Drug Repurposing: Design, Emulation and Analysis of Synthetic In-Silico Clinical Trials Using Electronic Health Records and Modern Data Analytics

by

Shenbo Xu

B.S., Dalian University of Technology (2016)

Submitted to the Department of Mechanical Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Mechanical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2019

Author . . . . Signature redacted . . . .
Department of Mechanical Engineering
January 22, 2018

Certified by . . . . Signature redacted . . . .
Roy Welsch
Professor of Statistics and Management Science
Thesis Supervisor

Certified by . . . . Signature redacted . . . .
Stan Finkelstein
Senior Research Scientist, Institute for Data, Systems, and Society
Thesis Supervisor

Certified by . . . . . Signature redacted . . . .
Pierre Lermusiaux
Professor of Mechanical Engineering
MechE Thesis Reader

Accepted by . . . . Signature redacted . . . .
Nicolas Hadjiconstantinou
Professor of Mechanical Engineering
Chairman, MechE Committee of Graduate Students

Drug Repurposing: Design, Emulation and Analysis of Synthetic In-Silico Clinical Trials Using Electronic Health Records and Modern Data Analytics

by

Shenbo Xu

Submitted to the Department of Mechanical Engineering
on January 22, 2018, in partial fulfillment of the
requirements for the degree of
Master of Science in Mechanical Engineering

Abstract

Cancer has been a worldwide health issue, and its burden is considered to increase in the future. For most cancer disorders, the success with current therapies has been limited. Even after huge investments in drug development, the need for therapeutic advances remains high. As effective anti-cancer drugs are in high demands, drug repurposing, using existing drugs for other diseases has sparked a growing interest. Drug repurposing presents a striking opportunity and potentially significant cost-saving in the future treatment of cancer.

The cost and complexity of conducting randomized clinical trials (RCT), the growth of electronic health record (EHR) sources, and the thriving technological advances in modern data analytics create an unparalleled opportunity to develop a systematic approach for drug repurposing,using EHR data and sophisticated analytical methods.

In this thesis, by leveraging enriched high dimensional EHR data with diagnosis, drug prescription and lab test information, we aim to develop a systematic approach to emulate clinical trials regarding various drugs and diseases based on modern data analytics. Specifically, we take a data-driven approach to repurpose anti-diabetic drugs for several types of cancer incidence and mortality risks among the aging population, through the lenses of optimization, statistics, and machine learning.

We start by introducing background knowledge for this study including cancer, drug repurposing, anti-diabetic drugs and clinical trials in Chapter 1. In Chapter 2, we describe the UK primary care database Clinical Practice Research Datalink (CPRD) along with its data structure for data preprocessing. Methods and mechanisms for missing data in clinical studies are also discussed as they will influence model robustness, statistical significance and directional results. In Chapter 3, we

discuss alternative frameworks for survival analysis and causal inference with emphasis on modelling the behavior of how physicians prescribe drugs, using propensity scores. Several Cox regression based semi-parametric methods are also reviewed for survival analysis. Chapter 4 offers baseline characteristics for a comprehensive in-silico randomized controlled trial with a total of 640 model specifications. Chapter 5 presents numerical risk ratio results for 10 sub-studies and discussions of covariate balance evaluation and sensitivity analyses among 64 schemes within each sub-study.

Through this work, we have made preliminary contributions to repurposing anti-diabetic drugs for cancer incidence and mortality risks. More importantly, we have offered a systematic approach that has the potential to be used to repurpose drugs for other diseases that are of interest. This use of modern data analytics offers tremendous potential to meet healthcare challenges in this era of rapid technological change.

Thesis Supervisor: Roy Welsch
Title: Professor of Statistics and Management Science

Thesis Supervisor: Stan Finkelstein
Title: Senior Research Scientist, Institute for Data, Systems, and Society

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my thesis advisor Professor Roy Welsch and Professor Stan Finkelstein for their guidance and support during my graduate study at MIT. I am truly grateful for the opportunity to work with and learn from such inspiring mentors. I especially thank them for bringing me into the world of statistics and machine learning.

I would like to express my special thanks to Imperial College team members, Bowen Su, Bang Zheng, Melek Somai, Lefkos Middleton, Ioanna Tzoulaki, with whom we have collaborated over the past year. Their patience and positive attitude helped me markedly on getting familiar with CPRD data and statistical methods that I have never stepped my foot in before. I would also like to thank MIT and Harvard group members, Marie-Laure Charpignon, Aamna Al-Shehhi, Yi-Han Sheu, Colin Magdamo and Bella Vakulenko-Lagun, who brought rigorous discussion on inspiring research topics.

In addition, I would like to thank several professors who have been part of my excellent MIT experience: Andrew Lo (Sloan), Colin Fogarty (Sloan) and Paul Mende (Sloan). I would also like to thank MechE staff members, Professor Rohan Abeyaratne and Leslie Regan, for their kind guidance and suggestions along my graduate study.

Many thanks to Chiang Chen Industrial Charity Foundation, Department of Mechanical Engineering for fellowships supporting my first-year graduate study, Professor Welsch for Sloan fellowship and teaching assistantships, and Professor Finkelstein for additional support.

Finally, to my family, especially my mum, for her incessant spiritual encouragement.

Contents

List of Figures

List of Tables

13

16

19

Chapter 1

Introduction

Cancer continues to be a major health issue worldwide. In 2012, more than 14 million people were diagnosed with cancer and 8.2 million were estimated to die from it. According to global statistics, it is estimated that more than 20 million people will be diagnosed with cancer in 2025 (Sleire et al., 2017). Correspondingly, the global economic burden of cancer treatments is expected to soar in the coming years. Effective, safe and economically viable cancer drug development is thus an imperative demand worldwide.

Drug development has always been circuitous. Particularly, discovering drugs from scratch to governmental approval is an expensive, time-consuming and risky process. According to estimates, the total capitalized cost of developing new drugs varies from 161 to 1800 million dollar per drug (Adams and Brantner, 2006). The average time span from initial discovery to approval varies between 11.4 to 13.5 years (Paul et al., 2010).

Traditionally, cancer drug discovery and development involves a time-consuming process, starts from identification and optimization of lead compounds, followed by pre-clinical research on microorganisms and animals and three to four phases of clinical trials on humans to detect and identify pharmacological features, pharmacokinetics, anti-tumor effects and toxicity.

Even though with more than 10,000 current clinical trials for cancer underway, only a limited number of candidates entered the next phase (Hay et al., 2014). The approval rate for cancer drugs passing phase I trials can be as low as 5% (Kola and Landis, 2004). The number of approvals declined yearly, both from the U.S. Food and Drug Administration (FDA) and from the European Medicines Agency (EMA). For example, FDA approved 22 new drugs in 2016 compared to 45 in 2015 (Mullard, 2017).

Even if cancer drugs ultimately receive clinical approval, the increasing prices become a heavy burden on both patients' families and national health economies. Although it is understandable that the high price intends to cover overall investments both for failed and successful drug candidates, it is unacceptable and unaffordable for many patients. These challenges have inspired great interests in searching for alternative approaches to improve success rates, shorten processing time and cut costs in cancer drug development.

Drug repurposing or repositioning offers an alternative way of finding effective drugs for cancer treatment. Drug repurposing refers to the use of an approved drug for a different indication than that for which it was originally developed, while drug repositioning refers to the novel use of a drug that was previously discontinued for development. Sometimes, drug repurposing and repositioning are used interchangeably (of Medicine, 2014). Since great amounts of time and money can be saved compared with the process of developing a drug de novo, drug repurposing or repositioning have received increasing attention among academics, clinical practitioners and pharmaceutical companies.

Large databases, such as those containing genome-based information have enabled advances in drug development. During traditional drug discovery and development processes for gaining approval for a specific indication, the safety, efficacy, and toxicity of the drug was extensively studied and large amounts of data were accumulated

and stored electronically. The available recorded data present a major opportunity for drug repurposing, as it could reduce the need for additional studies to investigate pharmacokinetic properties and toxicity and increases the chance of success. Besides, high failure rates of clinical trials and the fact that most drugs have multiple effects offer strong justification for drug repurposing.

The development of new tools and technologies adds momentum for repurposing and repositioning. Drug repurposing requires extensive exploration on existing large databases obtained from preclinical experiments, clinical trials and observational studies to find effective anti-cancer drugs which have already been approved for other indications to treat cancer. Recent advances in statistics and machine learning could make it feasible to reduce the complexity of traditional drug discovery by using advanced data analytics to explore the enriched data sources from electronic health records. We hypothesize that by emulating clinicals via construction of synthetic or in-silico clin- ical trials using electronic health records and sophisticated modern data analytics, it is possible to develop a systematic approach for drug repurposing.

This chapter will briefly introduce background knowledge including drug repurposing and repositioning methods, promising drugs for cancer treatment and potential mechanism of anti-diabetic drugs.

1.1 Drug Repurposing Methods

Drug repurposing provides an amazing opportunity and significant cost saving potential. Because toxicity testing was done in previous phase I clinical trials, drug repurposing has received increasing interest as an alternative strategy to de novo drug development. Drug repurposing relies on the extensive data obtained by randomized controlled trial, observational studies and meta-analysis, and requires employing reliable algorithms, statistical survival analysis, machine learning methods and big data techniques to analyze these data. In-silico drug repurposing is an emerging method

for drug repurposing based on available EHR data.

1.1.1 Randomized Controlled Trial

A randomized controlled trial (RCT) aims to reduce bias when testing a new treatment. In a RCT, two or more groups of people participating in the trial are randomly allocated as receiving the treatment under investigation and receiving standard treatment (or placebo treatment), respectively. The latter group acts as the control group. With all the other variables kept constant, the effects of the treatment can be compared with the control group.

Even with ultra-high cost to conduct randomized clinical trials and frequently encountered difficulties in designing them, they are still believed to be the best way to estimate of the outcomes of given treatments with minimal bias (Lilford et al., 1995; Schulz and Grimes, 2005). Strict implementation of RCT minimizes confounding bias which may affect outcomes and distort significant treatment outcomes. Other potential benefits are standardized protocols, improved supportive care, inclusion of low-risk patients and increased efforts on treatment hazard prevention.

Despite the benefits of randomized clinical trials, a variety of factors prevent physicians from conducting useful RCTs, including patients, healthcare providers, relatively low occurrence of disease, small sample sizes of targets (Fung and Lore, 2002; Abraham et al., 2006; Solomon and McLeod, 1995). Because of the Hawthorne or placebo effect, patients who participate in the trial can also observe different outcomes, which may distort the obvious therapeutic effect and harm the effectiveness of RCT. Besides, misleading information might be lead by improper testing and deficient reporting.

Nevertheless, RCT is still the gold standard for assessing the effectiveness of interventions. And physicians should not be prevented from conducting randomized clinical trials by the above reasons.

1.1.2 Observational Survival Analysis

In drug discovery and development, questions regarding comparative effectiveness, efficacy, or safety of a new treatment are preferably answered by appropriately designed and conducted randomized controlled trials. However, when it's not feasible, ethical, and timely to carry out randomized experiments, observational studies are conducted instead.

In contrast with randomized controlled trials, where person participating in the trial is randomly assigned to a treated group or a control group, an observational study draws inferences from a sample to a population where the independent variable is not under the control of the researcher. A major challenge in conducting observational studies is to draw inferences without biases (Hernán and Robins, 2016).

Retrospective or prospective historical observational cohort studies, as well as those with the most complex designs and analyses, can potentially control known confounding factors and lead to useful findings. Unknown confounding factors often can't be explained totally. A rudimentary criticism of observational studies is that they might lead to biased estimates of treatment outcomes (Sacks et al., 1982; Kunz et al., 2007), and many medical and surgical researchers believe that observational studies add in confusion and can be ineffective or even detrimental.

Treatment outcome estimates can be partially improved by novel statistical methods due to known and unknown confounders in non-randomized designs (Baggs et al., 1999; Sturmer et al., 2005). Conventionally, non-random or observational studies adjust for known confounders to model the mathematical relationship between one or more predictors and estimates isolated effects of each variable.

Large observational databases, i.e. electronic health record data can be used to

answer questions on comparative effectiveness or safety. These databases usually include a number of variables measured in many people and possess "big data" characteristics (Hernán and Robins, 2016). Since decisions need to be made in the absence of randomized trials, it is important to employ sound approaches to design and analyze observational studies to draw inferences without biases.

Causal inference is the process of identifying the cause or causes of a phenomenon. It infers causation by establishing covariation of cause and effect and eliminate plausible alternative explanations. In epidemiological studies or clinic trials, the disease in defined populations are studied, and the evidence of risk factors and their effects are collected and measured. An association between a putative risk factor and the disease may be suggested, however, it is not equivalent to causality as correlation does't indicate causation, and causation may be able to be be inferred by further methods.

Causal inference from a large observational database can be regarded as an attempt to emulate a randomized experiment, or a target trial. Formal counterfactual theory of causality is consistent with the target trial approach (Robins, 1986). If observational analysis can be emulated into a particular target trial, effect estimates can be obtained from observational data. However, an ideal trial can be rarely emulated, since a number of compromises, including eligibility criteria, treatment strategies to be compared, assignment procedures, outcome of interest, etc., have to be made while emulating the target trial from observational database. Observational data has the potential to generate useful effective estimate if target trial emulation is successful (Hernán and Robins, 2016).

An organizing principle for causal inference is provided by a target trial approach that relies on counterfactual reasoning implicitly. By outlining a protocol and a flow chart and using suitable analytic methods, the observational dataset can be used to emulate the target trial (Hernán, 2011; Schulz et al., 2010).

The big data feature of large observational databases facilitates the emulation of target trials. However, big data should not be regarded as an alternative to randomized trials (Hernán, 2011). Large observational databases have limitations. The certification of big data for research generally requires procedures for harmonization and standardization. Besides, it can require an in-depth understanding of the dataset, expensive validation, comprehensive internal consistency checks and cross-dataset comparisons (Hernán and Robins, 2016).

Survival analysis is an important sub-area of statistics. It involves the modelling of time to event data to analyze subsequent events of interest. Here events refer to disease occurrence, disease recurrence, recovery, or death in biological organisms, and time refers to the time from the beginning of an observation, e.g. beginning of treatment, to an event, or to the end of the study, or to the loss of contact or withdrawal from the study.

One of the main challenges in survival analysis is dealing with censoring. Censoring occurs if a subject does not have an event during the observation time. Furthermore, censoring is common in large observational databases (Hernán, 2011; Hernán and Robins, 2016). It is a form of a missing data problem in which an event is not observed due to the early termination of study, or loss of follow-up during the observation period.

Traditionally, statistical approaches have been widely developed to overcome censoring issues. However, applying predictive algorithms directly to analyze survival data using standard statistical and machine learning methods is not always appropriate. Apart from the complexity of processing censored data, challenges remain in survival data predictive modeling. In Chapter 3, we present a detailed introduction to techniques and representative statistical methods for survival analysis.

1.1.3 Meta-analysis

Meta-analysis integrates the statistical analysis of multiple scientific studies. The critical assumption behind meta-analyses is a common principle behind all analogous research with a certain error or bias associated with individual studies. Meta-analysis aims to conclude a weighted expectation to reduce the uncertainty around the estimate. Apart from that, meta-analysis can compare the results from different studies to recognize the patterns among individual results.

Meta-analysis is one of the key components of systematic literature review. For instance, based on several clinical trials of a medical treatment, meta-analysis can offer a better understanding regarding the effect of certain therapy. Meta-analyses can improve the confidence by offering a higher statistical power and a more robust estimate than individual studies. In performing meta-analysis, the search process for studies, objective criteria, methods dealing with missing data, analytical methods and associated bias can affect the results in a broad sense (Walker et al., 2008).

1.2 Promising Candidates for Cancer Treatment

Most drugs have multiple effects. Substantial published studies of preclinical experiments, clinical trials, and observational studies have also demonstrated that a wide range of drug classes have anti-tumor efficacy apart from their primary function. Some of the drugs suppress different aspects of cancer cell behavior or induce cancer cell death, others may prevent cancer development. Therefore, these licensed drugs have potential to be repurposed as anti-cancer drugs both for cancer prevention and for cancer therapy.

In this section, we will briefly introduce drugs for which studies have indicated effectiveness in cancer treatment from clinical, epidemiological and laboratory research. The drugs include aspirin, statins, selective estrogen receptor modulators, cardiovascular drugs, antipsychotic drugs, antidepressants, microbiological agents, anti-viral drugs, antibiotics, and nonsteroidal anti-inflammatory drugs (Sleire et al., 2017). One such drug, metformin, which is the main focus of the thesis, will be introduced in the next section.

Aspirin is a non-steroidal anti-inflammatory drug (NSAID) and widely used due to its analgesic, and antipyretic properties (Mcquay and Moore, 2007; Aronoff and Neilson, 2001). Currently, it is also used for prevention of thromboembolism in patients with manifest or increased risk of cardiovascular disease (Miner, 2007). Numerous experimental and observational studies have established a close correlation between inflammation and cancer. For certain cancer types including colrectal cancer and liver cancer the inflammatory process is considered a main driver of carcinogenesis. Therefore, the anti-inflammatory properties of aspirin underly its cancer protective effects. Data from in vitro and in vivo experiments, observational studies and prospective trials have convincingly confirmed the anti-neoplastic effects of aspirin (Cuzick et al., 2009; Phillips et al., 2013; Elder et al., 1996). It is reported that regular use of aspirin significantly reduced the incidence of colorectal, esophageal, gastric, biliary

31

and breast cancer after systematically comparing randomized trials with cohort and case-control studies (Algra and Rothwell, 2012).

Statins are among the most commonly prescribed drugs used to treat lipid disorders, as they can effectively prevent the development of cardiovascular diseases. So far, the association between statin use and cancer incidence is inconclusive overall. Some observational studies and meta-analyses suggest a positive correlation with reduced incidence of gastric cancers (Wu et al., 2013), esophageal cancer (Leonard et al., 2013) and hepatocarcinoma (Zhong et al., 2016), while several cohort and case-control studies as well as meta-analyses reported only weak or no significant link between reduced incidence of prevalent cancer types including breast (Borgquist et al., 2016) and colorectal cancer (Lytras et al., 2014). A possible role for statins in cancer prevention might best be determined through carefully designed RCTs with a sufficiently long follow-up and use cancer incidence rather than cardiovascular disease as a primary endpoint.

Selective estrogen receptor modulators (SERMs) are drugs that act on the estrogen receptor (ER) and mediate different effects depending on the organs or tissue. Tamoxifen, one drug belonging to SERMs, originally developed as a fertility drug, has been known to have anti-cancer properties since the 1970s, and approved as a preventive drug against breast cancer recurrence in 1998 by FDA (Li et al., 2016). Raloxifene, another drug belonging to this class, initially approved for treatment of osteoporosis inmenopausal and post-menopausal women, received approval in the US for breast cancer prevention in potmenopausal women in 2007 (Sleire et al., 2017).

Cardiac glycosides such as digoxin and digitoxin are compounds found in plants and animals that are used to treat different cardiac conditions (Elbaz et al., 2012). Already in 1967 Shiratori reported growth inhibitory effects on cancer cells from prostate cancer (Mcconkey et al., 2000) and breast cancer (Bielawski et al., 2006).

Antipsychotic drugs are commonly used to treat psychosis and schizophrenia. Studies have indicated that some of these drugs may reduce the risk of certain cancers (Dalton et al., 2006). For example, chlorpromazine might decrease risk of developing prostate cancer (Mortensen, 1992) for schizophrenic male user.

Tricyclic antidepressants are used to treat clinical depression and other mood disorders. Several studies have reported antidepressants, such as Lithium (LiCl), have anti-neoplastic effects on certain cancers, like prostate cancer and colon cancer (Tutton and Barkla, 1982; Sun et al., 2007; Maeng et al., 2016)

Microbiological agents, such as artemisinins, a traditional Chinese medicine for the treatment of malaria infections, are also reported having anti-angiogenic effects for renal cancer and hepatocellular carcinoma with reduced tumor growth in vivo. Growth inhibition in vitro and in vivo in cell lines are also noticed from other cancer types including lung cancer (Sasaki et al., 2002), colon cancer (Nygren et al., 2013), melanoma (Doudican et al., 2008, 2013) and glioblastoma (Bai et al., 2011).

Anti-viral drugs, such as ritonavir and nelfinavir, are protease inhibitors against HIV. Substantial data shows that ritonavir can inhibit cell cycle progression, induce apoptosis in ovarian, pancreatic and breast cancer cells (Kumar et al., 2009; Srirangam et al., 2006).

An antibiotic, such as doxycycline, is effective against a range of infectious diseases. Some tetracyclines were found effective in inhibition of angiogenesis (Tamargo et al., 1991), while doxycycline was found to have growth inhibitory effects in osteosarcoma, prostate cancer and mesothelioma cells (Fife et al., 1998; Rubins et al., 2001). Their proapoptotic effects on pancreatic cells pancreatic cells (Mouratidis et al., 2007) and leukemic cells (Song et al., 2014) have been observed lately.

Nonsteroidal anti-inflammatory drugs (NSAIDs), including ibuprofen, acetyl sali-

cylic acid, naproxen, and diclofenac, have analgesic antipyretic, and anti-inflammatory effects (Buer, 2014). Epidemiological studies and recent data suggest that NSAIDs may be effective in the treatment of established tumors (Pantziarka et al., 2016). For example, Ibuprofen can inhibit the growth of prostate cancer cells (Kim and Chung, 2007). Naproxen has demonstrated anti-neoplastic properties in vitro and in vivo in leukemic, breast, colon, bladder and osteosarcoma cell lines (Sleire et al., 2017).

1.3 Potential Mechanisms of Anti-diabetic Drugs

Metformin is a first-line oral medication for the treatment of type II diabetes. It decreases high blood sugar by suppressing liver glucose release (hepatic gluconeogenesis) and reducing sugar absorption from food in the gut. Metformin also increases the insulin sensitivity of body tissues, including skeletal muscle, adipose tissue, endothelium, ovary and liver (Diamanti-Kandarakis et al., 2010; Lord et al., 2003).

Epidemiological studies have suggested that diabetic patients have an increased risk of developing several types of cancer (Wu et al., 2018). A meta-analysis showed that diabetic women have a high risk of breast cancer and cancer related death. Many studies have found that diabetes, especially type II diabetes, has a close relationship with the development of non-Hodgkin's lymphoma, pancreatic, colorectal, bladder and endometrial cancer (Sleire et al., 2017).

Like many other drugs, the anticancer effect of metformin was discovered by serendipity. Considerable epidemiological evidence suggests taking metformin may reduced cancer risk. Compared with patients using other anti-diabetic drugs, diabetic patients using metformin have potentially reduced both incidence and mortality. From a meta-analysis on the impact of several anti-diabetic drugs, researchers found that metformin users reduced cancer incidence and mortality by 14% and 30% respectively. On the other hand, other anti-diabetic drugs, like insulin, increased risk and mortality of cancer (Sleire et al., 2017). After using metformin, diabetes patients with established cancer demonstrated a favorable response to treatment, and survival was increased for hepatocellular carcinoma (Chen et al., 2011), colorectal (Garrett et al., 2012), prostate (He et al., 2011), breast (He et al., 2012), ovian (Romero et al., 2012), pancreas (Sadeghi et al., 2012), esophageal (Skinner et al., 2013b) and rectal cancer (Skinner et al., 2013a). Several reviews suggest that in diabetic patients, using metformin is associated with lower incidence of colorectal, hepatocellular, pancreas, stomach, liver, esophagus and lung cancer incidence (Noto et al., 2012; Franciosi et al.,

2013; Tsilidis et al., 2014a).

Even with these findings based on observational studies and meta-analyses, no RCT has been reported about the effect of metformin on cancer incidence and mortality. However, a recent multicenter double-blind, placebo-controlled, randomized phase III trial found that both the occurrence and number of adenomas/polyps were reduced in the patients using metformin compared to the control group (Lytras et al., 2014).

Diabetes is a complex metabolic syndrome characterized by long-term high blood glucose levels and life-threatening complications (Wu et al., 2018). Diabetes and cancer may be linked by high glucose levels. Therefore, any measure that improves glycemic level may be expected to prevent cancer development. Its mechanism is thought to be inhibition of oxidative phosphorylation, causing energetic stress and inhibition of gluconeogenesis, especially in liver cells (Pollak, 2014).

Metformin is also believed to have a direct anti-proliferative effect because it activates the liver enzyme AMPK, with the effect of inhibiting cancer cell growth as reported by several in vivo and in vitro studies. Protein kinase LKB1, the upstream regulator of AMPK, is a recognized tumor suppressor. Nonetheless, the mechanism on how metformin improves cancer survival directly (insulin independent) or indirectly (insulin dependent) is still not fully understood (Sleire et al., 2017).

More recently, researchers found that metformin, in combination with the antihypertensive drug syrosingopine, can stop cancer tumors from growing by cutting energy supply to cancer cells. Cancer cells need energy to grow and spread. A molecule called NAD+ can turn nutrients into energy. Many cancer cells rely on glycolysis to break sugar down into lactate and generate NAD+ from nicotinamide adenine dinucleotides (NADH) in their metabolism. Too much lactate will block glycolytic pathways. It is found that two key lactate transporters are blocked by syrosingopine, resulting in

high intracellular lactate levels. Metformin, meanwhile, blocks the pathways that help NAD+ regenerate from NADH. The combined metformin and syrosingopine treatment results in glycolytic blockade, leading to a shortage of energy supply, which ultimately cause the death of cancer cells. The two drug combination may provide a viable cancer treatment strategy (Benjamin et al., 2018). Future studies are required to unravel the biological mechanisms underlying their anti-neoplastic effects.

Although the treatment of multiple cancers suggests a generalized anticancer effect of metformin and a great interest in repositioning metformin as an anticancer drug, there are also some studies indicating no or weak correlation between metformin use and cancer risk (Kordes et al., 2015; Tsilidis et al., 2014b). It also seems biologically implausible that the reduction in risk with the use of metformin starts from the first year of follow-up (Currie et al., 2009; Onnelly, 2009). For some epidemiologic studies demonstrating association with a reduced risk of cancer incidence or mortality (Currie et al., 2009; Onnelly, 2009; Ateren, 2010; Evans, 2005; Bodmer, 2010; Bowker, 2006; Hall, 2005; Yeung, 2009), studies focused on several types of specific cancer, and most of these findings compared metformin with other anti-diabetes drugs without taking severity into account as an active factor.

Since most studies are retrospective and the patient selection is heterogeneous, there is some degree of conflicting in these findings. This thesis, therefore, will focus on an investigation of anticancer effects of metformin on incidence and mortality of cancer in general, and, specifically breast cancer, prostate cancer, bowel cancer and lung cancer, with an aim to obtain conclusive results regarding its role on cancer treatment.

1.4 Main Contribution

In this thesis, we want to answer this question: what is the effect of anti-diabetic drugs on cancer incidence and mortality. The initial null hypothesis is that the effect of various anti-diabetic drugs on cancer incidence and mortality are the same. Many studies confuse cancer incidence with cancer mortality. For example, an event of interest for cancer incidence is the diagnosis of cancer one year after the initial anti-diabetic drug prescription. For cancer mortality, the corresponding event of interest is death within one year of initial anti-diabetic drug prescription after cancer diagnosis. Researchers also debate the effect of anti-diabetic drugs on various types of cancer incidence and mortality and most of them draw either of two different conclusions: different drugs have similar effect, or metformin is more protective. The contribution I aim to make in this thesis is to develop a standardized and comprehensive analytical framework to address the issues. By leveraging a high-dimensional EHR database with abundant diagnostic, drug and lab test information, we proposed a systematic approach to emulate randomized clinical trials of various drugs and diseases in silico. Specifically, we used a data-driven approach to reposition anti-diabetic drugs for cancer incidence and mortality risks using causal inference setting.

Chapter 1 introduces background knowledge for this study, including general drug repurposing methods, promising candidate for drug repurposing for cancer treatment, repositioning of anti-diabetic drugs and in-silico randomized clinical trials using a large dataset. In Chapter 2, we study the data structure of the UK primary care database Clinical Practice Research Data Link (CPRD). Methods and mechanisms for missing data in clinical studies are discussed as they will affect the robustness, statistical significance and directional results of various drugs. In Chapter 3, we discuss survival analysis and causal inference, with a focus on modeling physician behavior by propensity scores. Several semiparametric methods based on Cox regression were also evaluated. Chapter 4 provides baseline characteristics of a comprehensive randomized controlled trial emulation with a total of 640 specifications.

Chapter 5 presents numerical results for the emulated 640 randomized controlled trials in 10 sub-studies to detect the signal of two general anti-diabetic drugs on incidence and mortality risks for general cancer, breast cancer, prostate cancer, bowel cancer, and lung cancer. Within each sub-study, a total of 64 in-silico RCTs were conducted by semi-parametric conventional Cox regression, Cox regression in alternative weighting schemes, with 16 cases correspondingly. Chapter 6 presents the conclusion of the study.

I examine the impact of inverting treatment and control in a statistical setting when both groups are medically treated by anti-diabetic drugs. Sixteen studies were divided into two groups - metformin/sulfonylureas as control/treatment or sulfonylureas/metformin as control/treatment. Four methods of dealing with missing data, including using fewer variables, complete case analysis, treating missing as a separate category and inverse probability weighting, further divide 8 RCTs into a group of two. A cutoff of initial anti-diabetes prescriptions prior to year 2000 were considered to form the final cohort. Covariate balance were evaluated based on Somer's D. Sensitivity of all 640 trials was analyzed using weighted summary statistics.

Through this work, we have made preliminary contributions to repurpose anti-diabetic drugs for cancer treatment. More importantly, we offer a systematic approach to reposition many diseases based on all kinds of drugs. This has driven the rapid development of modern data analysis that has shown the potential for enormous medical challenges in this era.

Chapter 2

Data

2.1 Clinical Practice Research Datalink (CPRD)

2.1.1 Data Description

The Clinical Practice Research Datalink (CPRD) is an ongoing general practice's (GP) primary care database covering more than 11.3 million patients among 674 practices in UK. Data from a total of 4.4 million active (alive, currently enrolled) patients meet quality standards, with approximately 6.9% of the UK population being included, and patients are representative of the general population in terms of age, gender and ethnicity in UK (Su et al., 2018).

In this study, there are around 6 million participants in this CPRD cohort by April 30th, 2018. We use this cohort to analyze the effect of anti-diabetic drugs on cancer incidence risks and mortality risks. Only participants whose ages were over 40 at the first prescription of anti-diabetic drugs were included in this analysis. Time scale and age can affect conclusions, and this is discussed in section 4.4.

To maximize the number of participants and validate diagnoses in primary care EHR, we hope to establish a consistent cancer diagnoses using data both from Office for National Statistics (ONS) and Hospital Episode Statistics (HES). Two recent

studies on the consistency between primary care and HES in coronary heart disease (CHD) diagnoses revealed that the inclusion of HES data will increase CHD diagnoses from primary care EHR by 17%. Another comparison showed that the average incidence of community-acquired pneumonia among people using HES-linked data was 39% higher than independent EHR data. A similar increase is discovered by leveraging HES Chronic Obstructive Pulmonary Disease (COPD) diagnoses.

2.1.2   Data Structure

The main CPRD dataset consists of the following parts (Su et al., 2018):

1. Clinical dataset: mainly contains diagnostics;

2. Other data sets: smoking, alcohol consumption and other past medical history (other information about cancer diagnoses by patient identifier);

3. Consultation dataset: each visit and consultation for every patient;

4. Referral dataset: from primary care to secondary care;

5. Laboratory and medical tests: results with event date;

6. Treatment dataset: medication (with dosage).

In this analysis, a new CPRD dataset was extracted with follow-up updated to April 2018. Participants/patients were included if they meet the following standards. The start of follow-up was the latest of

1. January 1st, 1987;

2. the year that each individual turned 50 years old;

3. one year following CPRD registration date including an additional year to account for baseline risk factors;

4. one year after the practice achieved "up to standard" data quality status

Participants were followed up until the earliest of

1. patient death;

2. the individual leaving practice/CPRD database;

3. the practice last data collection date;

4. the end date (April 30th, 2018) of data inclusion.

