2	0.0620	2066.3	18056.1	3202.9
10%	200	-30	0.3681	0.0617	0.0007	0.0038	-4.4	0.0601	2127.2	17923.7	3335.3
10%	200	-10	0.3687	0.0621	0.0013	0.0038	-5.7	0.0584	2188.8	17791.8	3467.2
10%	200	0	0.3669	0.0604	-0.0005	0.0036	0.0	0.0576	2219.4	17724.9	3534.1
10%	200	10	0.3671	0.0528	-0.0003	0.0028	23.5	0.0534	2404.4	17307.7	3951.3
10%	200	30	0.3672	0.0472	-0.0002	0.0022	38.9	0.0473	2770.5	16471.5	4787.5
10%	200	50	0.3670	0.0431	-0.0005	0.0019	49.0	0.0429	3137.6	15635.5	5623.5

*displayed as the log(HR) **percent change

F-2 Plots at given prevalence for different effect sizes

IPTW for ATE

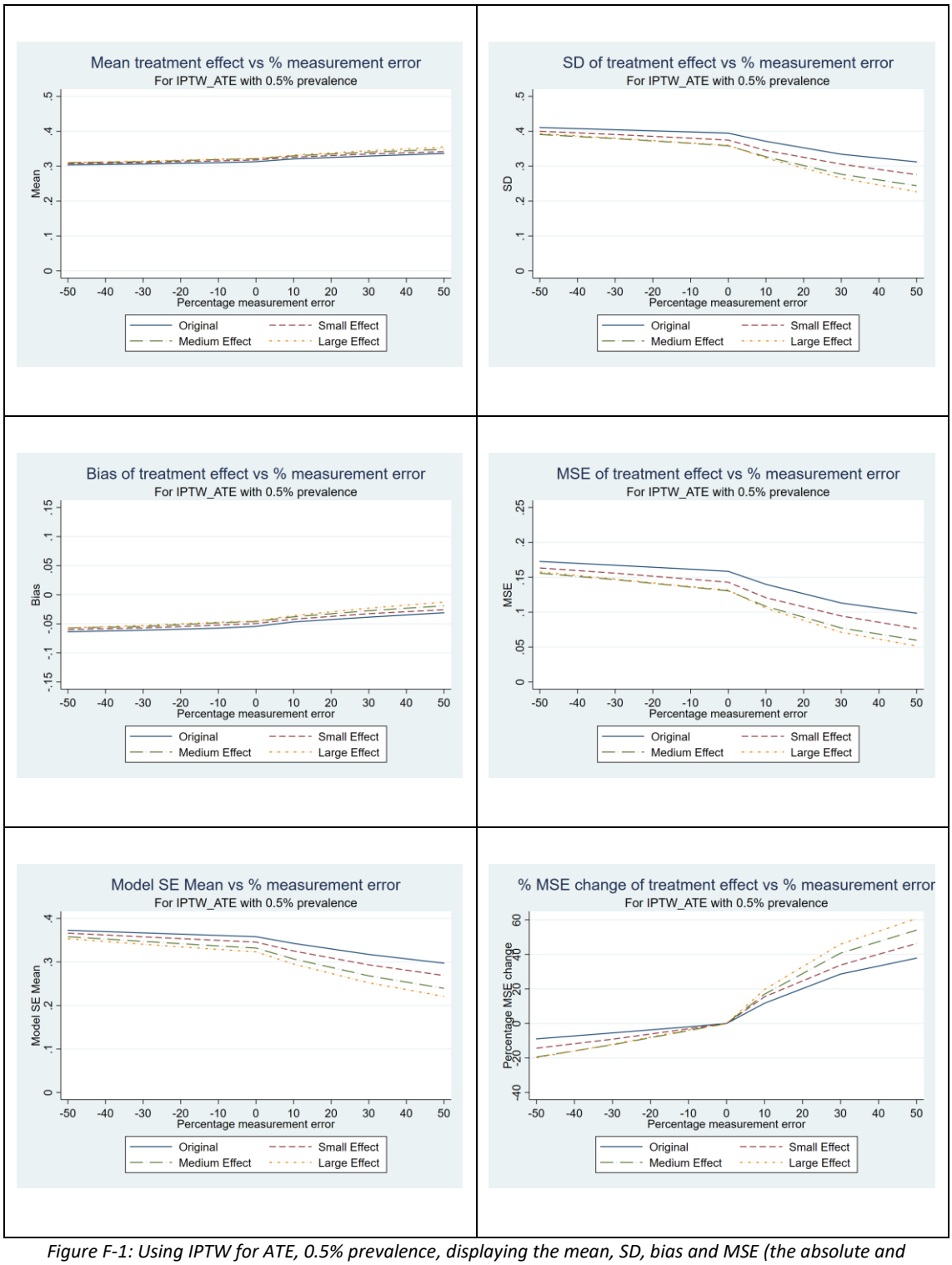


Figure F-1: Using IPTW for ATE, 0.5% prevalence, displaying the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean for different effect sizes.

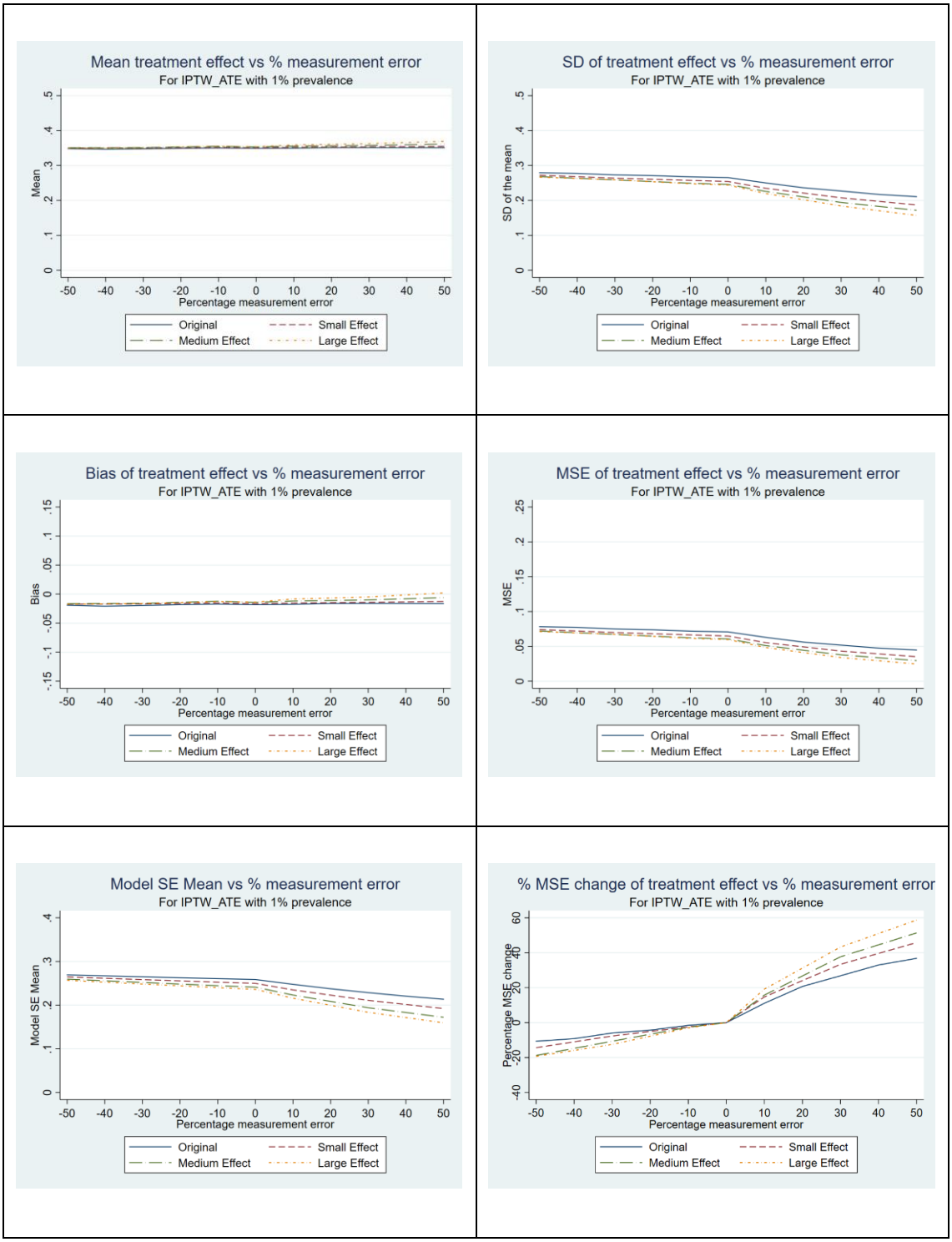


Figure F-2: IPTW for ATE, 1% prevalence, displaying the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean for different effect sizes.

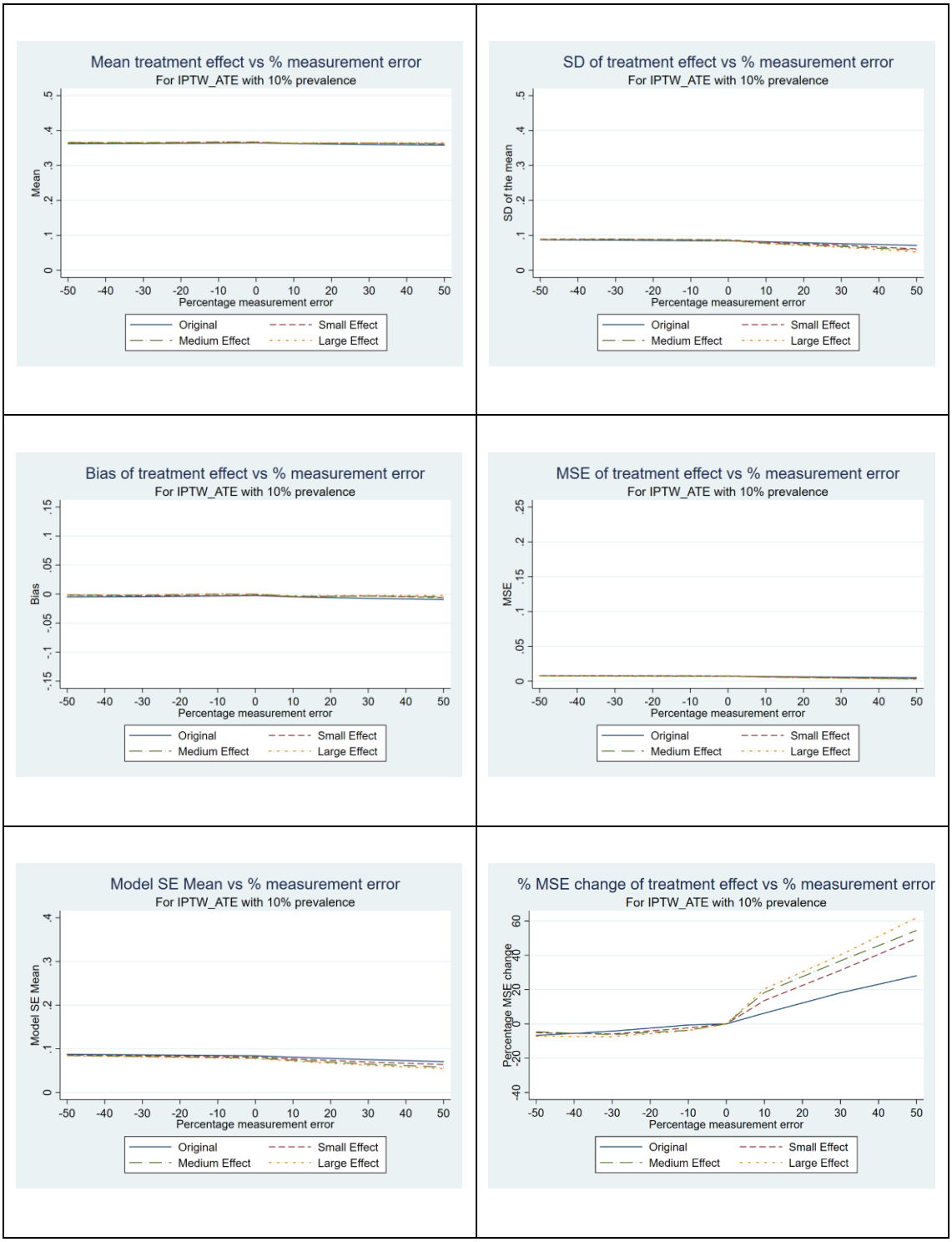


Figure F-3: IPTW for ATE, 10% prevalence, displaying the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean for different effect sizes.

IPTW for ATT

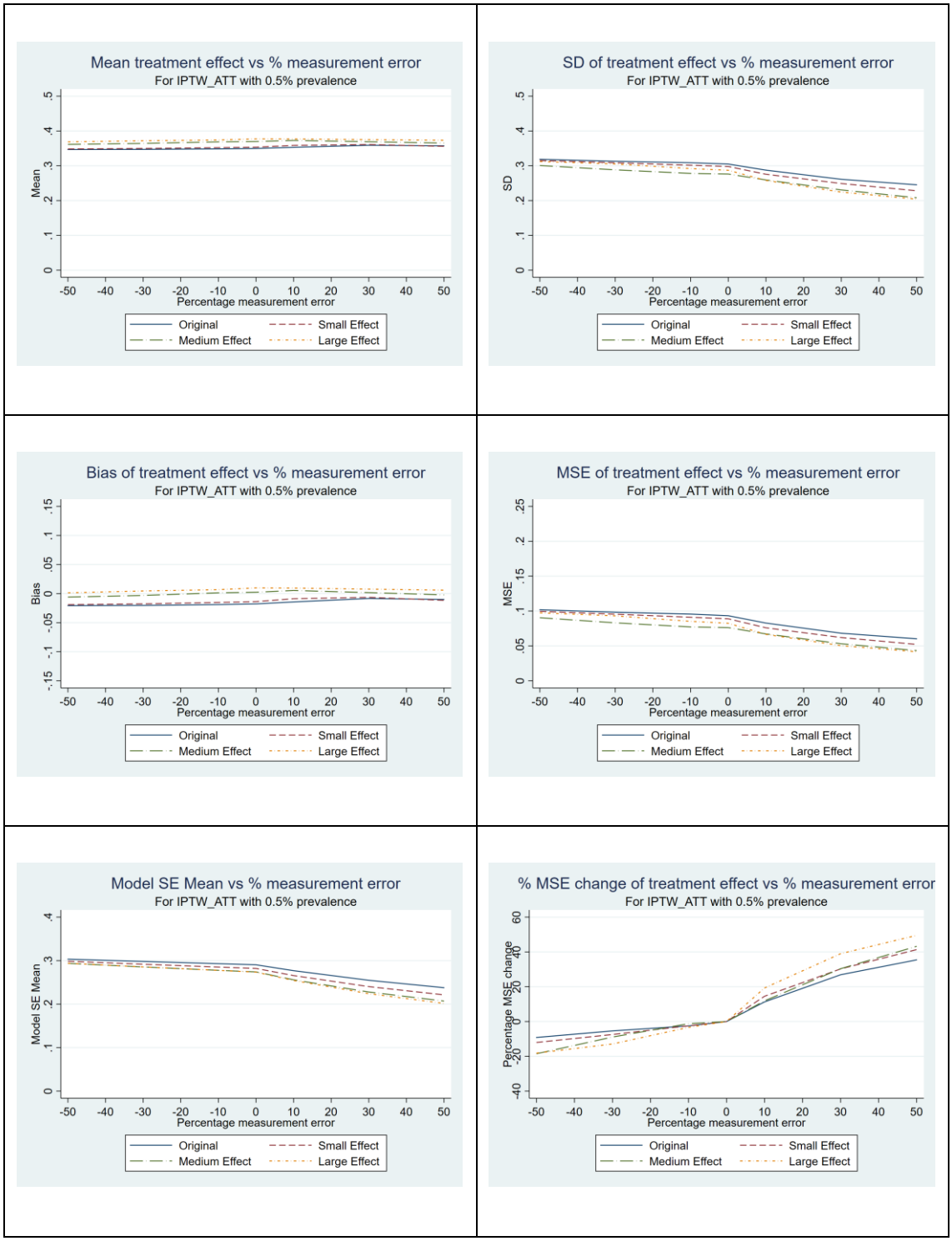


Figure F-4: IPTW for ATT, 0.5% prevalence, displaying the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean for different effect sizes.