2.2 Data Preparation

2.2.1 Study Population

We extracted data from CPRD based on the following criteria (Su et al., 2018):

1. All participants with type II diabetes aged over 50 between 1st January, 1987 and April 30 2018;

2. Prescribed at least one anti-diabetes agent between 1st January, 1987 and April 30, 2018;

3. First prescription dated at least 12 months after CPRD registration date (to ensure that most of the diabetic patients included are new users of anti-diabetes drugs and prevalent users with unknown type and duration of treatment are excluded);

4. Age at initial anti-diabetes prescription over 40;

5. Diabetes mellitus type 2 (DMT2) based on CPRD codes;

6. Diabetic complications or hospitalizations any time before index date from Hospital Episode Statistics.

2.2.2 Exposure Assessment

Anti-diabetes drugs can be categorized as (Su et al., 2018)

1. Metformin hydrochloride (Biguanides): recommended as the first choice for initial treatment.

2. Sulfonylureas, including first generation (tolbutamide, chlorpropamide, tolazamide, acetohexamide) and second generation (gliclazide, glibenclamide, glipizide, glimepiride, gliquidone, glibornuride, glymidine sodium).

3. Thiazolidinediones (rosiglitazone, pioglitazone).

4. α-glucosidase inhibitors (acarbose).

5. Meglitinide analogs (nateglinide, repaglinide).

6. The dipeptidyl peptidase-4 inhibitors (gliptins), alogliptin, linagliptin, sitagliptin, saxagliptin, and vildagliptin.

7. The sodium glucose co-transporter 2 inhibitors, canagliflozin, dapagliflozin, and empagliflozin.

8. The glucagon-like peptide-1 receptor agonists, albiglutide, dulaglutide, exenatide, liraglutide and lixisenatide, should be reserved for combination therapy when other treatment options have failed.

9. Insulin.

Based on prescription records in CPRD, individual anti-diabetic treatment is classified into one of the following mutually exclusive groups within initial 12-month period.

1. Monotherapy with metformin: excluding those prescribed with combination therapy in a single prescription (i.e. Metformin with rosiglitazone 500mg + 1mg Tablet) or through separate prescriptions during the initial 12 months;

2. Monotherapy with first generation of sulfonylureas;

3. Monotherapy with second generation of sulfonylureas;

4. Monotherapy with rosiglitazone;

5. Monotherapy with pioglitazone;

6. Monotherapy with other orally administered antihyperglycemic agents (OHAs), including acarbose, meglitinide (nateglinide, repaglinide), the dipeptidyl peptidase-4 inhibitors, the sodium glucose co-transporter 2 inhibitors, and the glucagon-like peptide-1 receptor agonists;

7. Monotherapy with insulin;

8. Combination therapies with metformin if the participants were simultaneously exposed to two or more OHAs including metformin during the initial 12-month treatment period;

9. Combination therapies with insulin;

10. Other combination therapies.

2.2.3 Outcome Assessment

For cancer incidence risks (Tsilidis et al., 2014a),

1. Malignant neoplasms National Health Service readcode in CPRD;

2. Exclude nonmelanoma skin cancers;

3. No previous history of cancer before index date;

4. Cancer diagnosed within the first year of initial anti-diabetic prescription is excluded;

For cancer mortality, outcome is death when deathdate is present. Censoring is assumed when deathdate is not recorded.

2.2.4 Covariate Selection

Based on empirical drug research, we try to find out all the true confounding factors, which are the covariates contributing to treatment outcomes. In addition, the propensity score model can also include predictors of outcomes unrelated to treatment, as these covariates will increase the ability to test therapeutic effect. (Brookhart et al., 2006)

Start by an apriori-defined list of confounders including demographic information, lifestyle characteristics, lab tests, medical diagnoses, and prescription history, we list

Table 2.1: Covariate selection and attributes

Covariates	Type	Attributes
Demographic variables		
Multiple Deprivation Index (IMD)	Ordinal	1, 2, 3, 4, 5
Gender	Nominal	Male or female
Year of index date	Numeric	
Year of birth (YOB)	Numeric	
Lifestyle variables		
Smoking	Nominal	Never, former, current
		5-year latest before index date
Laboratory tests		
Body Mass Index (BMI)	Numeric	5-year latest before index date
Glycated Hemoglobin (HbA1c)	Numeric	1-year mean/1-year latest level
		before index date
Medical diagnoses		
Heart Failure (HF)	Categorical	Yes/No
Coronary Heart Disease (CHD)	Categorical	Yes/No
Atrial Fibrillation (AF)	Categorical	Yes/No
Peripheral Vascular Disease (PVD)	Categorical	Yes/No
Chronic Kidney Disease (CKD)	Categorical	Yes/No
Chronic Obstructive	Categorical	Yes/No
Pulmonary Disease (COPD)		

out covariates by name, type and attributes in Table 2.1. Age and sex were recorded
at approximately the time of the first anti-diabetic drug prescription (index date) (Su
et al., 2018).

2.3 Missing Data

Missing data are ubiquitous in both clinical observational studies and experimental research. It refers to the phenomenon that no data is stored for a variable in a dataset. Missing data may introduce bias into the analysis and have a significant influence on the conclusions to be drawn from the data. Hence, dealing with missing data has always been considered critical and well-studied in academia (Molenberghs et al., 2014).

2.3.1 Missing Mechanisms

Understanding the nature of missing data mechanisms can facilitate correct handling the incomplete data in the dataset to obtain valid inferences. Missing data occur probably due to nonresponse, participants early drop out, one or more missing data, improper data collection, or mistakes in data entry. They are roughly classified as three types, missing completely at random, missing at random, and missing not at random (Molenberghs et al., 2014). The different forms of missingness have different impacts on the validity of research conclusions.

Assume for each of N individuals, response variable on the same individual is measured n_i times repeatedly. A subject with a complete set of responses has an $n_i \times 1$ response vector of $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})^T$, where Y_{ij} is the jth response for the ith subject at time t_{ij}, which is associated with an $n_i \times p$ matrix of covariates, X_i. Let Res_i be an $n_i \times 1$ vector of response indicators, $Res_i = (Res_{i1}, Res_{i2}, ..., Res_{in_i})^T$. The complete data, $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})^T$, can be partitioned into two subvectors Y_i^o and Y_i^m, where, Y_i^o is the observed data, representing the vector of observed responses on the ith subject and contains those Y_{ij} for which $Res_{ij} = 1$; Y_i^m is the missing data, indicating the complementary set of responses that are missing where $Res_{ij} = 0$. Here, $i = 1, 2, ..., N$ and $j = 1, 2, ..., n_i$ (Molenberghs et al., 2014).

The missing data mechanism describes the probability that a response is observed

48

or missing. It specifies the probability distribution of the response indicators, Res_i, conditional on Y_i^o, Y_i^m, and X_i. According to the dependent methods of response indicators, Res_i on the response, Y_i and covariates, X_i, the missing data mechanism can be classified into three basic categories, missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) (Molenberghs et al., 2014).

Missing completely at random (MCAR): the probability of missing is independent of both observed subvectors Y_i^o and unobserved subvectors Y_i^m. Therefore, missingness in Y_i is purely the result of a chance mechanism unrelated to any study variable. According to (Little and Rubin, 2002), the term MCAR is defined to the case where

$$\mathbb{P}(Res_i | Y_i^o, Y_i^m, X_i) = \mathbb{P}(Res_i) \qquad (2.1)$$

An MCAR mechanism has important consequences for the analysis of data. If data is MCAR, the observed responses are likely a random sample of the complete data. The analysis performed on the MCAR data is unbiased, and most standard approaches of analysis are applicable to yield valid inferences. However, data are rarely MCAR in clinical trials (Molenberghs et al., 2014).

Missing at Random (MAR): the probability of missing depends on the observed responses, but unrelated to unobserved responses. Both MCAR and MAR are often referred to as ignorable mechanisms. In particular, data are MAR when Res_i is conditionally independent of Y_i^m, given Y_i^o, as

$$\mathbb{P}(Res_i | Y_i^o, Y_i^m, X_i) = \mathbb{P}(Res_i | Y_i^o, X_i) \qquad (2.2)$$

If data is MAR, the observed responses are a random sample of the sampled values within a subclass defined by the observed data rather than the sample of the complete data. The MAR implies that the completers (i.e. those subjects with no missing data) are a biased sample from the target population, consequently, an analysis restricted

to the "completers" is not valid.

Not Missing at Random (NMAR): the probability of missing is related to both observed and unobserved values (Molenberghs et al., 2014). Or the missing data is related to the reason it's missing.

NMAR is often referred to as non-ignorable mechanisms as the analysis aims to make inferences about the distribution of the complete data. When data are NMAR, almost all standard methods of analysis are invalid. However, alternative methods, such as weighting, can be used for handling NMAR (Molenberghs et al., 2014).

According to missing data mechanisms, missing data can be dealt with by complete case analysis, weighting methods and imputation methods.

2.3.2 Complete Case Analysis

Complete case analysis is the most common methods of dealing with missing data. Complete case are the cases whose subjects have no missing data. During the complete case analysis, only those observations with complete data are kept, while cases with a missing value are deleted. Take a longitudinal study as an example, only those patients who were observed a response at each predetermined time point are included in the complete case analysis (Mayers, 2000).

A distinct advantage of this analysis is that it is straightforward and easy to implement. In addition, it provides effective results in the case of missing completely at random (MCAR), while for other missing data mechanism, the analysis may produce biased treatment comparisons. Besides, complete case approach may provide an inefficient estimate with low statistical power due to the reduced sample size. Finally, it is often not a good practice to discard data for the measurements of longitudinal clinical trials. (Weber et al., 2017; Nakai and Ke, 2011)

2.3.3 Inverse Probability Weighting

Another strategy of dealing with missing data is weighting methods. The main idea of weighting methods is to build weights for complete cases to reduce or eliminate bias. Inverse probability weighting is a statistical technique, usually employed to standardize from a sampled population in which the data was collected (i.e. the subjects with non-missing data) to a target population. Weighting, when correctly applied, can reduce the bias of unweighted estimators (Newson, 2013).

Weights are usually characterized as sampling probability. With various combo of covariates, the sampling probability is calculated by the ratio between the frequency of the same observation among overall population and the frequency of the same observation among sample population. In a treatment control study, the sampled population has a list of demographic information, diseases, therapies and labtests both in treatment and control population with/without the drug (disease).

Inverse probability weighting is also used to deal with missing data when subjects with missing data cannot be included in the analysis. Here, the inverse probability weight is defined as a completeness-propensity score (Mayers, 2000). Typically, we have a list of observed responses Y^o, and a list of missing data Y^m. Let "completeness" denote the dummy variable indicating all possible missing variables Y^m are present. Then, we apply logistic regression on "completeness" regarding complete variables Y^o. The predicted probability is considered to be the probability of being complete for each observation. The inverse probability weight is the reciprocal of the predicted probability for each observation. Observations with higher probability of being complete are down-weighted while instances with lower probability of being complete are up-weighted. As such, inverse probability weighting can be used to inflate the weight for subjects under-represented due to a large degree of missing data.

In this thesis, we require both a primary target population for handling miss-

ingness and a secondary target population for covariate balancing. To standardize from sampled population to the secondary target population, we multiply primary weight and secondary weight for the final weight. The primary weight in the product stands for a completeness-propensity weight, standardizing the sampled population to primary target population. While the secondary weight in the product depends on drug/disease status, being different in various schemes.

2.3.4 Imputation

Imputation is any method of replacing missing data with reasonable estimates. Once all the missing data have been imputed and a complete dataset is generated, standard statistical methods can be applied to analyze the dataset. Commonly used imputation methods with reasonable estimation of missing data include mean imputation, regression imputation, last observation carried forward, stochastic imputation, and multiple imputation.

Last observation carried forward (LOCF) method is a common imputation method. When observing the longitudinal measurement of each patient, LOCF obtains the last available response and replaces all subsequent missing values with it. However, LOCF may give a biased treatment comparison if different dropout rates or different drop time occur between treatment groups (Mayers, 2000).

One particular imputation method receiving a lot of attention recently is multiple imputation. Multiple imputation is to impute more than one value for the missing item from an appropriate distribution for the missing values. This will generate two or more complete datasets. Contrast to filling missing data with mean value directly, which usually underestimates variability, multiple imputation considers the uncertainty among missing values. Multiple imputation methodology relies on the MAR assumption.

In this thesis, multiple imputation, especially multiple imputation using chained equation (van Buuren and Groothuis-Oudshoorn, 2011), is replaced by inverse probability weighting due to their similar assumption on MAR and multiple imputations computational infeasibility on such a large dataset ($\geq$20,000 observations) in the current study.

Chapter 3

Statistical Methods for Survival Analysis

Survival analysis is an important sub-area of statistics. It involves the modelling of time to event T_i to analyze the occurrence of specific events of interest at future time points with feature predictors X_i. One of the main challenges in survival analysis is censoring, i.e. events of interest are not identified during observation, or the value of measurement or observation is only partially known. Censoring can be effectively processed using survival analysis techniques by either statistical methods or machine learning techniques (Wang et al., 2017).

In this chapter, we start with a brief introduction on basic concepts to facilitate understanding of survival analysis algorithms. Representative traditional statistical methods for survival analysis, especially semi-parametric models, will be presented next, followed by an introduction to causal inference.

3.1 Concepts and Methods

3.1.1 Survival Data and Censoring

Most statistical and supervised machine learning problems can be formulated as either classification or regression. Classification adopts binary categorical data of either 1, or 0 as response, while regression takes quantitative variable as response. Survival analysis can be used to describe the effect of categorical or quantitative variable on survival. However, standard classification formulation does not take time to event as outcome. Although standard regression formulation does take time to event into consideration, the outcome would confuse the impact of confounders in the regression model (Marubini and Valsecchi, 2004).

Events refer to disease incidence, disease recurrence, disease progression, death or other experience of interest. Time is defined as the time from the beginning of an observation to an event, or to an end of study, or to the loss of contact or withdraw from the study. Survival analysis incorporates both time to event information and event of interest.

Censoring is a main challenge in survival analysis. It is a form of missing data problem in which an instance does not have an event during the observation time. Censoring occurs probably due to limited observation time window or other events lead to missingness (Klein and Moeschberger, 2003).

Censoring can be roughly divide into three groups. The most commonly encountered type is right-censoring, where the observed survival time is less than the real survival time. Left-censoring has the observed survival time greater than the real survival time. When the event occurs within a given time interval, it is interval censoring. However, within all the three cases, the actual event occurrence time is unknown (Lee and Wang, 2003).

In survival analysis, only those instances of events which occur during the study have the precise time to the event of interest (T). For the rest of the instances, the censored time (C) may be the time of withdrawal, loss or end of observation due to lost of track during observation or because the events occur longer than the observation time. For any given instance i, either survival time T_i or censored time C_i can be observed. Right-censored happens if $y_i = min(T_i, C_i)$, of which survival time is a random variable since events may randomly terminate the study (Wang et al., 2017).

For a given instance i, represented by a triplet (X_i, y_i, δ_i), where $X_i \in \mathbb{R}^{1 \times P}$ is the feature vector of instance, δ_i is defined as

$$\delta_i = \begin{cases} 1 & \text{uncensored instance} \\ 0 & \text{censored instance} \end{cases} \tag{3.1}$$

where y_i denotes observed time which equals to survival time T_i for uncensored instances and C_i for censored instances, i.e.,

$$y_i = \begin{cases} T_i & \text{if } \delta_i = 1 \\ C_i & \text{if } \delta_i = 0 \end{cases} \tag{3.2}$$

The objective of survival analysis is to predict time of event of interest T_j for a new instance j with the feature predictor X_j. For a typical survival analysis problem, the value of T_j will be non-negative and continuous (Wang et al., 2017). Details of symbols are listed in Appendix B.

3.1.2 Survival and Hazard Function

Two important functions in survival analysis are survival function and hazard function. The survival function represents the probability that an event of interest survives longer than the specified time t, denoted by S, as

$$S(t) = \mathbb{P}(T \geq t) \tag{3.3}$$

where $\mathbb{P}$ stands for probability, T is a random variable denoting the time of event of interest (Marubini and Valsecchi, 2004; Klein and Moeschberger, 2003). The survival function is monotonically decreasing with t. Its initial value is 1, indicating that 100% of the observed subjects survive at the beginning of observation, i.e. no events of interest occurs at $t = 0$.

The cumulative death distribution function $F(t)$ indicates that the event of interest occurs earlier, defined as the complement of the survival function, i.e.

$$F(t) = 1 - S(t) \tag{3.4}$$

and the death density function, which is the rate of death per unit time, is defined as

$$f(t) = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \tag{3.5}$$

by continuity, we have

$$f(t) = \frac{d}{dt} F(t) \tag{3.6}$$

where Δt denotes a small time interval in discrete cases.

A hazard function represents the probability of event in which no event occurred before time t, which is defined as (Dunn and Clark, 2009)

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t \cdot S(t)} = \frac{f(t)}{S(t)} \tag{3.7}$$

and therefore

$$f(t) = -\frac{d}{dt} S(t) \tag{3.8}$$

58

the hazard function can be expressed as

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}S(t)\frac{1}{S(t)} = -\frac{d}{dt}[ln(S(t))] \tag{3.9}$$

Thus, the survival function can be rewritten as

$$S(t) = e^{-H(t)} \tag{3.10}$$

where the cumulative hazard function (CHF) (Klein and Moeschberger, 2003) is

$$H(t) = \int_0^t h(u)du \tag{3.11}$$

To sum up, the survival function $S(t)$, the cumulative hazard function $H(t)$, the death density function $f(t)$, and the hazard function $h(t)$ are related via

$$S(t) = e^{-H(t)} = \frac{f(t)}{h(t)}, t \geq 0 \tag{3.12}$$

3.1.3   Survival Analysis Methods

Survival analysis methods include statistical methods and machine learning methods. Both of them aim to predict survival time and estimate the probability of survival for estimated survival time. However, statistical methods focus on characterizing the distribution of event times and the statistical properties of parameter estimates by estimating survival curves for low-dimensional data, while machine learning methods concern with predicting event occurrences at a given point in time for high-dimensional settings. Besides, machine learning methods uses the latest developments and optimization to learn the dependencies between covariates and lifetimes (Klein and Moeschberger, 2003).

Within statistical methods, three models, i.e. non-parametric, semi-parametric, and parametric, are used to estimate the survival/hazard functions. Non-parametric methods are employed when no suitable theoretical distributions for the event time are

known. When using non-parametric methods, an empirical estimate of the survival function can be obtained by Kaplan-Meier (KM) method, Nelson-Aalen estimator (NA) or Life-Table (LT). These methods are efficient, yet difficult to interpret, and easy to produce inaccurate estimates. When the time to the event of interest follows a specific distribution, non-parametric methods are less efficient compared to the parametric methods (Lee and Wang, 2003).

Parametric methods are used when the time to event of interest follows a specific distribution, typically the normal, exponential, weibull, logistic, log-logistic and log-normal distributions. Parametric methods include linear regression and accelerated failure time (AFT) methods. The linear regression model includes Tobit, Buckley-James regression and the penalized regression models, which are the most commonly used linear parametric models for survival analysis. If the *log(survival time)* for all instances follow the above specific distributions, then AFT taking survival time as a function of covariates can be employed (Klein and Moeschberger, 2003). Parametric methods are simple, efficient and accurate in predicting the time to event of interest when survival time follows some pre-defined distribution. They are considered as essential alternatives to semi-parametric models including Cox-based regression. However, when the distribution assumption is violated, the survival estimates obtained by the parametric survival models can be inconsistent with a theoretical survival distribution. Also, if there is no suitable theoretical distribution available, non-parametric methods can be more efficient than parametric methods.

The semi-parametric method can be regarded as a hybrid of the parametric and non-parametric methods. It does not require the knowledge of the underlying distribution of survival times. The semi-parametric method can yield more consistent estimators among other conditions compared to parametric models, and a more accurate estimator than non-parametric methods (Powell, 1994). Semi-parametric method includes basic Cox, and its different variants and extensions, such as regularized Cox, CoxBoost, time-dependent Cox models. Although based on parametric regression

model, the event distribution is not known and not easy to interpret (Wang et al., 2017).

Machine learning algorithms, such as survival trees, Bayesian methods, neural networks and support vector machines, have gained popularity in survival analysis. Currently, several advanced machine learning methods, such as ensemble learning, active learning, transfer learning, and multitasking learning, also emerge in the survival analysis (Wang et al., 2017).

3.2 Semi-parametric Statistical Methods

Semi-parametric models are a class of survival models in statistics. This section will introduce two semi-parametric models, i.e. conventional Cox regression (CCR) model and weighted Cox regression model, which will be used in the thesis.

3.2.1 Conventional Cox Regression

The Cox model (Cox, 1972) is the most commonly adopted semi-parametric methods for survival analysis. Based on the proportional hazards assumption, it optimizes partial log likelihood function for parameter estimation.

For a given instance i, represented by a triplet (X_i, y_i, δ_i), hazard function $h(t, X_i)$ in a Cox model follows the proportional hazards assumption given by

$$h(t, X_i) = h_0(t)e^{X_i\beta} \tag{3.13}$$

where the baseline hazard function $h_0(t)$ describes risk change per unit time for baseline covariates. It can be an arbitrary nonnegative function of time. $X_i = (x_{i1}, x_{i2}, \cdots, x_{iP})$ is the covariate vector for instance i. In the thesis, the covariates include treatment assignment, as well as patient characteristics such as age, gender, and the presence of other diseases at start of study, etc. $\beta = (\beta_1, \beta_1, \cdots, \beta_P)^T$ is the coefficient vector. The method represents the effects of covariate vector as a multiplier of a common baseline hazard function, $h_0(t)$.

The hazard ratio between instance X_1 and X_2 can expressed as

$$\frac{h(t, X_1)}{h(t, X_2)} = \frac{h_0(t)e^{X_1\beta}}{h_0(t)e^{X_2\beta}} = e^{(X_1-X_2)\beta} \tag{3.14}$$

Therefore, the hazard ratio is constant and independent of baseline hazard function. Since all subjects share the same baseline hazard function in the Cox propor-

tional hazard model, the survival function can be calculated as

$$S(t) = e^{-H_0(t)e^{X\beta}} = S_0(t)^{e^{X\beta}} \tag{3.15}$$

where $H_0(t)$ represents the cumulative baseline hazard function, and $S_0(t) = e^{-H_0(t)}$ is the baseline survival function.

The Breslow's estimator (Breslow, 1972) is most widely employed to estimate $H_0(t)$,

$$\hat{H}_0(t) = \sum_{t_i \leq t} \hat{h}_0(t_i) \tag{3.16}$$

if t_i is event time

$$\hat{h}_0(t_i) = \frac{1}{\sum\limits_{j \in RK_i} e^{X\beta}} \tag{3.17}$$

otherwise $\hat{h}_0(t_i) = 0$. Here, RK_i stands for the subject set at risk at time t_i.

Since the baseline hazard function $h_0(t)$ in the Cox model is not specified, the standard likelihood function cannot be used to determine the model. In other words, the hazard function $h_0(t)$ is an annoying, and the coefficients β are of interest in the model.

To estimate $\hat{\beta}$, Cox proposed partial likelihood (Lee and Wang, 2003), where the covariates can be estimated without taking time-varying hazard into account.

Let $j = 1, 2, ..., I$ be the total number of events of interest occurred during observation, and $T_1 < T_2 < \cdots < T_j$ be the time for the event of interest. Suppose X_j is the corresponding covariate vector at T_j and R_j be the instances set at risk at T_j.

The individual probability corresponding to the covariate X_j occurred at time T_j

can be expressed as

$$L_i(\beta) = \Big[\frac{h(T_j, X_j)dt}{\sum_{i \in RK_j} h(T_j, X_i)dt}\Big]^{\delta_j} \qquad (3.18)$$

The partial likelihood is the product of the probability of each instance, defined as

$$L(\beta) = \prod_{j=1}^{N} \Big[\frac{e^{X_j\beta}}{\sum_{i \in RK_j} e^{X_i\beta}}\Big]^{\delta_j} \qquad (3.19)$$

When the event occurs, $\delta_j = 1$ in the jth term represents conditional probability. If censored, i.e. $\delta_j = 0$, the corresponding term equals to 1, implying no influence on the partial likelihood.

The corresponding log partial likelihood is

$$log(L(\beta)) = -\sum_{j=1}^{N} \delta_j\Big\{X_j\beta - log\Big[\sum_{i \in RK_j} e^{X_i\beta}\Big]\Big\} \qquad (3.20)$$

The coefficient vector $\hat{\beta}$ is estimated by maximizing the partial likelihood, or equivalently, the negative log-partial likelihood is minimized to improving efficiency (Cox, 1972, 1975). Using numerical Newton-Raphson method (Kelley, 1999), regression coefficients estimates $\hat{\beta}$ can be calculated iteratively by

$$\frac{\partial log(L(\beta))}{\partial \beta} = -\sum_{j=1}^{N} \delta_j\Big\{X_j - \frac{\sum_{i \in RK_j} X_i e^{X_i\beta}}{\sum_{i \in RK_j} e^{X_i\beta}}\Big\} = 0 \qquad (3.21)$$

3.2.2 Weighted Cox Regression

Cox's regression for survival analysis relies on the proportional hazards assumption. When this assumption is violated, an alternative approach, a weighted Cox regression model can be used instead.

In a weighted Cox regression model, except for $\delta_j = w_j$, where w_j is individual

weight, all else will be the same as conventional Cox regression. Using numerical Newton-Raphson method, weighted regression coefficients estimates $\hat{\boldsymbol{\beta}}$ can be calculated iteratively by

$$\frac{\partial log(L(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta}} = -\sum_{j=1}^{N} w_j \left\{ X_j - \frac{\sum_{i \in RK_j} X_i e^{X_i \boldsymbol{\beta}}}{\sum_{i \in RK_j} e^{X_i \boldsymbol{\beta}}} \right\} = 0 \tag{3.22}$$

3.3 Causal Inference

Causal inference is the process of drawing a conclusion about a causal connection based on the occurrence condition of an effect. It attempts to estimate the treatment effects by accounting for the covariates that predict receiving the treatment (Imbens and Rubin, 2015).

In randomized controlled trials, the randomization, which implies that treatment groups will be balanced on average for each covariate, enables unbiased estimation of treatment effects. However, in observational studies, the treatment assignment of participants is typically not random. Due to non-random covariates differences for treated and untreated groups, treatment effect estimates might be biased. It is, therefore, ideally to mimic randomization by creating a treated group with comparable covariates to a control group to reduce bias in treatment effect estimates (Leite, 2017).

In observational studies, treatment effects are commonly estimated by propensity score analysis methods. In simple cases, the treatment and control groups can be easily matched on single characteristics. However, in complex cases which contain many covariates, it is hard to find an appropriate match for each participant with respect to all covariates. (Imbens and Rubin, 2015) This difficulty of multivariate matching can be solved by propensity scores (Rubin, 1973).

Propensity scores are probabilities of treatment assignment that can be used to reduce selection bias. They simplify analysis by reducing all the information in the predictors to one number. Rubin causal model is one of the most commonly used model to obtain propensity score for causal inference. Rosenbaum and Rubin have shown that adjustment for the propensity score is sufficient to remove all bias related to covariates (Rosenbaum and Rubin, 1983). Bias is removed by balancing covariates between treated and untreated groups. It will be much more straightforward to match

instances with different combinations of covariates once covariates are balanced.

3.3.1 Causal Models

Rosenbaum and Rubin proposed the use of propensity scores to reduce selection bias due to confounding variables (Rosenbaum and Rubin, 1983). The propensity score method was connected to matching methods for selecting an untreated group that was similar to the treated group with respect to covariates (Rubin, 1973). It has been wide applied to many fields involving causal inference, such as statistics, sociology, education, economics, psychology, and epidemiology.

In Rubin causal model, all participants have potential outcomes associated with either the presence of treatment or in the absence of treatment. As illustrated in 3.1, an individual i participates in the treatment $Z_i = 1$ has a potential outcome Y_i^1 which is only observed in the presence of the treatment condition; while an individual does not participate $Z_i = 0$ has a potential outcome Y_i^0 which is only observed in the absence of the treatment. Therefore, the treatment effect for each individual is $\tau_i = Y_i^1 - Y_i^0$.

Table 3.1: Potential outcome in Rubin's causal model

	Outcome for Treatment (Y_i^1)	Outcome for Control (Y_i^0)
Treatment ($Z_i = 1$)	$Y_i^1 \vert Z_i = 1$	$Y_i^0 \vert Z_i = 1$
Control ($Z_i = 0$)	$Y_i^1 \vert Z_i = 0$	$Y_i^0 \vert Z_i = 0$

In Table 3.1, the outcomes ($Y_i^1 \vert Z_i = 1$) and ($Y_i^0 \vert Z_i = 0$) are observed, while the outcomes ($Y_i^1 \vert Z_i = 0$) and ($Y_i^0 \vert Z_i = 1$) are missing.

Accordingly, three different types of treatment effects can be defined:

(1) Average treatment effect (ATE): is defined as the difference between the expected values of the potential outcomes of all individuals in the treated and untreated

conditions.

$$ATE = \mathbb{E}(Y_i^1) - \mathbb{E}(Y_i^0) \tag{3.23}$$

(2) The average treatment effect on the treated (ATT): ATT is defined as the difference between the expected values of the potential outcomes of treated individuals.

$$ATT = \mathbb{E}(Y_i^1 | Z_i = 1) - \mathbb{E}(Y_i^0 | Z_i = 1) \tag{3.24}$$

(3) The average treatment effect on the untreated (ATU): ATU is defined as the difference between the expected values of the potential outcomes of the untreated individuals.

$$ATU = \mathbb{E}(Y_i^1 | Z_i = 0) - \mathbb{E}(Y_i^0 | Z_i = 0) \tag{3.25}$$

The treatment effects are selected according to different criteria, such as research question (risk difference, ratio difference), or whether assumptions are met for the treatment effect of interest.

Naive average treatment effect: a simple difference in mean outcomes (SDO) is the difference between the population average outcome for the treatment and control groups.

$$SDO = \mathbb{E}(Y_i^1 | Z_i = 1) - \mathbb{E}(Y_i^0 | Z_i = 0) \tag{3.26}$$

Unobserved counterfactual outcome, defined as

$$Y_i^{CF} = (1 - Z_i)Y_i^1 + Z_i Y_i^0 \tag{3.27}$$

In RCTs, ATE equals to the ATT and ATU because random assignment of participants to conditions implies that they are exchangeable and therefore

$$\mathbb{E}(Y_i^1 | Z_i = 1) = \mathbb{E}(Y_i^0 | Z_i = 0) \tag{3.28}$$

68

and

$$\mathbb{E}(Y_i^0|Z_i = 1) = \mathbb{E}(Y_i^1|Z_i = 0) \qquad (3.29)$$

While in observational studies, the ATE, ATT, and ATU could differ substantially.

While using Rubin's causal model to estimate unbiased treatment effects, it is usually assumed that treatment assignment has strong ignorability, a stable unit treatment value, and adequate balance of covariate distributions between treated and untreated groups (Leite, 2017).

3.3.2 Propensity Score Estimation

The objective of propensity score estimation is to obtain propensity scores for treated and untreated individuals. A propensity score is the probability of an individual given some observed covariates belonging to the treatment group (Rosenbaum and Rubin, 1983).

$$e(X) = \mathbb{P}(Z = 1|X) \qquad (3.30)$$

Propensity score methods aim to reduce selection bias by balancing covariates between treatment and control. With propensity scores, each individual has a unique score that indicates the relationship between covariates and the treatment assignment. Therefore, matching participants according to multiple covariates can be simplified to matching by the propensity score (Rubin, 1973).

If treatment is independent of potential outcomes Y^0 and Y^1 given a set of observed covariates X, potential outcomes are also independent of treatment given propensity score $e(X)$ which is a function of these covariates. Treatment is also independent of covariates with propensity score, i.e. (Rosenbaum and Rubin, 1983)

$$if \ (Y^0, Y^1) \perp Z|X, \ then \qquad (3.31)$$

$$(Y^0, Y^1) \perp Z|X \text{ and } Z \perp X|e(X) \tag{3.32}$$

Since the mean difference between treated and untreated outcomes at a specific propensity score is the average treatment effect at that propensity score due to the balancing characteristic of propensity score, matching, weighting, and stratification based on the propensity score can provide unbiased estimates of the treatment effect (Imbens and Rubin, 2015).

Propensity scores can be estimated by parametric models and data mining methods. Parametric models include logistic regression, probit regression, and discriminant function analysis, while classification trees and random forests belong to a class of methods known as data mining, or machine learning (Berk, 2006). However, the selection of propensity score estimation methods is less important than the controlling for the right covariates in reducing bias (Setoguchi et al., 2008; Westreich et al., 2010).

A successful estimation of propensity scores should produce adequate balance of covariate distributions between treated and untreated groups when estimated in combination with a matching, stratification, or weighting strategy (Ho et al., 2007). In this thesis, we will only fit a logistic regression model to the data to predict treatment assignment while maximizing covariate balance by propensity score weighting.