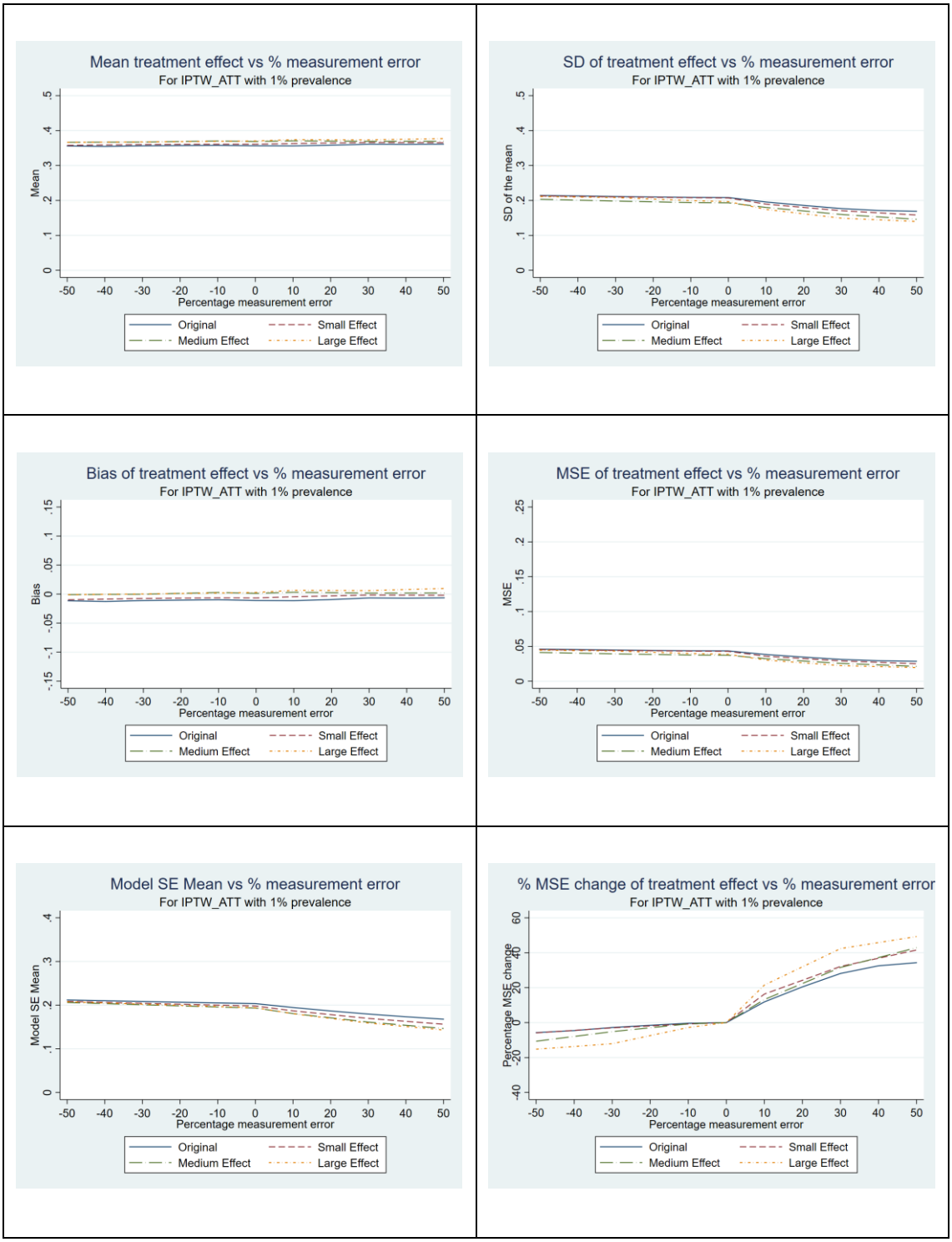


Figure F-5: IPTW for ATT, 1% prevalence, displaying the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean for different effect sizes.

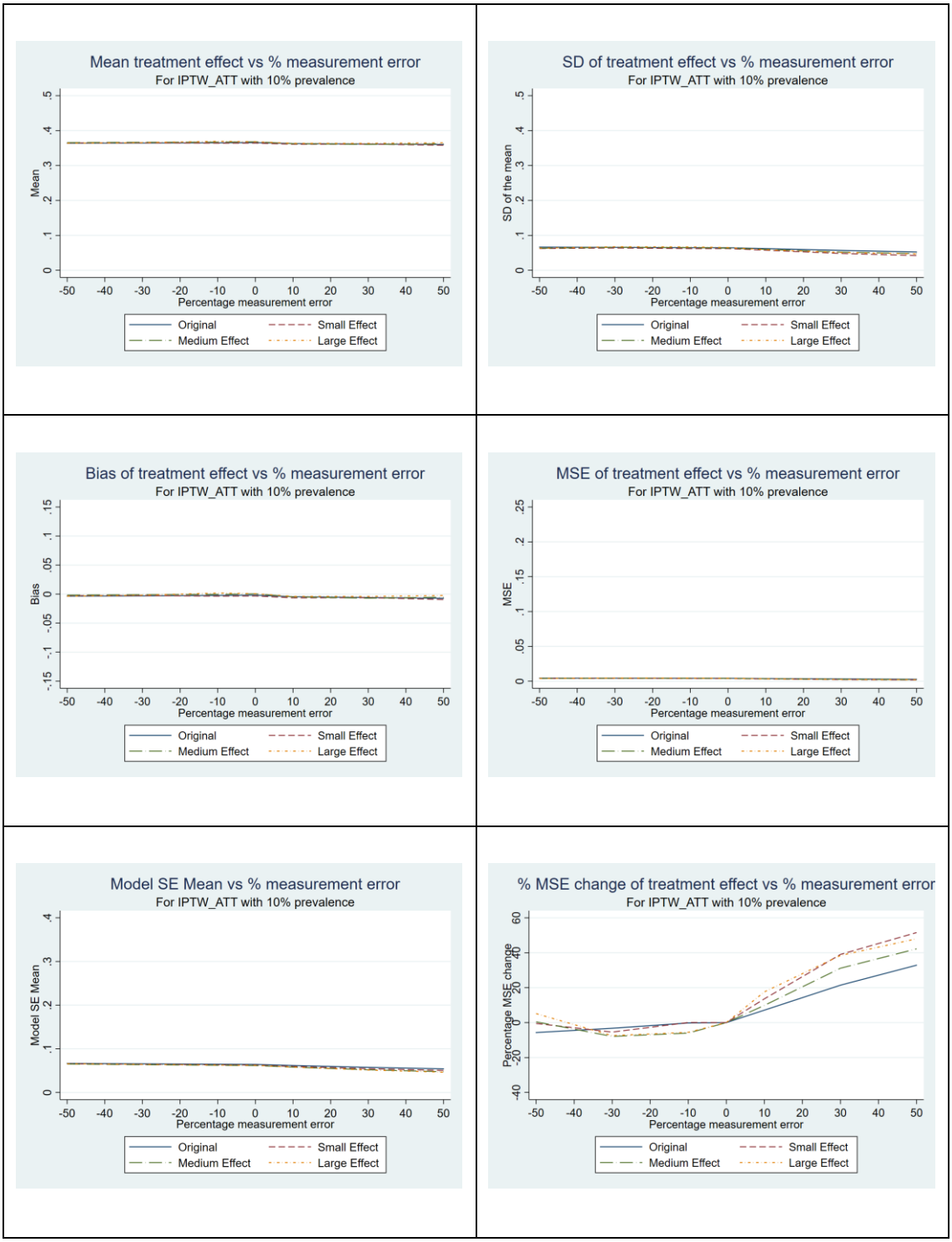


Figure F-6: IPTW for ATT, 10% prevalence, displaying the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean for different effect sizes.

3to1 PS matching

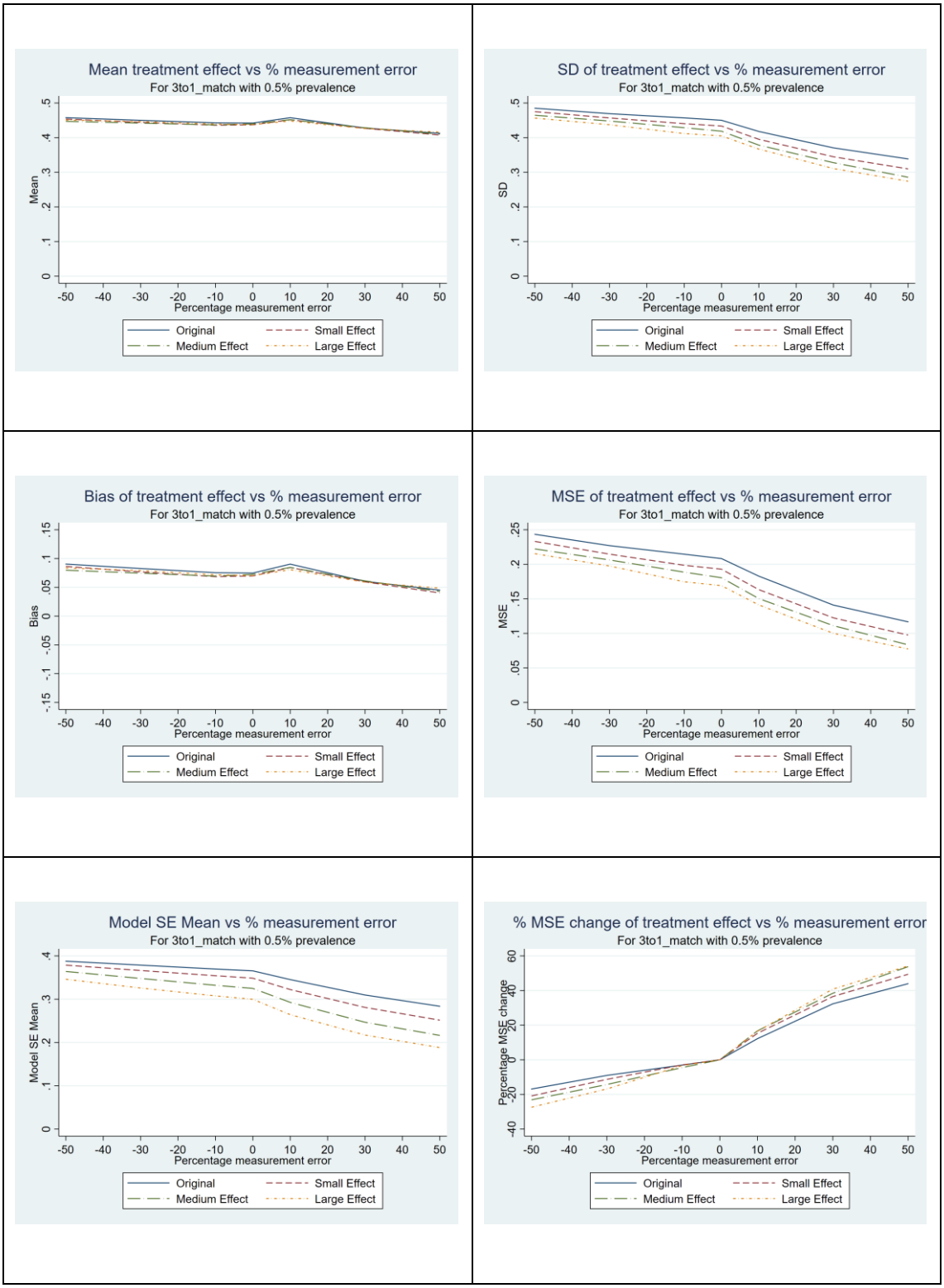


Figure F-7: 3to1 PS matching, 0.5% prevalence, displaying the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean for different effect sizes.

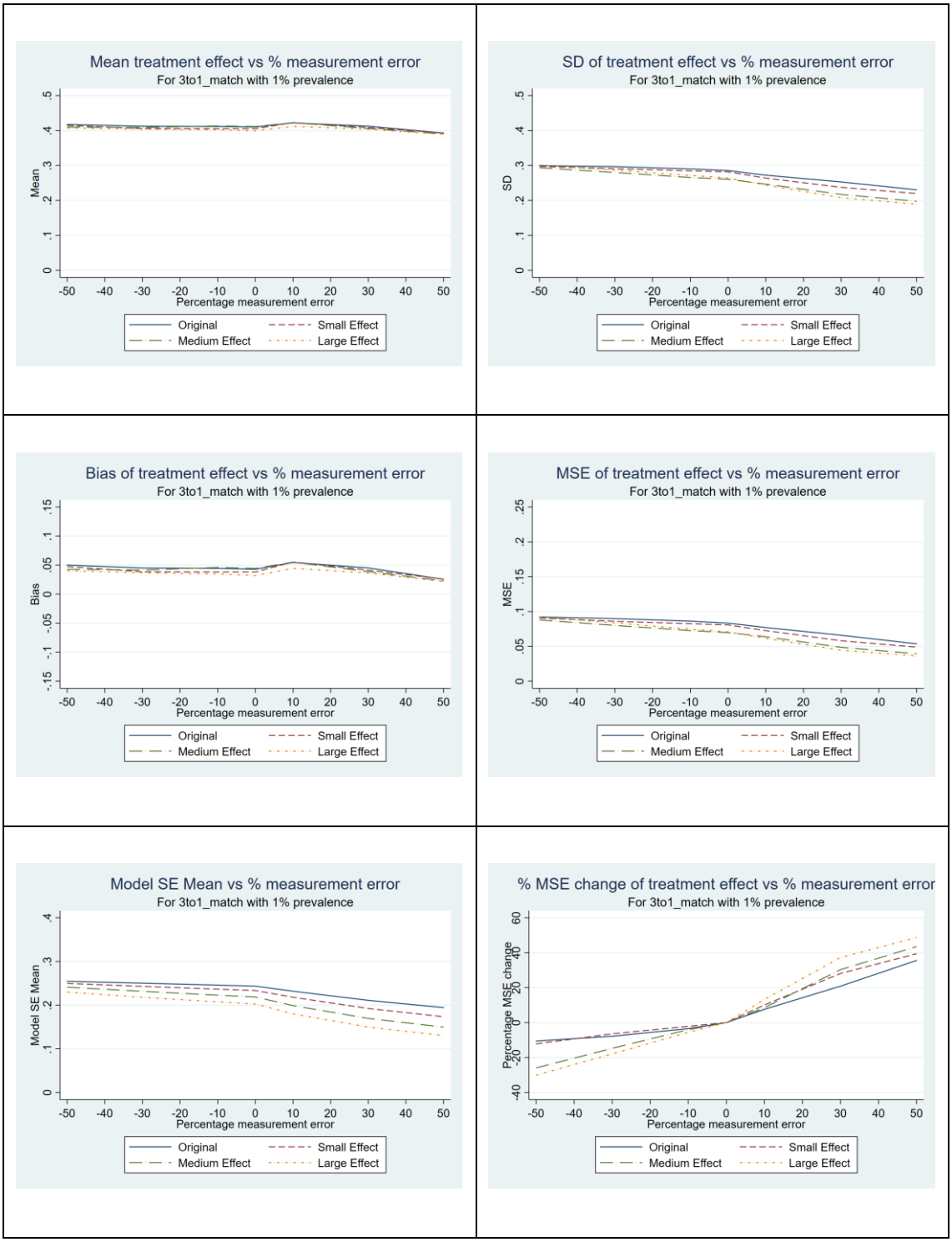


Figure F-8: 3to1 PS matching, 1% prevalence, displaying the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean for different effect sizes.