Logistic regression is a commonly used model propensity score estimation and the obtained propensity scores are similar to those acquired by probit regression or discriminant function analysis. A basic logistic regression model for estimating propensity scores is expressed as

$$logit(Z_i = 1|X) = \beta_0 + \beta_1 X_{1,i} + ... + \beta_k X_{k,i} \tag{3.33}$$

Covariates $X_1, X_2, ..., X_k$ are considered to be either true confounders or predictors of the outcome. Higher order polynomial terms (e.g. X_k^2, X_k^3) and/or interaction

terms (e.g. $X_1 \times X_2$) can be added to the model. The coefficients in logistic regression is optimized by maximum likelihood estimation. The propensity scores are estimated probabilities of treatment assignment, expressed in terms of estimated logits,

$$e_i(X) = \frac{exp(logit(Z_i = 1|X))}{1 + exp(logit(Z_i = 1|X))} \tag{3.34}$$

The selection of covariates for the propensity score model is critical. The propensity score model should contain all true confounders, the covariates that affect the treatment assignment and the outcome. If an important true confounder is not added, there will be substantial bias remaining even if covariates are well balanced. The propensity score model may also include predictors of the outcome that are unrelated to treatment assignment, but not covariates that are related to treatment assignment but not the outcome (Brookhart et al., 2006).

3.3.3 Propensity Score Weighting

This section discusses treatment effects estimation by using propensity score weighting (PSW) by different types of treatment effect of interest. By using propensity scores, we are able to model the behavior of primary care physicians by balancing covariates among different drug groups.

For average treatment effect (ATE) weights (Robins et al., 2000)

$$w_{ATE} = \begin{cases} \frac{1}{e(X)} & Z_i = 1 \\ \frac{1}{1-e(X)} & Z_i = 0 \end{cases} \tag{3.35}$$

In the ATE analysis, we are actually inflating numbers of participants in both cases reaching to the number of participants level n.

ATT weight is estimated by the odds (Harder et al., 2010),

$$w_{ATT} = \begin{cases} 1 & Z_i = 1 \\ \frac{e(X)}{1-e(X)} & Z_i = 0 \end{cases} \tag{3.36}$$

Under ATT weighting scheme, the number of untreated participants is adjusted which will lead to a bias.

ATU weight is estimated by the odds inverse,

$$w_{ATU} = \begin{cases} \frac{1-e(X)}{e(X)} & Z_i = 1 \\ 1 & Z_i = 0 \end{cases} \tag{3.37}$$

Under ATU weighting scheme, the number of treated participants is adjusted which will lead to a bias.

It is easy to find out that

$$w_{ATT} + w_{ATU} = w_{ATE} \tag{3.38}$$

Though we will inflate both anti-diabetes cases to the population level applying ATE weights, the benefit is obvious: we will have reciprocal directional results if we inverse "treatment and control" leading to the same conclusion.

3.3.4 Covariate Balance Evaluation

The objective of propensity score weighting is to balance covariates and therefore to remove selection bias. In this work, Somers' D $D(Y|X)$ is employed to assess covariate balance. X are covariates, Y is a confounder or a propensity score (Newson, 2006).

Assume pairs (X_i, Y_i) and (X_j, Y_j) are sampled under a specified sampling scheme from a population of bivariate pairs (X, Y).

Kendall's tau τ_a can be defined as

$$\tau_{XY} = \mathbb{E}[sign(X_i - X_j)sign(Y_i - Y_j)] \tag{3.39}$$

Kendall's tau can be interpreted as the expectation of concordance or discordance between two (X, Y) pairs. A pair of (X, Y) is concordant if the ranks of both elements agree, e.g. the larger X value is paired with the larger Y value, and is discordant if the ranks of both elements disagree, e.g. the larger X value is paired with the smaller Y value (Newson, 2006).

Somer's D $D(X|Y)$ is defined as

$$D(X|Y) = \frac{\tau_{XY}}{\tau_{XX}} \tag{3.40}$$

Somer's D can be interpreted as the asymmetrical ratio between two conditional expectation of concordance or discordance. Definition and calculation can be extended to censored cases and also to treatment effect of X on Y (Newson, 2006).

Specifically, if X and Y are binary variables, Somers' D can be calculated by the difference between proportions (Newson, 2014)

$$D_{XY} = \mathbb{P}(Y = 1|X = 1) - \mathbb{P}(Y = 1|X = 0) \tag{3.41}$$

If X is binary, Y^1 and Y^0 corresponds to a sample from $X = 1$ and $X = 0$ (Newson, 2014)

$$D_{XY} = \mathbb{P}(Y^1 > Y^0) - \mathbb{P}(Y^0 > Y^1) \tag{3.42}$$

3.3.5 Sensitivity Analysis

Sensitivity analysis aims to identify the impact of unobserved confounding on the significance test of the treatment effect (Rosenbaum, 2010; Rosenbaum and Rubin, 1983). It can improve the understanding of the relationships between input and output variables in a model. By focusing on the sensitive parameters obtained from sensitivity analysis, important connections between observations, inputs, and predictions, can be identified, leading to the development of robust models.

Usually, the larger the sensitivity parameter, the greater the impact of the unobserved confounding is. Therefore, in practical analysis, small sensitivity parameter attributes are removed, and researches are focus on the large sensitivity parameter inputs that cause significant uncertainty in the output to increase robustness. This can greatly simplify the model and reduce the workload of data analysis and processing.

Chapter 4

Study Design

4.1 Exploratory Data Analysis

Included were 294,701 participants with type II diabetes and at least one anti-diabetes prescription recorded in CPRD. Among these diabetes patients, 148,983 individuals (50.6%) started taking metformin monotherapy, 61,741 (21.0%) started sulfonylureas monotherapy, and some smaller number started taking other drugs as shown in Table 4.1. Figure 4-1 and Figure 4-2 reveals the trend of all types of initial anti-diabetic drug prescription in number and in percentage by prescription year respectively. Since synthetic RCT can only include two drugs as "treatment" and "control", combination therapies are dropped due to our focus on intention-to-treat (ITT) analysis and its hybrid feature. Furthermore, due to similar clinical indications and number of patients with initial anti-diabetic drug prescription, we limit our horizon within metformin and sulfonylureas monotherapy initiators as listed in Table 4.2. Figure 4-3 and Figure 4-4 reveal the trend of initial metformin and sulfonylureas prescription in number and in percentage by prescription year, respectively.

Table 4.1: Number of anti-diabetes initiators by drug classes

Model	Number of observations	%
Metformin	148,983	50.6%
Sulfonylureas	61,741	21.0%
Glitazone	519	1.8%
Insulin	15,793	5.4%
Combination	67,665	23.0%
Total	294,701	100%

Table 4.2: Number of metformin and sulfonylureas initiators

Model	Number of observations	%
Metformin	148,983	70.7%
Sulfonylureas	61,741	29.3%
Total	210,724	100%
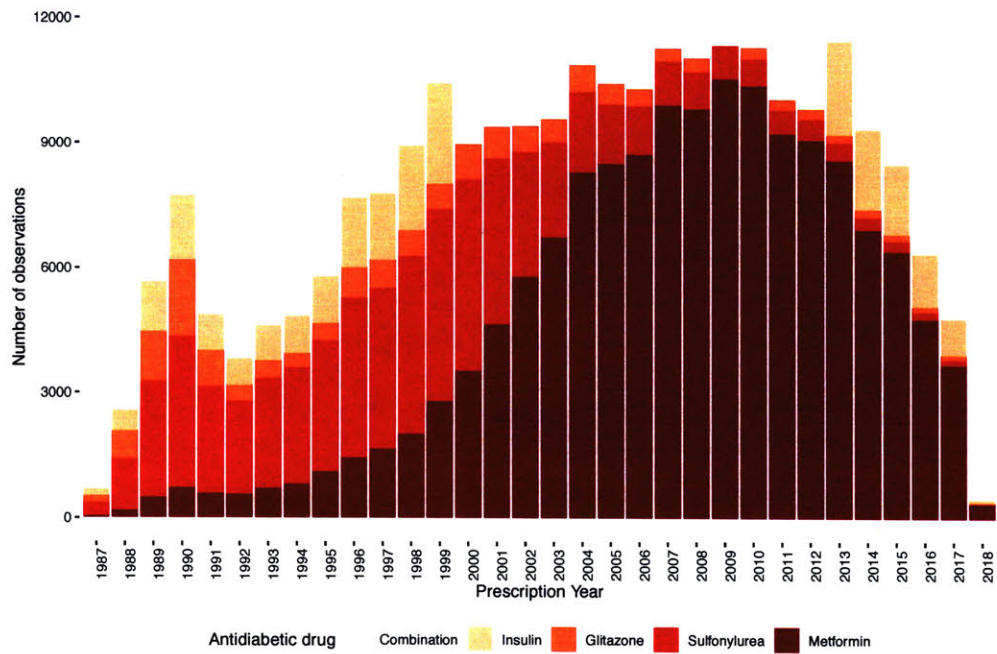
Figure 4-1: Number of new initial anti-diabetes prescription on all drug classes by calendar year
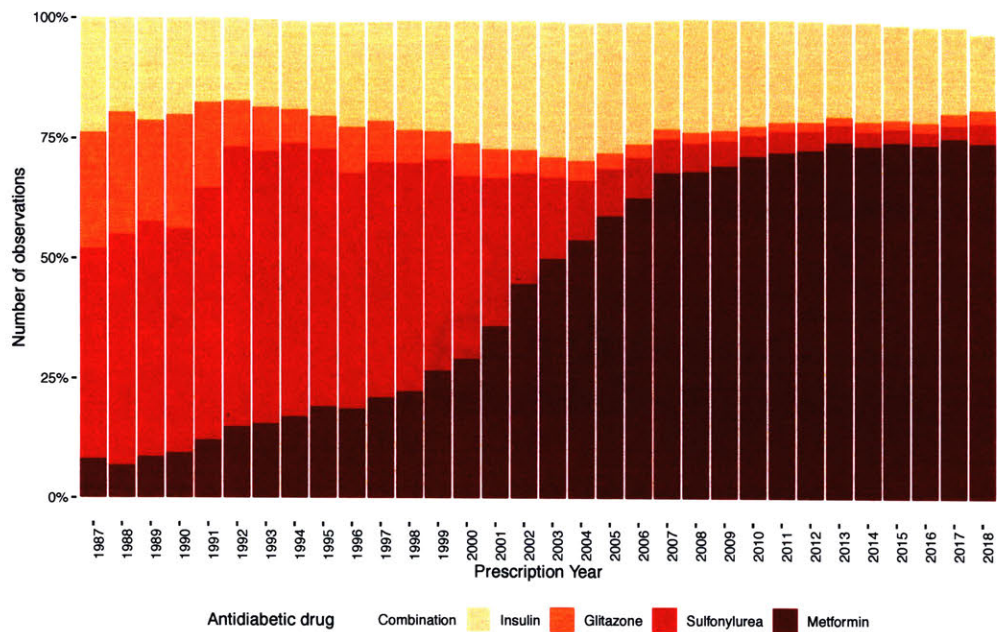
Figure 4-2: Percentage of new initial anti-diabetes prescription on all drug classes by calendar year
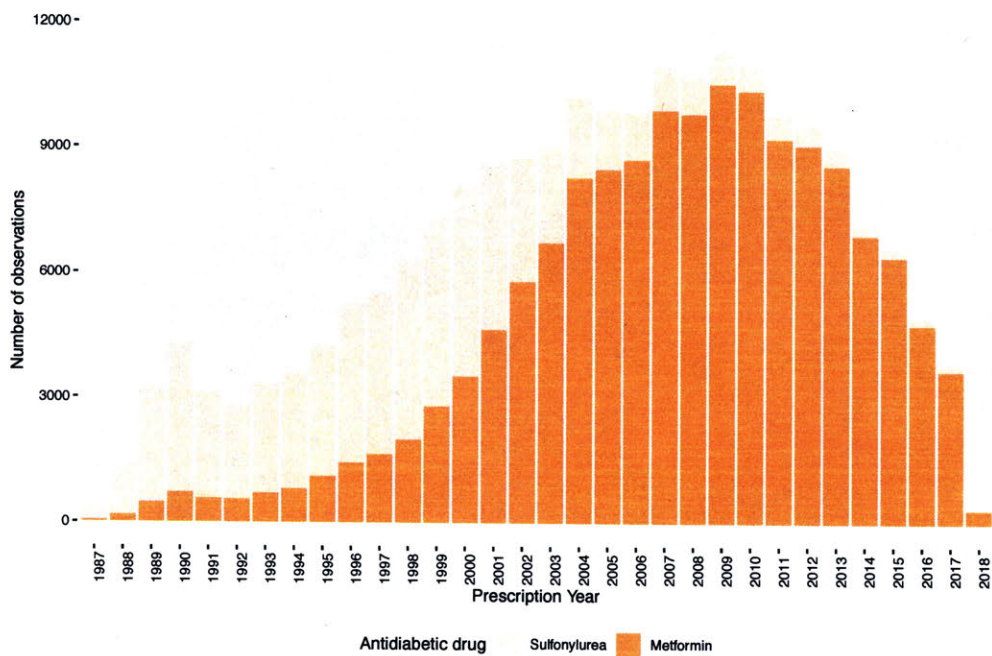
Figure 4-3: Number of new initial anti-diabetes prescription on metformin and sulfonylureas by calendar year
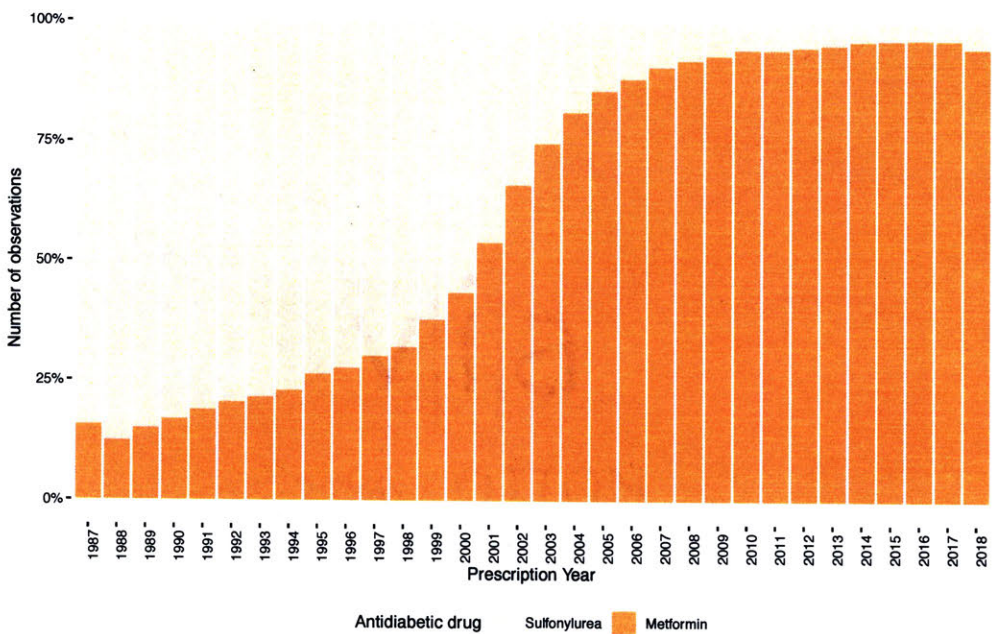
Figure 4-4: Percentage of new initial anti-diabetes prescription on metformin and sulfonylureas by calendar year

Table 4.3 and Table 4.4 report the baseline demographic and region characteristics according to initial metformin and sulfonylureas. Compared with those who started using sulfonylureas monotherapy, those who started using metformin monotherapy were, on average, younger (mean age: 63.45 vs. 68.50 years), had a higher BMI (mean: 32.31 vs. 27.52 kg/m^2), their median year of the first prescription was more recent (2008 vs. 1998), with a much smaller proportion of missing on smoking.

Table 4.3: Baseline characteristics by initial anti-diabetes prescription on metformin and sulfonylureas

Covariates	Metformin	%	Sulfonylureas	%	Total
Gender					
Male	83,143	70.7%	34,392	29.3%	117,535
Female	65,840	70.7%	27,349	29.3%	93,189
Prescripton age					
35 − 49	20,594	81.9%	4,549	18.1%	25,143
50 − 64	64,580	76.9%	19,417	23.1%	83,997
65 − 79	53,380	65.6%	27,942	34.4%	81,322
≥ 80	10,429	51.5%	9,833	48.5%	20,262
Prescripton year					
< 1990	722	14.4%	4,306	85.6%	5,028
1990 − 1999	15,257	30.8%	34,353	69.2%	49,610
2000 − 2009	83,318	83.0%	17,082	17.0%	100,400
≥ 2010	48,954	95.3%	2,395	4.7%	51,349
Region					
1	2,054	67.9%	973	32.1%	3,027
2	16,976	68.9%	7,661	31.1%	24,637
3	4,041	57.9%	2,941	42.1%	6,982
4	4,120	58.8%	2,887	41.2%	7,007
5	14,975	72.6%	5,664	27.4%	20,639
6	11,199	66.3%	5,681	33.7%	16,880
7	12,241	69.5%	5,381	30.5%	17,622
8	14,143	73.4%	5,137	26.6%	19,280
9	16,866	73.5%	6,084	26.5%	22,950
10	14,817	74.1%	5,189	25.9%	20,006
11	4,725	74.0%	1,659	26.0%	6,384
12	13,832	73.2%	5,073	26.8%	18,905
13	18,994	71.9%	7,411	28.1%	26,405

Table 4.5 and Table 4.6 reflects number of overall commorbidites and commorbidites after initial metformin and sulfonylureas prescription, respectively. CHD and hypertension dropped far more than other commorbities (50%). Also, the proportion of CHD and hypertension changed quite a bit compared with other commorbities.

Speaking of cases of interest, there are a total of 42,898 first incident cancers identified, of which 4,929 were postmenopausal breast cancers, 5,195 were prostate cancers, 4,876 were bowel cancers, and 3,328 were lung cancers. After first metformin and sulfonylureas prescription, the number of first incident cancers identified dropped to 23,466, of which 2,023 were postmenopausal breast cancers, 2,952 were prostate cancers, 2,928 were bowel cancers, and 2,776 were lung cancers.

Table 4.4: Baseline characteristics by initial anti-diabetes prescription on metformin and sulfonylureas, continued

Covariates	Metformin	%	Sulfonylureas	%	Total
BMI, kg/m^2					
< 18.5	206	33.3%	412	66.7%	618
$18.5 - 24.9$	10,038	46.7%	11,453	53.3%	21,491
$25 - 29.9$	41,017	71.8%	16,108	28.2%	57,125
≥ 30	80,720	89.3%	9,636	10.7%	90,356
Missing	17,002	41.3%	24,132	58.7%	41,134
IMD, kg/m^2					
Least deprived	26,854	72.0%	10,455	28.0%	37,309
2	25,322	69.4%	11,163	30.6%	36,485
3	29,508	70.5%	12,326	29.5%	41,834
4	30,147	70.4%	12,680	29.6%	42,827
Most deprived	27,219	69.3%	12,072	30.7%	39,291
Missing	9,933	76.5%	3,045	23.5%	12,978
Smoking					
Non-smoker	69,784	71.5%	27,767	28.5%	97,551
Ex-smoker	25,524	73.9%	8,996	26.1%	34,520
Smoker	48,794	82.9%	10,801	18.1%	59,595
Missing	4,881	25.6%	14,177	74.4%	19,058
HbA1c					
$< 7\%$	18,961	85.7%	3,168	14.3%	22,129
$7\% - 10\%$	76,497	84.8%	13,697	15.2%	90,194
$> 10\%$	17,192	75.0%	5,729	25.0%	22,921
Missing	36,333	48.1%	39,147	51.9%	75,480

Table 4.5: Overall comorbidity characteristics by metformin and sulfonylureas

Covariates	Metformin	%	Sulfonylureas	%	Total
HF	10,825	48.4%	11,535	51.6%	22,360
CHD	77,375	70.5%	32,320	29.5%	109,695
AF	15,620	63.9%	8,812	36.1%	24,432
Stroke	17,368	57.0%	13,095	43.0%	30,463
Hypertension	144,643	71.6%	57,490	28.4%	202,133
PVD	46,266	74.9%	15,482	25.1%	61,748
CKD	32,700	69.2%	14,564	30.8%	47,264
COPD	11,482	73.1%	4,215	26.9%	15,697
Cancer	28,290	65.9%	14,608	34.1%	42,898
Breast cancer	3,605	73.1%	1,324	26.9%	4,929
Prostate cancer	3,464	66.7%	1,731	33.3%	5,195
Bowel cancer	3,026	62.1%	1,850	37.9%	4,876
Lung cancer	1,949	58.6%	1,379	41.4%	3,328

Table 4.6: Comorbidity characteristics after initial metformin and sulfonylureas prescription

Covariates	Metformin	%	Sulfonylureas	%	Total
HF	6,043	46.0%	7,081	54.0%	13,124
CHD	17,594	54.1%	14,945	45.9%	32,539
AF	7,596	61.8%	4,688	38.2%	12,284
Stroke	6,956	48.5%	7,389	51.5%	14,345
Hypertension	7,224	39.8%	10,911	60.2%	18,135
PVD	26,233	69.7%	11,396	30.3%	37,629
CKD	21,854	63.6%	12,510	36.4%	34,364
COPD	4,941	65.3%	2,628	34.7%	7,569
Cancer	14,338	61.1%	9,128	38.9%	23,466
Breast cancer	1,381	68.3%	642	31.7%	2,023
Prostate cancer	1,836	62.2%	1,116	37.8%	2,952
Bowel cancer	1,763	60.2%	1,165	39.8%	2,928
Lung cancer	1,645	59.5%	1,131	40.7%	2,776

For all of the covariates included in this thesis, four of them have missing data problems - BMI, IMD, Smoking and HbA1c. The quality of existing data depends on first anti-diabetes prescription year as in Table 4.7. We figured out that the proportion of missing data is lower when first anti-diabetes prescription happens later than 2000, and given the large number of observations with initial anti-diabetes prescription year later than 2000, we decided to add a constriant on initial anti-diabetes prescription at 2000 for subgroup analyses, as shown in Table 4.17.

Table 4.7: BMI, IMD, smoking and HbA1c summary statistics by first anti-diabetes prescription year

Covariates	<2000	%	≥2000	%	Total
BMI, kg/m^2					
< 18.5	155	25.1%	463	74.9%	618
18.5 − 25	5,852	27.2%	15,639	72.8%	21,491
25 − 30	11,321	19.8%	45,804	80.2%	57,125
≥ 30	10,435	11.5%	79,921	88.5%	90,356
Missing	23,096	56.1%	18,038	43.8%	41,134
IMD, kg/m^2					
Least deprived	8,419	22.6%	28,890	77.4%	37,309
2	9,012	32.8%	27,473	67.2%	36,485
3	10,214	24.4%	31,620	75.6%	41,834
4	10,439	24.4%	32,388	75.6%	42,827
Most deprived	10,577	26.9%	28,714	73.1%	39,291
Missing	2,198	16.9%	10,780	83.1%	12,978
Smoking					
Non-smoker	22,661	23.2%	74,890	76.8%	97,551
Ex-smoker	6,874	19.9%	27,646	80.1%	34,520
Smoker	6,205	10.4%	53,390	89.6%	59,595
Missing	15,119	79.3%	3,939	20.7%	19,058
HbA1c					
< 7%	1,808	8.2%	20,321	91.8%	22,129
7% − 10%	6,559	7.3%	83,635	92.7%	90,194
> 10%	3,518	15.3%	19,403	84.7%	22,921
Missing	38,974	51.6%	36,506	48.4%	75,480

Table 4.8, 4.9 and 4.10 presents average initial metformin and sulfonylureas pre-scription age, average cancer incidence age and average all-cause death age among all metformin and sulfonylureas initiators, respectively. Average initial metformin or sulfonylureas prescription age and average cancer diagnosis age are close, while average all-cause death age are approximately 10 years after. The proportion of death for metformin initiators dropped to 50% compared with the proportion of metformin initiators among metformin and sulfonylureas initiators which is around 70% from Table 4.2.

Table 4.11, 4.12 and 4.13 presents average initial metformin and sulfonylureas prescription age, average cancer incidence age and average all-cause death age among those who prescribed metformin or sulfonylureas before cancer diagnosis, which are designed for cancer incidence risks analyses. Since the proportion of cancer diagnosis for both drugs are similar among the population level as in Table 4.2, metformin and sulfonylureas might have similar risks in cancer incidence.

Table 4.14, 4.15 and 4.16 presents average initial metformin and sulfonylureas prescription age, average cancer incidence age and average all-cause death age among those who prescribed metformin or sulfonylureas after cancer diagnosis, which are designed for cancer mortality risks analyses. Since the proportion of all-cause death for metformin initiators dropped from from 70% (Table 4.2) to around 50%, this may contribute to the relative protective effect of metformin on cancer mortality.

Table 4.8: Average initial metformin/sulfonylureas prescription age

Disease	Met	N	%	Sulf	N	%	All	N
Cancer	68.3	28,290	65.9%	70.2	14,608	34.1%	69.0	42,898
Breast cancer	66.8	3,605	73.1%	70.4	1,324	26.9%	67.8	4,929
Prostate cancer	70.5	3,464	66.7%	71.5	1,731	33.3%	70.8	5,195
Bowel cancer	68.7	3,026	62.1%	71.2	1,850	37.9%	69.3	4,876
Lung cancer	68.3	1,949	58.6%	68.2	1,379	41.4%	68.2	3,328

Table 4.9: Average cancer incidence age by initial metformin/sulfonylureas prescription

Disease	Met	N	%	Sulf	N	%	All	N
Cancer	67.1	28,268	66.0%	72.0	14,587	34.0%	68.8	42,855
Breast cancer	64.0	3,600	73.2%	69.7	1,318	26.8%	65.5	4,918
Prostate cancer	71.0	3,462	66.7%	74.8	1,726	33.3%	72.2	5,188
Bowel cancer	68.7	3,023	62.1%	72.2	1,846	37.9%	70.1	4,869
Lung cancer	72.1	1,948	58.6%	73.5	1,377	41.4%	72.7	3,325

Table 4.10: Average all-cause death age by initial metformin/sulfonylureas prescription

Disease	Met	N	%	Sulf	N	%	All	N
Cancer	77.0	9,101	50.0%	78.9	9,101	50.0%	78.0	18,202
Breast cancer	77.2	799	52.8%	80.1	713	47.2%	78.5	1,512
Prostate cancer	79.9	959	48.6%	81.1	1,016	51.4%	80.5	1,975
Bowel cancer	77.0	1,143	48.1%	78.7	1,234	51.9%	77.9	2,377
Lung cancer	74.2	1,422	54.3%	74.8	1,198	45.7%	74.5	2,620

Table 4.11: Average initial metformin/sulfonylureas prescription age when prescription happens before cancer diagnosis

Disease	Met	N	%	Sulf	N	%	All	N
Cancer	66.2	14,338	61.1%	68.0	9,128	38.9%	66.9	23,466
Breast cancer	64.3	1,381	68.3%	68.1	642	31.7%	65.5	2,023
Prostate cancer	67.6	1,836	62.2%	68.7	1,116	37.8%	68.0	2,952
Bowel cancer	66.5	1,763	60.2%	68.0	1,165	39.8%	67.1	2,928
Lung cancer	67.8	1,645	59.3%	67.6	1,131	40.7%	67.7	2,776

Table 4.12: Average cancer incidence age when initial metformin/sulfonylureas prescription happens before cancer diagnosis

Disease	Met	N	%	Sulf	N	%	All	N
Cancer	71.5	14,338	61.1%	75.0	9,128	38.9%	72.9	23,466
Breast cancer	69.4	1,381	68.3%	74.9	642	31.7%	71.1	2,023
Prostate cancer	72.7	1,836	62.2%	76.1	1,116	37.8%	74.0	2,952
Bowel cancer	72.2	1,763	60.2%	75.3	1,165	39.8%	73.4	2,928
Lung cancer	73.2	1,645	59.3%	74.5	1,131	40.7%	73.8	2,776

Table 4.13: Average all-cause death age when initial metformin/sulfonylureas prescription happens before cancer diagnosis

Disease	Met	N	%	Sulf	N	%	All	N
Cancer	75.8	5,768	50.3%	78.5	5,690	49.7%	77.2	11,458
Breast cancer	77.3	365	52.5%	81.3	330	47.5%	79.2	695
Prostate cancer	79.0	507	45.6%	81.1	605	54.4%	80.1	1,112
Bowel cancer	76.2	788	50.2%	78.6	781	49.8%	77.4	1,569
Lung cancer	74.3	1,277	56.4%	75.3	989	43.6%	74.7	2,266

Table 4.14: Average initial metformin/sulfonylureas prescription age when prescription happens after cancer diagnosis

Disease	Met	N	%	Sulf	N	%	All	N
Cancer	70.5	13,900	71.9%	73.9	5,422	28.1%	71.5	19,322
Breast cancer	68.2	2,213	76.6%	72.6	676	23.4%	69.3	2,889
Prostate cancer	73.7	1,619	72.7%	76.5	607	27.3%	74.5	2,226
Bowel cancer	71.8	1,259	65.0%	73.8	679	35.0%	72.5	1,938
Lung cancer	70.7	301	55.9%	71.0	237	44.1%	70.8	538

Table 4.15: Average cancer incidence age when initial metformin/sulfonylureas prescription happens after cancer diagnosis

Disease	Met	N	%	Sulf	N	%	All	N
Cancer	62.5	13,900	71.9%	67.0	5,422	28.1%	63.8	19,322
Breast cancer	60.6	2,213	76.6%	64.8	676	23.4%	61.6	2,889
Prostate cancer	69.0	1,619	72.7%	72.5	607	27.3%	69.9	2,226
Bowel cancer	63.9	1,259	65.0%	66.9	679	35.0%	65.0	1,938
Lung cancer	65.8	301	55.9%	68.9	237	44.0%	67.1	538

Table 4.16: Average all-cause death age when initial metformin/sulfonylureas prescription happens after cancer diagnosis

Disease	Met	N	%	Sulf	N	%	All	N
Cancer	79.2	3,315	49.6%	79.5	3,367	50.4%	79.3	6,682
Breast cancer	77.1	430	53.2%	79.1	379	46.8%	78.0	809
Prostate cancer	80.8	445	52.3%	81.1	406	47.7%	80.9	851
Bowel cancer	78.8	353	44.1%	78.9	448	55.9%	78.8	801
Lung cancer	73.8	143	41.6%	72.7	201	58.4%	73.1	344

4.2 Synthetic Randomized Controlled Trial Design

Table 4.17 reveals the design of randomized controlled trial emulation in this thesis. Particularly, 10 sub-studies are composed of general cancer, breast cancer, prostate cancer, bowel cancer and lung cancer together with incidence and mortality risks. Within in each study, we investigated CCR, ATE, ATT and ATU weighting schemes with metformin/sulfonylureas and sulfonylureas/metformin as treatment and control. Fewer variables, complete case analysis, treating missing as a separate category and inverse probability weighting are employed to enclose various missing mechanisms on missing data. From Table 4.7, we add a final constraint on initial anti-diabetes prescription at 2000 to avoid massive missing data problem. All the cases formulated 640 emulated RCTs so as to take all factors and cases into consideration.

Table 4.17: Synthetic RCT study design, 640 in total (+: include possible cases above)

Multiples	5	2	4	2	4	2	Model
Sub-study 1	Cancer	Incidence	CCR	Ref=Met	Fewer	All	Model 1
Sub-study 1	Cancer	Incidence	CCR	Ref=Met	Fewer	≥2000	Model 2
Sub-study 1	Cancer	Incidence	CCR	Ref=Met	Complete	All	Model 3
Sub-study 1	Cancer	Incidence	CCR	Ref=Met	Complete	≥2000	Model 4
Sub-study 1	Cancer	Incidence	CCR	Ref=Met	Separate	All	Model 5
Sub-study 1	Cancer	Incidence	CCR	Ref=Met	Separate	≥2000	Model 6
Sub-study 1	Cancer	Incidence	CCR	Ref=Met	IPW	All	Model 7
Sub-study 1	Cancer	Incidence	CCR	Ref=Met	IPW	≥2000	Model 8
Sub-study 1	Cancer	Incidence	CCR	Ref=Sulf	Fewer	All	Model 9
Sub-study 1	Cancer	Incidence	CCR	Ref=Sulf	Fewer	≥2000	Model 10
Sub-study 1	Cancer	Incidence	CCR	Ref=Sulf	Complete	All	Model 11
Sub-study 1	Cancer	Incidence	CCR	Ref=Sulf	Complete	≥2000	Model 12
Sub-study 1	Cancer	Incidence	CCR	Ref=Sulf	Separate	All	Model 13
Sub-study 1	Cancer	Incidence	CCR	Ref=Sulf	Separate	≥2000	Model 14
Sub-study 1	Cancer	Incidence	CCR	Ref=Sulf	IPW	All	Model 15
Sub-study 1	Cancer	Incidence	CCR	Ref=Sulf	IPW	≥2000	Model 16
Sub-study 1	Cancer	Incidence	ATE	+	+	+	+
Sub-study 1	Cancer	Incidence	ATT	+	+	+	+
Sub-study 1	Cancer	Incidence	ATU	+	+	+	+
Sub-study 2	Cancer	Mortality	CCR	+	+	+	+
Sub-study 2	Cancer	Mortality	ATE	+	+	+	+
Sub-study 2	Cancer	Mortality	ATT	+	+	+	+
Sub-study 2	Cancer	Mortality	ATU	+	+	+	+
Sub-study 3	Breast	Incidence	+	+	+	+	+
Sub-study 4	Breast	Mortality	+	+	+	+	+
Sub-study 5	Prostate	Incidence	+	+	+	+	+
Sub-study 6	Prostate	Mortality	+	+	+	+	+
Sub-study 7	Bowel	Incidence	+	+	+	+	+
Sub-study 8	Bowel	Mortality	+	+	+	+	+
Sub-study 9	Lung	Incidence	+	+	+	+	+
Sub-study 10	Lung	Mortality	+	+	+	+	+

4.3 Model Specifications

Within each scheme, 16 Models on the rightmost column tabulated in 4.17 are specified as follows,

Model 1: metformin as control; adjust for fewer variables including gender, initial anti-diabetes prescription year, region and stratified on prescription age (5-year span).

Model 2: metformin as control; restrict initial anti-diabetes prescription year from 2000 onwards; adjust for fewer variables including gender, prescription year, region and stratified on prescription age (5-year span).

Model 3: metformin as control; complete case analysis; adjusted for gender, initial anti-diabetes prescription year, region, stratified on prescription age (5-year span), IMD, BMI, HbA1c, smoking, and comorbidities (HF, CHD, AF, stroke, hypertension, PVD, CKD, COPD).

Model 4: metformin as control; restrict initial anti-diabetes prescription year from 2000 onwards; complete case analysis; adjusted for gender, initial anti-diabetes prescription year, region, stratified on prescription age (5-year span), IMD, BMI, HbA1c, smoking, and comorbidities (HF, CHD, AF, stroke, hypertension, PVD, CKD, COPD).

Model 5: metformin as control; missing data as a separate category; adjusted for gender, initial anti-diabetes prescription year, region, stratified on prescription age (5-year span), IMD, BMI, HbA1c, smoking, and comorbidities (HF, CHD, AF, stroke, hypertension, PVD, CKD, COPD).

Model 6: metformin as control; restrict initial anti-diabetes prescription year from 2000 onwards; missing data as a separate category; adjusted for gender, initial anti-

diabetes prescription year, region, stratified on prescription age (5-year span), IMD, BMI, HbA1c, smoking, and comorbidities (HF, CHD, AF, stroke, hypertension, PVD, CKD, COPD).

Model 7: metformin as control; inverse probability weighting on complete cases; adjusted for gender, initial anti-diabetes prescription year, region, stratified on prescription age (5-year span), IMD, BMI, HbA1c, smoking, and comorbidities (HF, CHD, AF, stroke, hypertension, PVD, CKD, COPD).

Model 8: metformin as control; restrict initial anti-diabetes prescription year from 2000 onwards; inverse probability weighting on complete cases; adjusted for gender, initial anti-diabetes prescription year, region, stratified on prescription age (5-year span), IMD, BMI, HbA1c, smoking, and comorbidities (HF, CHD, AF, stroke, hypertension, PVD, CKD, COPD).

Model 9: sulfonylureas as control; adjust for fewer variables including gender, initial anti-diabetes prescription year, region and stratified on prescription age (5-year span).

Model 10: sulfonylureas as control; restrict initial anti-diabetes prescription year from 2000 onwards; adjust for fewer variables including gender, prescription year, region and stratified on prescription age (5-year span).

Model 11: sulfonylureas as control; complete case analysis; adjusted for gender, initial anti-diabetes prescription year, region, stratified on prescription age (5-year span), IMD, BMI, HbA1c, smoking, and comorbidities (HF, CHD, AF, stroke, hypertension, PVD, CKD, COPD).

Model 12: sulfonylureas as control; restrict initial anti-diabetes prescription year from 2000 onwards; complete case analysis; adjusted for gender, initial anti-diabetes

prescription year, region, stratified on prescription age (5-year span), IMD, BMI, HbA1c, smoking, and comorbidities (HF, CHD, AF, stroke, hypertension, PVD, CKD, COPD).

Model 13: sulfonylureas as control; missing data as a separate category; adjusted for gender, initial anti-diabetes prescription year, region, stratified on prescription age (5-year span), IMD, BMI, HbA1c, smoking, and comorbidities (HF, CHD, AF, stroke, hypertension, PVD, CKD, COPD).

Model 14: sulfonylureas as control; restrict initial anti-diabetes prescription year from 2000 onwards; missing data as a separate category; adjusted for gender, initial anti-diabetes prescription year, region, stratified on prescription age (5-year span), IMD, BMI, HbA1c, smoking, and comorbidities (HF, CHD, AF, stroke, hypertension, PVD, CKD, COPD).

Model 15: sulfonylureas as control; inverse probability weighting on complete cases; adjusted for gender, initial anti-diabetes prescription year, region, stratified on prescription age (5-year span), IMD, BMI, HbA1c, smoking, and comorbidities (HF, CHD, AF, stroke, hypertension, PVD, CKD, COPD).

Model 16: sulfonylureas as control; restrict initial anti-diabetes prescription year from 2000 onwards; inverse probability weighting on complete cases; adjusted for gender, initial anti-diabetes prescription year, region, stratified on prescription age (5-year span), IMD, BMI, HbA1c, smoking, and comorbidities (HF, CHD, AF, stroke, hypertension, PVD, CKD, COPD).

Gender is excluded in breast cancer and prostate cancer cases by limiting gender to female and male correspondingly.

4.4 Positive Study Settings

Dependent variable settings

As comorbidities/complications are potential confounders, we take comorbidities as 1 when it happens before initial anti-diabetes prescription. Comorbidities after initial anti-diabetes prescription are assigned to 0 even if they take place afterwards. Under this agreement, we delete those with comorbidities but without corresponding eventdate.

In a Cox model, stratification will offer the same number of hazard functions as the number of strata. Hazard ratios (e^{β}) will be optimized for each strata before fitting.

In this study, the model will output a hazard ratio for initial anti-diabetes prescription age in the presence of 10 hazards intrinsic to the levels of initial anti-diabetes prescription age. If initial anti-diabetes prescription age violats the proportional hazards assumption, generating strata may potentially make the proportional hazard assumption satisfied and make the estimates for initial anti-diabetes prescription age valid. The effect of initial anti-diabetes prescription age is not explicitly provided as a hazard ratio.

Both smoking and BMI (weight) are imported by 5-year latest record before initial anti-diabetic drug prescription. If BMI is not recorded in the database directly, we take the weight from $[-5, +1]$ with respect to initial anti-diabetic drug prescription year, and height is considered as a constant after age 30.

Those who are not in this range or missing from the database is regarded as missing. For HbA1c, the time window is 2 years since it might deviate from original level quite a bit. Also, as socioeconomic status is quite stable in UK, we take the latest IMD from the database.

Since there are only around 1000 observations when $BMI < 18.5$, we combine those BMI under 25 as one category instead of two.

Without stratification, initial anti-diabetes prescription age will output a hazard ratio for itself, assuming that the hazard for different levels of initial anti-diabetes prescription age are the same. Initial anti-diabetes prescription age is stratified into 10 layers, starting from 40 to over 85, with 5 years in each level.

Independent variable settings

Key components for a typical survival object are presented in Table 4.18. Origin time is when an individual begins at risk. For cancer incidence, origin time is 0 since patients become at risk of cancer when they are born. While for cancer mortality, patients become at risk of death after cancer diagnosis. Start date is when an individual enters survival study. Start date is initial anti-diabetes prescription date for both incidence and mortality. For incidence risks, we choose age as time scale considering $\mathbb{P}(cancer\ age = 50|prescription\ age = 40) \neq \mathbb{P}(cancer\ age = 60|prescription\ age = 50)$. However, average age for all-cause mortality after cancer diagnosis among diabetes patients is 11.69 years, $\mathbb{P}(death\ age = 50|cancer\ age = 40) \approx \mathbb{P}(death\ age = 60|cancer\ age = 50)$, age after cancer diagnosis should be chosen instead.

Table 4.18: Survival object settings

Case	Incidence	Mortality
Origin time	0	Cancer age
Start time	Prescription age	Prescription age
End time	$min(icd\_n, t_c, t_e, t_d) - YOB$	$min(icd\_n, t_e, t_d) - YOB$
Study population	DMT2 patients	DMT2 & cancer patients
Event of Interest	Cancer	Death

where icd_n is latest GP data upload date, t_c is cancer eventdate, t_e is data extraction date $(2018 - 06 - 01)$, t_d is death date, and YOB is year of birth.

Other settings

Inverse probability weights for missing adjustment are calculated after deleting incomplete cases. Propensity score weighting is calculated by specific models and schemes, accordingly. Only instantaneous risk ratio "hazard ratio" is considered thoughout this study.

Chapter 5

Results and Analysis

5.1 Sub-study 1: Cancer Incidence Risks

In this section, we compare cancer incidence risks between metformin and sulfonylureas under CCR, ATE, ATT and ATU schemes along with four distinct methods dealing with missing data, including adjusting for fewer variables, complete case analysis, treating missing as a separate category and inverse probability weighting. Tabulated in Table 5.1, 5.2, 5.3 and 5.4, we found out that both drugs reveal similar risks on cancer incidence since p-values exceed 0.05 and 95% CIs include 1 among 64 sub-analyses.

Table 5.1: Conventional Cox regression comparing cancer incidence risks between metformin and sulfonylureas

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	0.976 (0.944 − 1.009)	0.156	184,573	21,881
Fewer, ≥2000	0.979 (0.937 − 1.024)	0.362	137,261	14,092
Complete	1.021 (0.968 − 1.077)	0.437	99,620	10,658
Complete, ≥2000	1.035 (0.973 − 1.100)	0.274	91,297	9,110
Separate	1.002 (0.968 − 1.038)	0.898	184,573	21,881
Separate, ≥2000	1.010 (0.964 − 1.059)	0.665	137,261	14,092
IPW	1.028 (0.993 − 1.065)	0.121	99,620	10,658
IPW, ≥2000	1.038 (0.983 − 1.096)	0.180	91,297	9,110

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	1.025 (0.991 − 1.060)	0.156	184,573	21,881
Fewer, ≥2000	1.021 (0.976 − 1.068)	0.362	137,261	14,092
Complete	0.979 (0.928 − 1.033)	0.437	99,620	10,658
Complete, ≥2000	0.966 (0.909 − 1.027)	0.274	91,297	9,110
Separate	0.998 (0.964 − 1.033)	0.898	184,573	21,881
Separate, ≥2000	0.990 (0.945 − 1.037)	0.665	137,261	14,092
IPW	0.973 (0.939 − 1.007)	0.121	99,620	10,658
IPW, ≥2000	0.963 (0.913 − 1.017)	0.180	91,297	9,110

Table 5.2: Propensity score analysis comparing cancer incidence risks between metformin and sulfonylureas using ATE weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.019 (0.970 − 1.072)	0.449	184,573	21,881
Fewer, ≥2000	1.028 (0.948 − 1.115)	0.502	137,261	14,092
Complete	1.006 (0.886 − 1.141)	0.931	99,620	10,658
Complete, ≥2000	0.978 (0.851 − 1.124)	0.754	91,297	9,110
Separate	1.041 (0.983 − 1.103)	0.168	184,573	21,881
Separate, ≥2000	1.048 (0.955 − 1.151)	0.325	137,261	14,092
IPW	1.005 (0.909 − 1.111)	0.924	99,620	10,658
IPW, ≥2000	0.980 (0.860 − 1.118)	0.766	91,297	9,110
Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.981 (0.933 − 1.031)	0.449	184,573	21,881
Fewer, ≥2000	0.973 (0.897 − 1.055)	0.502	137,261	14,092
Complete	0.994 (0.877 − 1.128)	0.931	99,620	10,658
Complete, ≥2000	1.023 (0.890 − 1.175)	0.754	91,297	9,110
Separate	0.960 (0.907 − 1.017)	0.168	184,573	21,881
Separate, ≥2000	0.954 (0.869 − 1.048)	0.325	137,261	14,092
IPW	0.995 (0.900 − 1.101)	0.924	99,620	10,658
IPW, ≥2000	1.020 (0.895 − 1.163)	0.766	91,297	9,110

Table 5.3: Propensity score analysis comparing cancer incidence risks between metformin and sulfonylureas using ATT weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.007 (0.954 − 1.064)	0.791	184,573	21,881
Fewer, ≥2000	0.991 (0.947 − 1.038)	0.705	137,261	14,092
Complete	1.031 (0.959 − 1.110)	0.406	99,620	10,658
Complete, ≥2000	1.052 (0.985 − 1.124)	0.133	91,297	9,110
Separate	1.024 (0.963 − 1.089)	0.443	184,573	21,881
Separate, ≥2000	1.019 (0.968 − 1.074)	0.465	137,261	14,092
IPW	1.028 (0.906 − 1.167)	0.665	99,620	10,658
IPW, ≥2000	1.050 (0.981 − 1.123)	0.161	91,297	9,110

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.981 (0.917 − 1.050)	0.587	184,573	21,881
Fewer, ≥2000	0.963 (0.874 − 1.062)	0.452	137,261	14,092
Complete	1.005 (0.862 − 1.172)	0.949	99,620	10,658
Complete, ≥2000	1.037 (0.883 − 1.216)	0.660	91,297	9,110
Separate	0.961 (0.887 − 1.041)	0.326	184,573	21,881
Separate, ≥2000	0.948 (0.846 − 1.062)	0.354	137,261	14,092
IPW	1.008 (0.891 − 1.139)	0.903	99,620	10,658
IPW, ≥2000	1.034 (0.888 − 1.204)	0.665	91,297	9,110

Table 5.4: Propensity score analysis comparing cancer incidence risks between metformin and sulfonylureas using ATU weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.019 (0.952 − 1.091)	0.587	184,573	21,881
Fewer, ≥2000	1.038 (0.942 − 1.144)	0.452	137,261	14,092
Complete	0.995 (0.853 − 1.160)	0.949	99,620	10,658
Complete, ≥2000	0.965 (0.822 − 1.132)	0.66	91,297	9,110
Separate	1.041 (0.961 − 1.127)	0.326	184,573	21,881
Separate, ≥2000	1.055 (0.942 − 1.182)	0.354	137,261	14,092
IPW	0.992 (0.878 − 1.122)	0.903	99,620	10,658
IPW, ≥2000	0.967 (0.831 − 1.126)	0.665	91,297	9,110

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.993 (0.940 − 1.048)	0.791	184,573	21,881
Fewer, ≥2000	1.009 (0.964 − 1.056)	0.705	137,261	14,092
Complete	0.970 (0.901 − 1.043)	0.406	99,620	10,658
Complete, ≥2000	0.951 (0.890 − 1.016)	0.133	91,297	9,110
Separate	0.976 (0.918 − 1.038)	0.443	184,573	21,881
Separate, ≥2000	0.981 (0.931 − 1.033)	0.465	137,261	14,092
IPW	0.972 (0.857 − 1.104)	0.665	99,620	10,658
IPW, ≥2000	0.953 (0.890 − 1.019)	0.161	91,297	9,110

5.2 Sub-study 2: Cancer Mortality Risks

In this section, we compare cancer mortality risks between metformin and sulfonylureas under CCR, ATE, ATT and ATU schemes with four methods dealing with missing data accordingly. Tabulated in Table 5.5, 5.6, 5.7 and 5.8, we found out that metformin is protective over sulfonylureas on cancer mortality since p-values are highly significant and 95% CIs do not include 1 among 64 sub-analyses.

Table 5.5: Conventional Cox regression comparing cancer mortality risks between metformin and sulfonylureas

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.465 (1.376 − 1.561)	<0.001	18,041	5,897
Fewer, ≥2000	1.558 (1.452 − 1.673)	<0.001	15,594	4,285
Complete	1.376 (1.253 − 1.510)	<0.001	11,176	2,835
Complete, ≥2000	1.414 (1.283 − 1.559)	<0.001	10,697	2,549
Separate	1.385 (1.297 − 1.480)	<0.001	18,041	5,897
Separate, ≥2000	1.439 (1.337 − 1.550)	<0.001	15,594	4,285
IPW	1.325 (1.238 − 1.418)	<0.001	11,176	2,835
IPW, ≥2000	1.396 (1.276 − 1.526)	<0.001	10,697	2,549

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.682 (0.641 − 0.729)	<0.001	18,041	5,897
Fewer, ≥2000	0.642 (0.598 − 0.689)	<0.001	15,594	4,285
Complete	0.727 (0.662 − 0.798)	<0.001	11,176	2,835
Complete, ≥2000	0.707 (0.642 − 0.780)	<0.001	10,697	2,549
Separate	0.722 (0.676 − 0.771)	<0.001	18,041	5,897
Separate, ≥2000	0.695 (0.645 − 0.748)	<0.001	15,594	4,285
IPW	0.755 (0.705 − 0.808)	<0.001	11,176	2,835
IPW, ≥2000	0.717 (0.655 − 0.783)	<0.001	10,697	2,549

Table 5.6: Propensity score analysis comparing cancer incidence risks between metformin and sulfonylureas using ATE weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.673 (1.542 − 1.816)	<0.001	18,041	5,897
Fewer, ≥2000	1.897 (1.734 − 2.074)	<0.001	15,594	4,285
Complete	1.869 (1.541 − 2.266)	<0.001	11,176	2,835
Complete, ≥2000	1.873 (1.509 − 2.325)	<0.001	10,697	2,549
Separate	1.577 (1.433 − 1.736)	<0.001	18,041	5,897
Separate, ≥2000	1.800 (1.616 − 2.005)	<0.001	15,594	4,285
IPW	1.699 (1.446 − 1.997)	<0.001	11,176	2,835
IPW, ≥2000	1.835 (1.502 − 2.242)	<0.001	10,697	2,549

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.598 (0.551 − 0.649)	<0.001	18,041	5,897
Fewer, ≥2000	0.527 (0.482 − 0.577)	<0.001	15,594	4,285
Complete	0.535 (0.441 − 0.649)	<0.001	11,176	2,835
Complete, ≥2000	0.534 (0.430 − 0.663)	<0.001	10,697	2,549
Separate	0.634 (0.576 − 0.698)	<0.001	18,041	5,897
Separate, ≥2000	0.556 (0.499 − 0.619)	<0.001	15,594	4,285
IPW	0.588 (0.501 − 0.692)	<0.001	11,176	2,835
IPW, ≥2000	0.545 (0.446 − 0.666)	<0.001	10,697	2,549

Table 5.7: Propensity score analysis comparing cancer incidence risks between metformin and sulfonylureas using ATT weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.241 (1.127 − 1.366)	<0.001	18,041	5,897
Fewer, ≥2000	1.384 (1.291 − 1.484)	<0.001	15,594	4,285
Complete	1.223 (1.114 − 1.343)	<0.001	11,176	2,835
Complete, ≥2000	1.267 (1.148 − 1.397)	<0.001	10,697	2,549
Separate	1.172 (1.043 − 1.317)	0.008	18,041	5,897
Separate, ≥2000	1.264 (1.167 − 1.370)	<0.001	15,594	4,285
IPW	1.167 (1.049 − 1.298)	0.004	11,176	2,835
IPW, ≥2000	1.257 (1.137 − 1.389)	<0.001	10,697	2,549

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.496 (0.450 − 0.546)	<0.001	18,041	5,897
Fewer, ≥2000	0.474 (0.428 − 0.526)	<0.001	15,594	4,285
Complete	0.473 (0.378 − 0.593)	<0.001	11,176	2,835
Complete, ≥2000	0.482 (0.376 − 0.618)	<0.001	10,697	2,549
Separate	0.516 (0.461 − 0.577)	<0.001	18,041	5,897
Separate, ≥2000	0.489 (0.429 − 0.556)	<0.001	15,594	4,285
IPW	0.517 (0.425 − 0.630)	<0.001	11,176	2,835
IPW, ≥2000	0.492 (0.390 − 0.622)	<0.001	10,697	2,549

Table 5.8: Propensity score analysis comparing cancer incidence risks between metformin and sulfonylureas using ATU weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	2.017 (1.831 − 2.222)	<0.001	18,041	5,897
Fewer, ≥2000	2.109 (1.901 − 2.339)	<0.001	15,594	4,285
Complete	2.113 (1.686 − 2.648)	<0.001	11,176	2,835
Complete, ≥2000	2.073 (1.617 − 2.658)	<0.001	10,697	2,549
Separate	1.939 (1.733 − 2.170)	<0.001	18,041	5,897
Separate, ≥2000	2.047 (1.798 − 2.330)	<0.001	15,594	4,285
IPW	1.934 (1.588 − 2.355)	<0.001	11,176	2,835
IPW, ≥2000	2.031 (1.609 − 2.563)	<0.001	10,697	2,549

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.806 (0.732 − 0.887)	<0.001	18,041	5,897
Fewer, ≥2000	0.723 (0.674 − 0.775)	<0.001	15,594	4,285
Complete	0.818 (0.745 − 0.898)	<0.001	11,176	2,835
Complete, ≥2000	0.789 (0.716 − 0.871)	<0.001	10,697	2,549
Separate	0.853 (0.759 − 0.959)	0.008	18,041	5,897
Separate, ≥2000	0.791 (0.730 − 0.857)	<0.001	15,594	4,285
IPW	0.857 (0.771 − 0.953)	0.004	11,176	2,835
IPW, ≥2000	0.796 (0.720 − 0.879)	<0.001	10,697	2,549

5.3 Sub-study 3: Breast Cancer Incidence Risks

In this section, we compare breast cancer incidence risks between metformin and sulfonylureas under CCR, ATE, ATT and ATU schemes with four methods dealing with missing data accordingly. Tabulated in Table 5.9, 5.10, 5.11 and 5.12, we found out that both drugs reveal similar risks on breast cancer incidence since p-values exceed 0.05 and 95% CIs include 1 among 64 sub-analyses.

Table 5.9: Conventional Cox regression comparing breast cancer incidence risks between metformin and sulfonylureas

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	0.899 (0.797 − 1.013)	0.081	87,335	1,845
Fewer, ≥2000	0.899 (0.767 − 1.053)	0.186	64,375	1,279
Complete	0.977 (0.808 − 1.182)	0.814	46,130	940
Complete, ≥2000	1.013 (0.816 − 1.259)	0.905	42,291	818
Separate	0.950 (0.839 − 1.075)	0.412	87,335	1,845
Separate, ≥2000	0.952 (0.807 − 1.123)	0.561	64,375	1,279
IPW	1.072 (0.949 − 1.210)	0.266	46,130	940
IPW, ≥2000	1.033 (0.853 − 1.251)	0.738	42,291	818

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	1.113 (0.987 − 1.255)	0.081	87,335	1,845
Fewer, ≥2000	1.113 (0.950 − 1.304)	0.186	64,375	1,279
Complete	1.023 (0.846 − 1.237)	0.814	46,130	940
Complete, ≥2000	0.987 (0.795 − 1.226)	0.905	42,291	818
Separate	1.053 (0.931 − 1.191)	0.412	87,335	1,845
Separate, ≥2000	1.050 (0.890 − 1.239)	0.561	64,375	1,279
IPW	0.933 (0.826 − 1.054)	0.266	46,130	940
IPW, ≥2000	0.968 (0.799 − 1.172)	0.738	42,291	818

Table 5.10: Propensity score analysis comparing breast cancer incidence risks between metformin and sulfonylureas using ATE weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	0.926 (0.804 − 1.066)	0.286	87,335	1,845
Fewer, ≥2000	0.824 (0.667 − 1.017)	0.072	64,375	1,279
Complete	1.213 (0.794 − 1.852)	0.372	46,130	940
Complete, ≥2000	1.128 (0.720 − 1.768)	0.598	42,291	818
Separate	1.056 (0.888 − 1.257)	0.537	87,335	1,845
Separate, ≥2000	1.023 (0.788 − 1.327)	0.865	64,375	1,279
IPW	1.248 (0.909 − 1.715)	0.171	46,130	940
IPW, ≥2000	1.125 (0.737 − 1.716)	0.585	42,291	818
Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	1.080 (0.938 − 1.243)	0.286	87,335	1,845
Fewer, ≥2000	1.214 (0.983 − 1.498)	0.072	64,375	1,279
Complete	0.824 (0.540 − 1.259)	0.372	46,130	940
Complete, ≥2000	0.886 (0.566 − 1.389)	0.598	42,291	818
Separate	0.947 (0.796 − 1.126)	0.537	87,335	1,845
Separate, ≥2000	0.978 (0.753 − 1.269)	0.865	64,375	1,279
IPW	0.801 (0.583 − 1.101)	0.171	46,130	940
IPW, ≥2000	0.889 (0.583 − 1.356)	0.585	42,291	818

Table 5.11: Propensity score analysis comparing breast cancer incidence risks between metformin and sulfonylureas using ATT weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.093 (0.926 − 1.290)	0.293	87,335	1,845
Fewer, ≥2000	0.961 (0.817 − 1.131)	0.633	64,375	1,279
Complete	1.082 (0.877 − 1.335)	0.464	46,130	940
Complete, ≥2000	1.128 (0.700 − 1.415)	0.297	42,291	818
Separate	1.115 (0.922 − 1.349)	0.260	87,335	1,845
Separate, ≥2000	0.997 (0.835 − 1.191)	0.974	64,375	1,279
IPW	1.242 (0.945 − 1.631)	0.120	46,130	940
IPW, ≥2000	1.150 (0.913 − 1.447)	0.235	42,291	818

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	1.198 (1.002 − 1.433)	0.047	87,335	1,845
Fewer, ≥2000	1.262 (0.990 − 1.610)	0.061	64,375	1,279
Complete	0.815 (0.504 − 1.319)	0.405	46,130	940
Complete, ≥2000	0.890 (0.543 − 1.458)	0.644	42,291	818
Separate	0.985 (0.786 − 1.234)	0.894	87,335	1,845
Separate, ≥2000	0.971 (0.719 − 1.312)	0.850	64,375	1,279
IPW	0.806 (0.553 − 1.176)	0.263	46,130	940
IPW, ≥2000	0.895 (0.559 − 1.430)	0.642	42,291	818

Table 5.12: Propensity score analysis comparing breast cancer incidence risks between metformin and sulfonylureas using ATU weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	0.835 (0.698 − 0.998)	0.047	87,335	1,845
Fewer, ≥2000	0.792 (0.621 − 1.010)	0.061	64,375	1,279
Complete	1.227 (0.758 − 1.985)	0.405	46,130	940
Complete, ≥2000	1.124 (0.686 − 1.841)	0.644	42,291	818
Separate	1.015 (0.810 − 1.272)	0.894	87,335	1,845
Separate, ≥2000	1.029 (0.762 − 1.390)	0.850	64,375	1,279
IPW	1.240 (0.851 − 1.809)	0.263	46,130	940
IPW, ≥2000	1.118 (0.699 − 1.788)	0.642	42,291	818
Reference (sulfonylureas)	**HR (95% CI)**	**p-value**	**N**	**I**
Fewer	0.915 (0.775 − 1.080)	0.293	87,335	1,845
Fewer, ≥2000	1.040 (0.884 − 1.224)	0.633	64,375	1,279
Complete	0.924 (0.749 − 1.141)	0.464	46,130	940
Complete, ≥2000	0.887 (0.707 − 1.112)	0.297	42,291	818
Separate	0.897 (0.742 − 1.084)	0.260	87,335	1,845
Separate, ≥2000	1.003 (0.840 − 1.198)	0.974	64,375	1,279
IPW	0.805 (0.613 − 1.058)	0.120	46,130	940
IPW, ≥2000	0.870 (0.691 − 1.095)	0.235	42,291	818

5.4 Sub-study 4: Breast Cancer Mortality Risks

In this section, we compare breast cancer mortality risks between metformin and sulfonylureas under CCR, ATE, ATT and ATU schemes with four methods dealing with missing data accordingly. Tabulated in Table 5.13, 5.14, 5.15 and 5.16, we found out that metformin is protective over sulfonylureas on breast cancer mortality since p-values are significant and 95% CIs do not include 1 among most sub-analyses.

Table 5.13: Conventional Cox regression comparing breast cancer mortality risks between metformin and sulfonylureas

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.542 (1.280 − 1.858)	<0.001	2,669	709
Fewer, ≥2000	1.661 (1.337 − 2.064)	<0.001	2,341	508
Complete	1.455 (1.072 − 1.977)	0.016	1,677	315
Complete, ≥2000	1.539 (1.111 − 2.131)	0.009	1,616	280
Separate	1.581 (1.297 − 1.927)	<0.001	2,669	709
Separate, ≥2000	1.587 (1.254 − 2.008)	<0.001	2,341	508
IPW	1.374 (1.102 − 1.714)	0.005	1,677	315
IPW, ≥2000	1.500 (1.120 − 2.010)	0.007	1,616	280

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.648 (0.538 − 0.781)	<0.001	2,669	709
Fewer, ≥2000	0.602 (0.484 − 0.748)	<0.001	2,341	508
Complete	0.687 (0.506 − 0.933)	0.016	1,677	315
Complete, ≥2000	0.650 (0.469 − 0.900)	0.009	1,616	280
Separate	0.633 (0.519 − 0.771)	<0.001	2,669	709
Separate, ≥2000	0.630 (0.498 − 0.798)	<0.001	2,341	508
IPW	0.728 (0.584 − 0.908)	0.005	1,677	315
IPW, ≥2000	0.667 (0.497 − 0.893)	0.007	1,616	280

Table 5.14: Propensity score analysis comparing breast cancer mortality risks between metformin and sulfonylureas using ATE weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.987 (1.504 − 2.625)	<0.001	2,669	709
Fewer, ≥2000	2.328 (1.697 − 3.193)	<0.001	2,341	508
Complete	3.239 (1.942 − 5.402)	<0.001	1,677	315
Complete, ≥2000	3.550 (2.163 − 5.827)	<0.001	1,616	280
Separate	2.109 (1.486 − 2.995)	<0.001	2,669	709
Separate, ≥2000	2.376 (1.552 − 3.638)	<0.001	2,341	508
IPW	2.757 (1.641 − 4.631)	<0.001	1,677	315
IPW, ≥2000	3.329 (2.017 − 5.492)	<0.001	1,616	280
Reference (sulfonylureas)	**HR (95% CI)**	**p-value**	**N**	**I**
Fewer	0.503 (0.381 − 0.665)	<0.001	2,669	709
Fewer, ≥2000	0.430 (0.313 − 0.589)	<0.001	2,341	508
Complete	0.309 (0.185 − 0.515)	<0.001	1,677	315
Complete, ≥2000	0.282 (0.172 − 0.462)	<0.001	1,616	280
Separate	0.474 (0.334 − 0.673)	<0.001	2,669	709
Separate, ≥2000	0.421 (0.275 − 0.644)	<0.001	2,341	508
IPW	0.363 (0.216 − 0.609)	<0.001	1,677	315
IPW, ≥2000	0.300 (0.182 − 0.496)	<0.001	1,616	280

Table 5.15: Propensity score analysis comparing breast cancer mortality risks between metformin and sulfonylureas using ATT weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.283 (1.023 − 1.610)	0.031	2,669	709
Fewer, ≥2000	1.425 (1.147 − 1.771)	0.001	2,341	508
Complete	1.377 (1.012 − 1.873)	0.042	1,677	315
Complete, ≥2000	1.458 (1.054 − 2.016)	0.023	1,616	280
Separate	1.406 (1.052 − 1.879)	0.021	2,669	709
Separate, ≥2000	1.472 (1.134 − 1.909)	0.003	2,341	508
IPW	1.338 (0.958 − 1.867)	0.087	1,677	315
IPW, ≥2000	1.436 (1.033 − 1.996)	0.031	1,616	280

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.406 (0.294 − 0.562)	<0.001	2,669	709
Fewer, ≥2000	0.378 (0.267 − 0.535)	<0.001	2,341	508
Complete	0.268 (0.162 − 0.445)	<0.001	1,677	315
Complete, ≥2000	0.250 (0.154 − 0.405)	<0.001	1,616	280
Separate	0.388 (0.258 − 0.582)	<0.001	2,669	709
Separate, ≥2000	0.375 (0.234 − 0.602)	<0.001	2,341	508
IPW	0.311 (0.183 − 0.529)	<0.001	1,677	315
IPW, ≥2000	0.266 (0.162 − 0.435)	<0.001	1,616	280

Table 5.16: Propensity score analysis comparing breast cancer mortality risks between metformin and sulfonylureas using ATU weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	2.461 (1.780 − 3.401)	<0.001	2,669	709
Fewer, ≥2000	2.645 (1.869 − 3.744)	<0.001	2,341	508
Complete	3.726 (2.249 − 6.173)	<0.001	1,677	315
Complete, ≥2000	4.000 (2.469 − 6.480)	<0.001	1,616	280
Separate	2.581 (1.719 − 3.876)	<0.001	2,669	709
Separate, ≥2000	2.667 (1.662 − 4.280)	<0.001	2,341	508
IPW	3.211 (1.891 − 5.452)	<0.001	1,677	315
IPW, ≥2000	3.763 (2.297 − 6.164)	<0.001	1,616	280

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.779 (0.621 − 0.977)	0.031	2,669	709
Fewer, ≥2000	0.702 (0.565 − 0.872)	0.001	2,341	508
Complete	0.726 (0.534 − 0.988)	0.042	1,677	315
Complete, ≥2000	0.686 (0.496 − 0.949)	0.023	1,616	280
Separate	0.711 (0.532 − 0.951)	0.021	2,669	709
Separate, ≥2000	0.680 (0.524 − 0.882)	0.004	2,341	508
IPW	0.748 (0.536 − 1.043)	0.087	1,677	315
IPW, ≥2000	0.696 (0.501 − 0.968)	0.031	1,616	280

5.5 Sub-study 5: Prostate Cancer Incidence Risks

In this section, we compare prostate cancer incidence risks between metformin and sulfonylureas under CCR, ATE, ATT and ATU schemes with four methods dealing with missing data accordingly. Tabulated in Table 5.17, 5.18, 5.19 and 5.20, we found out that sulfonylureas reveals protective effect on prostate cancer incidence in around half of the cases, especially under the ATE weighting scheme. The other half revealed similar risks on prostate cancer incidence risks. Since the number of participants and events of interest are not sufficient to reach significance level, we can't tell whether sulfonylureas is more protective or not.