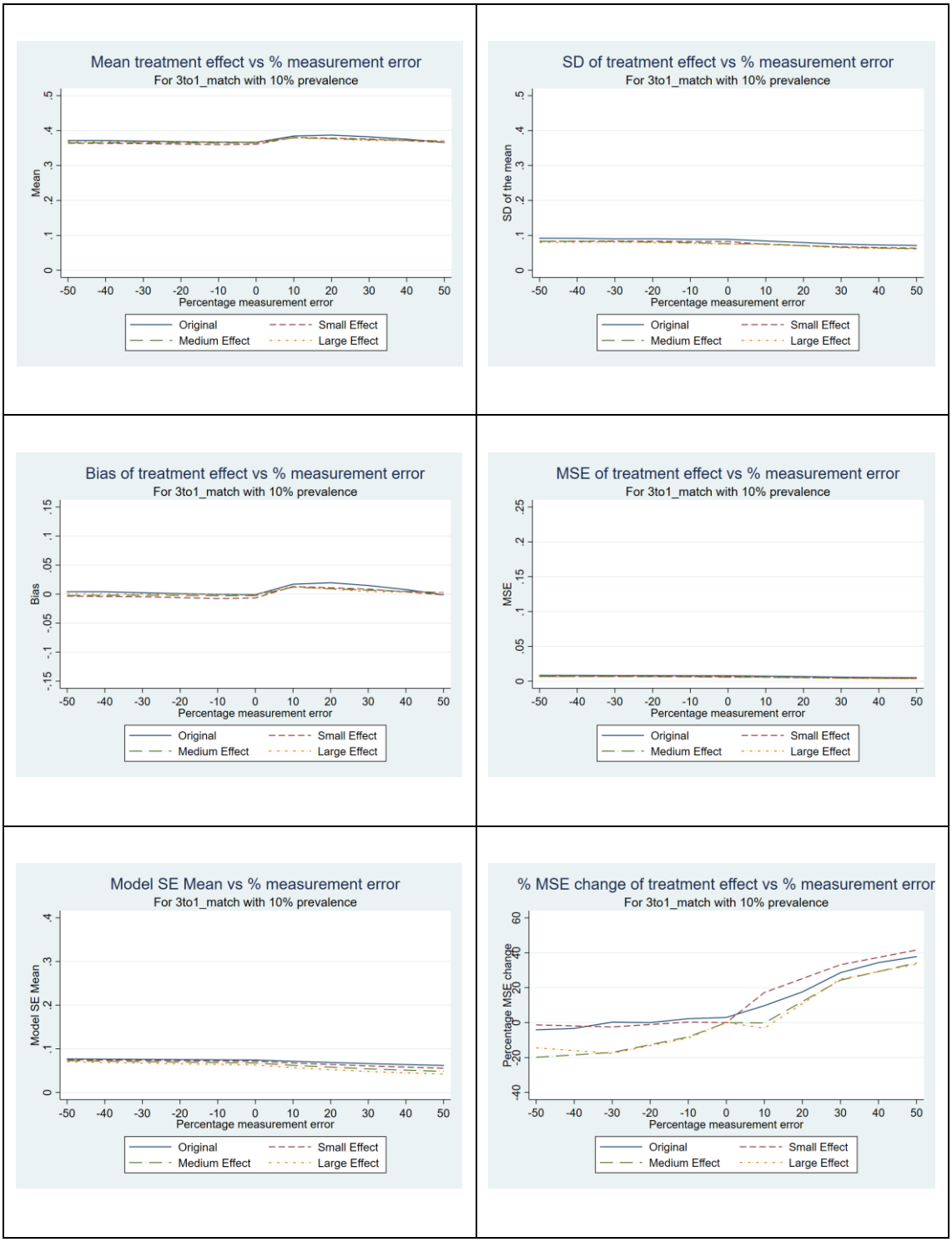


Figure F-9: 3to1 PS matching, 10% prevalence, displaying the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean for different effect sizes.

PS Stratification

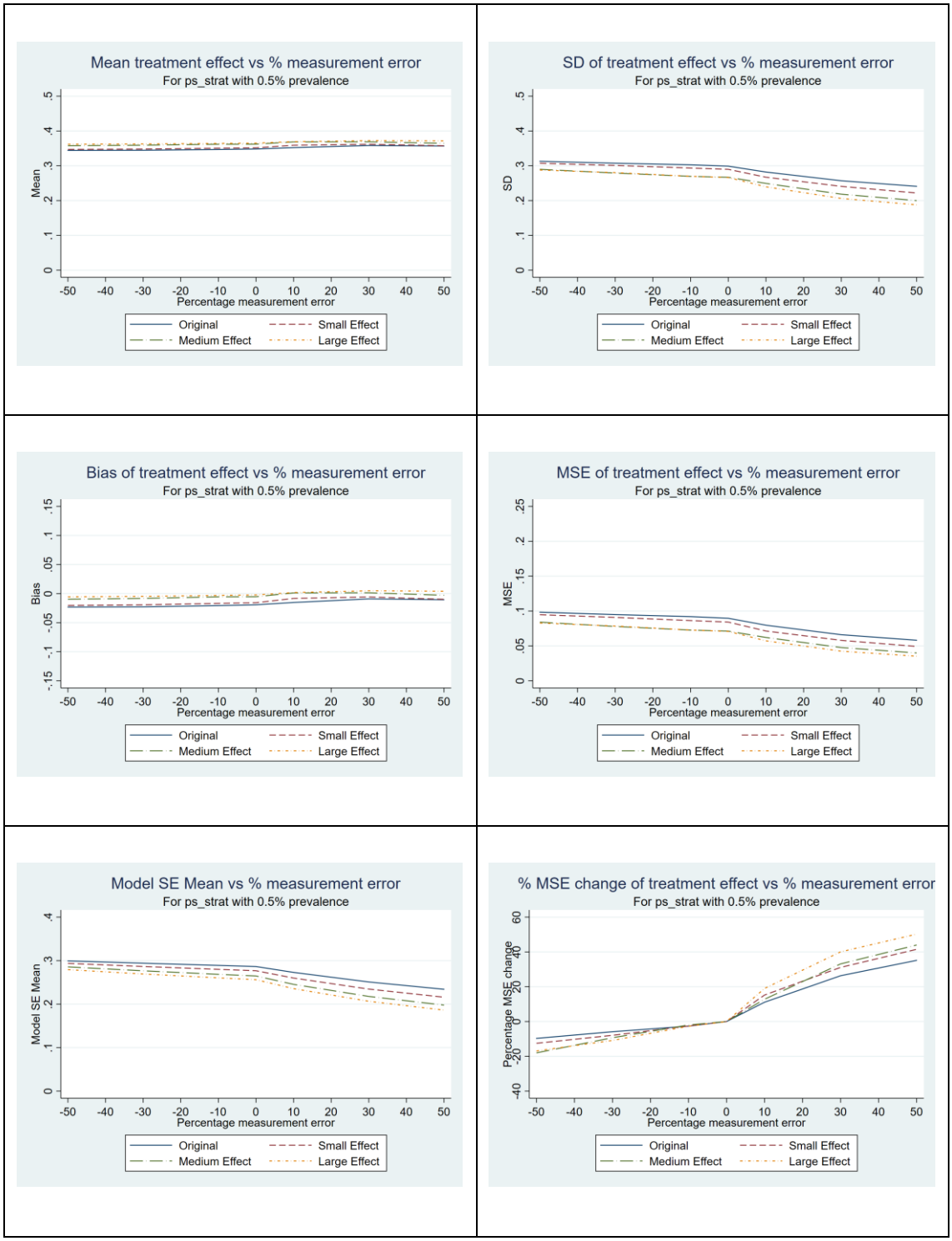


Figure F-10: PS stratification, 0.5% prevalence, displaying the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean for different effect sizes.

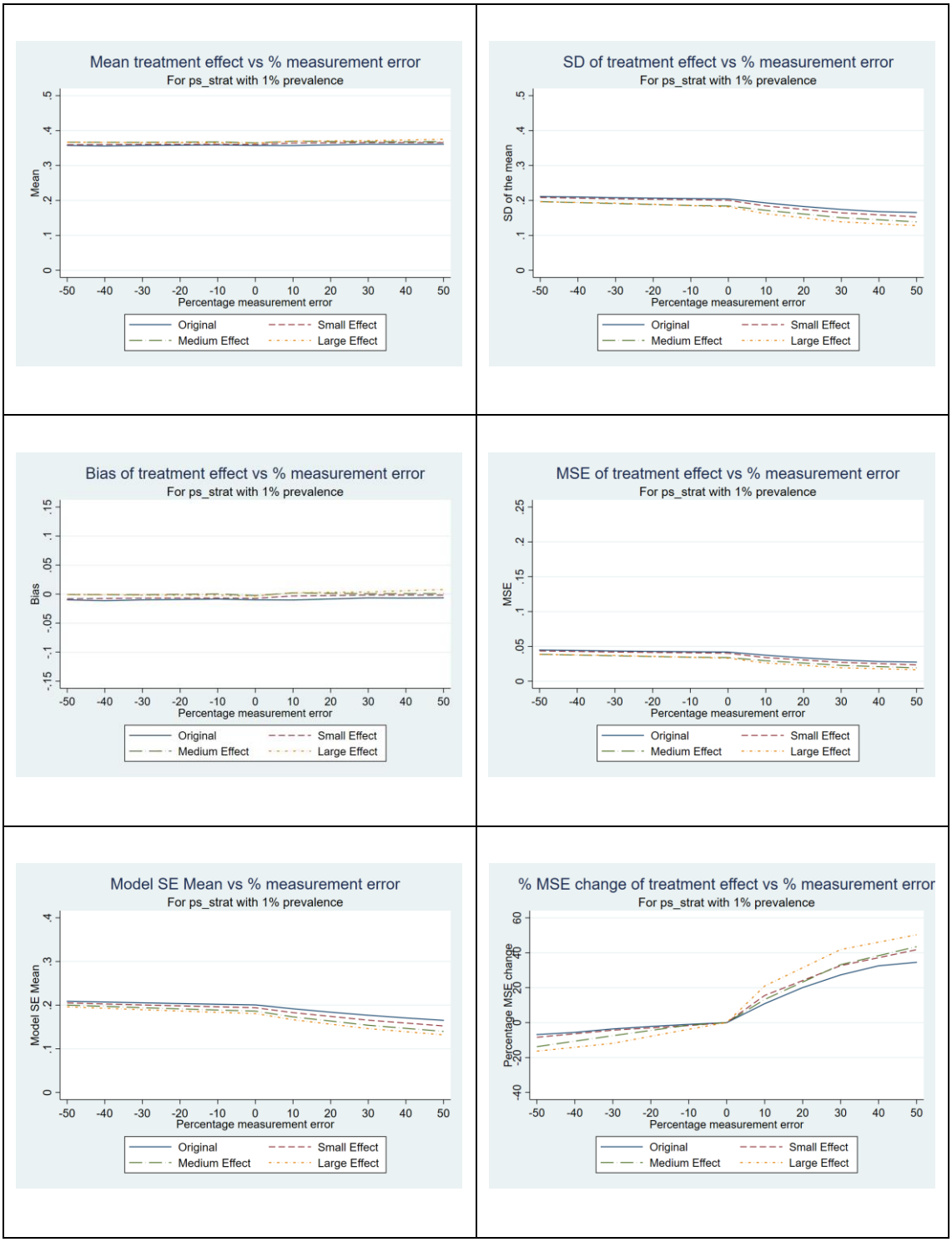


Figure F-11: PS stratification, 1% prevalence, displaying the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean for different effect sizes.

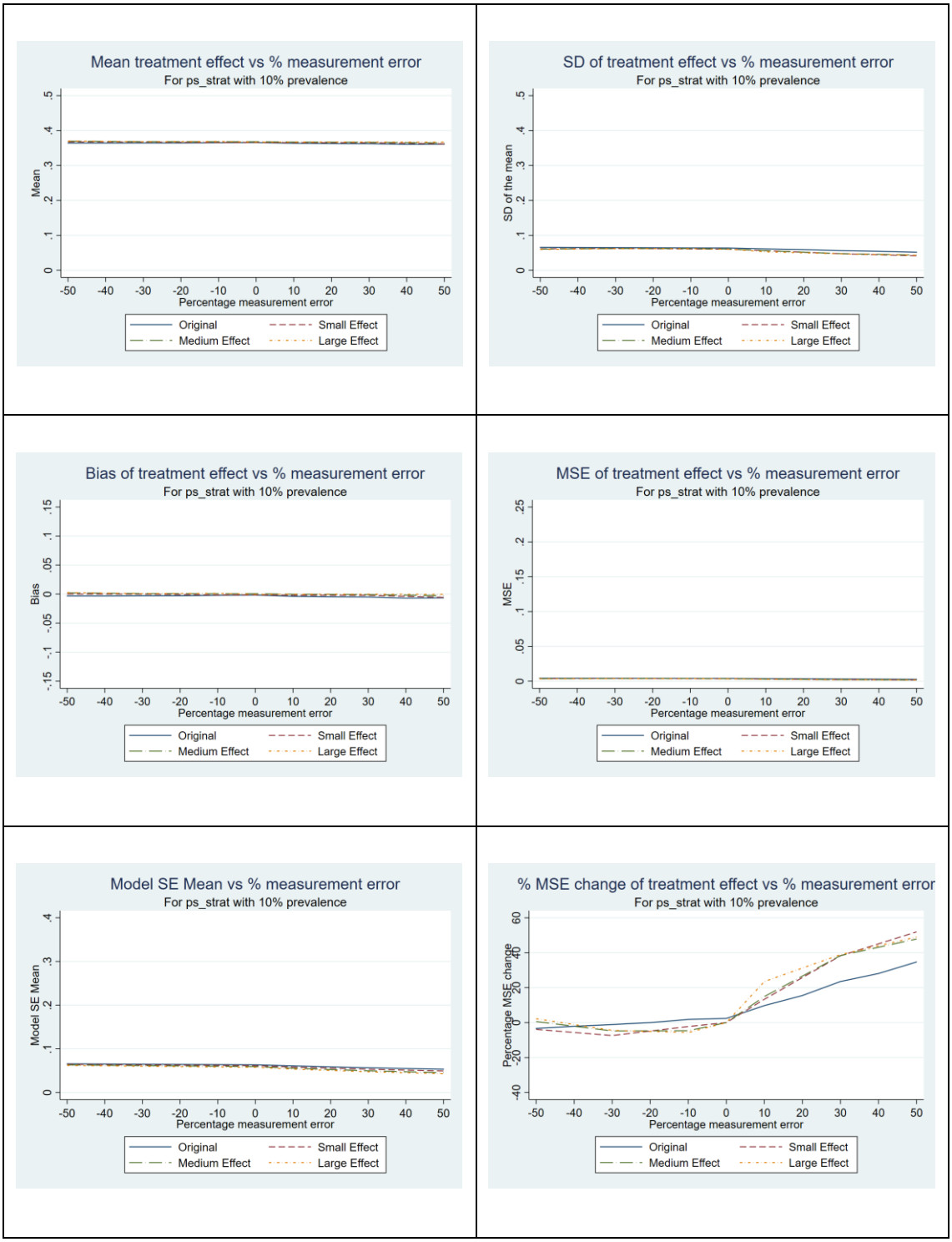


Figure F-12: PS stratification, 10% prevalence, displaying the mean, SD, bias and MSE (the absolute and percentage change) of the estimated treatment effect displayed as log(HR) and the model SE mean for different effect sizes.