Table 5.17: Conventional Cox regression comparing prostate cancer incidence risks between metformin and sulfonylureas

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	0.917 (0.833 − 1.009)	0.077	111,562	2,733
Fewer, ≥2000	0.913 (0.807 − 1.032)	0.146	85,091	1,840
Complete	0.919 (0.794 − 1.063)	0.253	62,194	1,419
Complete, ≥2000	0.897 (0.758 − 1.061)	0.204	57,273	1,215
Separate	0.921 (0.834 − 1.016)	0.100	111,562	2,733
Separate, ≥2000	0.922 (0.811 − 1.047)	0.211	85,091	1,840
IPW	0.948 (0.859 − 1.046)	0.284	62,194	1,419
IPW, ≥2000	0.895 (0.770 − 1.040)	0.147	57,273	1,215

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	1.091 (0.991 − 1.201)	0.077	111,562	2,733
Fewer, ≥2000	1.096 (0.969 − 1.239)	0.146	85,091	1,840
Complete	1.089 (0.941 − 1.259)	0.253	62,194	1,419
Complete, ≥2000	1.115 (0.943 − 1.320)	0.204	57,273	1,215
Separate	1.086 (0.984 − 1.199)	0.100	111,562	2,733
Separate, ≥2000	1.085 (0.955 − 1.233)	0.211	85,091	1,840
IPW	1.055 (0.957 − 1.164)	0.284	62,194	1,419
IPW, ≥2000	1.118 (0.962 − 1.299)	0.147	57,273	1,215

Table 5.18: Propensity score analysis comparing prostate cancer incidence risks between metformin and sulfonylureas using ATE weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	0.923 (0.794 − 1.073)	0.296	111,562	2,733
Fewer, ≥2000	0.902 (0.707 − 1.151)	0.408	85,091	1,840
Complete	0.648 (0.526 − 0.797)	<0.001	62,194	1,419
Complete, ≥2000	0.610 (0.469 − 0.794)	<0.001	57,273	1,215
Separate	0.867 (0.757 − 0.994)	0.041	111,562	2,733
Separate, ≥2000	0.798 (0.654 − 0.974)	0.027	85,091	1,840
IPW	0.759 (0.632 − 0.910)	0.003	62,194	1,419
IPW, ≥2000	0.631 (0.485 − 0.822)	<0.001	57,273	1,215

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	1.083 (0.932 − 1.259)	0.296	111,562	2,733
Fewer, ≥2000	1.108 (0.869 − 1.414)	0.408	85,091	1,840
Complete	1.544 (1.255 − 1.900)	<0.001	62,194	1,419
Complete, ≥2000	1.639 (1.259 − 2.134)	<0.001	57,273	1,215
Separate	1.153 (1.006 − 1.322)	0.041	111,562	2,733
Separate, ≥2000	1.253 (1.026 − 1.529)	0.027	85,091	1,840
IPW	1.318 (1.098 − 1.582)	0.003	62,194	1,419
IPW, ≥2000	1.585 (1.217 − 2.063)	<0.001	57,273	1,215

Table 5.19: Propensity score analysis comparing prostate cancer incidence risks between metformin and sulfonylureas using ATT weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	0.975 (0.822 − 1.157)	0.774	111,562	2,733
Fewer, ≥2000	0.962 (0.848 − 1.091)	0.545	85,091	1,840
Complete	1.078 (0.911 − 1.275)	0.382	62,194	1,419
Complete, ≥2000	1.006 (0.846 − 1.198)	0.943	57,273	1,215
Separate	0.994 (0.830 − 1.191)	0.948	111,562	2,733
Separate, ≥2000	0.969 (0.842 − 1.115)	0.662	85,091	1,840
IPW	1.165 (0.934 − 1.453)	0.176	62,194	1,419
IPW, ≥2000	1.001 (0.839 − 1.194)	0.992	57,273	1,215

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	1.145 (0.925 − 1.417)	0.214	111,562	2,733
Fewer, ≥2000	1.131 (0.832 − 1.538)	0.432	85,091	1,840
Complete	1.861 (1.418 − 2.441)	<0.001	62,194	1,419
Complete, ≥2000	1.855 (1.346 − 2.557)	<0.001	57,273	1,215
Separate	1.307 (1.091 ⁻ 1.566)	0.004	111,562	2,733
Separate, ≥2000	1.338 (1.039 − 1.723)	0.024	85,091	1,840
IPW	1.634 (1.290 − 2.069)	<0.001	62,194	1,419
IPW, ≥2000	1.789 (1.292 − 2.476)	<0.001	57,273	1,215

Table 5.20: Propensity score analysis comparing prostate cancer incidence risks between metformin and sulfonylureas using ATU weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	0.874 (0.706 − 1.081)	0.214	111,562	2,733
Fewer, ≥2000	0.884 (0.650 − 1.202)	0.432	85,091	1,840
Complete	0.537 (0.410 − 0.705)	<0.001	62,194	1,419
Complete, ≥2000	0.539 (0.391 − 0.743)	<0.001	57,273	1,215
Separate	0.765 (0.639 − 0.916)	0.004	111,562	2,733
Separate, ≥2000	0.747 (0.580 − 0.963)	0.024	85,091	1,840
IPW	0.612 (0.483 − 0.775)	<0.001	62,194	1,419
IPW, ≥2000	0.559 (0.404 − 0.774)	<0.001	57,273	1,215

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	1.025 (0.864 − 1.217)	0.774	111,562	2,733
Fewer, ≥2000	1.040 (0.917 − 1.179)	0.545	85,091	1,840
Complete	0.928 (0.784 − 1.098)	0.382	62,194	1,419
Complete, ≥2000	0.994 (0.835 − 1.183)	0.943	57,273	1,215
Separate	1.006 (0.840 − 1.206)	0.948	111,562	2,733
Separate, ≥2000	1.032 (0.897 − 1.188)	0.662	85,091	1,840
IPW	0.859 (0.688 − 1.071)	0.176	62,194	1,419
IPW, ≥2000	0.999 (0.837 − 1.192)	0.992	57,273	1,215

5.6 Sub-study 6: Prostate Cancer Mortality Risks

In this section, we compare prostate cancer mortality risks between metformin and sulfonylureas under CCR, ATE, ATT and ATU schemes with four methods dealing with missing data accordingly. Tabulated in Table 5.21, 5.22, 5.23 and 5.24, we found out that metformin is protective over sulfonylureas on prostate cancer mortality risks in most circumstances. ATT and ATU weighting schemes tend not to offer stable hazard ratio if treatment and control are inverted. This implies that the ATT and ATU weighting schemes are not suitable for synthetic randomized controlled trials

Table 5.21: Conventional Cox regression comparing prostate cancer mortality risks between metformin and sulfonylureas

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.389 (1.163 − 1.659)	<0.001	2,088	775
Fewer, ≥2000	1.459 (1.206 − 1.765)	<0.001	1,870	608
Complete	1.258 (0.982 − 1.610)	0.069	1,378	424
Complete, ≥2000	1.292 (0.997 − 1.673)	0.052	1,340	401
Separate	1.379 (1.143 − 1.664)	<0.001	2,088	775
Separate, ≥2000	1.446 (1.181 − 1.769)	<0.001	1,870	608
IPW	1.245 (1.035 − 1.498)	0.020	1,378	424
IPW, ≥2000	1.284 (1.009 − 1.633)	0.042	1,340	401

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.720 (0.603 − 0.860)	<0.001	2,088	775
Fewer, ≥2000	0.685 (0.566 − 0.829)	<0.001	1,870	608
Complete	0.795 (0.621 − 1.018)	0.069	1,378	424
Complete, ≥2000	0.774 (0.598 − 1.003)	0.052	1,340	401
Separate	0.725 (0.601 − 0.875)	<0.001	2,088	775
Separate, ≥2000	0.692 (0.565 − 0.847)	<0.001	1,870	608
IPW	0.803 (0.668 − 0.966)	0.020	1,378	424
IPW, ≥2000	0.779 (0.612 − 0.991)	0.042	1,340	401

Table 5.22: Propensity score analysis comparing prostate cancer mortality risks between metformin and sulfonylureas using ATE weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.573 (1.279 − 1.936)	<0.001	2,088	775
Fewer, ≥2000	1.754 (1.405 − 2.190)	<0.001	1,870	608
Complete	2.092 (1.532 − 2.856)	<0.001	1,378	424
Complete, ≥2000	2.151 (1.518 − 3.048)	<0.001	1,340	401
Separate	1.609 (1.276 − 2.030)	<0.001	2,088	775
Separate, ≥2000	1.784 (1.387 − 2.294)	<0.001	1,870	608
IPW	1.832 (1.359 − 2.469)	<0.001	1,378	424
IPW, ≥2000	2.097 (1.491 − 2.950)	<0.001	1,340	401

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.636 (0.517 − 0.782)	<0.001	2,088	775
Fewer, ≥2000	0.570 (0.457 − 0.712)	<0.001	1,870	608
Complete	0.478 (0.350 − 0.653)	<0.001	1,378	424
Complete, ≥2000	0.465 (0.328 − 0.659)	<0.001	1,340	401
Separate	0.621 (0.493 − 0.784)	<0.001	2,088	775
Separate, ≥2000	0.561 (0.436 − 0.721)	<0.001	1,870	608
IPW	0.546 (0.405 − 0.736)	<0.001	1,378	424
IPW, ≥2000	0.477 (0.339 − 0.671)	<0.001	1,340	401

Table 5.23: Propensity score analysis comparing prostate cancer mortality risks between metformin and sulfonylureas using ATT weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.162 (0.944 − 1.432)	0.157	2,088	775
Fewer, ≥2000	1.334 (1.098 − 1.622)	0.004	1,870	608
Complete	1.072 (0.841 − 1.366)	0.574	1,378	424
Complete, ≥2000	1.132 (0.885 − 1.447)	0.325	1,340	401
Separate	1.123 (0.871 − 1.449)	0.371	2,088	775
Separate, ≥2000	1.257 (1.021 − 1.546)	0.031	1,870	608
IPW	1.010 (0.769 − 1.328)	0.941	1,378	424
IPW, ≥2000	1.127 (0.879 − 1.445)	0.346	1,340	401

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.533 (0.417 − 0.682)	<0.001	2,088	775
Fewer, ≥2000	0.515 (0.400 − 0.663)	<0.001	1,870	608
Complete	0.387 (0.276 − 0.543)	<0.001	1,378	424
Complete, ≥2000	0.388 (0.268 − 0.564)	<0.001	1,340	401
Separate	0.495 (0.378 − 0.650)	<0.001	2,088	775
Separate, ≥2000	0.485 (0.361 − 0.653)	<0.001	1,870	608
IPW	0.435 (0.313 − 0.605)	<0.001	1,378	424
IPW, ≥2000	0.398 (0.275 − 0.574)	<0.001	1,340	401

Table 5.24: Propensity score analysis comparing prostate cancer mortality risks between metformin and sulfonylureas using ATU weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.876 (1.466 − 2.400)	<0.001	2,088	775
Fewer, ≥2000	1.941 (1.507 − 2.498)	<0.001	1,870	608
Complete	2.582 (1.841 − 3.620)	<0.001	1,378	424
Complete, ≥2000	2.576 (1.775 − 3.739)	<0.001	1,340	401
Separate	2.019 (1.538 − 2.649)	<0.001	2,088	775
Separate, ≥2000	2.061 (1.532 − 2.773)	<0.001	1,870	608
IPW	2.299 (1.653 − 3.198)	<0.001	1,378	424
IPW, ≥2000	2.515 (1.742 − 3.632)	<0.001	1,340	401

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.860 (0.699 − 1.060)	0.157	2,088	775
Fewer, ≥2000	0.749 (0.617 − 0.911)	0.004	1,870	608
Complete	0.933 (0.732 − 1.189)	0.574	1,378	424
Complete, ≥2000	0.884 (0.691 − 1.130)	0.325	1,340	401
Separate	0.890 (0.690 − 1.149)	0.371	2,088	775
Separate, ≥2000	0.798 (0.647 − 0.979)	0.031	1,870	608
IPW	0.990 (0.753 − 1.300)	0.941	1,378	424
IPW, ≥2000	0.887 (0.692 − 1.138)	0.346	1,340	401

5.7 Sub-study 7: Bowel Cancer Incidence Risks

In this section, we compare bowel cancer incidence risks between metformin and sulfonylureas under CCR, ATE, ATT and ATU schemes with four methods dealing with missing data accordingly. Tabulated in Table 5.25, 5.26, 5.27 and 5.28, we found out that both drugs reveal similar risks on bowel cancer incidence since p-values exceed 0.05 and 95% CIs include 1 among 64 sub-analyses.

Table 5.25: Conventional Cox regression comparing bowel cancer incidence risks between metformin and sulfonylureas

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	0.927 (0.844 − 1.019)	0.115	202,087	2,753
Fewer, ≥2000	0.944 (0.833 − 1.070)	0.368	152,379	1,725
Complete	1.013 (0.870 − 1.179)	0.871	110,443	1,254
Complete, ≥2000	1.054 (0.886 − 1.255)	0.550	101,641	1,064
Separate	0.950 (0.862 − 1.047)	0.302	202,087	2,753
Separate, ≥2000	0.969 (0.850 − 1.105)	0.637	152,379	1,725
IPW	1.027 (0.928 − 1.136)	0.606	110,443	1,254
IPW, ≥2000	1.076 (0.924 − 1.254)	0.345	101,641	1,064

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	1.079 (0.982 − 1.185)	0.115	202,087	2,753
Fewer, ≥2000	1.059 (0.934 − 1.201)	0.368	152,379	1,725
Complete	0.987 (0.848 − 1.150)	0.871	110,443	1,254
Complete, ≥2000	0.948 (0.797 − 1.128)	0.550	101,641	1,064
Separate	1.053 (0.955 − 1.160)	0.302	202,087	2,753
Separate, ≥2000	1.032 (0.905 − 1.177)	0.637	152,379	1,725
IPW	0.974 (0.880 − 1.077)	0.606	110,443	1,254
IPW, ≥2000	0.929 (0.797 − 1.082)	0.345	101,641	1,064

Table 5.26: Propensity score analysis comparing bowel cancer incidence risks between metformin and sulfonylureas using ATE weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	0.942 (0.808 − 1.098)	0.443	202,087	2,753
Fewer, ≥2000	0.987 (0.773 − 1.261)	0.916	152,379	1,725
Complete	1.090 (0.721 − 1.647)	0.683	110,443	1,254
Complete, ≥2000	1.108 (0.710 − 1.729)	0.652	101,641	1,064
Separate	0.928 (0.801 − 1.075)	0.317	202,087	2,753
Separate, ≥2000	0.945 (0.773 − 1.155)	0.582	152,379	1,725
IPW	1.077 (0.774 − 1.500)	0.660	110,443	1,254
IPW, ≥2000	1.141 (0.739 − 1.760)	0.552	101,641	1,064

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	1.062 (0.911 − 1.238)	0.443	202,087	2,753
Fewer, ≥2000	1.013 (0.793 − 1.295)	0.916	152,379	1,725
Complete	0.918 (0.607 − 1.387)	0.683	110,443	1,254
Complete, ≥2000	0.903 (0.579 − 1.409)	0.652	101,641	1,064
Separate	1.078 (0.930 − 1.249)	0.317	202,087	2,753
Separate, ≥2000	1.058 (0.866 − 1.293)	0.582	152,379	1,725
IPW	0.928 (0.667 − 1.293)	0.660	110,443	1,254
IPW, ≥2000	0.877 (0.568 − 1.353)	0.552	101,641	1,064

Table 5.27: Propensity score analysis comparing bowel cancer incidence risks between metformin and sulfonylureas using ATT weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	0.883 (0.763 − 1.021)	0.093	202,087	2,753
Fewer, ≥2000	0.952 (0.837 − 1.082)	0.451	152,379	1,725
Complete	1.026 (0.841 − 1.251)	0.804	110,443	1,254
Complete, ≥2000	1.053 (0.864 − 1.283)	0.607	101,641	1,064
Separate	0.881 (0.731 − 1.061)	0.182	202,087	2,753
Separate, ≥2000	0.970 (0.824 − 1.141)	0.711	152,379	1,725
IPW	1.056 (0.808 − 1.379)	0.691	110,443	1,254
IPW, ≥2000	1.076 (0.879 − 1.317)	0.479	101,641	1,064

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	1.007 (0.803 − 1.262)	0.955	202,087	2,753
Fewer, ≥2000	1.003 (0.739 − 1.361)	0.985	152,379	1,725
Complete	0.902 (0.543 − 1.499)	0.690	110,443	1,254
Complete, ≥2000	0.894 (0.535 − 1.493)	0.668	101,641	1,064
Separate	1.032 (0.861 − 1.238)	0.733	202,087	2,753
Separate, ≥2000	1.068 (0.837 − 1.361)	0.598	152,379	1,725
IPW	0.921 (0.596 − 1.424)	0.713	110,443	1,254
IPW, ≥2000	0.867 (0.522 − 1.437)	0.578	101,641	1,064

Table 5.28: Propensity score analysis comparing bowel cancer incidence risks between metformin and sulfonylureas using ATU weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	0.993 (0.793 − 1.245)	0.955	202,087	2,753
Fewer, $\geq$2000	0.997 (0.735 − 1.353)	0.985	152,379	1,725
Complete	1.109 (0.667 − 1.843)	0.690	110,443	1,254
Complete, $\geq$2000	1.119 (0.670 − 1.869)	0.668	101,641	1,064
Separate	0.969 (0.808 − 1.162)	0.733	202,087	2,753
Separate, $\geq$2000	0.937 (0.735 − 1.194)	0.598	152,379	1,725
IPW	1.085 (0.702 − 1.677)	0.713	110,443	1,254
IPW, $\geq$2000	1.154 (0.696 − 1.915)	0.578	101,641	1,064

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	1.133 (0.980 − 1.310)	0.093	202,087	2,753
Fewer, $\geq$2000	1.051 (0.924 − 1.194)	0.451	152,379	1,725
Complete	0.975 (0.780 − 1.189)	0.804	110,443	1,254
Complete, $\geq$2000	0.949 (0.779 − 1.157)	0.607	101,641	1,064
Separate	1.135 (0.942 − 1.368)	0.182	202,087	2,753
Separate, $\geq$2000	1.031 (0.876 − 1.214)	0.711	152,379	1,725
IPW	0.947 (0.725 − 1.237)	0.691	110,443	1,254
IPW, $\geq$2000	0.930 (0.759 − 1.138)	0.479	101,641	1,064

5.8 Sub-study 8: Bowel Cancer Mortality Risks

In this section, we compare bowel cancer mortality risks between metformin and sulfonylureas under CCR, ATE, ATT and ATU schemes with four methods dealing with missing data accordingly. Tabulated in Table 5.29, 5.30, 5.31 and 5.32, we found out that metformin is protective over metformin on bowel cancer mortality risks in most circumstances. ATT and ATU weighting schemes tend not to offer stable hazard ratio if treatment and control are inverted. This implies that the ATT and ATU weighting schemes are not suitable for synthetic randomized controlled trials.

Table 5.29: Conventional Cox regression comparing bowel cancer mortality risks between metformin and sulfonylureas

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.497 (1.244 − 1.802)	<0.001	1,786	697
Fewer, ≥2000	1.500 (1.213 − 1.854)	<0.001	1,491	491
Complete	1.503 (1.128 − 2.004)	0.005	1,051	316
Complete, ≥2000	1.516 (1.120 − 2.052)	0.007	992	277
Separate	1.511 (1.241 − 1.839)	<0.001	1,786	697
Separate, ≥2000	1.492 (1.193 − 1.866)	<0.001	1,491	491
IPW	1.577 (1.276 − 1.948)	<0.001	1,051	316
IPW, ≥2000	1.506 (1.139 − 1.991)	0.004	992	277

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.668 (0.555 − 0.804)	<0.001	1,786	697
Fewer, ≥2000	0.667 (0.540 − 0.824)	<0.001	1,491	491
Complete	0.665 (0.499 − 0.887)	0.005	1,051	316
Complete, ≥2000	0.660 (0.487 − 0.893)	0.007	992	277
Separate	0.662 (0.544 − 0.806)	<0.001	1,786	697
Separate, ≥2000	0.670 (0.536 − 0.838)	<0.001	1,491	491
IPW	0.634 (0.513 − 0.784)	<0.001	1,051	316
IPW, ≥2000	0.664 (0.502 − 0.878)	0.004	992	277

Table 5.30: Propensity score analysis comparing bowel cancer mortality risks between metformin and sulfonylureas using ATE weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.582 (1.269 − 1.974)	<0.001	1,786	697
Fewer, ≥2000	1.619 (1.297 − 2.021)	<0.001	1,491	491
Complete	1.365 (0.973 − 1.915)	0.072	1,051	316
Complete, ≥2000	1.400 (0.988 − 1.984)	0.059	992	277
Separate	1.398 (1.073 − 1.822)	0.013	1,786	697
Separate, ≥2000	1.551 (1.232 − 1.952)	<0.001	1,491	491
IPW	1.372 (1.022 − 1.841)	0.035	1,051	316
IPW, ≥2000	1.421 (1.015 − 1.991)	0.041	992	277

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.632 (0.507 − 0.788)	<0.001	1,786	697
Fewer, ≥2000	0.618 (0.495 − 0.771)	<0.001	1,491	491
Complete	0.733 (0.522 − 1.028)	0.072	1,051	316
Complete, ≥2000	0.714 (0.504 − 1.012)	0.059	992	277
Separate	0.715 (0.549 − 0.932)	0.013	1,786	697
Separate, ≥2000	0.645 (0.512 − 0.812)	<0.001	1,491	491
IPW	0.729 (0.543 − 0.979)	0.035	1,051	316
IPW, ≥2000	0.704 (0.502 − 0.986)	0.041	992	277

Table 5.31: Propensity score analysis comparing bowel cancer mortality risks between metformin and sulfonylureas using ATT weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.442 (1.067 − 1.949)	0.017	1,786	697
Fewer, ≥2000	1.390 (1.144 − 1.688)	<0.001	1,491	491
Complete	1.270 (0.961 − 1.679)	0.093	1,051	316
Complete, ≥2000	1.344 (1.016 − 1.778)	0.038	992	277
Separate	1.244 (0.864 − 1.790)	0.240	1786	697
Separate, ≥2000	1.277 (1.021 − 1.597)	0.032	1,491	491
IPW	1.216 (0.892 − 1.658)	0.217	1051	316
IPW, ≥2000	1.328 (0.999 − 1.765)	0.051	992	277

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.592 (0.467 − 0.751)	<0.001	1,786	697
Fewer, ≥2000	0.581 (0.450 − 0.751)	<0.001	1,491	491
Complete	0.719 (0.476 − 1.086)	0.117	1,051	316
Complete, ≥2000	0.705 (0.471 − 1.054)	0.089	992	277
Separate	0.658 (0.514 − 0.842)	<0.001	1,786	697
Separate, ≥2000	0.594 (0.454 − 0.777)	<0.001	1,491	491
IPW	0.694 (0.486 − 0.991)	0.044	1,051	316
IPW, ≥2000	0.689 (0.467 − 1.017)	0.061	992	277

Table 5.32: Propensity score analysis comparing bowel cancer mortality risks between metformin and sulfonylureas using ATU weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.689 (1.332 − 2.142)	<0.001	1,786	697
Fewer, ≥2000	1.720 (1.332 − 2.220)	<0.001	1,491	491
Complete	1.391 (0.921 − 2.103)	0.117	1,051	316
Complete, ≥2000	1.419 (0.948 − 2.123)	0.089	992	277
Separate	1.520 (1.188 − 1.946)	<0.001	1,786	697
Separate, ≥2000	1.685 (1.287 − 2.204)	<0.001	1,491	491
IPW	1.441 (1.009 − 2.058)	0.044	1,051	316
IPW, ≥2000	1.451 (0.983 − 2.140)	0.061	992	277

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.693 (0.513 − 0.937)	0.017	1,786	697
Fewer, ≥2000	0.719 (0.592 − 0.874)	<0.001	1,491	491
Complete	0.788 (0.596 − 1.041)	0.093	1,051	316
Complete, ≥2000	0.744 (0.562 − 0.984)	0.038	992	277
Separate	0.804 (0.559 − 1.157)	0.240	1,786	697
Separate, ≥2000	0.783 (0.626 − 0.980)	0.032	1,491	491
IPW	0.822 (0.603 − 1.121)	0.217	1,051	316
IPW, ≥2000	0.753 (0.566 − 1.001)	0.051	992	277

5.9 Sub-study 9: Lung Cancer Incidence Risks

In this section, we compare lung cancer incidence risks between metformin and sulfonylureas under CCR, ATE, ATT and ATU schemes with four methods dealing with missing data accordingly. Tabulated in Table 5.33, 5.34, 5.35 and 5.36, we found out that both drugs reveal similar risks on lung cancer incidence since p-values exceed 0.05 and 95% CIs include 1 among most 64 sub-analyses.

Table 5.33: Conventional Cox regression comparing lung cancer incidence risks between metformin and sulfonylureas

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.076 (0.976 − 1.187)	0.143	203,483	2,554
Fewer, ≥2000	1.137 (1.003 − 1.288)	0.044	153,538	1,667
Complete	1.011 (0.865 − 1.180)	0.893	111,294	1,213
Complete, ≥2000	1.045 (0.877 − 1.244)	0.623	102,444	1,043
Separate	1.027 (0.928 − 1.137)	0.606	203,483	2,554
Separate, ≥2000	1.083 (0.951 − 1.234)	0.229	153,538	1,667
IPW	1.028 (0.927 − 1.140)	0.599	111,294	1,213
IPW, ≥2000	1.048 (0.897 − 1.225)	0.553	102,444	1,043

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.929 (0.843 − 1.025)	0.143	203,483	2,554
Fewer, ≥2000	0.880 (0.777 − 0.997)	0.044	153,538	1,667
Complete	0.989 (0.847 − 1.156)	0.893	111,294	1,213
Complete, ≥2000	0.957 (0.804 − 1.140)	0.623	102,444	1,043
Separate	0.974 (0.880 − 1.078)	0.606	203,483	2,554
Separate, ≥2000	0.923 (0.810 − 1.051)	0.229	153,538	1,667
IPW	0.973 (0.877 − 1.079)	0.560	111,294	1,213
IPW, ≥2000	0.954 (0.816 − 1.115)	0.553	102,444	1,043

Table 5.34: Propensity score analysis comparing lung cancer incidence risks between metformin and sulfonylureas using ATE weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.058 (0.940 − 1.191)	0.351	203,483	2,554
Fewer, ≥2000	1.076 (0.923 − 1.255)	0.349	153,538	1,667
Complete	0.815 (0.642 − 1.033)	0.091	111,294	1,213
Complete, ≥2000	0.811 (0.620 − 1.061)	0.126	102,444	1,043
Separate	0.994 (0.877 − 1.128)	0.932	203,483	2,554
Separate, ≥2000	0.938 (0.802 − 1.097)	0.424	153,538	1,667
IPW	0.862 (0.699 − 1.064)	0.167	111,294	1,213
IPW, ≥2000	0.830 (0.642 − 1.072)	0.154	102,444	1,043
Reference (sulfonylureas)	**HR (95% CI)**	**p-value**	**N**	**I**
Fewer	0.945 (0.840 − 1.064)	0.351	203,483	2,554
Fewer, ≥2000	0.929 (0.797 − 1.084)	0.349	153,538	1,667
Complete	1.228 (0.968 − 1.557)	0.091	111,294	1,213
Complete, ≥2000	1.234 (0.943 − 1.614)	0.126	102,444	1,043
Separate	1.006 (0.887 − 1.140)	0.932	203,483	2,554
Separate, ≥2000	1.066 (0.912 − 1.247)	0.424	153,538	1,667
IPW	1.160 (0.940 − 1.431)	0.167	111,294	1,213
IPW, ≥2000	1.205 (0.933 − 1.558)	0.154	102,444	1,043

Table 5.35: Propensity score analysis comparing lung cancer incidence risks between metformin and sulfonylureas using ATT weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.038 (0.890 − 1.210)	0.637	203,483	2,554
Fewer, $\geq$2000	1.137 (1.000 − 1.291)	0.050	153,538	1,667
Complete	0.976 (0.815 − 1.170)	0.795	111,294	1,213
Complete, $\geq$2000	1.044 (0.867 − 1.256)	0.649	102,444	1,043
Separate	1.043 (0.884 − 1.231)	0.619	203,483	2,554
Separate, $\geq$2000	1.110 (0.964 − 1.278)	0.149	153,538	1,667
IPW	0.973 (0.766 − 1.235)	0.821	111,294	1,213
IPW, $\geq$2000	1.044 (0.865 − 1.260)	0.657	102,444	1,043

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.938 (0.817 − 1.076)	0.358	203,483	2,554
Fewer, $\geq$2000	0.946 (0.792 − 1.129)	0.537	153,538	1,667
Complete	1.301 (0.967 − 1.751)	0.082	111,294	1,213
Complete, $\geq$2000	1.305 (0.952 − 1.789)	0.098	102,444	1,043
Separate	1.056 (0.909 − 1.227)	0.475	203,483	2,554
Separate, $\geq$2000	1.132 (0.941 − 1.362)	0.190	153,538	1,667
IPW	1.228 (0.943 − 1.598)	0.127	111,294	1,213
IPW, $\geq$2000	1.274 (0.941 − 1.725)	0.117	102,444	1,043

Table 5.36: Propensity score analysis comparing lung cancer incidence risks between metformin and sulfonylureas using ATU weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.067 (0.930 − 1.224)	0.358	203,483	2,554
Fewer, ≥2000	1.058 (0.885 − 1.263)	0.537	153,538	1,667
Complete	0.769 (0.571 − 1.034)	0.082	111,294	1,213
Complete, ≥2000	0.766 (0.559 − 1.050)	0.098	102,444	1,043
Separate	0.947 (0.815 − 1.100)	0.475	203,483	2,554
Separate, ≥2000	0.884 (0.734 − 1.063)	0.190	153,538	1,667
IPW	0.815 (0.626 − 1.060)	0.127	111,294	1,213
IPW, ≥2000	0.785 (0.580 − 1.062)	0.117	102,444	1,043

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.964 (0.826 − 1.124)	0.637	203,483	2,554
Fewer, ≥2000	0.880 (0.774 − 0.100)	0.050	153,538	1,667
Complete	1.024 (0.855 − 1.227)	0.795	111,294	1,213
Complete, ≥2000	0.958 (0.796 − 1.153)	0.649	102,444	1,043
Separate	0.959 (0.812 − 1.132)	0.619	203,483	2,554
Separate, ≥2000	0.901 (0.783 − 1.038)	0.149	153,538	1,667
IPW	1.028 (0.809 − 1.305)	0.821	111,294	1,213
IPW, ≥2000	0.958 (0.794 − 1.157)	0.657	102,444	1,043

5.10 Sub-study 10: Lung Cancer Incidence Risks

In this section, we compare lung cancer mortality risks between metformin and sulfonylureas under CCR, ATE, ATT and ATU schemes with four methods dealing with missing data accordingly. Tabulated in Table 5.37, 5.38, 5.39 and 5.40, we found out there are very limited cases and events of interest. We have very few patients with lung cancer diagnoses after initial anti-diabetes prescription and death after lung cancer diagnoses. As such, we can't draw conlusions on lung cancer mortality risks between metformin and sulfonylureas within this cohort.

Table 5.37: Conventional Cox regression comparing lung cancer mortality risks between metformin and sulfonylureas

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.673 (1.218 − 2.300)	0.002	408	232
Fewer, ≥2000	1.796 (1.243 − 2.596)	0.002	349	185
Complete	1.660 (0.890 − 3.097)	0.111	214	109
Complete, ≥2000	2.158 (1.108 − 4.205)	0.024	205	102
Separate	1.503 (1.040 − 2.172)	0.030	408	232
Separate, ≥2000	1.830 (1.200 − 2.792)	0.005	349	185
IPW	1.858 (1.176 − 2.935)	0.008	214	109
IPW, ≥2000	2.155 (1.159 − 4.004)	0.015	205	102

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.598 (0.435 − 0.821)	0.002	408	232
Fewer, ≥2000	0.557 (0.385 − 0.805)	0.002	349	185
Complete	0.602 (0.323 − 1.124)	0.111	214	109
Complete, ≥2000	0.463 (0.238 − 0.903)	0.024	205	102
Separate	0.665 (0.461 − 0.961)	0.030	408	232
Separate, ≥2000	0.546 (0.358 − 0.834)	0.005	349	185
IPW	0.538 (0.341 − 0.850)	0.008	214	109
IPW, ≥2000	0.464 (0.250 − 0.863)	0.015	205	102

Table 5.38: Propensity score analysis comparing lung cancer mortality risks between metformin and sulfonylureas using ATE weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.332 (1.007 − 1.761)	0.045	408	232
Fewer, ≥2000	1.462 (1.070 − 1.995)	0.017	349	185
Complete	1.279 (0.756 − 2.163)	0.359	214	109
Complete, ≥2000	1.920 (1.326 − 2.780)	<0.001	205	102
Separate	1.027 (0.723 − 1.461)	0.880	408	232
Separate, ≥2000	1.314 (0.905 − 1.908)	0.152	349	185
IPW	1.194 (0.744 − 1.918)	0.462	214	109
IPW, ≥2000	1.919 (1.336 − 2.757)	<0.001	205	102

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.751 (0.568 − 0.993)	0.045	408	232
Fewer, ≥2000	0.684 (0.501 − 0.934)	0.017	349	185
Complete	0.782 (0.462 − 1.322)	0.359	214	109
Complete, ≥2000	0.521 (0.360 − 0.754)	<0.001	205	102
Separate	0.973 (0.685 − 1.384)	0.880	408	232
Separate, ≥2000	0.761 (0.524 − 1.106)	0.152	349	185
IPW	0.837 (0.522 − 1.344)	0.462	214	109
IPW, ≥2000	0.521 (0.363 − 0.748)	<0.001	205	102

Table 5.39: Propensity score analysis comparing lung cancer mortality risks between metformin and sulfonylureas using ATT weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.157 (0.835 − 1.602)	0.380	408	232
Fewer, ≥2000	1.406 (0.969 − 2.038)	0.073	349	185
Complete	1.185 (0.850 − 1.652)	0.317	214	109
Complete, ≥2000	1.650 (1.058 − 2.572)	0.027	205	102
Separate	0.954 (0.669 − 1.360)	0.795	408	232
Separate, ≥2000	1.194 (0.769 − 1.855)	0.430	349	185
IPW	1.026 (0.732 − 1.439)	0.882	214	109
IPW, ≥2000	1.665 (1.078 − 2.571)	0.022	205	102

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.679 (0.507 − 0.908)	0.009	408	232
Fewer, ≥2000	0.674 (0.489 − 0.929)	0.016	349	185
Complete	0.796 (0.427 − 1.485)	0.474	214	109
Complete, ≥2000	0.499 (0.344 − 0.724)	<0.001	205	102
Separate	0.908 (0.582 − 1.415)	0.670	408	232
Separate, ≥2000	0.723 (0.469 − 1.113)	0.141	349	185
IPW	0.830 (0.460 − 1.497)	0.536	214	109
IPW, ≥2000	0.503 (0.349 − 0.726)	<0.001	205	102

Table 5.40: Propensity score analysis comparing lung cancer mortality risks between metformin and sulfonylureas using ATU weights

Reference (metformin)	HR (95% CI)	p-value	N	I
Fewer	1.473 (1.101 − 1.972)	0.009	408	232
Fewer, ≥2000	1.483 (1.077 − 2.044)	0.016	349	185
Complete	1.256 (0.674 − 2.341)	0.474	214	109
Complete, ≥2000	2.003 (1.381 − 2.905)	<0.001	205	102
Separate	1.101 (0.707 − 1.717)	0.670	408	232
Separate, ≥2000	1.384 (0.898 − 2.131)	0.141	349	185
IPW	1.205 (0.668 − 2.172)	0.536	214	109
IPW, ≥2000	1.987 (1.378 − 2.864)	<0.001	205	102

Reference (sulfonylureas)	HR (95% CI)	p-value	N	I
Fewer	0.864 (0.624 − 1.197)	0.380	408	232
Fewer, ≥2000	0.712 (0.491 − 1.032)	0.073	349	185
Complete	0.844 (0.605 − 1.177)	0.317	214	109
Complete, ≥2000	0.606 (0.389 − 0.945)	0.027	205	102
Separate	1.048 (0.735 − 1.495)	0.795	408	232
Separate, ≥2000	0.838 (0.539 − 1.301)	0.430	349	185
IPW	0.975 (0.695 − 1.367)	0.882	214	109
IPW, ≥2000	0.601 (0.389 − 0.928)	0.022	205	102

5.11 Covariate Balance Evaluation

As in section 3.3.4, we plan to evaluate covariate balance by comparing the difference between before-weighting Somers' D and after-weighting Somers' D. Covariate balance evaluation is the main indicator of the success of the propensity score analysis. It provides evidence that the distribution of each covariate for treated and untreated individuals is similar, and therefore selection bias due to the covariate has been removed.

Since there are 480 summary tables in total, we analyze covariate balance by generalizing information in the big picture instead of detailing local facts among sub-analyses.

From all 640 tables, we found out that in general, prescription age and prescription year are always not balanced before weighting. As shown in Fig. 4-3 and Fig. 4-4, anti-diabetes treatment is strongly related to prescription year. On the other hand, anti-diabetes treatment is less related to prescription age due to year of birth. Both prescription age and prescription year are balanced significantly after weighting. Moreover, prescription year is less balanced compared with prescription age before weighting, and after weighting. For further analysis, even after introducing propensity scores, we may still want to adjust for prescription year for those model with Somer's D larger than 0.1.

In terms of other covariates, some covariates (no coherently along 640 sub-analyses) might be not balanced before weighting. But after weighting, these covariates are well balanced.

5.12 Sensitivity Analysis

Propensity score weighting may result in extreme weights, which inflate the standard errors of the treatment estimates (Robins et al., 2000) and may also increase bias (Harder et al., 2010). In this section, we perform sensitivity analysis by looking into fantasy population and weight summary statistics in Appendix C. Analysis follows the study design sequence as in Table 4.17.

Table 5.41: Number of "fantasy" observations between metformin and sulfonylureas under different schemes

Ref=Met	Metformin Participants	Sulfonylureas Participants	Total Participants
CCR	N_{met} (real)	N_{sulf} (real)	N (real)
ATE	N (fantasy)	N (fantasy)	$2N$ (fantasy)
ATT	N_{met} (real)	N_{met} (fantasy)	$2N_{met}$ (fantasy)
ATU	N_{sulf} (fantasy)	N_{sulf} (real)	$2N_{sulf}$ (fantasy)

Ref=Sulf	Metformin Participants	Sulfonylureas Participants	Total Participants
CCR	N_{met} (real)	N_{sulf} (real)	N (real)
ATE	N (fantasy)	N (fantasy)	$2N$ (fantasy)
ATT	N_{sulf} (fantasy)	N_{sulf} (real)	$2N_{sulf}$ (fantasy)
ATU	N_{met} (real)	N_{met} (fantasy)	$2N_{met}$ (fantasy)

Table 5.41 reveals the number of metformin participants, sulfonylureas participants and overall participants for all the analysis in this thesis. For CCR, all three numbers are actual participants. By balancing two groups of interest, weighting scheme would always distort the number of participants for all ATE, ATT and ATU cases. Since ATE and CCR have the same number of participants if treatment/control inverted, and thanks to the homogeneity within the partial log likelihood function for optimization, inverting treatment/control will yield reciprocal hazard ratios. Take treatment/control into consideration, the conclusion will always be the same, and thus, CCR and ATE are considered optimal for comparing ratio risks among survival analyses.

By multiplying $0 < e(X) < 1$ (ATT) and $0 < 1 - e(X) < 1$ (ATU) based on ATE weights, ATE weights are more extreme compared with ATT weights and ATU weights (Harder et al., 2010). Though Equation 3.36 and 3.37 looks heterogeneous, the idea is essentially the same. By substituting $e(X)_{ref=met} + e(X)_{ref=sulf} = 1$, we have

$$\frac{1 - e(X)_{ref=sulf}}{e(X)_{ref=sulf}} = \frac{e(X)_{ref=met}}{1 - e(X)_{ref=met}} \tag{5.1}$$

thus,

$$w_{ATT,ref=met} = w_{ATU,ref=sulf}, w_{ATT,ref=sulf} = w_{ATU,ref=met} \tag{5.2}$$

this eventually leads to

$$N_{ATT,ref=met} = N_{ATU,ref=sulf}, N_{ATT,ref=sulf} = N_{ATU,ref=met} \tag{5.3}$$

Weighting relationships regarding methods of dealing with missing data are expressed in Equation 5.4. Since inverse probability weighting and complete case analysis are based on complete records, propensity score generally upweights both methods to the original population level as in "fewer" and "separate". Inverse probability weighting further adjusts the weights by inverting the possibility of being complete, leading to larger overall weights. Moreover, treating missing as a separate category increases the number of categories for BMI, IMD, smoking and HbA1c, and this result in fewer common cases among the population which direct to larger propensity weights on being a control or treatment, resulting in larger weights compared with only adjusting for fewer variables. Last, initial anti-diabetes prescription year does not have a large impact on weighting.

$$w_{IPW} > w_{Complete} \gg w_{Seperate} > w_{Fewer} \tag{5.4}$$

Chapter 6

Conclusions

6.1 Discussions

Cancer has always been a worldwide health issue. With limited success and huge investment in drug development, effective anti-cancer drugs were inspired by using drugs already approved for other indications. Based on high dimensional UK primary care EHR system CPRD, we emulated a total of 640 RCTs among 10 sub-studies to test the effect of two universal anti-diabetic drugs — metformin and sulfonylureas on various cancer incidence and mortality risks among the aging population. Sub-studies are formulated according to incidence/mortality risks and 5 cancer types — general cancer, breast cancer, prostate cancer, bowel cancer and lung cancer.

Within each sub-study, a total of 64 in-silico RCTs is emulated by semi-parametric conventional Cox regression, weighted Cox regression in ATE, ATT and ATU schemes, 16 survival analysis each, correspondingly. To distinguish two treatments for risk ratios in the medical sense, 16 studies within each scheme are initially classified by inverting treatment and control. Four methods of dealing with missing data, including, using fewer variables, complete case analysis, treating missing as a separate category and inverse probability weighting further divided 8 RCTs into a group of 2. Final division is added on anti-diabetic drug prescription year due to poor data quality before 2000.

Conclusions on various cancer incidence and mortality risks are tabulated in Table 6.1, 6.2 and 6.3 for clarification. In the rightmost column, " + ", " − ", "0" means metformin is more protective, sulfonylureas is more protective and neither is protective, respectively. In general, metformin is protective over sulfonylureas on cancer mortality risk, whereas the effect of the two drugs on cancer incidence risks is similar.

Within each sub-study, we found that models are quite robust within CCR, ATE, ATT and ATU schemes. However, risk among CCR, ATE, ATT and ATU schemes revealed differences. Specifically, CCR/ATE, ATT/ATU gives similar results respectively. Inversing treatment and control in CCR/ATE does not change significance nor hazard ratio. On the other hand, inverting treatment/control groups in ATT/ATU will exaggerate one case, leaving the other case much less significant. From section 3.3.3, we demonstrated that ATT/ATU weights are biased when the reference is a drug treatment instead of placebo, thus ATT/ATU weights are not considered optimal for comparing the effect between two drugs. Recommended approach would be applying four schemes including conventional Cox and conduct sensitivity analysis afterwards.

In terms of sensitivity analysis based on weight summary attached, we found that combining IPW for missing data and PSW are always larger than original PSW weights as IPW will always up-weight observations to some extent. Also, weight summary statistics including Min, 1st Qu., Median, Mean, 3rd Qu. Max for ATE are greater than ATU/ATT, while ATU weights and ATT weights from Appendix C are similar and they are smaller than 1 in most cases. ATE weights are always larger than 1, while ATU/ATT weights approach to zero. Generally speaking, though extreme weights are less common by looking into 3rd Qu, extreme weights for ATE are larger than ATT/ATU weights.

Table 6.1: Conclusions regarding general cancer and breast cancer incidence/mortality risks between metformin and sulfonylureas initiators

Sub-study	Scheme	Reference	Effect
Cancer incidence	CCR	Metformin	0
Cancer incidence	CCR	Sulfonylureas	0
Cancer incidence	ATE	Metformin	0
Cancer incidence	ATE	Sulfonylureas	0
Cancer incidence	ATT	Metformin	0
Cancer incidence	ATT	Sulfonylureas	0
Cancer incidence	ATU	Metformin	0
Cancer incidence	ATU	Sulfonylureas	0
Cancer mortality	CCR	Metformin	+
Cancer mortality	CCR	Sulfonylureas	+
Cancer mortality	ATE	Metformin	+
Cancer mortality	ATE	Sulfonylureas	+
Cancer mortality	ATT	Metformin	+
Cancer mortality	ATT	Sulfonylureas	+
Cancer mortality	ATU	Metformin	+
Cancer mortality	ATU	Sulfonylureas	+
Breast cancer incidence	CCR	Metformin	0
Breast cancer incidence	CCR	Sulfonylureas	0
Breast cancer incidence	ATE	Metformin	0
Breast cancer incidence	ATE	Sulfonylureas	0
Breast cancer incidence	ATT	Metformin	0
Breast cancer incidence	ATT	Sulfonylureas	0
Breast cancer incidence	ATU	Metformin	0
Breast cancer incidence	ATU	Sulfonylureas	0
Breast cancer mortality	CCR	Metformin	+
Breast cancer mortality	CCR	Sulfonylureas	+
Breast cancer mortality	ATE	Metformin	+
Breast cancer mortality	ATE	Sulfonylureas	+
Breast cancer mortality	ATT	Metformin	+
Breast cancer mortality	ATT	Sulfonylureas	+
Breast cancer mortality	ATU	Metformin	+
Breast cancer mortality	ATU	Sulfonylureas	+

Table 6.2: Conclusions regarding prostate cancer and bowel cancer incidence/mortality risks between metformin and sulfonylureas initiators

Sub-study	Scheme	Reference	Protective effect
Prostate cancer incidence	CCR	Metformin	0
Prostate cancer incidence	CCR	Sulfonylureas	0
Prostate cancer incidence	ATE	Metformin	−
Prostate cancer incidence	ATE	Sulfonylureas	−
Prostate cancer incidence	ATT	Metformin	0
Prostate cancer incidence	ATT	Sulfonylureas	−
Prostate cancer incidence	ATU	Metformin	−
Prostate cancer incidence	ATU	Sulfonylureas	0
Prostate cancer mortality	CCR	Metformin	+
Prostate cancer mortality	CCR	Sulfonylureas	+
Prostate cancer mortality	ATE	Metformin	+
Prostate cancer mortality	ATE	Sulfonylureas	+
Prostate cancer mortality	ATT	Metformin	0
Prostate cancer mortality	ATT	Sulfonylureas	+
Prostate cancer mortality	ATU	Metformin	+
Prostate cancer mortality	ATU	Sulfonylureas	0
Bowel cancer incidence	CCR	Metformin	0
Bowel cancer incidence	CCR	Sulfonylureas	0
Bowel cancer incidence	ATE	Metformin	0
Bowel cancer incidence	ATE	Sulfonylureas	0
Bowel cancer incidence	ATT	Metformin	0
Bowel cancer incidence	ATT	Sulfonylureas	0
Bowel cancer incidence	ATU	Metformin	0
Bowel cancer incidence	ATU	Sulfonylureas	0
Bowel cancer mortality	CCR	Metformin	+
Bowel cancer mortality	CCR	Sulfonylureas	+
Bowel cancer mortality	ATE	Metformin	+
Bowel cancer mortality	ATE	Sulfonylureas	+
Bowel cancer mortality	ATT	Metformin	0
Bowel cancer mortality	ATT	Sulfonylureas	+
Bowel cancer mortality	ATU	Metformin	+
Bowel cancer mortality	ATU	Sulfonylureas	0

Table 6.3: Conclusions regarding lung cancer incidence/mortality risks between metformin and sulfonylureas initiators

Sub-study	Scheme	Reference	Protective effect
Lung cancer incidence	CCR	Metformin	0
Lung cancer incidence	CCR	Sulfonylureas	0
Lung cancer incidence	ATE	Metformin	0
Lung cancer incidence	ATE	Sulfonylureas	0
Lung cancer incidence	ATT	Metformin	0
Lung cancer incidence	ATT	Sulfonylureas	0
Lung cancer incidence	ATU	Metformin	0
Lung cancer incidence	ATU	Sulfonylureas	0
Lung cancer mortality	CCR	Metformin	+
Lung cancer mortality	CCR	Sulfonylureas	+
Lung cancer mortality	ATE	Metformin	0
Lung cancer mortality	ATE	Sulfonylureas	0
Lung cancer mortality	ATT	Metformin	0
Lung cancer mortality	ATT	Sulfonylureas	0
Lung cancer mortality	ATU	Metformin	0
Lung cancer mortality	ATU	Sulfonylureas	0

By looking into Somer's D for all 640 RCTs, we found out that before balancing, prescription year and prescription age have high Somer's D, indicating balancing these two variables is the main objective. After weighting, prescription age is well-balanced. While with a Somer's D between 0.1 and 0.2, prescription year is somewhat balanced but not well-balanced in most cases. We may also want to control it beyond anti-diabetic prescription in weighted Cox regression.

In this study, missing mechanisms and methods dealing with missing data revealed very limited impact on significance level and confidence interval. From results with and without constraint on calendar year of initial anti-diabetic drug prescription, data quality with prescription year later than 2000 is much better despite a decline in observation and events. If the drop in observation and events is not too sharp, it's a good idea to add a limit on prescription year.

As such, we draw a final conclusion that metformin is protective over sulfonylureas on cancer mortality risk, whereas the effect of the two drugs on cancer incidence risks is similar. In terms of methodologies, less variable or inverse probability weighting, conventional Cox regression and propensity score analysis in ATE scheme, together with prescription year over 2000 should be adopted.

6.2 Future Directions

In this thesis, after investigating the effect of anti-diabetic drugs on cancer incidence and mortality risks with multiple schemes and settings, we come to a systematic way to repurpose not only incidence/mortality risks, but also onset/progression risks and even prevention/development risks for many other types of diseases based on different group drugs.

Apart from simply applying this framework directly, we can also investigate more on methods dealing with missing data, especially under MNAR scenarios. Many other machine learning methods to generate propensity score can be adopted to further balance prescription year etc. Furthermore, though not highly recommended due to reduction in observations, propensity score stratification and propensity score matching can also be added. Besides graphical model especially casual diagrams such as directed acyclic graph (DAG) and undirected graph for clarification can be integrated for variational inference. Stepwise Cox and penalty methods are a potential approach to justify covariate selection. Supervised machine learning techniques including survival trees, Bayesian method, deep learning, support vector machines, transfer learning, ensemble learning, multi-task learning, reinforcement learning and active learning will be the next tipping point for survival analysis.

Broadly speaking, applying and combining advances in the areas of optimization, statistics, and machine learning, with new and interesting sources of data, we have the potential to make substantive improvements in finding the effective treatments for some disease without therapy or affordable drugs.

Appendix A

Abbreviations

Table A.1: Abbreviation and full form in alphabetic order

Abbreviation	Full Form
AF	Atrial Fibrillation
AFT	Accelerated Time to Failure
ATE	Average Treatment Effect
ATT	Average Treatment Effect on the Treated
ATU	Average Treatment Effect on the Untreated
AUROC	Area under Receiver Operating Characteristic
BMI	Body Mass Index
CCR	Conventional Cox Regression
CHD	Coronary Heart Disease
CHF	Cumulative Hazard Function
CI	Confidence Interval
CKD	Chronic Kidney Disease
COPD	Chronic Obstructive Pulmonary Disease
CPRD	Clinical Practice Research Datalink
DAG	Directed Acyclic Graph
DMT2	Diabetes Mellitus Type 2
EHR	Electronic Health Record
FDA	Food and Drug Administration
GBM	Generalized Boosted Model
GP	General Practice
HbA1c	Glycated Hemoglobin
HES	Hospital Episode Statistics
HF	Heart Failure
HR	Hazard Ratio

Table A.2: Abbreviation and full form in alphabetic order, continued

IMD	Multiple Deprivation Index
IPW	Inverse Probability Weighting
ITT	Intention-To-Treat
LOCF	Last Observation Carried Forward
MAR	Missing at Random
MCAR	Missing Completely at Random
Met	Metformin
NMAR	Not Missing at Random
MPLE	Maximum Partial Likelihood Estimation
OHAs	Orally administered antihyperglycemic agents
ONS	Office for National Statistics
PSW	Propensity Score Weighting
PVD	Peripheral Vascular Disease
RCT	Randomized Controlled Trials
Ref	Reference
ROC	Receiver Operating Characteristic
SDO	Simple Difference in Mean Outcomes
Sulf	Sulfonylurea
YOB	Year of birth

Appendix B

Notations

Table B.1: Symbol notation in alphabetic order

Notation	Description
C	$\mathbb{R}^{N \times 1}$ vector of last follow up time
C_i	Individual last follow up time
I	Number of events of interest
i	$i = 1, 2, ...N^{th}$ observation
j	$j = 1, 2..., n_i^{th}$ variable
N	Number of observations
n	Number of features
n_i	Repeated measures of the response variable on the same individual
RK_i	Subjects at risk at time t_i
Res_i	Response indicators
T	$\mathbb{R}^{N \times 1}$ vector of event time
T_i	Individual event time
t	Time
X	$\mathbb{R}^{N \times P}$ feature covariate matrix
X_i	$\mathbb{R}^{1 \times P}$ covariate vector of observation i
Y_i^{CF}	Unobserved counterfactual outcome
Y^m	Missing responses
Y_i^m	Missing responses of ith object
Y^o	Observed responses
Y_i^o	Observed responses of ith object
Y^0	Outcome for control
Y_i^0	Individual outcome for control
Y_1	Outcome for treatment
Y_i^1	Individual outcome for treatment
Z_i	Treatment($=1$),control($=0$)
z	Mean difference in units of the common standard deviation

154

Table B.2: Greek notation in alphabetic order

Notation	Description
β	$\mathbb{R}^{P \times 1}$ coefficient
$\hat{\beta}$	Estimated coefficient
$\boldsymbol{\beta}$	Coefficient vector
$\hat{\boldsymbol{\beta}}$	Estimated coefficient vector
e^{β}	Hazard ratio
$e^{\boldsymbol{\beta}}$	Hazard ratio for all covariates
Δt	Small time interval
δ	$N \times 1$ binary vector for event
δ_i	Individual event/outcome
σ	Standard deviation
τ_a	Kendall's tau
τ_i	Individual treatment effect
Φ	Standard Normal distribution

Table B.3: Function notation in alphabetic order

Notation	Description
D	Somer's D
$\mathbb{E}$	Expectation
$\mathbb{E}(\cdot\|\cdot)$	Conditional expectation
$e(X)$	Propensity score
$F(t)$	Cumulative event probability function
$f(t)$	Death density function
$H(t)$	Cumulative hazard function
$H_0(t)$	Baseline cumulative hazard function
$\hat{H}_0(t)$	Estimated baseline cumulative hazard function
$h(t)$	Hazard function
$h_0(t)$	Baseline hazard function
$\hat{h}_0(t)$	Estimated baseline hazard function
$L(\beta)$	Partial likelihood
$L_i(\beta)$	Individual partial likelihood
y	$\mathbb{R}^{N\times 1}$ vector of observed time which is equal to $min(T, C)$
y_i	Individual observed time equal to $min(T_i, C_i)$
$S(t)$	Survival probability function
$\mathbb{P}$	Probability
$\mathbb{P}(\cdot\|\cdot)$	Conditional probability
w	Weight
w_{ATE}	ATE weight
w_{ATT}	ATT weight
w_{ATU}	ATU weight
$w_{Complete}$	Weight for complete case analysis
w_{Fewer}	Weight for using fewer variables
w_{IPW}	Weight for IPW on missing data
$w_{Seperate}$	Weight for treat missing as a separate category
$\|\cdot\|$	Norm

Appendix C

Covariate Balance Evaluation

C.1 Weights summary for cancer incidence risks

Table C.1: ATE weights summary for cancer incidence risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	1.003	1.050	1.159	2.079	1.573	229.991
Fewer, met	1.003	1.035	1.096	1.491	1.292	84.058
Fewer, sulf	1.011	1.196	1.587	3.486	2.588	229.991
Fewer, $\geq$2000, overall	1.002	1.029	1.093	2.195	1.363	340.178
Fewer, $\geq$2000, met	1.002	1.024	1.067	1.163	1.189	7.454
Fewer, $\geq$2000, sulf	1.169	2.152	3.193	8.363	6.255	340.178
Complete, overall	1.000	1.015	1.062	2.487	1.278	2,170.904
Complete, met	1.000	1.012	1.041	1.174	1.140	51.206
Complete, sulf	1.008	1.486	2.230	9.849	4.680	2,170.904
Complete, $\geq$2000, overall	1.000	1.013	1.048	2.409	1.194	1,974.650
Complete, $\geq$2000, met	1.000	1.011	1.037	1.116	1.116	24.369
Complete, $\geq$2000, sulf	1.062	2.070	3.337	13.120	7.550	1,974.650
Separate, overall	1.002	1.038	1.124	2.054	1.503	327.708
Separate, met	1.002	1.026	1.073	1.496	1.253	98.844
Separate, sulf	1.006	1.145	1.474	3.389	2.376	327.708
Separate, $\geq$2000, overall	1.001	1.020	1.068	2.126	1.305	550.730
Separate, $\geq$2000, met	1.001	1.017	1.048	1.165	1.149	17.051
Separate, $\geq$2000, sulf	1.052	1.742	2.684	7.873	5.670	550.730
IPW, overall	1.118	1.334	1.597	4.411	2.496	3,009.869
IPW, met	1.118	1.304	1.497	1.927	1.884	346.112
IPW, sulf	2.214	4.889	6.454	18.334	11.778	3,009.869
IPW, $\geq$2000, overall	1.018	1.089	1.160	2.790	1.379	2,317.062
IPW, $\geq$2000, met	1.018	1.083	1.140	1.242	1.268	30.436
IPW, $\geq$2000, sulf	1.309	2.899	4.503	15.616	9.636	2,317.062

Table C.2: ATE weights summary for cancer incidence risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	1.003	1.050	1.159	2.079	1.573	229.991
Fewer, met	1.003	1.035	1.096	1.491	1.292	84.058
Fewer, sulf	1.011	1.196	1.587	3.486	2.588	229.991
Fewer, ≥2000, overall	1.002	1.029	1.093	2.195	1.363	340.178
Fewer, ≥2000, met	1.002	1.024	1.067	1.163	1.189	7.454
Fewer, ≥2000, sulf	1.169	2.152	3.193	8.363	6.255	340.178
Complete, overall	1.000	1.015	1.062	2.487	1.278	2,170.904
Complete, met	1.000	1.012	1.041	1.174	1.140	51.206
Complete, sulf	1.008	1.486	2.230	9.849	4.680	2,170.904
Complete, ≥2000, overall	1.000	1.013	1.048	2.409	1.194	1,974.650
Complete, ≥2000, met	1.000	1.011	1.037	1.116	1.116	24.369
Complete, ≥2000, sulf	1.062	2.070	3.337	13.120	7.550	1,974.650
Separate, overall	1.002	1.038	1.124	2.054	1.503	327.708
Separate, met	1.002	1.026	1.073	1.496	1.253	98.844
Separate, sulf	1.006	1.145	1.474	3.389	2.376	327.708
Separate, ≥2000, overall	1.001	1.020	1.068	2.126	1.305	550.730
Separate, ≥2000, met	1.001	1.017	1.048	1.165	1.149	17.051
Separate, ≥2000, sulf	1.052	1.742	2.684	7.873	5.670	550.730
IPW, overall	1.118	1.334	1.597	4.411	2.496	3,009.869
IPW, met	1.118	1.304	1.497	1.927	1.884	346.112
IPW, sulf	2.214	4.889	6.454	18.334	11.778	3,009.869
IPW, ≥2000, overall	1.018	1.089	1.160	2.790	1.379	2,317.062
IPW, ≥2000, met	1.018	1.083	1.140	1.242	1.268	30.436
IPW, ≥2000, sulf	1.309	2.899	4.503	15.616	9.636	2,317.062

Table C.3: ATT weights summary for cancer incidence risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.003	0.054	0.237	0.641	1	83.058
Fewer, met	0.003	0.035	0.096	0.491	0.292	83.058
Fewer, sulf	1	1	1	1	1	1
Fewer, $\geq$2000, overall	0.002	0.029	0.093	0.283	0.366	6.454
Fewer, $\geq$2000, met	0.002	0.024	0.067	0.163	0.189	6.454
Fewer, $\geq$2000, sulf	1	1	1	1	1	1
Complete, overall	0.00005	0.015	0.062	0.299	0.340	50.206
Complete, met	0.00005	0.012	0.041	0.174	0.140	50.206
Complete, sulf	1	1	1	1	1	1
Complete, $\geq$2000, overall	0.00004	0.013	0.048	0.211	0.195	23.369
Complete, $\geq$2000, met	0.00004	0.011	0.037	0.116	0.116	23.369
Complete, $\geq$2000, sulf	1	1	1	1	1	1
Separate, overall	0.002	0.040	0.201	0.645	1	97.844
Separate, met	0.002	0.026	0.073	0.496	0.253	97.844
Separate, sulf	1	1	1	1	1	1
Separate, $\geq$2000, overall	0.001	0.020	0.068	0.284	0.330	16.051
Separate, $\geq$2000, met	0.001	0.017	0.048	0.165	0.149	16.051
Separate, $\geq$2000, sulf	1	1	1	1	1	1
IPW, overall	0.0001	0.021	0.092	0.794	0.605	335.406
IPW, met	0.0001	0.016	0.059	0.398	0.226	335.406
IPW, sulf	1.197	2.058	2.679	3.014	3.425	26.426
IPW, $\geq$2000, overall	0.00004	0.014	0.053	0.265	0.223	29.187
IPW, $\geq$2000, met	0.00004	0.012	0.041	0.135	0.131	29.187
IPW, $\geq$2000, sulf	1.027	1.197	1.296	1.339	1.430	3.226