APPENDIX G – ADDITIONAL TABLES AND GRAPHS

G-1 Tables and graphs for 5% prevalence simulations

In the headers in all tables in this Appendix, *Prevalence* is Outcome Prevalence, *Num Events* is the Number of Future Strokes, *Num WA* is the Number of participants prescribed Warfarin and *Num RI* is the Number of participants prescribed Rivaroxaban.

5% prevalence runs for IPTW for ATE

Table G-1: 5% prevalence simulation runs for IPTW for ATE.

Effect size	Preva-lence	N	% m error	Mean*	SE*	Bias*	MSE*	MSE* %**	Num event	Num WA	Num RI
ORIG	5%	250	-50	0.3594	0.1094	-0.0080	0.0120	-14.9	1184.7	18367.3	2891.7
ORIG	5%	250	-30	0.3596	0.1069	-0.0079	0.0115	-9.7	1218.5	18359.5	2899.5
ORIG	5%	250	-10	0.3632	0.1040	-0.0042	0.0108	-3.3	1252.5	18351.8	2907.2
ORIG	5%	250	0	0.3643	0.1023	-0.0031	0.0104	0.0	1268.8	18347.8	2911.1
ORIG	5%	250	10	0.3628	0.1008	-0.0047	0.0101	2.9	1367.9	18323.0	2936.0
ORIG	5%	250	30	0.3606	0.0944	-0.0068	0.0089	14.5	1566.2	18273.6	2985.4
ORIG	5%	250	50	0.3617	0.0878	-0.0058	0.0077	26.2	1767.8	18223.5	3035.5
SMALL	5%	250	-50	0.3591	0.1100	-0.0083	0.0121	-13.1	1186.4	18300.0	2959.0
SMALL	5%	250	-30	0.3593	0.1078	-0.0081	0.0116	-8.6	1220.7	18264.4	2994.6
SMALL	5%	250	-10	0.3611	0.1045	-0.0063	0.0109	-1.9	1255.2	18229.7	3029.3
SMALL	5%	250	0	0.3612	0.1035	-0.0062	0.0107	-0.0	1271.7	18212.3	3046.7
SMALL	5%	250	10	0.3615	0.0959	-0.0059	0.0092	14.2	1373.0	18100.8	3158.2
SMALL	5%	250	30	0.3634	0.0848	-0.0041	0.0072	33.0	1577.6	17873.8	3385.2
SMALL	5%	250	50	0.3635	0.0757	-0.0040	0.0057	46.6	1786.0	17647.6	3611.4
MED	5%	250	-50	0.3603	0.1075	-0.0072	0.0116	-13.9	1188.9	18188.6	3070.4
MED	5%	250	-30	0.3610	0.1049	-0.0064	0.0110	-8.4	1224.2	18109.4	3149.6
MED	5%	250	-10	0.3635	0.1024	-0.0039	0.0105	-3.0	1260.3	18029.6	3229.4
MED	5%	250	0	0.3633	0.1009	-0.0041	0.0102	-0.0	1277.2	17989.6	3269.4
MED	5%	250	10	0.3610	0.0910	-0.0065	0.0083	18.3	1382.4	17737.2	3521.8
MED	5%	250	30	0.3620	0.0797	-0.0054	0.0064	37.4	1596.6	17227.7	4031.3
MED	5%	250	50	0.3651	0.0693	-0.0023	0.0048	52.8	1815.8	16715.6	4543.4
HIGH	5%	250	-50	0.3601	0.1076	-0.0073	0.0116	-11.1	1192.3	18057.7	3201.3
HIGH	5%	250	-30	0.3623	0.1049	-0.0051	0.0110	-5.3	1229.0	17925.7	3333.3
HIGH	5%	250	-10	0.3644	0.1028	-0.0030	0.0105	-1.0	1265.9	17793.3	3465.7
HIGH	5%	250	0	0.3637	0.1023	-0.0038	0.0104	0.0	1283.6	17726.8	3532.2
HIGH	5%	250	10	0.3620	0.0903	-0.0055	0.0082	21.8	1394.5	17309.1	3949.9
HIGH	5%	250	30	0.3646	0.0754	-0.0028	0.0057	45.6	1618.6	16474.3	4784.7
HIGH	5%	250	50	0.3678	0.0630	0.0004	0.0039	62.1	1851.0	15636.8	5622.2

*displayed as the log(HR) **percent change

5% prevalence runs for IPTW for ATE

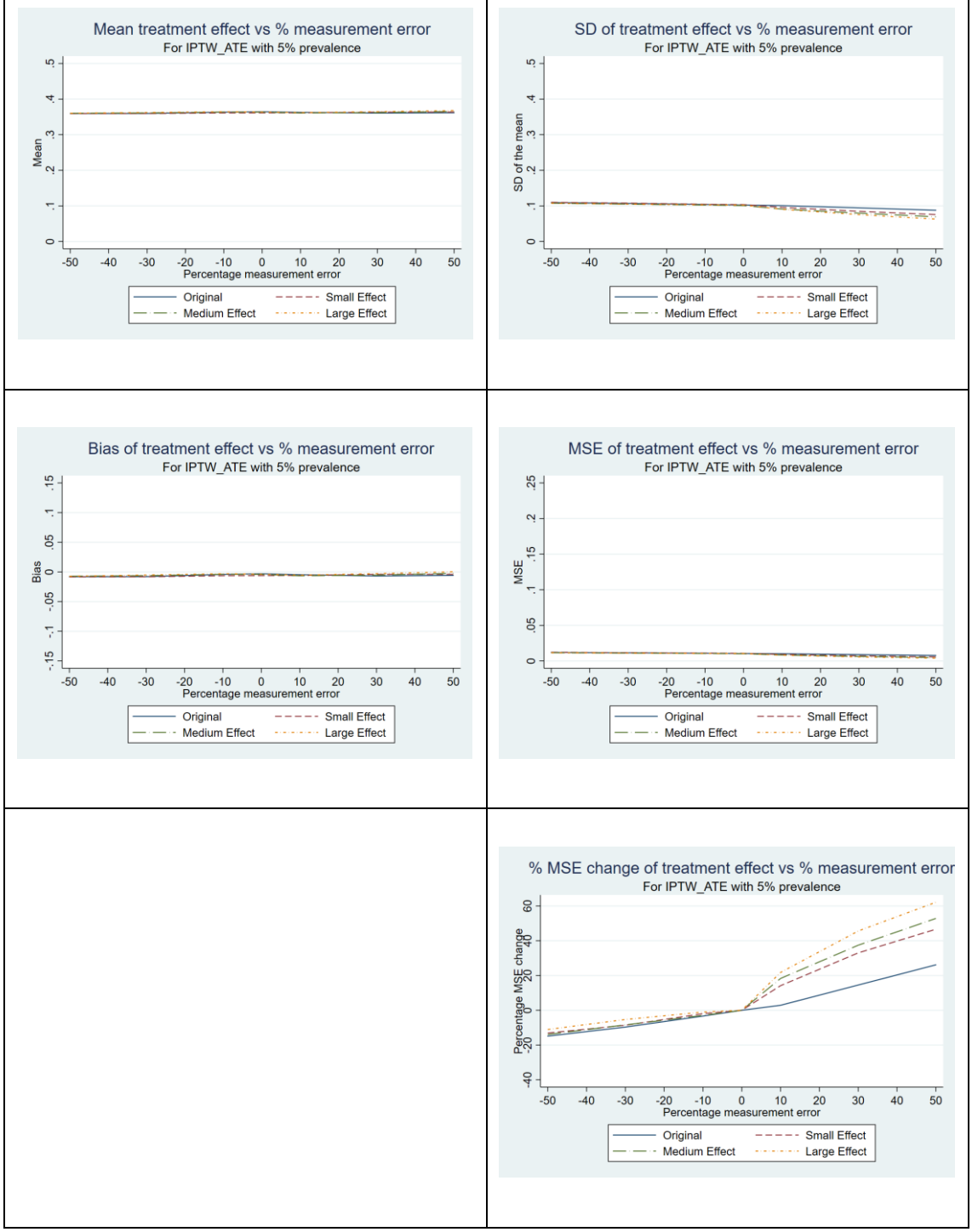


Figure G-1:IPTW for ATE, 5% prevalence - the mean, SD, bias and MSE (absolute & % change) of the estimated treatment effect displayed as log(HR).

G-2 Graphs plotting all PS methods

10% prevalence

10% prevalence with Original effect size

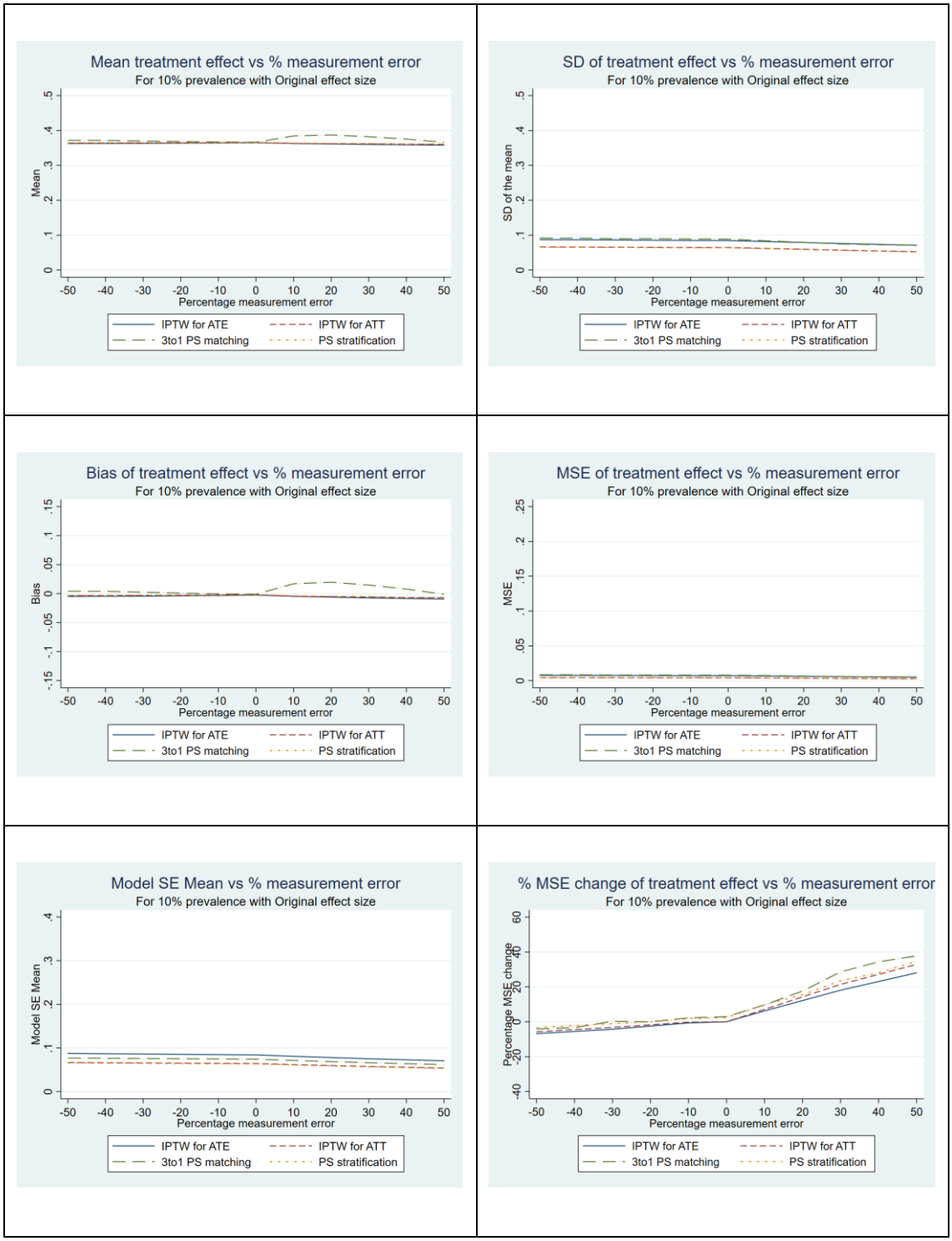


Figure G-2: Study PS methods, 10% prevalence, original effect size - the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect displayed as log(HR).

10% prevalence with Small effect size

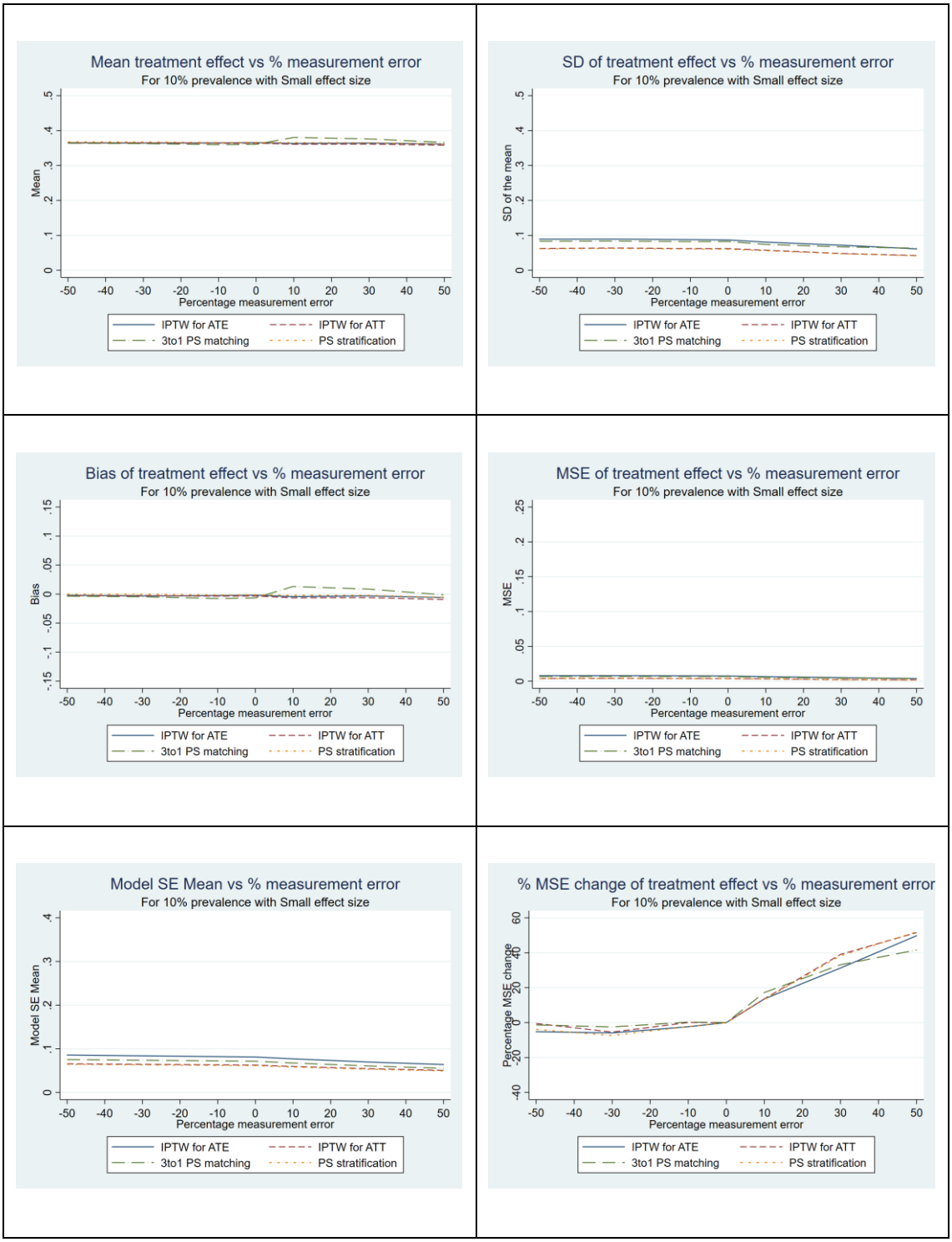


Figure G-3: Study PS methods, 10% prevalence, Small effect size - the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect displayed as log(HR).