Table C.4: ATT weights summary for cancer incidence risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.011	1	1	1.438	1	228.991
Fewer, met	1	1	1	1	1	1
Fewer, sulf	0.011	0.196	0.587	2.486	1.588	228.991
Fewer, ≥2000, overall	0.169	1	1	1.912	1	339.178
Fewer, ≥2000, met	1	1	1	1	1	1
Fewer, ≥2000, sulf	0.169	1.152	2.193	7.363	5.255	339.178
Complete, overall	0.008	1	1	2.188	1	2,169.904
Complete, met	1	1	1	1	1	1
Complete, sulf	0.008	0.486	1.230	8.849	3.680	2,169.904
Complete, ≥2000, overall	0.062	1	1	2.198	1	1,973.650
Complete, ≥2000, met	1	1	1	1	1	1
Complete, ≥2000, sulf	0.062	1.070	2.337	12.120	6.550	1,973.650
Separate, overall	0.006	1	1	1.410	1	326.708
Separate, met	1	1	1	1	1	1
Separate, sulf	0.006	0.145	0.474	2.389	1.376	326.708
Separate, ≥2000, overall	0.052	1	1	1.842	1	549.730
Separate, ≥2000, met	1	1	1	1	1	1
Separate, ≥2000, sulf	0.052	0.742	1.684	6.873	4.670	549.730
IPW, overall	0.053	1.276	1.444	3.617	1.766	3,008.176
IPW, met	1.114	1.271	1.423	1.529	1.628	14.095
IPW, sulf	0.053	1.560	3.351	15.320	8.636	3,008.176
IPW, ≥2000, overall	0.100	1.063	1.095	2.525	1.153	2,315.889
IPW, ≥2000, met	1.015	1.062	1.090	1.107	1.133	2.146
IPW, ≥2000, sulf	0.100	1.483	3.144	14.277	8.395	2,315.889

Table C.5: ATU weights summary for cancer incidence risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.011	1	1	1.438	1	228.991
Fewer, met	1	1	1	1	1	1
Fewer, sulf	0.011	0.196	0.587	2.486	1.588	228.991
Fewer, ≥2000, overall	0.169	1	1	1.912	1	339.178
Fewer, ≥2000, met	1	1	1	1	1	1
Fewer, ≥2000, sulf	0.169	1.152	2.193	7.363	5.255	339.178
Complete, overall	0.008	1	1	2.188	1	2,169.904
Complete, met	1	1	1	1	1	1
Complete, sulf	0.008	0.486	1.230	8.849	3.680	2,169.904
Complete, ≥2000, overall	0.062	1	1	2.198	1	1,973.650
Complete, ≥2000, met	1	1	1	1	1	1
Complete, ≥2000, sulf	0.062	1.070	2.337	12.120	6.550	1,973.650
Separate, overall	0.006	1	1	1.410	1	326.708
Separate, met	1	1	1	1	1	1
Separate, sulf	0.006	0.145	0.474	2.389	1.376	326.708
Separate, ≥2000, overall	0.052	1	1	1.842	1	549.730
Separate, ≥2000, met	1	1	1	1	1	1
Separate, ≥2000, sulf	0.052	0.742	1.684	6.873	4.670	549.730
IPW, overall	0.053	1.276	1.444	3.617	1.766	3,008.176
IPW, met	1.114	1.271	1.423	1.529	1.628	14.095
IPW, sulf	0.053	1.560	3.351	15.320	8.636	3,008.176
IPW, ≥2000, overall	0.100	1.063	1.095	2.525	1.153	2,315.889
IPW, ≥2000, met	1.015	1.062	1.090	1.107	1.133	2.146
IPW, ≥2000, sulf	0.100	1.483	3.144	14.277	8.395	2,315.889

Table C.6: ATU weights summary for cancer incidence risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.003	0.054	0.237	0.641	1	83.058
Fewer, met	0.003	0.035	0.096	0.491	0.292	83.058
Fewer, sulf	1	1	1	1	1	1
Fewer, ≥2000, overall	0.002	0.029	0.093	0.283	0.366	6.454
Fewer, ≥2000, met	0.002	0.024	0.067	0.163	0.189	6.454
Fewer, ≥2000, sulf	1	1	1	1	1	1
Complete, overall	0.00005	0.015	0.062	0.299	0.340	50.206
Complete, met	0.00005	0.012	0.041	0.174	0.140	50.206
Complete, sulf	1	1	1	1	1	1
Complete, ≥2000, overall	0.00004	0.013	0.048	0.211	0.195	23.369
Complete, ≥2000, met	0.00004	0.011	0.037	0.116	0.116	23.369
Complete, ≥2000, sulf	1	1	1	1	1	1
Separate, overall	0.002	0.040	0.201	0.645	1	97.844
Separate, met	0.002	0.026	0.073	0.496	0.253	97.844
Separate, sulf	1	1	1	1	1	1
Separate, ≥2000, overall	0.001	0.020	0.068	0.284	0.330	16.051
Separate, ≥2000, met	0.001	0.017	0.048	0.165	0.149	16.051
Separate, ≥2000, sulf	1	1	1	1	1	1
IPW, overall	0.0001	0.021	0.092	0.794	0.605	335.406
IPW, met	0.0001	0.016	0.059	0.398	0.226	335.406
IPW, sulf	1.197	2.058	2.679	3.014	3.425	26.426
IPW, ≥2000, overall	0.00004	0.014	0.053	0.265	0.223	29.187
IPW, ≥2000, met	0.00004	0.012	0.041	0.135	0.131	29.187
IPW, ≥2000, sulf	1.027	1.197	1.296	1.339	1.430	3.226

C.2 Weights summary for cancer mortality risks

Table C.7: ATE weights summary for cancer mortality risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	1.006	1.075	1.205	2.109	1.638	138.614
Fewer, met	1.006	1.059	1.133	1.380	1.324	41.775
Fewer, sulf	1.018	1.353	1.902	4.081	3.508	138.614
Fewer, $\geq$2000, overall	1.006	1.069	1.176	2.110	1.573	146.011
Fewer, $\geq$2000, met	1.006	1.057	1.123	1.228	1.275	5.985
Fewer, $\geq$2000, sulf	1.177	2.036	2.987	5.875	5.820	146.011
Complete, overall	1.000	1.037	1.115	2.474	1.405	850.639
Complete, met	1.000	1.029	1.079	1.194	1.208	9.275
Complete, sulf	1.024	1.684	2.610	8.626	5.619	850.639
Complete, $\geq$2000, overall	1.000	1.034	1.101	2.373	1.348	881.207
Complete, $\geq$2000, met	1.000	1.028	1.074	1.170	1.190	10.866
Complete, $\geq$2000, sulf	1.081	2.032	3.218	9.279	6.339	881.207
Separate, overall	1.004	1.057	1.163	2.084	1.558	157.505
Separate, met	1.004	1.044	1.105	1.401	1.281	50.836
Separate, sulf	1.013	1.251	1.717	3.932	3.144	157.505
Separate, $\geq$2000, overall	1.004	1.049	1.138	2.057	1.498	174.741
Separate, $\geq$2000, met	1.004	1.039	1.093	1.233	1.230	17.016
Separate, $\geq$2000, sulf	1.065	1.694	2.564	5.574	5.128	174.741
IPW, overall	1.138	1.355	1.596	4.036	2.479	1,137.134
IPW, met	1.138	1.325	1.494	1.754	1.832	21.298
IPW, sulf	2.293	4.476	6.042	15.001	10.658	1,137.134
IPW, $\geq$2000, overall	1.020	1.109	1.210	2.705	1.543	977.351
IPW, $\geq$2000, met	1.020	1.098	1.170	1.286	1.322	13.711
IPW, $\geq$2000, sulf	1.402	2.728	4.201	10.850	7.530	977.351

Table C.8: ATE weights summary for cancer mortality risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	1.006	1.075	1.205	2.109	1.638	138.614
Fewer, met	1.006	1.059	1.133	1.380	1.324	41.775
Fewer, sulf	1.018	1.353	1.902	4.081	3.508	138.614
Fewer, $\geq$2000, overall	1.006	1.069	1.176	2.110	1.573	146.011
Fewer, $\geq$2000, met	1.006	1.057	1.123	1.228	1.275	5.985
Fewer, $\geq$2000, sulf	1.177	2.036	2.987	5.875	5.820	146.011
Complete, overall	1.000	1.037	1.115	2.474	1.405	850.639
Complete, met	1.000	1.029	1.079	1.194	1.208	9.275
Complete, sulf	1.024	1.684	2.610	8.626	5.619	850.639
Complete, $\geq$2000, overall	1.000	1.034	1.101	2.373	1.348	881.207
Complete, $\geq$2000, met	1.000	1.028	1.074	1.170	1.190	10.866
Complete, $\geq$2000, sulf	1.081	2.032	3.218	9.279	6.339	881.207
Separate, overall	1.004	1.057	1.163	2.084	1.558	157.505
Separate, met	1.004	1.044	1.105	1.401	1.281	50.836
Separate, sulf	1.013	1.251	1.717	3.932	3.144	157.505
Separate, $\geq$2000, overall	1.004	1.049	1.138	2.057	1.498	174.741
Separate, $\geq$2000, met	1.004	1.039	1.093	1.233	1.230	17.016
Separate, $\geq$2000, sulf	1.065	1.694	2.564	5.574	5.128	174.741
IPW, overall	1.138	1.355	1.596	4.036	2.479	1,137.134
IPW, met	1.138	1.325	1.494	1.754	1.832	21.298
IPW, sulf	2.293	4.476	6.042	15.001	10.658	1,137.134
IPW, $\geq$2000, overall	1.020	1.109	1.210	2.705	1.543	977.351
IPW, $\geq$2000, met	1.020	1.098	1.170	1.286	1.322	13.711
IPW, $\geq$2000, sulf	1.402	2.728	4.201	10.850	7.530	977.351

Table C.9: ATT weights summary for cancer mortality risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.006	0.080	0.250	0.548	1	40.775
Fewer, met	0.006	0.059	0.133	0.380	0.324	40.775
Fewer, sulf	1	1	1	1	1	1
Fewer, $\geq$2000, overall	0.006	0.069	0.176	0.375	0.651	4.985
Fewer, $\geq$2000, met	0.006	0.057	0.123	0.228	0.275	4.985
Fewer, $\geq$2000, sulf	1	1	1	1	1	1
Complete, overall	0.0003	0.037	0.116	0.333	0.495	8.275
Complete, met	0.0003	0.029	0.079	0.194	0.208	8.275
Complete, sulf	1	1	1	1	1	1
Complete, $\geq$2000, overall	0.0003	0.034	0.101	0.293	0.363	9.866
Complete, $\geq$2000, met	0.0003	0.028	0.074	0.170	0.190	9.866
Complete, $\geq$2000, sulf	1	1	1	1	1	1
Separate, overall	0.004	0.062	0.212	0.563	1	49.836
Separate, met	0.004	0.044	0.105	0.401	0.281	49.836
Separate, sulf	1	1	1	1	1	1
Separate, $\geq$2000, overall	0.004	0.049	0.139	0.379	0.692	16.016
Separate, $\geq$2000, met	0.004	0.039	0.093	0.233	0.230	16.016
Separate, $\geq$2000, sulf	1	1	1	1	1	1
IPW, overall	0.0004	0.049	0.162	0.676	0.809	18.449
IPW, met	0.0004	0.038	0.109	0.324	0.309	18.449
IPW, sulf	1.301	1.747	2.140	2.369	2.731	12.885
IPW, $\geq$2000, overall	0.0003	0.036	0.111	0.353	0.410	12.449
IPW, $\geq$2000, met	0.0003	0.030	0.080	0.192	0.212	12.449
IPW, $\geq$2000, sulf	1.033	1.138	1.218	1.275	1.351	2.368

Table C.10: ATT weights summary for cancer mortality risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.018	1	1	1.561	1	137.614
Fewer, met	1	1	1	1	1	1
Fewer, sulf	0.018	0.353	0.902	3.081	2.508	137.614
Fewer, ≥2000, overall	0.177	1	1	1.735	1	145.011
Fewer, ≥2000, met	1	1	1	1	1	1
Fewer, ≥2000, sulf	0.177	1.036	1.987	4.875	4.820	145.011
Complete, overall	0.024	1	1	2.141	1	849.639
Complete, met	1	1	1	1	1	1
Complete, sulf	0.024	0.684	1.610	7.626	4.619	849.639
Complete, ≥2000, overall	0.081	1	1	2.080	1	880.207
Complete, ≥2000, met	1	1	1	1	1	1
Complete, ≥2000, sulf	0.081	1.032	2.218	8.279	5.339	880.207
Separate, overall	0.013	1	1	1.521	1	156.505
Separate, met	1	1	1	1	1	1
Separate, sulf	0.013	0.251	0.717	2.932	2.144	156.505
Separate, ≥2000, overall	0.065	1	1	1.678	1	173.741
Separate, ≥2000, met	1	1	1	1	1	1
Separate, ≥2000, sulf	0.065	0.694	1.564	4.574	4.128	173.741
IPW, overall	0.133	1.268	1.401	3.360	1.628	1,135.786
IPW, met	1.133	1.263	1.364	1.430	1.508	4.738
IPW, sulf	0.133	1.777	3.569	12.633	8.496	1,135.786
IPW, ≥2000, overall	0.144	1.054	1.085	2.352	1.145	976.242
IPW, ≥2000, met	1.011	1.052	1.078	1.094	1.117	1.662
IPW, ≥2000, sulf	0.144	1.379	2.885	9.575	6.334	976.242

Table C.11: ATU weights summary for cancer mortality risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.018	1	1	1.561	1	137.614
Fewer, met	1	1	1	1	1	1
Fewer, sulf	0.018	0.353	0.902	3.081	2.508	137.614
Fewer, ≥2000, overall	0.177	1	1	1.735	1	145.011
Fewer, ≥2000, met	1	1	1	1	1	1
Fewer, ≥2000, sulf	0.177	1.036	1.987	4.875	4.820	145.011
Complete, overall	0.024	1	1	2.141	1	849.639
Complete, met	1	1	1	1	1	1
Complete, sulf	0.024	0.684	1.610	7.626	4.619	849.639
Complete, ≥2000, overall	0.081	1	1	2.080	1	880.207
Complete, ≥2000, met	1	1	1	1	1	1
Complete, ≥2000, sulf	0.081	1.032	2.218	8.279	5.339	880.207
Separate, overall	0.013	1	1	1.521	1	156.505
Separate, met	1	1	1	1	1	1
Separate, sulf	0.013	0.251	0.717	2.932	2.144	156.505
Separate, ≥2000, overall	0.065	1	1	1.678	1	173.741
Separate, ≥2000, met	1	1	1	1	1	1
Separate, ≥2000, sulf	0.065	0.694	1.564	4.574	4.128	173.741
IPW, overall	0.133	1.268	1.401	3.360	1.628	1,135.786
IPW, met	1.133	1.263	1.364	1.430	1.508	4.738
IPW, sulf	0.133	1.777	3.569	12.633	8.496	1,135.786
IPW, ≥2000, overall	0.144	1.054	1.085	2.352	1.145	976.242
IPW, ≥2000, met	1.011	1.052	1.078	1.094	1.117	1.662
IPW, ≥2000, sulf	0.144	1.379	2.885	9.575	6.334	976.242

Table C.12: ATU weights summary for cancer mortality risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.006	0.080	0.250	0.548	1	40.775
Fewer, met	0.006	0.059	0.133	0.380	0.324	40.775
Fewer, sulf	1	1	1	1	1	1
Fewer, ≥2000, overall	0.006	0.069	0.176	0.375	0.651	4.985
Fewer, ≥2000, met	0.006	0.057	0.123	0.228	0.275	4.985
Fewer, ≥2000, sulf	1	1	1	1	1	1
Complete, overall	0.0003	0.037	0.116	0.333	0.495	8.275
Complete, met	0.0003	0.029	0.079	0.194	0.208	8.275
Complete, sulf	1	1	1	1	1	1
Complete, ≥2000, overall	0.0003	0.034	0.101	0.293	0.363	9.866
Complete, ≥2000, met	0.0003	0.028	0.074	0.170	0.190	9.866
Complete, ≥2000, sulf	1	1	1	1	1	1
Separate, overall	0.004	0.062	0.212	0.563	1	49.836
Separate, met	0.004	0.044	0.105	0.401	0.281	49.836
Separate, sulf	1	1	1	1	1	1
Separate, ≥2000, overall	0.004	0.049	0.139	0.379	0.692	16.016
Separate, ≥2000, met	0.004	0.039	0.093	0.233	0.230	16.016
Separate, ≥2000, sulf	1	1	1	1	1	1
IPW, overall	0.0004	0.049	0.162	0.676	0.809	18.449
IPW, met	0.0004	0.038	0.109	0.324	0.309	18.449
IPW, sulf	1.301	1.747	2.140	2.369	2.731	12.885
IPW, ≥2000, overall	0.0003	0.036	0.111	0.353	0.410	12.449
IPW, ≥2000, met	0.0003	0.030	0.080	0.192	0.212	12.449
IPW, ≥2000, sulf	1.033	1.138	1.218	1.275	1.351	2.368

C.3 Weights summary for breast cancer incidence risks

Table C.13: ATE weights summary for breast cancer incidence risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	1.004	1.051	1.165	2.053	1.586	211.118
Fewer, met	1.004	1.035	1.098	1.487	1.299	55.563
Fewer, sulf	1.010	1.209	1.593	3.404	2.623	211.118
Fewer, ≥2000, overall	1.002	1.029	1.093	2.142	1.360	331.844
Fewer, ≥2000, overall	1.002	1.029	1.093	2.142	1.360	331.844
Fewer, ≥2000, overall	1.002	1.029	1.093	2.142	1.360	331.844
Fewer, ≥2000, met	1.002	1.024	1.067	1.165	1.186	6.285
Fewer, ≥2000, sulf	1.130	2.084	3.194	7.926	6.373	331.844
Complete, overall	1.000	1.015	1.060	2.420	1.273	1,367.363
Complete, met	1.000	1.011	1.040	1.167	1.139	46.051
Complete, sulf	1.016	1.498	2.253	9.607	4.895	1,367.363
Complete, ≥2000, overall	1.000	1.013	1.046	2.359	1.189	1,535.591
Complete, ≥2000, met	1.000	1.011	1.036	1.115	1.114	13.739
Complete, ≥2000, sulf	1.074	2.034	3.342	12.776	7.801	1,535.591
Separate, overall	1.002	1.040	1.132	2.027	1.519	313.741
Separate, met	1.002	1.026	1.076	1.486	1.261	89.701
Separate, sulf	1.007	1.161	1.487	3.320	2.418	313.741
Separate, ≥2000, overall	1.001	1.021	1.069	2.082	1.301	462.960
Separate, ≥2000, met	1.001	1.017	1.049	1.166	1.148	10.663
Separate, ≥2000, sulf	1.063	1.712	2.670	7.504	5.753	462.960
IPW, overall	1.108	1.339	1.616	4.386	2.522	2,041.518
IPW, met	1.108	1.310	1.514	1.945	1.914	257.934
IPW, sulf	2.229	5.145	6.829	18.382	12.467	2,041.518
IPW, ≥2000, overall	1.015	1.085	1.158	2.738	1.382	1,704.006
IPW, ≥2000, met	1.015	1.080	1.138	1.245	1.269	17.486
IPW, ≥2000, sulf	1.357	2.934	4.572	15.234	9.865	1,704.006

Table C.14: ATE weights summary for breast cancer incidence risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	1.004	1.051	1.165	2.053	1.586	211.118
Fewer, met	1.004	1.035	1.098	1.487	1.299	55.563
Fewer, sulf	1.010	1.209	1.593	3.404	2.623	211.118
Fewer, ≥2000, overall	1.002	1.029	1.093	2.142	1.360	331.844
Fewer, ≥2000, met	1.002	1.024	1.067	1.165	1.186	6.285
Fewer, ≥2000, sulf	1.130	2.084	3.194	7.926	6.373	331.844
Complete, overall	1.000	1.015	1.060	2.420	1.273	1,367.363
Complete, met	1.000	1.011	1.040	1.167	1.139	46.051
Complete, sulf	1.016	1.498	2.253	9.607	4.895	1,367.363
Complete, ≥2000, overall	1.000	1.013	1.046	2.359	1.189	1,535.591
Complete, ≥2000, met	1.000	1.011	1.036	1.115	1.114	13.739
Complete, ≥2000, sulf	1.074	2.034	3.342	12.776	7.801	1,535.591
Separate, overall	1.002	1.040	1.132	2.027	1.519	313.741
Separate, met	1.002	1.026	1.076	1.486	1.261	89.701
Separate, sulf	1.007	1.161	1.487	3.320	2.418	313.741
Separate, ≥2000, overall	1.001	1.021	1.069	2.082	1.301	462.960
Separate, ≥2000, met	1.001	1.017	1.049	1.166	1.148	10.663
Separate, ≥2000, sulf	1.063	1.712	2.670	7.504	5.753	462.960
IPW, overall	1.108	1.339	1.616	4.386	2.522	2,041.518
IPW, met	1.108	1.310	1.514	1.945	1.914	257.934
IPW, sulf	2.229	5.145	6.829	18.382	12.467	2,041.518
IPW, ≥2000, overall	1.015	1.085	1.158	2.738	1.382	1,704.006
IPW, ≥2000, met	1.015	1.080	1.138	1.245	1.269	17.486
IPW, ≥2000, sulf	1.357	2.934	4.572	15.234	9.865	1,704.006

Table C.15: ATT weights summary for breast cancer incidence risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.004	0.055	0.240	0.638	1	54.563
Fewer, met	0.004	0.035	0.098	0.487	0.299	54.563
Fewer, sulf	1	1	1	1	1	1
Fewer, $\geq$2000, overall	0.002	0.029	0.093	0.286	0.367	5.285
Fewer, $\geq$2000, met	0.002	0.024	0.067	0.165	0.186	5.285
Fewer, $\geq$2000, sulf	1	1	1	1	1	1
Complete, overall	0.0001	0.015	0.061	0.291	0.328	45.051
Complete, met	0.0001	0.011	0.040	0.167	0.139	45.051
Complete, sulf	1	1	1	1	1	1
Complete, $\geq$2000, overall	0.0001	0.013	0.046	0.209	0.190	12.739
Complete, $\geq$2000, met	0.0001	0.011	0.036	0.115	0.114	12.739
Complete, $\geq$2000, sulf	1	1	1	1	1	1
Separate, overall	0.002	0.042	0.209	0.638	1	88.701
Separate, met	0.002	0.026	0.076	0.486	0.261	88.701
Separate, sulf	1	1	1	1	1	1
Separate, $\geq$2000, overall	0.001	0.021	0.069	0.287	0.329	9.663
Separate, $\geq$2000, met	0.001	0.017	0.049	0.166	0.148	9.663
Separate, $\geq$2000, sulf	1	1	1	1	1	1
IPW, overall	0.0001	0.020	0.091	0.801	0.598	252.333
IPW, met	0.0001	0.015	0.059	0.393	0.228	252.333
IPW, sulf	1.230	2.059	2.730	3.139	3.640	29.733
IPW, $\geq$2000, overall	0.0001	0.014	0.051	0.267	0.219	16.213
IPW, $\geq$2000, met	0.0001	0.011	0.039	0.135	0.129	16.213
IPW, $\geq$2000, sulf	1.031	1.197	1.303	1.364	1.469	3.635

Table C.16: ATT weights summary for breast cancer incidence risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.010	1	1	1.414	1	210.118
Fewer, met	1	1	1	1	1	1
Fewer, sulf	0.010	0.209	0.593	2.404	1.623	210.118
Fewer, ≥2000, overall	0.130	1	1	1.857	1	330.844
Fewer, ≥2000, met	1	1	1	1	1	1
Fewer, ≥2000, sulf	0.130	1.084	2.194	6.926	5.373	330.844
Complete, overall	0.016	1	1	2.129	1	1,366.363
Complete, met	1	1	1	1	1	1
Complete, sulf	0.016	0.498	1.253	8.607	3.895	1,366.363
Complete, ≥2000, overall	0.074	1	1	2.150	1	1,534.591
Complete, ≥2000, met	1	1	1	1	1	1
Complete, ≥2000, sulf	0.074	1.034	2.342	11.776	6.801	1,534.591
Separate, overall	0.007	1	1	1.390	1	312.741
Separate, met 1	1	1	1	1	1	
Separate, sulf	0.007	0.161	0.487	2.320	1.418	312.741
Separate, ≥2000, overall	0.063	1	1	1.795	1	461.960
Separate, ≥2000, met	1	1	1	1	1	1
Separate, ≥2000, sulf	0.063	0.712	1.670	6.504	4.753	461.960
IPW, overall	0.104	1.288	1.471	3.585	1.794	2,040.025
IPW, met	1.108	1.281	1.436	1.552	1.661	16.107
IPW, sulf	0.104	1.665	3.536	15.243	9.251	2,040.025
IPW, ≥2000, overall	0.120	1.061	1.095	2.472	1.158	1,702.896
IPW, ≥2000, met	1.012	1.060	1.090	1.110	1.136	2.511
IPW, ≥2000, sulf	0.120	1.476	3.156	13.870	8.560	1,702.896

Table C.17: ATU weights summary for breast cancer incidence risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.010	1	1	1.414	1	210.118
Fewer, met	1	1	1	1	1	1
Fewer, sulf	0.010	0.209	0.593	2.404	1.623	210.118
Fewer, ≥2000, overall	0.130	1	1	1.857	1	330.844
Fewer, ≥2000, met	1	1	1	1	1	1
Fewer, ≥2000, sulf	0.130	1.084	2.194	6.926	5.373	330.844
Complete, overall	0.016	1	1	2.129	1	1,366.363
Complete, met	1	1	1	1	1	1
Complete, sulf	0.016	0.498	1.253	8.607	3.895	1,366.363
Complete, ≥2000, overall	0.074	1	1	2.150	1	1,534.591
Complete, ≥2000, met	1	1	1	1	1	1
Complete, ≥2000, sulf	0.074	1.034	2.342	11.776	6.801	1,534.591
Separate, overall	0.007	1	1	1.390	1	312.741
Separate, met	1	1	1	1	1	1
Separate, sulf	0.007	0.161	0.487	2.320	1.418	312.741
Separate, ≥2000, overall	0.063	1	1	1.795	1	461.960
Separate, ≥2000, met	1	1	1	1	1	1
Separate, ≥2000, sulf	0.063	0.712	1.670	6.504	4.753	461.960
IPW, overall	0.104	1.288	1.471	3.585	1.794	2,040.025
IPW, met	1.108	1.281	1.436	1.552	1.661	16.107
IPW, sulf	0.104	1.665	3.536	15.243	9.251	2,040.025
IPW, ≥2000, overall	0.120	1.061	1.095	2.472	1.158	1,702.896
IPW, ≥2000, met	1.012	1.060	1.090	1.110	1.136	2.511
IPW, ≥2000, sulf	0.120	1.476	3.156	13.870	8.560	1,702.896

Table C.18: ATU weights summary for breast cancer incidence risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.004	0.055	0.240	0.638	1	54.563
Fewer, met	0.004	0.035	0.098	0.487	0.299	54.563
Fewer, sulf	1	1	1	1	1	1
Fewer, ≥2000, overall	0.002	0.029	0.093	0.286	0.367	5.285
Fewer, ≥2000, met	0.002	0.024	0.067	0.165	0.186	5.285
Fewer, ≥2000, sulf	1	1	1	1	1	1
Complete, overall	0.0001	0.015	0.061	0.291	0.328	45.051
Complete, met	0.0001	0.011	0.040	0.167	0.139	45.051
Complete, sulf	1	1	1	1	1	1
Complete, ≥2000, overall	0.0001	0.013	0.046	0.209	0.190	12.739
Complete, ≥2000, met	0.0001	0.011	0.036	0.115	0.114	12.739
Complete, ≥2000, sulf	1	1	1	1	1	1
Separate, overall	0.002	0.042	0.209	0.638	1	88.701
Separate, met	0.002	0.026	0.076	0.486	0.261	88.701
Separate, sulf	1	1	1	1	1	1
Separate, ≥2000, overall	0.001	0.021	0.069	0.287	0.329	9.663
Separate, ≥2000, met	0.001	0.017	0.049	0.166	0.148	9.663
Separate, ≥2000, sulf	1	1	1	1	1	1
IPW, overall	0.0001	0.020	0.091	0.801	0.598	252.333
IPW, met	0.0001	0.015	0.059	0.393	0.228	252.333
IPW, sulf	1.230	2.059	2.730	3.139	3.640	29.733
IPW, ≥2000, overall	0.0001	0.014	0.051	0.267	0.219	16.213
IPW, ≥2000, met	0.0001	0.011	0.039	0.135	0.129	16.213
IPW, ≥2000, sulf	1.031	1.197	1.303	1.364	1.469	3.635

C.4   Weights summary for breast cancer mortality risks

Table C.19: ATE weights summary for breast cancer mortality risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	1.005	1.058	1.158	2.228	1.526	138.261
Fewer, met	1.005	1.046	1.103	1.296	1.250	20.953
Fewer, sulf	1.021	1.432	2.124	5.414	3.898	138.261
Fewer, $\geq$2000, overall	1.006	1.048	1.124	2.298	1.406	158.404
Fewer, $\geq$2000, met	1.006	1.042	1.090	1.179	1.206	4.131
Fewer, $\geq$2000, sulf	1.154	2.131	3.227	8.276	6.663	158.404
Complete, overall	1.000	1.020	1.071	3.901	1.246	1,601.023
Complete, met	1.000	1.016	1.049	1.139	1.141	5.242
Complete, sulf	1.062	1.626	2.490	21.194	5.959	1,601.023
Complete, $\geq$2000, overall	1.000	1.017	1.060	3.816	1.223	1,495.470
Complete, $\geq$2000, met	1.000	1.015	1.045	1.121	1.138	4.428
Complete, $\geq$2000, sulf	1.128	1.886	3.160	23.921	7.837	1,495.470
Separate, overall	1.003	1.042	1.123	2.278	1.447	245.460
Separate, met	1.003	1.032	1.079	1.314	1.212	29.352
Separate, sulf	1.015	1.307	1.838	5.577	3.283	245.460
Separate, $\geq$2000, overall	1.002	1.031	1.093	2.317	1.328	248.952
Separate, $\geq$2000, met	1.002	1.025	1.065	1.179	1.172	8.083
Separate, $\geq$2000, sulf	1.079	1.691	2.552	8.397	5.549	248.952
IPW, overall	1.140	1.320	1.517	6.026	2.084	2,155.788
IPW, met	1.140	1.303	1.451	1.653	1.737	10.151
IPW, sulf	2.440	4.426	6.060	33.400	12.256	2,155.788
IPW, $\geq$2000, overall	1.020	1.097	1.169	4.343	1.390	1,596.462
IPW, $\geq$2000, met	1.020	1.091	1.146	1.236	1.270	5.100
IPW, $\geq$2000, sulf	1.608	2.708	4.095	27.521	9.799	1,596.462

Table C.20: ATE weights summary for breast cancer mortality risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	1.005	1.058	1.158	2.228	1.526	138.261
Fewer, met	1.005	1.046	1.103	1.296	1.250	20.953
Fewer, sulf	1.021	1.432	2.124	5.414	3.898	138.261
Fewer, ≥2000, overall	1.006	1.048	1.124	2.298	1.406	158.404
Fewer, ≥2000, met	1.006	1.042	1.090	1.179	1.206	4.131
Fewer, ≥2000, sulf	1.154	2.131	3.227	8.276	6.663	158.404
Complete, overall	1.000	1.020	1.071	3.901	1.246	1,601.023
Complete, met	1.000	1.016	1.049	1.139	1.141	5.242
Complete, sulf	1.062	1.626	2.490	21.194	5.959	1,601.023
Complete, ≥2000, overall	1.000	1.017	1.060	3.816	1.223	1,495.470
Complete, ≥2000, met	1.000	1.015	1.045	1.121	1.138	4.428
Complete, ≥2000, sulf	1.128	1.886	3.160	23.921	7.837	1,495.470
Separate, overall	1.003	1.042	1.123	2.278	1.447	245.460
Separate, met	1.003	1.032	1.079	1.314	1.212	29.352
Separate, sulf	1.015	1.307	1.838	5.577	3.283	245.460
Separate, ≥2000, overall	1.002	1.031	1.093	2.317	1.328	248.952
Separate, ≥2000, met	1.002	1.025	1.065	1.179	1.172	8.083
Separate, ≥2000, sulf	1.079	1.691	2.552	8.397	5.549	248.952
IPW, overall	1.140	1.320	1.517	6.026	2.084	2,155.788
IPW, met	1.140	1.303	1.451	1.653	1.737	10.151
IPW, sulf	2.440	4.426	6.060	33.400	12.256	2,155.788
IPW, ≥2000, overall	1.020	1.097	1.169	4.343	1.390	1,596.462
IPW, ≥2000, met	1.020	1.091	1.146	1.236	1.270	5.100
IPW, ≥2000, sulf	1.608	2.708	4.095	27.521	9.799	1,596.462