10% prevalence with Medium effect size

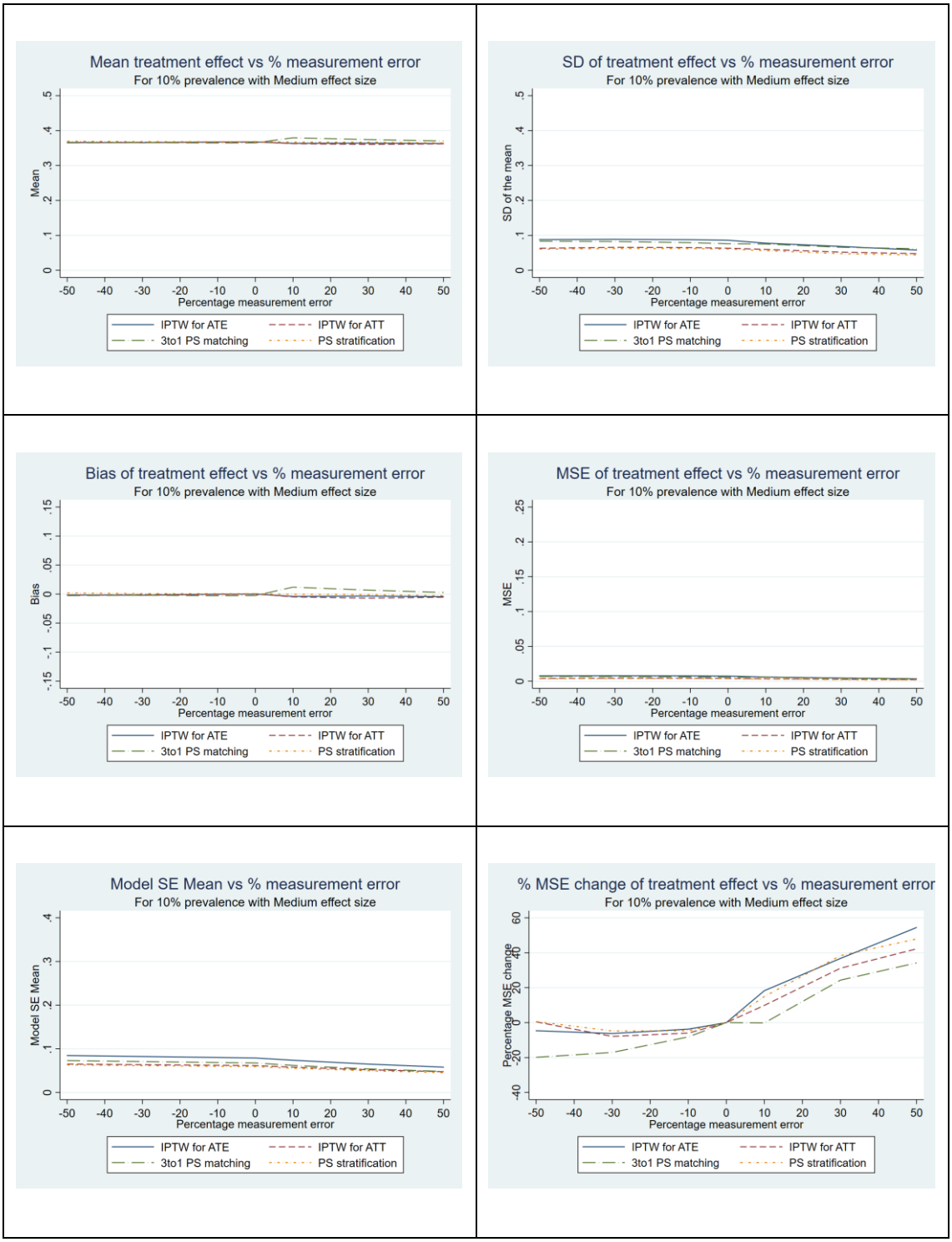


Figure G-4: Study PS methods, 10% prevalence, Medium effect size - the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect displayed as log(HR).

10% prevalence with High effect size

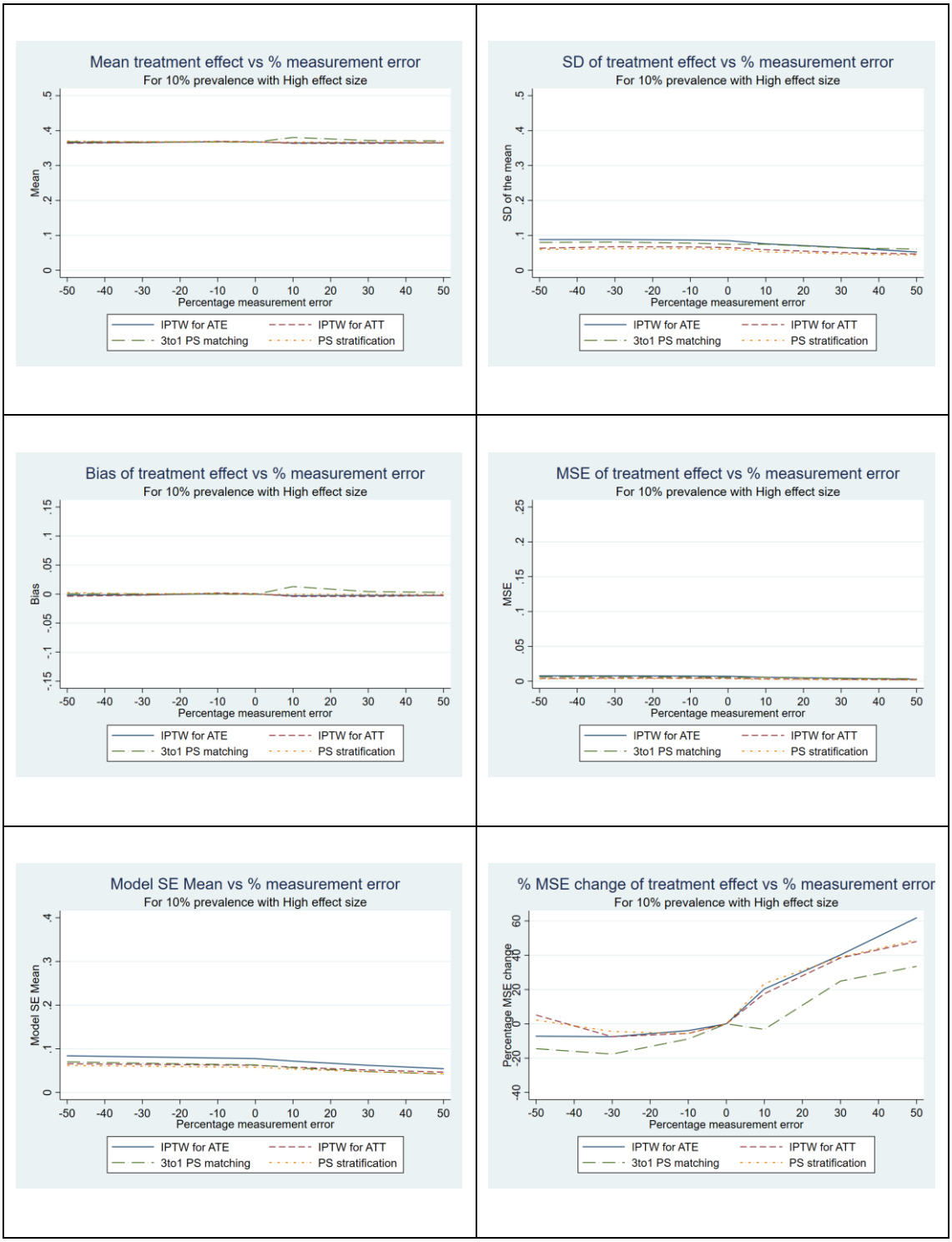


Figure G-5: Study PS methods, 10% prevalence, High effect size - the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect displayed as log(HR).

1% prevalence

1% prevalence for Original effect size

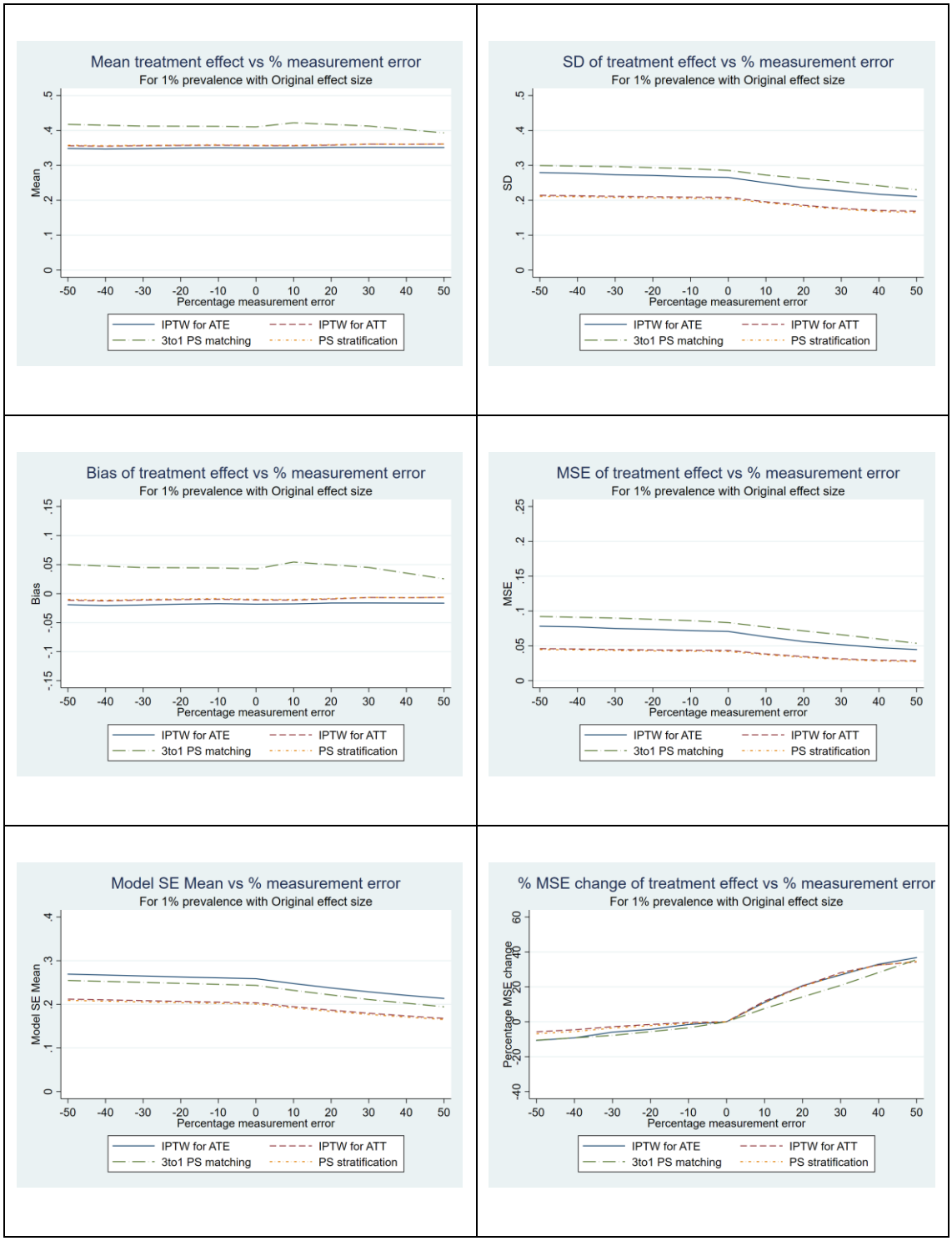


Figure G-6: Study PS methods, 1% prevalence, Original effect size - the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect displayed as log(HR).

1% prevalence with Small effect size

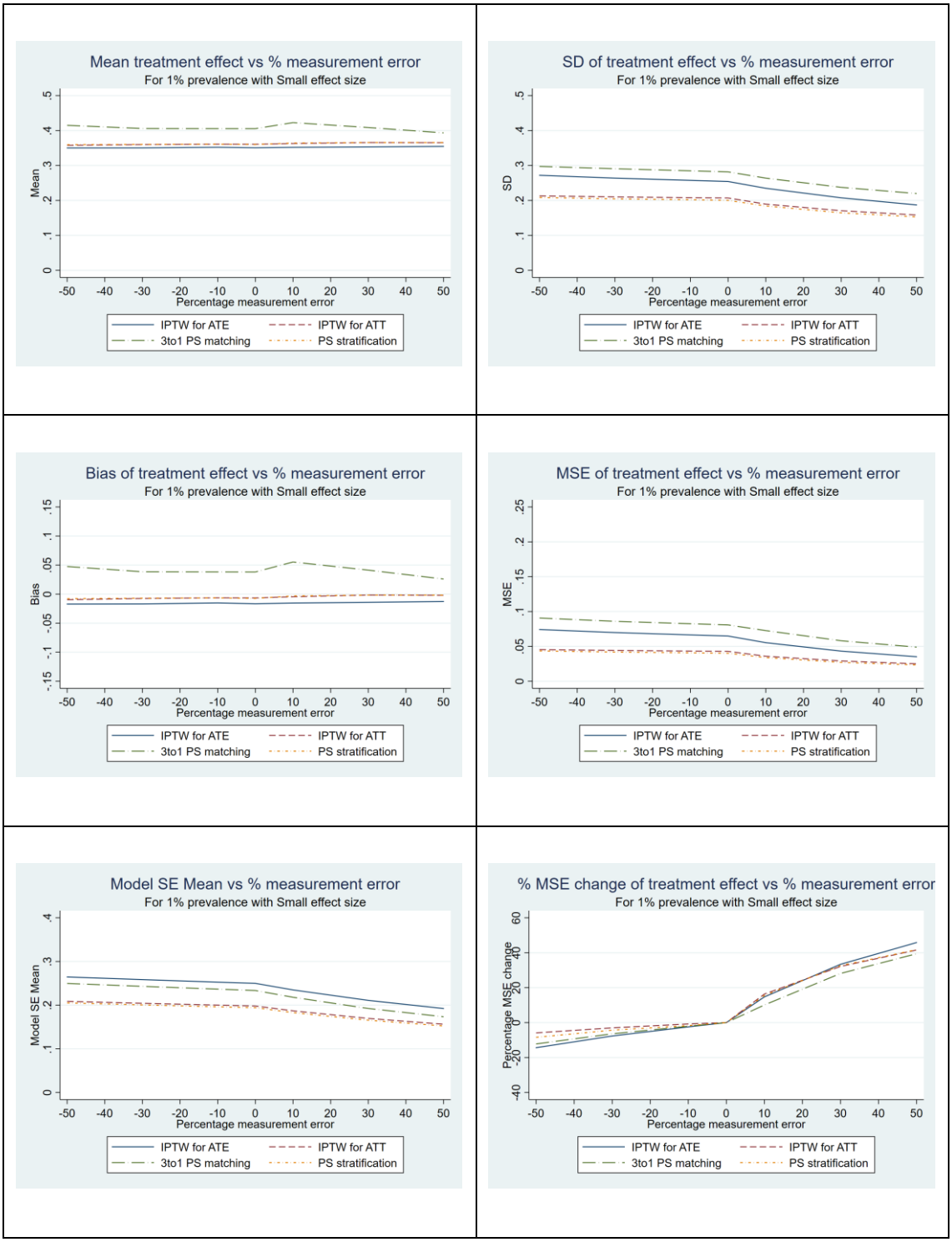


Figure G-7: Study PS methods, 1% prevalence, Small effect size - the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect displayed as log(HR).

1% prevalence with Medium effect size

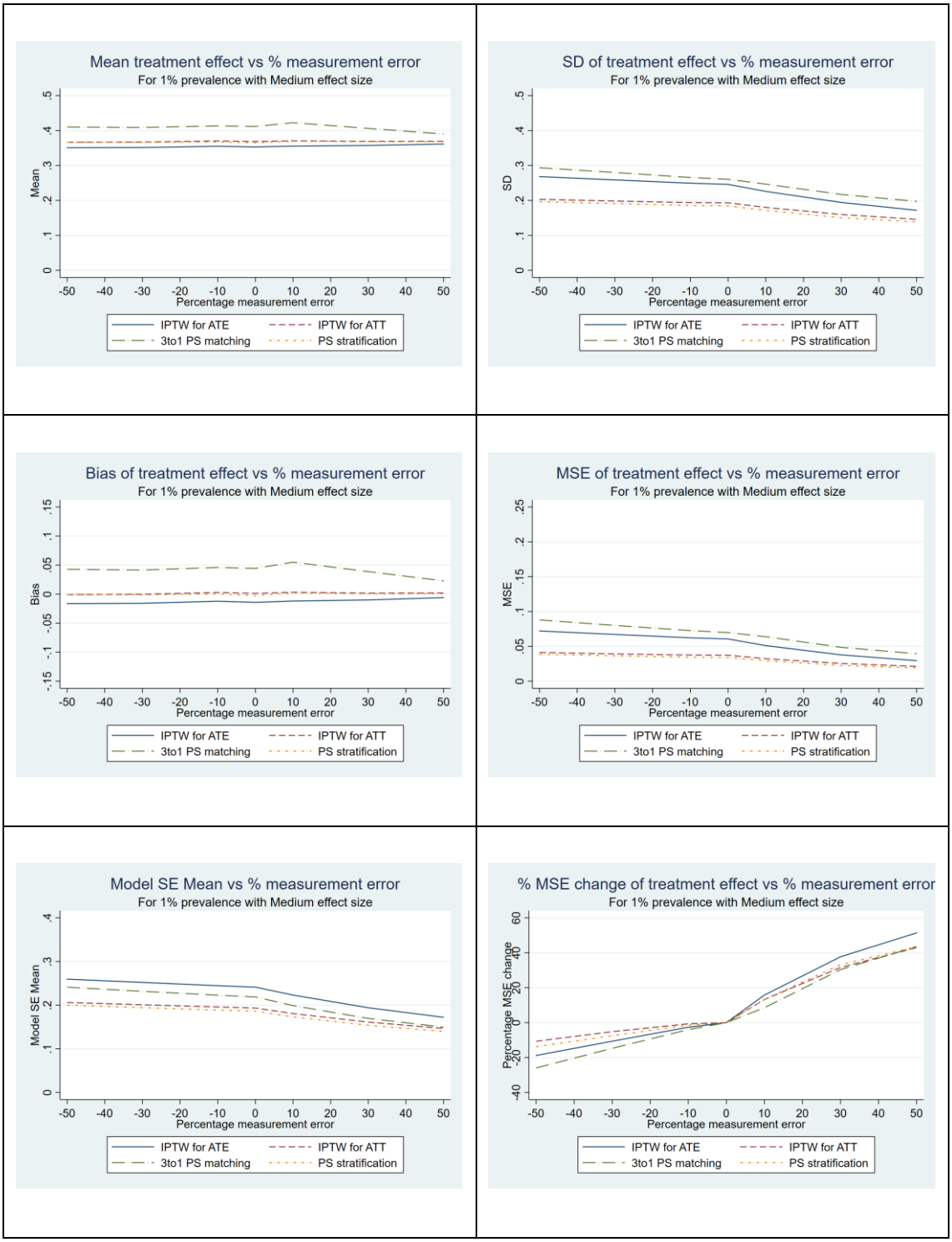


Figure G-8: Study PS methods, 1% prevalence, Medium effect size - the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect displayed as log(HR).

1% prevalence with High effect size

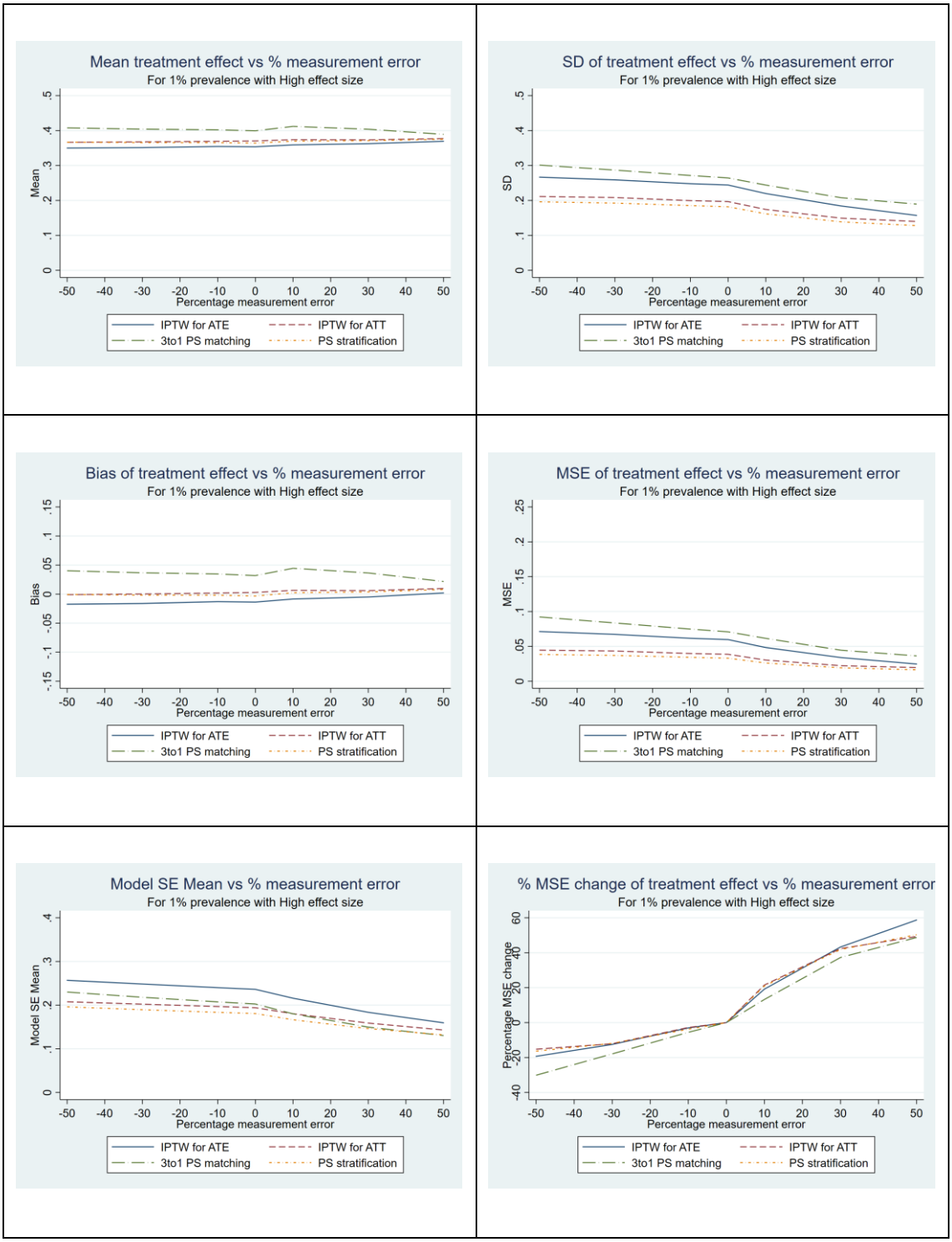


Figure G-9: Study PS methods, 1% prevalence, High effect size - the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect displayed as log(HR).

0.5% prevalence

0.5% prevalence with Original effect size

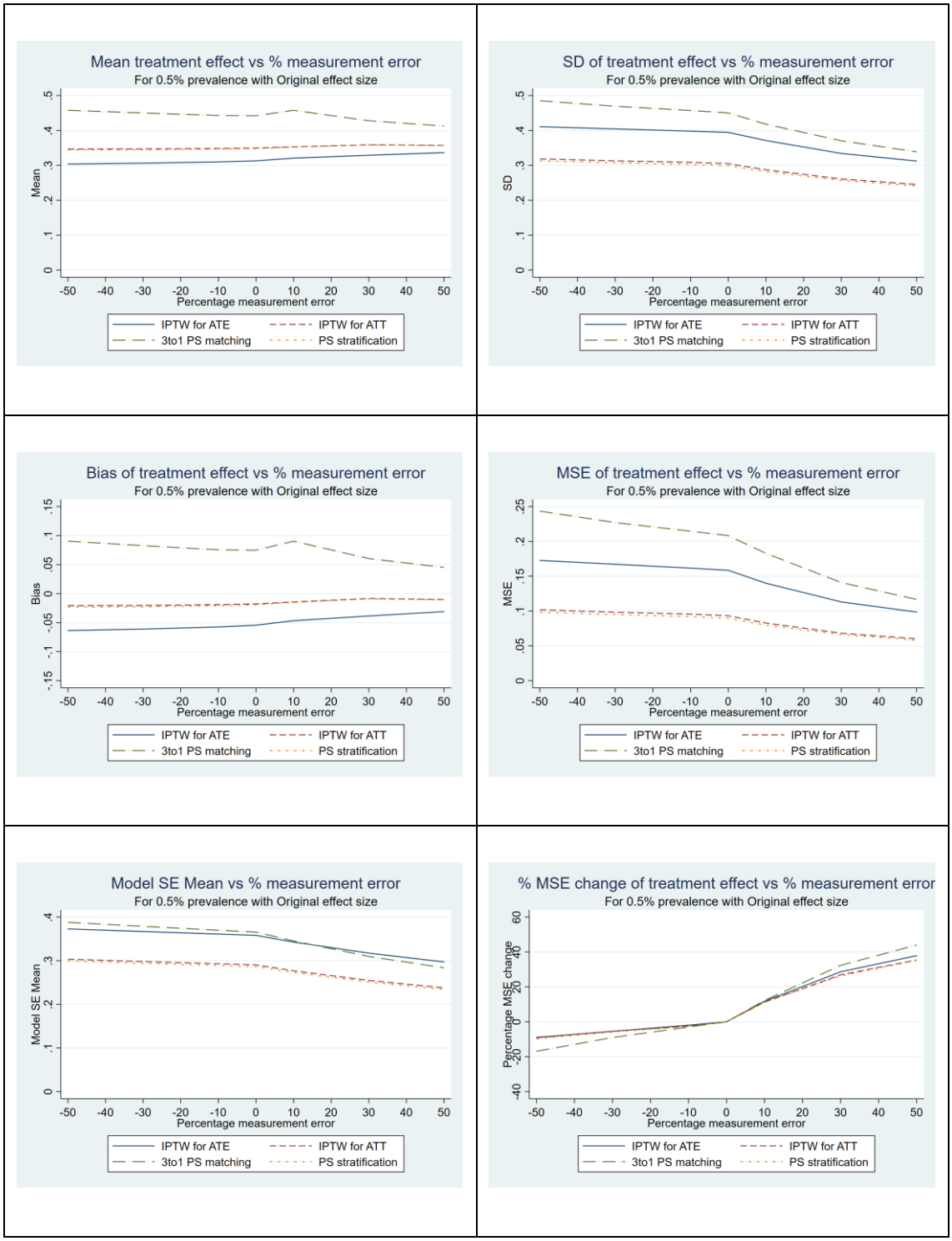


Figure G-10: Study PS methods, 0.5% prevalence, Original effect size - the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect displayed as log(HR).

0.5% prevalence for Small effect size

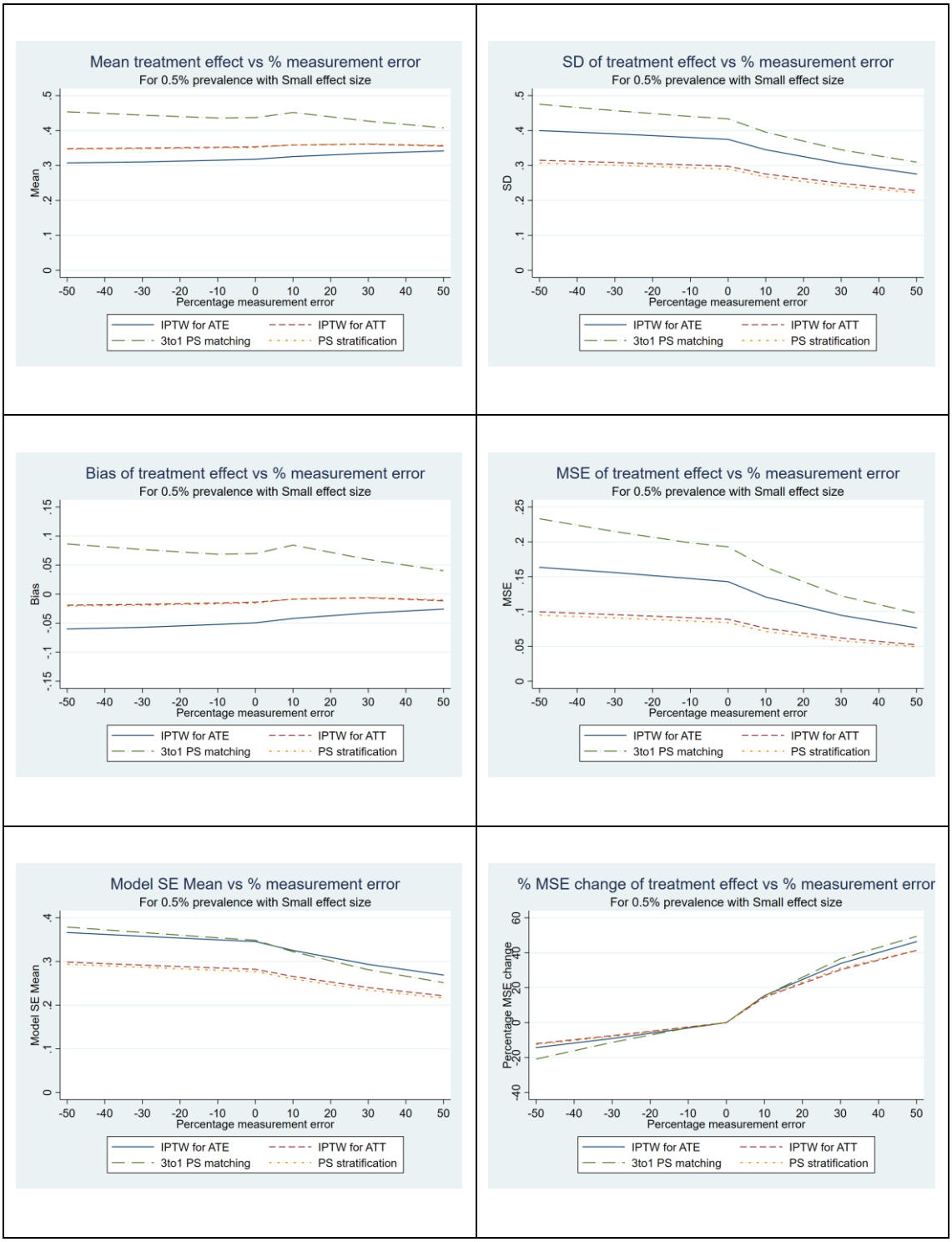


Figure G-11: Study PS methods, 0.5% prevalence, Small effect size - the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect displayed as log(HR).

0.5% prevalence with Medium effect size

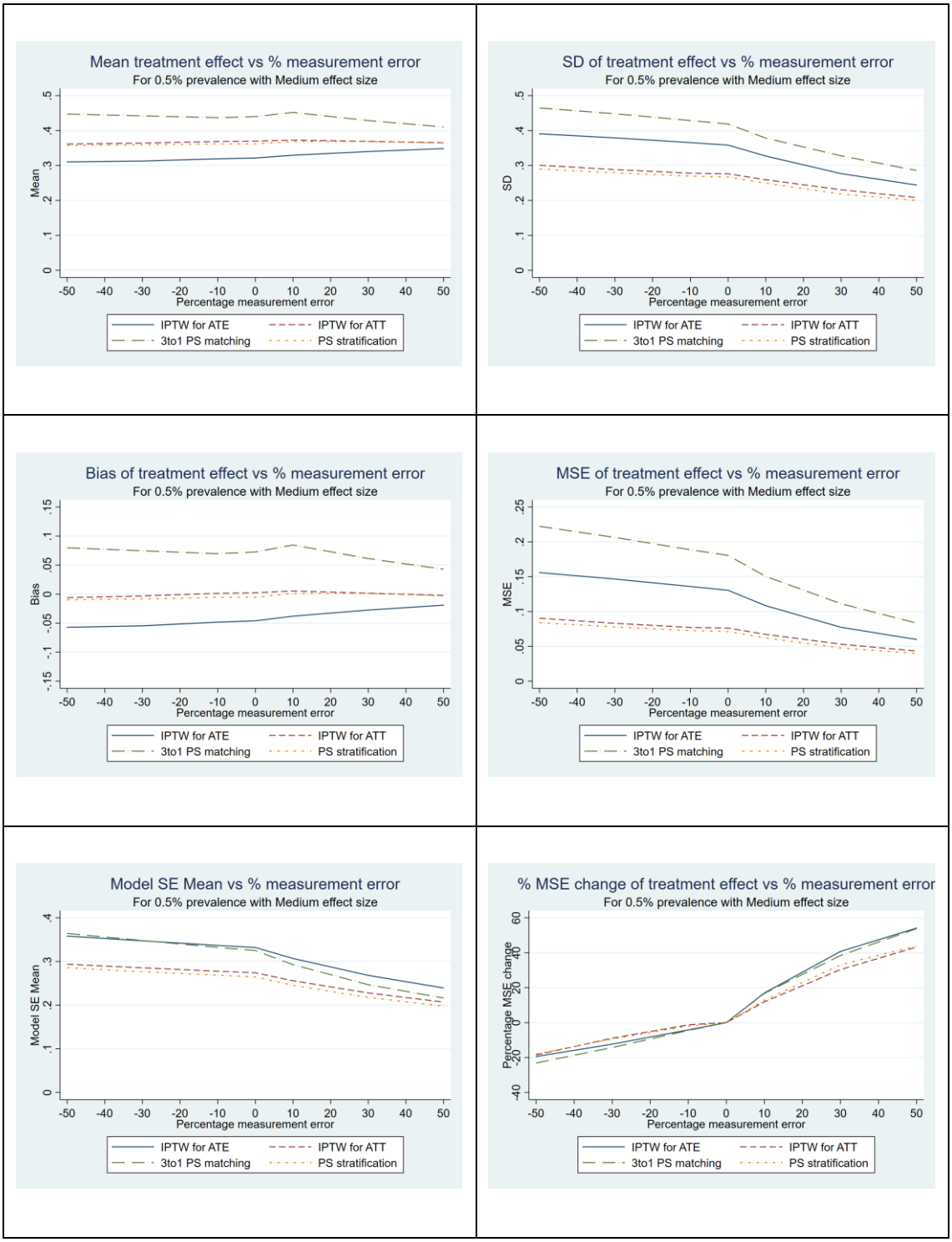


Figure G-12: Study PS methods, 0.5% prevalence, Medium effect size - the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect displayed as log(HR).

0.5% prevalence with High effect size

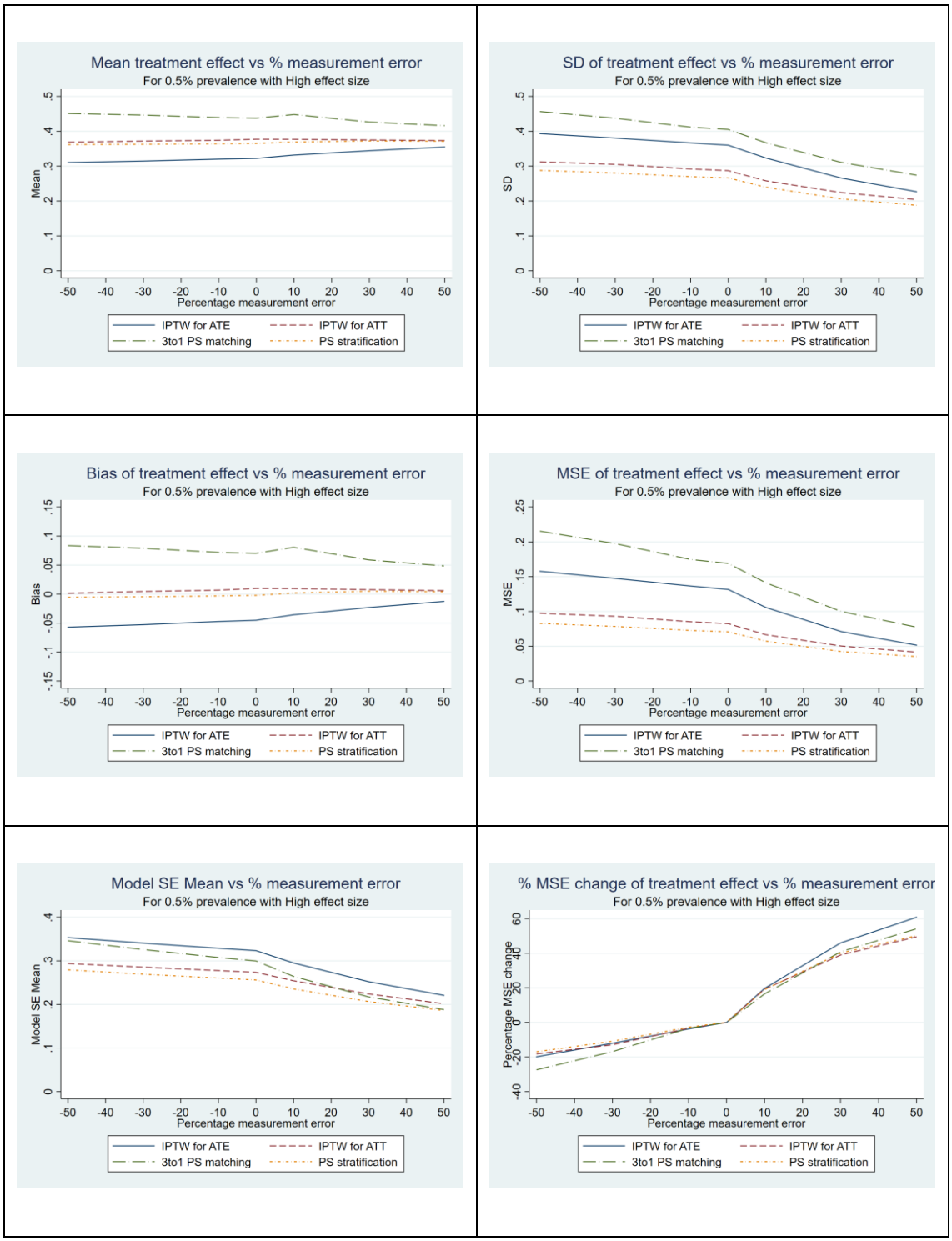


Figure G-13: Study PS methods, 0.5% prevalence, High effect size - the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect displayed as log(HR).

G-3 Graphs PS methods for ATE and PS methods for ATT

For ATE

ATE 10% prevalence

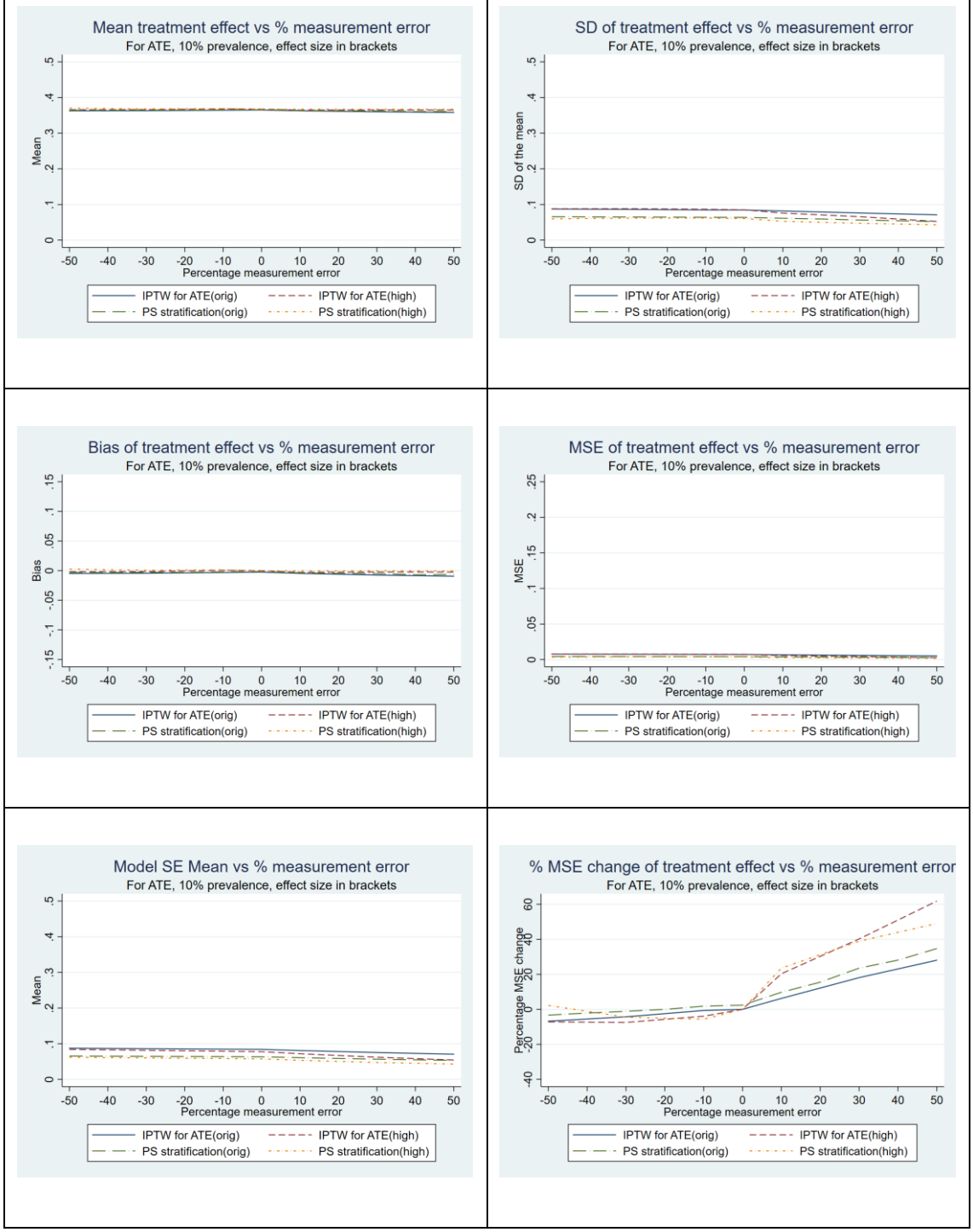


Figure G-14: Study PS methods for ATE, 10% prevalence - the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect displayed as log(HR).

ATE 1% prevalence

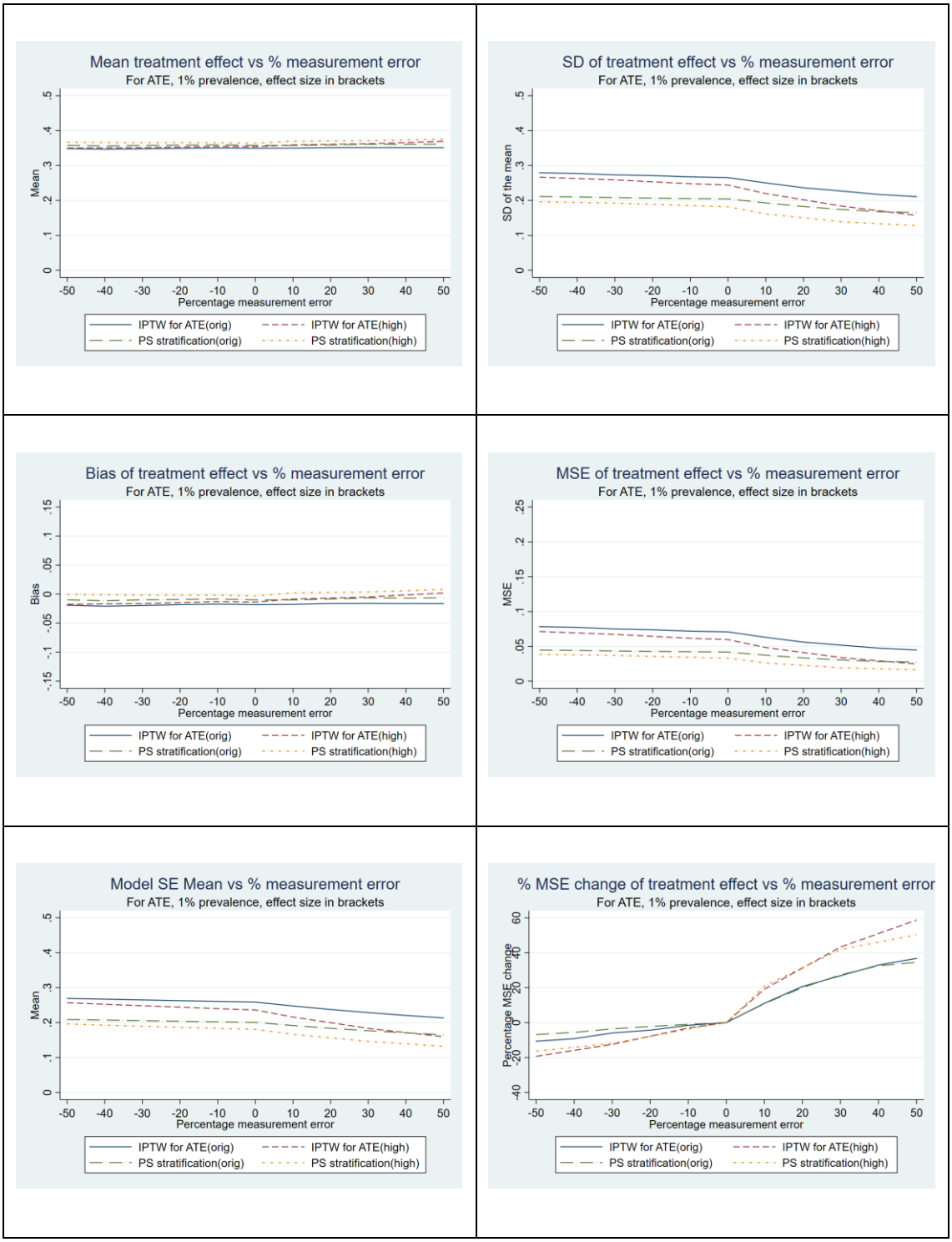


Figure G-15: Study PS methods for ATE, 1% prevalence - the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect displayed as log(HR).

ATE for 0.5% prevalence

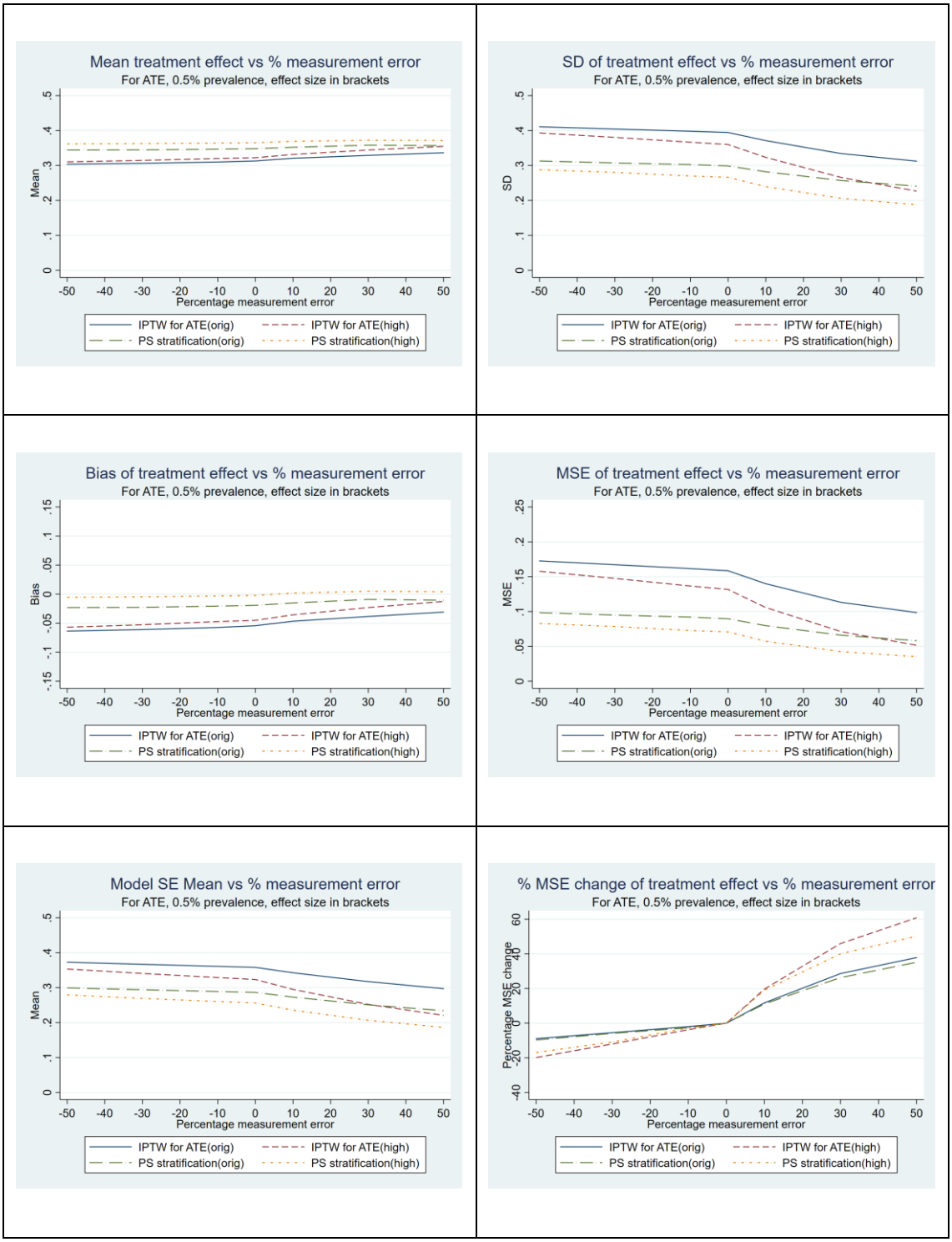


Figure G-16: Study PS methods for ATE, 0.5% prevalence - the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect displayed as log(HR).

For ATT

ATT 10% prevalence

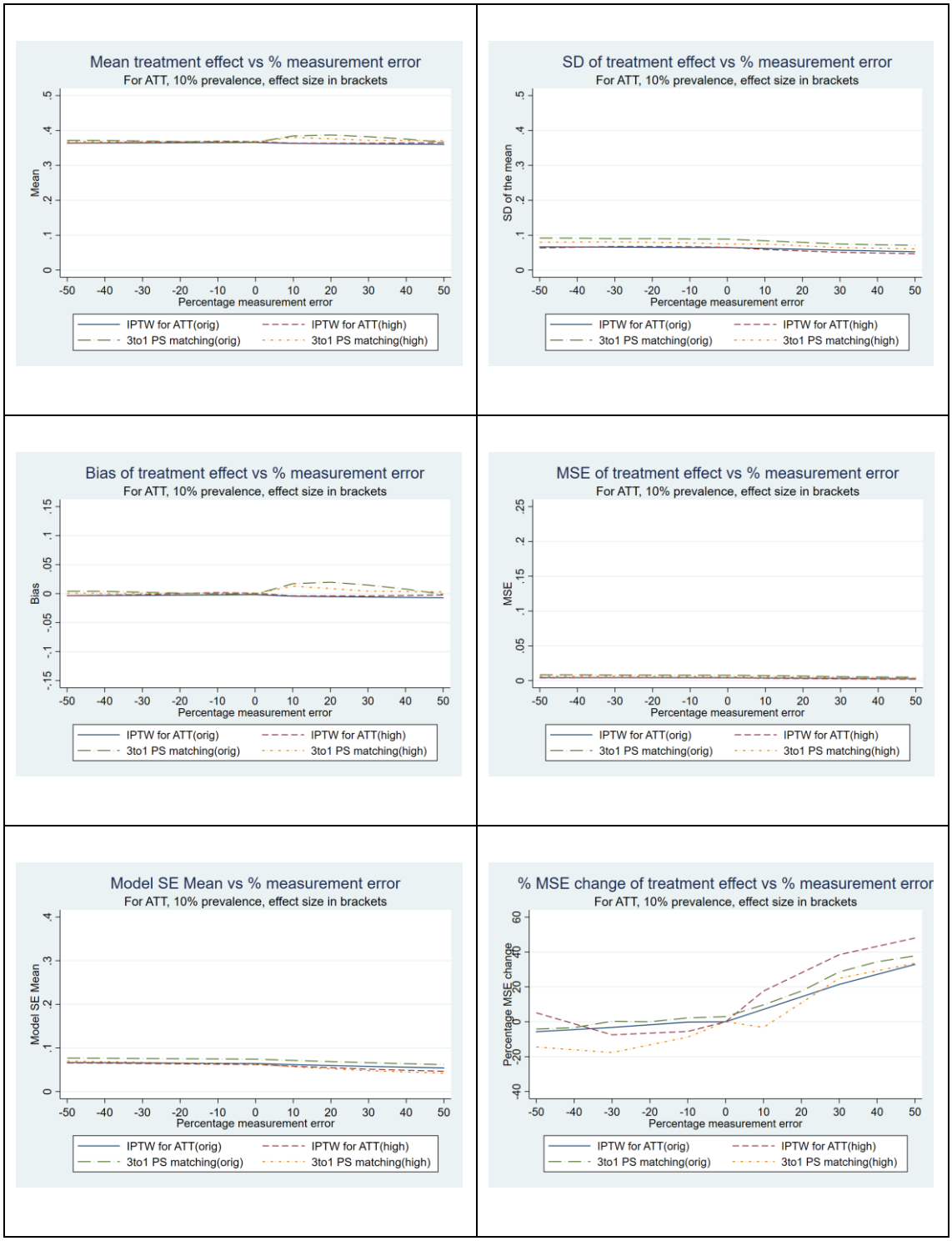


Figure G-17: Study PS methods for ATT, 10% prevalence - the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect displayed as log(HR).

ATT 1% prevalence

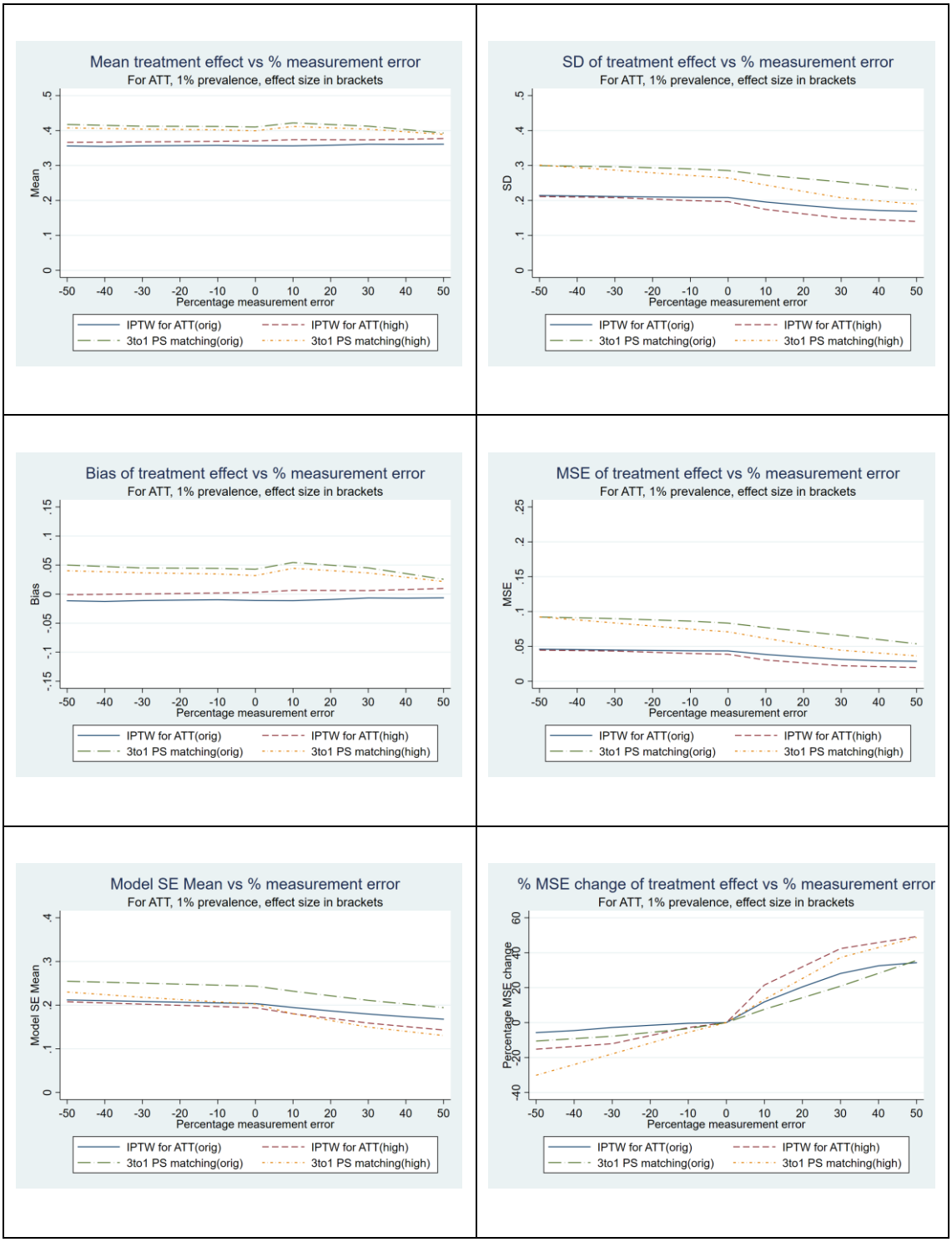


Figure G-18: Study PS methods for ATT, 1% prevalence - the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect displayed as log(HR).

ATT 0.5% prevalence

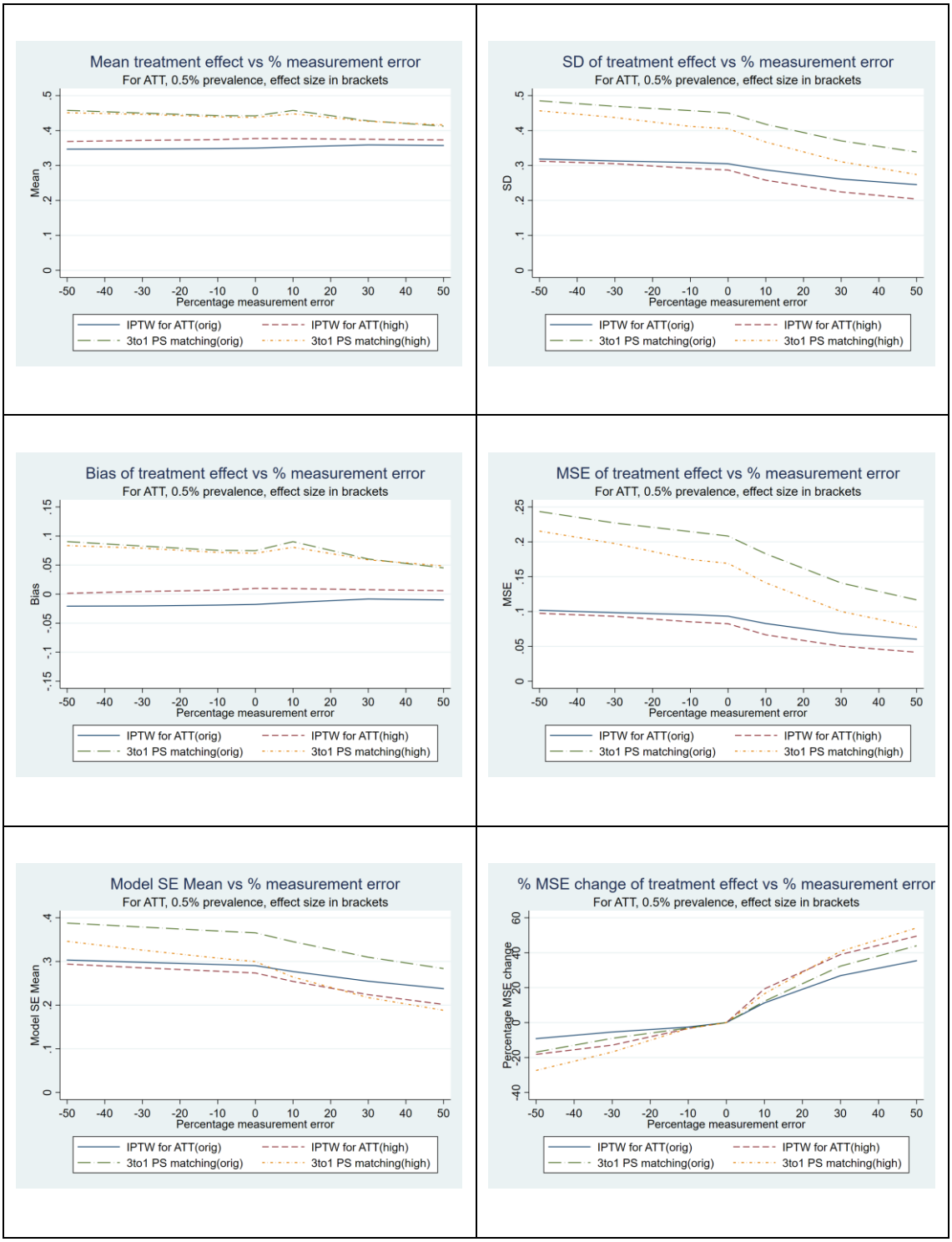


Figure G-19: Study PS methods for ATT, 0.5% prevalence - the mean, SD, bias, MSE (absolute and % change) and model SE mean of the estimated treatment effect displayed as log(HR).