Table C.21: ATT weights summary for breast cancer mortality risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.005	0.059	0.171	0.455	1	19.953
Fewer, met	0.005	0.046	0.103	0.296	0.250	19.953
Fewer, sulf	1	1	1	1	1	1
Fewer, ≥2000, overall	0.006	0.048	0.124	0.308	0.420	3.131
Fewer, ≥2000, met	0.006	0.042	0.090	0.179	0.206	3.131
Fewer, ≥2000, sulf	1	1	1	1	1	1
Complete, overall	0.0002	0.020	0.071	0.257	0.269	4.242
Complete, met	0.0002	0.016	0.049	0.139	0.141	4.242
Complete, sulf	1	1	1	1	1	1
Complete, ≥2000, overall	0.0001	0.017	0.060	0.225	0.225	3.428
Complete, ≥2000, met	0.0001	0.015	0.045	0.121	0.138	3.428
Complete, ≥2000, sulf	1	1	1	1	1	1
Separate, overall	0.003	0.042	0.135	0.469	1	28.352
Separate, met	0.003	0.032	0.079	0.314	0.212	28.352
Separate, sulf	1	1	1	1	1	1
Separate, ≥2000, overall	0.002	0.031	0.093	0.309	0.383	7.083
Separate, ≥2000, met	0.002	0.025	0.065	0.179	0.172	7.083
Separate, ≥2000, sulf	1	1	1	1	1	1
IPW, overall	0.0002	0.027	0.101	0.531	0.425	8.214
IPW, met	0.0002	0.022	0.067	0.227	0.211	8.214
IPW, sulf	1.316	1.754	2.417	2.430	2.805	7.340
IPW, ≥2000, overall	0.0001	0.018	0.066	0.282	0.255	3.949
IPW, ≥2000, met	0.0001	0.016	0.049	0.137	0.153	3.949
IPW, ≥2000, sulf	1.042	1.200	1.311	1.362	1.465	2.416

178

Table C.22: ATT weights summary for breast cancer mortality risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.021	1	1	1.773	1	137.261
Fewer, met	1	1	1	1	1	1
Fewer, sulf	0.021	0.432	1.124	4.414	2.898	137.261
Fewer, ≥2000, overall	0.154	1	1	1.989	1	157.404
Fewer, ≥2000, met	1	1	1	1	1	1
Fewer, ≥2000, sulf	0.154	1.131	2.227	7.276	5.663	157.404
Complete, overall	0.062	1	1	3.644	1	1,600.023
Complete, met	1	1	1	1	1	1
Complete, sulf	0.062	0.626	1.490	20.194	4.959	1,600.023
Complete, ≥2000, overall	0.128	1	1	3.591	1	1,494.470
Complete, ≥2000, met	1	1	1	1	1	1
Complete, ≥2000, sulf	0.128	0.886	2.160	22.921	6.837	1,494.470
Separate, overall	0.015	1	1	1.809	1	244.460
Separate, met	1	1	1	1	1	1
Separate, sulf	0.015	0.307	0.838	4.577	2.283	244.460
Separate, ≥2000, overall	0.079	1	1	2.008	1	247.952
Separate, ≥2000, met	1	1	1	1	1	1
Separate, ≥2000, sulf	0.079	0.691	1.552	7.397	4.549	247.952
IPW, overall	0.259	1.257	1.388	5.496	1.602	2,154.441
IPW, met	1.118	1.253	1.351	1.426	1.500	3.364
IPW, sulf	0.259	1.637	3.574	30.971	10.429	2,154.441
IPW, ≥2000, overall	0.216	1.062	1.090	4.061	1.138	1,595.395
IPW, ≥2000, met	1.015	1.062	1.085	1.099	1.122	1.447
IPW, ≥2000, sulf	0.216	1.260	2.869	26.158	8.558	1,595.395

Table C.23: ATU weights summary for breast cancer mortality risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.021	1	1	1.773	1	137.261
Fewer, met	1	1	1	1	1	1
Fewer, sulf	0.021	0.432	1.124	4.414	2.898	137.261
Fewer, ≥2000, overall	0.154	1	1	1.989	1	157.404
Fewer, ≥2000, met	1	1	1	1	1	1
Fewer, ≥2000, sulf	0.154	1.131	2.227	7.276	5.663	157.404
Complete, overall	0.062	1	1	3.644	1	1,600.023
Complete, met	1	1	1	1	1	1
Complete, sulf	0.062	0.626	1.490	20.194	4.959	1,600.023
Complete, ≥2000, overall	0.128	1	1	3.591	1	1,494.470
Complete, ≥2000, met	1.	1	1	1	1	1
Complete, ≥2000, sulf	0.128	0.886	2.160	22.921	6.837	1,494.470
Separate, overall	0.015	1	1	1.809	1	244.460
Separate, met	1	1	1	1	1	1
Separate, sulf	0.015	0.307	0.838	4.577	2.283	244.460
Separate, ≥2000, overall	0.079	1	1	2.008	1	247.952
Separate, ≥2000, met	1	1	1	1	1	1
Separate, ≥2000, sulf	0.079	0.691	1.552	7.397	4.549	247.952
IPW, overall	0.259	1.257	1.388	5.496	1.602	2,154.441
IPW, met	1.118	1.253	1.351	1.426	1.500	3.364
IPW, sulf	0.259	1.637	3.574	30.971	10.429	2,154.441
IPW, ≥2000, overall	0.216	1.062	1.090	4.061	1.138	1,595.395
IPW, ≥2000, met	1.015	1.062	1.085	1.099	1.122	1.447
IPW, ≥2000, sulf	0.216	1.260	2.869	26.158	8.558	1,595.395

Table C.24: ATU weights summary for breast cancer mortality risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.005	0.059	0.171	0.455	1	19.953
Fewer, met	0.005	0.046	0.103	0.296	0.250	19.953
Fewer, sulf	1	1	1	1	1	1
Fewer, $\geq$2000, overall	0.006	0.048	0.124	0.308	0.420	3.131
Fewer, $\geq$2000, met	0.006	0.042	0.090	0.179	0.206	3.131
Fewer, $\geq$2000, sulf	1	1	1	1	1	1
Complete, overall	0.0002	0.020	0.071	0.257	0.269	4.242
Complete, met	0.0002	0.016	0.049	0.139	0.141	4.242
Complete, sulf	1	1	1	1	1	1
Complete, $\geq$2000, overall	0.0001	0.017	0.060	0.225	0.225	3.428
Complete, $\geq$2000, met	0.0001	0.015	0.045	0.121	0.138	3.428
Complete, $\geq$2000, sulf	1	1	1	1	1	1
Separate, overall	0.003	0.042	0.135	0.469	1	28.352
Separate, met	0.003	0.032	0.079	0.314	0.212	28.352
Separate, sulf	1	1	1	1	1	1
Separate, $\geq$2000, overall	0.002	0.031	0.093	0.309	0.383	7.083
Separate, $\geq$2000, met	0.002	0.025	0.065	0.179	0.172	7.083
Separate, $\geq$2000, sulf	1	1	1	1	1	1
IPW, overall	0.0002	0.027	0.101	0.531	0.425	8.214
IPW, met	0.0002	0.022	0.067	0.227	0.211	8.214
IPW, sulf	1.316	1.754	2.417	2.430	2.805	7.340
IPW, $\geq$2000, overall	0.0001	0.018	0.066	0.282	0.255	3.949
IPW, $\geq$2000, met	0.0001	0.016	0.049	0.137	0.153	3.949
IPW, $\geq$2000, sulf	1.042	1.200	1.311	1.362	1.465	2.416

C.5 Weights summary for prostate cancer incidence risks

Table C.25: ATE weights summary for prostate cancer incidence risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	1.005	1.052	1.161	2.103	1.573	176.274
Fewer, met	1.005	1.038	1.099	1.482	1.295	78.364
Fewer, sulf	1.012	1.196	1.606	3.597	2.616	176.274
Fewer, ≥2000, overall	1.003	1.034	1.105	2.185	1.404	278.097
Fewer, ≥2000, met	1.003	1.029	1.075	1.172	1.205	6.249
Fewer, ≥2000, sulf	1.203	2.166	3.149	7.917	6.061	278.097
Complete, overall	1.000	1.019	1.070	2.519	1.306	3,166.679
Complete, met	1.000	1.015	1.047	1.183	1.153	33.651
Complete, sulf	1.007	1.500	2.250	9.647	4.683	3,166.679
Complete, ≥2000, overall	1.000	1.017	1.056	2.404	1.219	2,853.532
Complete, ≥2000, met	1.000	1.014	1.043	1.126	1.128	27.730
Complete, ≥2000, sulf	1.061	2.083	3.325	12.228	7.296	2,853.532
Separate, overall	1.002	1.039	1.125	2.076	1.499	316.688
Separate, met	1.002	1.027	1.076	1.492	1.253	68.977
Separate, sulf	1.006	1.142	1.478	3.482	2.411	316.688
Separate, ≥2000, overall	1.001	1.024	1.077	2.114	1.341	455.129
Separate, ≥2000, met	1.001	1.019	1.054	1.175	1.162	16.320
Separate, ≥2000, sulf	1.057	1.754	2.662	7.429	5.570	455.129
IPW, overall	1.126	1.329	1.585	4.330	2.517	3,988.474
IPW, met	1.126	1.300	1.482	1.893	1.864	284.413
IPW, sulf	2.077	4.643	6.176	17.331	11.236	3,988.474
IPW, ≥2000, overall	1.018	1.093	1.170	2.776	1.403	3,122.628
IPW, ≥2000, met	1.018	1.087	1.146	1.249	1.278	33.815
IPW, ≥2000, sulf	1.261	2.857	4.422	14.512	9.113	3,122.628

Table C.26: ATE weights summary for prostate cancer incidence risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	1.005	1.052	1.161	2.103	1.573	176.274
Fewer, met	1.005	1.038	1.099	1.482	1.295	78.364
Fewer, sulf	1.012	1.196	1.606	3.597	2.616	176.274
Fewer, $\geq$2000, overall	1.003	1.034	1.105	2.185	1.404	278.097
Fewer, $\geq$2000, met	1.003	1.029	1.075	1.172	1.205	6.249
Fewer, $\geq$2000, sulf	1.203	2.166	3.149	7.917	6.061	278.097
Complete, overall	1.000	1.019	1.070	2.519	1.306	3,166.679
Complete, met	1.000	1.015	1.047	1.183	1.153	33.651
Complete, sulf	1.007	1.500	2.250	9.647	4.683	3,166.679
Complete, $\geq$2000, overall	1.000	1.017	1.056	2.404	1.219	2,853.532
Complete, $\geq$2000, met	1.000	1.014	1.043	1.126	1.128	27.730
Complete, $\geq$2000, sulf	1.061	2.083	3.325	12.228	7.296	2,853.532
Separate, overall	1.002	1.039	1.125	2.076	1.499	316.688
Separate, met	1.002	1.027	1.076	1.492	1.253	68.977
Separate, sulf	1.006	1.142	1.478	3.482	2.411	316.688
Separate, $\geq$2000, overall	1.001	1.024	1.077	2.114	1.341	455.129
Separate, $\geq$2000, met	1.001	1.019	1.054	1.175	1.162	16.320
Separate, $\geq$2000, sulf	1.057	1.754	2.662	7.429	5.570	455.129
IPW, overall	1.126	1.329	1.585	4.330	2.517	3,988.474
IPW, met	1.126	1.300	1.482	1.893	1.864	284.413
IPW, sulf	2.077	4.643	6.176	17.331	11.236	3,988.474
IPW, $\geq$2000, overall	1.018	1.093	1.170	2.776	1.403	3,122.628
IPW, $\geq$2000, met	1.018	1.087	1.146	1.249	1.278	33.815
IPW, $\geq$2000, sulf	1.261	2.857	4.422	14.512	9.113	3,122.628

Table C.27: ATT weights summary for prostate cancer incidence risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.005	0.057	0.240	0.634	1	77.364
Fewer, met	0.005	0.038	0.099	0.482	0.295	77.364
Fewer, sulf	1	1	1	1	1	1
Fewer, $\geq$2000, overall	0.003	0.034	0.105	0.297	0.407	5.249
Fewer, $\geq$2000, met	0.003	0.029	0.075	0.172	0.205	5.249
Fewer, $\geq$2000, sulf	1	1	1	1	1	1
Complete, overall	0.0001	0.019	0.071	0.312	0.382	32.651
Complete, met	0.0001	0.015	0.047	0.183	0.153	32.651
Complete, sulf	1	1	1	1	1	1
Complete, $\geq$2000, overall	0.0001	0.017	0.056	0.227	0.221	26.730
Complete, $\geq$2000, met	0.0001	0.014	0.043	0.126	0.128	26.730
Complete, $\geq$2000, sulf	1	1	1	1	1	1
Separate, overall	0.002	0.042	0.200	0.641	1	67.977
Separate, met	0.002	0.042	0.200	0.641	1	67.977
Separate, sulf	1	1	1	1	1	1
Separate, $\geq$2000, overall	0.001	0.024	0.077	0.299	0.375	15.320
Separate, $\geq$2000, met	0.001	0.019	0.054	0.175	0.162	15.320
Separate, $\geq$2000, sulf	1	1	1	1	1	1
IPW, overall	0.0001	0.025	0.104	0.778	0.665	274.663
IPW, met	0.0001	0.019	0.067	0.392	0.241	274.663
IPW, sulf	1.211	1.966	2.535	2.839	3.175	22.719
IPW, $\geq$2000, overall	0.0001	0.018	0.062	0.280	0.253	32.595
IPW, $\geq$2000, met	0.0001	0.015	0.047	0.145	0.144	32.595
IPW, $\geq$2000, sulf	1.033	1.186	1.281	1.316	1.400	3.081

Table C.28: ATT weights summary for prostate cancer incidence risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.012	1	1	1.469	1	175.274
Fewer, met	1	1	1	1	1	1
Fewer, sulf	0.012	0.196	0.606	2.597	1.616	175.274
Fewer, $\geq$2000, overall	0.203	1	1	1.889	1	277.097
Fewer, $\geq$2000, met	1	1	1	1	1	1
Fewer, $\geq$2000, sulf	0.203	1.166	2.149	6.917	5.061	277.097
Complete, overall	0.007	1	1	2.207	1	3,165.679
Complete, met	1	1	1	1	1	1
Complete, sulf	0.007	0.500	1.250	8.647	3.683	3,165.679
Complete, $\geq$2000, overall	0.061	1	1	2.177	1	2,852.532
Complete, $\geq$2000, met	1	1	1	1	1	1
Complete, $\geq$2000, sulf	0.061	1.083	2.325	11.228	6.296	2,852.532
Separate, overall	0.006	1	1	1.435	1	315.688
Separate, met	1	1	1	1	1	1
Separate, sulf	0.006	0.142	0.478	2.482	1.411	315.688
Separate, $\geq$2000, overall	0.057	1	1	1.815	1	454.129
Separate, $\geq$2000, met	1	1	1	1	1	1
Separate, $\geq$2000, sulf	0.057	0.754	1.662	6.429	4.570	454.129
IPW, overall	0.042	1.274	1.433	3.552	1.715	3,987.215
IPW, met	1.120	1.271	1.395	1.501	1.604	10.014
IPW, sulf	0.042	1.525	3.297	14.492	8.230	3,987.215
IPW, $\geq$2000, overall	0.083	1.063	1.095	2.496	1.151	3,121.534
IPW, $\geq$2000, met	1.017	1.061	1.089	1.104	1.130	1.789
IPW, $\geq$2000, sulf	0.083	1.475	3.068	13.196	7.867	3,121.534

Table C.29: ATU weights summary for prostate cancer incidence risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.012	1	1	1.469	1	175.274
Fewer, met	1	1	1	1	1	1
Fewer, sulf	0.012	0.196	0.606	2.597	1.616	175.274
Fewer, ≥2000, overall	0.203	1	1	1.889	1	277.097
Fewer, ≥2000, met	1	1	1	1	1	1
Fewer, ≥2000, sulf	0.203	1.166	2.149	6.917	5.061	277.097
Complete, overall	0.007	1	1	2.207	1	3,165.679
Complete, met	1	1	1	1	1	1
Complete, sulf	0.007	0.500	1.250	8.647	3.683	3,165.679
Complete, ≥2000, overall	0.061	1	1	2.177	1	2,852.532
Complete, ≥2000, met	1	1	1	1	1	1
Complete, ≥2000, sulf	0.061	1.083	2.325	11.228	6.296	2,852.532
Separate, overall	0.006	1	1	1.435	1	315.688
Separate, met	1	1	1	1	1	1
Separate, sulf	0.006	0.142	0.478	2.482	1.411	315.688
Separate, ≥2000, overall	0.057	1	1	1.815	1	454.129
Separate, ≥2000, met	1	1	1	1	1	1
Separate, ≥2000, sulf	0.057	0.754	1.662	6.429	4.570	454.129
IPW, overall	0.042	1.274	1.433	3.552	1.715	3,987.215
IPW, met	1.120	1.271	1.395	1.501	1.604	10.014
IPW, sulf	0.042	1.525	3.297	14.492	8.230	3,987.215
IPW, ≥2000, overall	0.083	1.063	1.095	2.496	1.151	3,121.534
IPW, ≥2000, met	1.017	1.061	1.089	1.104	1.130	1.789
IPW, ≥2000, sulf	0.083	1.475	3.068	13.196	7.867	3,121.534

Table C.30: ATU weights summary for prostate cancer incidence risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.005	0.057	0.240	0.634	1	77.364
Fewer, met	0.005	0.038	0.099	0.482	0.295	77.364
Fewer, sulf	1	1	1	1	1	1
Fewer, ≥2000, overall	0.003	0.034	0.105	0.297	0.407	5.249
Fewer, ≥2000, met	0.003	0.029	0.075	0.172	0.205	5.249
Fewer, ≥2000, sulf	1	1	1	1	1	1
Complete, overall	0.0001	0.019	0.071	0.312	0.382	32.651
Complete, met	0.0001	0.015	0.047	0.183	0.153	32.651
Complete, sulf	1	1	1	1	1	1
Complete, ≥2000, overall	0.0001	0.017	0.056	0.227	0.221	26.730
Complete, ≥2000, met	0.0001	0.014	0.043	0.126	0.128	26.730
Complete, ≥2000, sulf	1	1	1	1	1	1
Separate, overall	0.002	0.042	0.200	0.641	1	67.977
Separate, met	0.002	0.027	0.076	0.492	0.253	67.977
Separate, sulf	1	1	1	1	1	1
Separate, ≥2000, overall	0.001	0.024	0.077	0.299	0.375	15.320
Separate, ≥2000, met	0.001	0.019	0.054	0.175	0.162	15.320
Separate, ≥2000, sulf	1	1	1	1	1	1
IPW, overall	0.0001	0.025	0.104	0.778	0.665	274.663
IPW, met	0.0001	0.019	0.067	0.392	0.241	274.663
IPW, sulf	1.211	1.966	2.535	2.839	3.175	22.719
IPW, ≥2000, overall	0.0001	0.018	0.062	0.280	0.253	32.595
IPW, ≥2000, met	0.0001	0.015	0.047	0.145	0.144	32.595
IPW, ≥2000, sulf	1.033	1.186	1.281	1.316	1.400	3.081

C.6 Weights summary for prostate cancer mortality risks

Table C.31: ATE weights summary for prostate cancer mortality risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	1.007	1.071	1.202	2.121	1.599	99.992
Fewer, met	1.007	1.056	1.131	1.351	1.305	18.374
Fewer, sulf	1.014	1.313	1.891	4.265	3.767	99.992
Fewer, ≥2000, overall	1.007	1.073	1.190	2.117	1.608	100.552
Fewer, ≥2000, met	1.007	1.058	1.124	1.242	1.279	4.701
Fewer, ≥2000, sulf	1.257	1.911	2.844	5.642	5.525	100.552
Complete, overall	1.001	1.040	1.119	2.025	1.422	98.188
Complete, met	1.001	1.031	1.077	1.209	1.218	6.373
Complete, sulf	1.063	1.601	2.678	5.799	5.322	98.188
Complete, ≥2000, overall	1.000	1.032	1.104	1.985	1.401	152.538
Complete, ≥2000, met	1.000	1.026	1.069	1.196	1.180	7.986
Complete, ≥2000, sulf	1.066	1.771	2.625	5.983	5.368	152.538
Separate, overall	1.004	1.056	1.159	2.019	1.526	54.482
Separate, met	1.004	1.042	1.103	1.369	1.271	40.035
Separate, sulf	1.011	1.235	1.711	3.829	3.318	54.482
Separate, ≥2000, overall	1.003	1.048	1.142	1.971	1.515	58.703
Separate, ≥2000, met	1.003	1.038	1.094	1.254	1.239	6.562
Separate, ≥2000, sulf	1.092	1.714	2.425	4.857	4.643	58.703
IPW, overall	1.145	1.303	1.532	3.187	2.499	182.082
IPW, met	1.145	1.274	1.431	1.706	1.756	15.479
IPW, sulf	2.232	3.908	5.597	10.038	9.886	182.082
IPW, ≥2000, overall	1.021	1.089	1.186	2.210	1.611	164.194
IPW, ≥2000, met	1.021	1.081	1.143	1.294	1.297	8.817
IPW, ≥2000, sulf	1.277	2.195	3.390	6.849	6.264	164.194

Table C.32: ATE weights summary for prostate cancer mortality risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	1.007	1.071	1.202	2.121	1.599	99.992
Fewer, met	1.007	1.056	1.131	1.351	1.305	18.374
Fewer, sulf	1.014	1.313	1.891	4.265	3.767	99.992
Fewer, $\geq$2000, overall	1.007	1.073	1.190	2.117	1.608	100.552
Fewer, $\geq$2000, met	1.007	1.058	1.124	1.242	1.279	4.701
Fewer, $\geq$2000, sulf	1.257	1.911	2.844	5.642	5.525	100.552
Complete, overall	1.001	1.040	1.119	2.025	1.422	98.188
Complete, met	1.063	1.601	2.678	5.799	5.322	98.188
Complete, sulf	1.001	1.031	1.077	1.209	1.218	6.373
Complete, $\geq$2000, overall	1.000	1.032	1.104	1.985	1.401	152.538
Complete, $\geq$2000, met	1.066	1.771	2.625	5.983	5.368	152.538
Complete, $\geq$2000, sulf	1.000	1.026	1.069	1.196	1.180	7.986
Separate, overall	1.004	1.056	1.159	2.019	1.526	54.482
Separate, met	1.011	1.235	1.711	3.829	3.318	54.482
Separate, sulf	1.004	1.042	1.103	1.369	1.271	40.035
Separate, $\geq$2000, overall	1.003	1.048	1.142	1.971	1.515	58.703
Separate, $\geq$2000, met	1.092	1.714	2.425	4.857	4.643	58.703
Separate, $\geq$2000, sulf	1.003	1.038	1.094	1.254	1.239	6.562
IPW, overall	1.145	1.303	1.532	3.187	2.499	182.082
IPW, met	1.145	1.274	1.431	1.706	1.756	15.479
IPW, sulf	2.232	3.908	5.597	10.038	9.886	182.082
IPW, $\geq$2000, overall	1.021	1.089	1.186	2.210	1.611	164.194
IPW, $\geq$2000, met	1.021	1.081	1.143	1.294	1.297	8.817
IPW, $\geq$2000, sulf	1.277	2.195	3.390	6.849	6.264	164.194

Table C.33: ATT weights summary for prostate cancer mortality risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.007	0.078	0.238	0.522	1	17.374
Fewer, met	0.007	0.056	0.131	0.351	0.305	17.374
Fewer, sulf	1	1	1	1	1	1
Fewer, ≥2000, overall	0.007	0.073	0.190	0.392	0.772	3.701
Fewer, ≥2000, met	0.007	0.058	0.124	0.242	0.279	3.701
Fewer, ≥2000, sulf	1	1	1	1	1	1
Complete, overall	0.001	0.040	0.121	0.349	0.562	5.373
Complete, met	0.001	0.031	0.077	0.209	0.218	5.373
Complete, sulf	1	1	1	1	1	1
Complete, ≥2000, overall	0.0004	0.032	0.104	0.328	0.466	6.986
Complete, ≥2000, met	0.0004	0.026	0.069	0.196	0.180	6.986
Complete, ≥2000, sulf	1	1	1	1	1	1
Separate, overall	0.004	0.061	0.201	0.536	1	39.035
Separate, met	0.004	0.042	0.103	0.369	0.271	39.035
Separate, sulf	1	1	1	1	1	1
Separate, ≥2000, overall	0.003	0.048	0.143	0.402	0.837	5.562
Separate, ≥2000, met	0.003	0.038	0.094	0.254	0.239	5.562
Separate, ≥2000, sulf	1	1	1	1	1	1
IPW, overall	0.001	0.050	0.163	0.656	0.884	13.050
IPW, met	0.001	0.039	0.101	0.340	0.306	13.050
IPW, sulf	1.246	1.620	1.977	2.118	2.486	6.453
IPW, ≥2000, overall	0.0004	0.034	0.111	0.382	0.517	7.713
IPW, ≥2000, met	0.0004	0.027	0.073	0.218	0.199	7.713
IPW, ≥2000, sulf	1.014	1.110	1.186	1.213	1.274	1.734

Table C.34: ATT weights summary for prostate cancer mortality risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.014	1	1	1.599	1	98.992
Fewer, met	1	1	1	1	1	1
Fewer, sulf	0.014	0.313	0.891	3.265	2.767	98.992
Fewer, ≥2000, overall	0.257	1	1	1.725	1	99.552
Fewer, ≥2000, met	1	1	1	1	1	1
Fewer, ≥2000, sulf	0.257	0.911	1.844	4.642	4.525	99.552
Complete, overall	0.063	1	1	1.675	1	97.188
Complete, met	1	1	1	1	1	1
Complete, sulf	0.063	0.601	1.678	4.799	4.322	97.188
Complete, ≥2000, overall	0.066	1	1	1.657	1	151.538
Complete, ≥2000, met	1	1	1	1	1	1
Complete, ≥2000, sulf	0.066	0.771	1.625	4.983	4.368	151.538
Separate, overall	0.011	1	1	1.483	1	53.482
Separate, met	1	1	1	1	1	1
Separate, sulf	0.011	0.235	0.711	2.829	2.318	53.482
Separate, ≥2000, overall	0.092	1	1	1.568	1	57.703
Separate, ≥2000, met	1	1	1	1	1	1
Separate, ≥2000, sulf	0.092	0.714	1.425	3.857	3.643	57.703
IPW, overall	0.173	1.222	1.327	2.531	1.553	180.228
IPW, met	1.098	1.219	1.311	1.366	1.438	3.786
IPW, sulf	0.173	1.491	3.593	7.920	7.795	180.228
IPW, ≥2000, overall	0.080	1.043	1.069	1.829	1.119	163.117
IPW, ≥2000, met	1.010	1.044	1.065	1.077	1.100	1.327
IPW, ≥2000, sulf	0.080	0.947	2.054	5.636	5.106	163.117

Table C.35: ATU weights summary for prostate cancer mortality risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.014	1	1	1.599	1	98.992
Fewer, met	1	1	1	1	1	1
Fewer, sulf	0.014	0.313	0.891	3.265	2.767	98.992
Fewer, ≥2000, overall	0.257	1	1	1.725	1	99.552
Fewer, ≥2000, met	1	1	1	1	1	1
Fewer, ≥2000, sulf	0.257	0.911	1.844	4.642	4.525	99.552
Complete, overall	0.063	1	1	1.675	1	97.188
Complete, met	1	1	1	1	1	1
Complete, sulf	0.063	0.601	1.678	4.799	4.322	97.188
Complete, ≥2000, overall	0.066	1	1	1.657	1	151.538
Complete, ≥2000, met	1	1	1	1	1	1
Complete, ≥2000, sulf	0.066	0.771	1.625	4.983	4.368	151.538
Separate, overall	0.011	1	1	1.483	1	53.482
Separate, met	1	1	1	1	1	1
Separate, sulf	0.011	0.235	0.711	2.829	2.318	53.482
Separate, ≥2000, overall	0.092	1	1	1.568	1	57.703
Separate, ≥2000, met	1	1	1	1	1	1
Separate, ≥2000, sulf	0.092	0.714	1.425	3.857	3.643	57.703
IPW, overall	0.173	1.222	1.327	2.531	1.553	180.228
IPW, met	1.098	1.219	1.311	1.366	1.438	3.786
IPW, sulf	0.173	1.491	3.593	7.920	7.795	180.228
IPW, ≥2000, overall	0.080	1.043	1.069	1.829	1.119	163.117
IPW, ≥2000, met	1.010	1.044	1.065	1.077	1.100	1.327
IPW, ≥2000, sulf	0.080	0.947	2.054	5.636	5.106	163.117

Table C.36: ATU weights summary for prostate cancer mortality risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.007	0.078	0.238	0.522	1	17.374
Fewer, met	0.007	0.056	0.131	0.351	0.305	17.374
Fewer, sulf	1	1	1	1	1	1
Fewer, ≥2000, overall	0.007	0.073	0.190	0.392	0.772	3.701
Fewer, ≥2000, met	0.007	0.058	0.124	0.242	0.279	3.701
Fewer, ≥2000, sulf	1	1	1	1	1	1
Complete, overall	0.001	0.040	0.121	0.349	0.562	5.373
Complete, met	0.001	0.031	0.077	0.209	0.218	5.373
Complete, sulf	1	1	1	1	1	1
Complete, ≥2000, overall	0.0004	0.032	0.104	0.328	0.466	6.986
Complete, ≥2000, met	0.0004	0.026	0.069	0.196	0.180	6.986
Complete, ≥2000, sulf	1	1	1	1	1	1
Separate, overall	0.004	0.061	0.201	0.536	1	39.035
Separate, met	0.004	0.042	0.103	0.369	0.271	39.035
Separate, sulf	1	1	1	1	1	1
Separate, ≥2000, overall	0.003	0.048	0.143	0.402	0.837	5.562
Separate, ≥2000, met	0.003	0.038	0.094	0.254	0.239	5.562
Separate, ≥2000, sulf	1	1	1	1	1	1
IPW, overall	0.001	0.050	0.163	0.656	0.884	13.050
IPW, met	0.001	0.039	0.101	0.340	0.306	13.050
IPW, sulf	1.246	1.620	1.977	2.118	2.486	6.453
IPW, ≥2000, overall	0.0004	0.034	0.111	0.382	0.517	7.713
IPW, ≥2000, met	0.0004	0.027	0.073	0.218	0.199	7.713
IPW, ≥2000, sulf	1.014	1.110	1.186	1.213	1.274	1.734

C.7 Weights summary for bowel cancer incidence risks

Table C.37: ATE weights summary for bowel cancer incidence risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	1.003	1.052	1.163	2.082	1.578	218.960
Fewer, met	1.003	1.037	1.099	1.479	1.294	84.099
Fewer, sulf	1.011	1.206	1.610	3.540	2.652	218.960
Fewer, $\geq$2000, overall	1.002	1.032	1.100	2.171	1.380	304.239
Fewer, $\geq$2000, met	1.002	1.027	1.072	1.169	1.197	7.775
Fewer, $\geq$2000, sulf	1.157	2.141	3.197	7.955	6.259	304.239
Complete, overall	1.000	1.017	1.066	2.483	1.291	2,533.461
Complete, met	1.000	1.013	1.045	1.175	1.147	48.014
Complete, sulf	1.008	1.506	2.271	9.716	4.837	2,533.461
Complete, $\geq$2000, overall	1.000	1.015	1.052	2.387	1.207	2,096.414
Complete, $\geq$2000, met	1.000	1.013	1.040	1.121	1.123	22.850
Complete, $\geq$2000, sulf	1.059	2.067	3.347	12.466	7.429	2,096.414
Separate, overall	1.002	1.040	1.128	2.056	1.508	310.178
Separate, met	1.002	1.027	1.076	1.485	1.256	96.764
Separate, sulf	1.006	1.152	1.490	3.437	2.432	310.178
Separate, $\geq$2000, overall	1.001	1.023	1.073	2.103	1.323	450.934
Separate, $\geq$2000, met	1.001	1.019	1.052	1.171	1.156	17.458
Separate, $\geq$2000, sulf	1.050	1.739	2.685	7.484	5.668	450.934
IPW, overall	1.118	1.334	1.598	4.360	2.507	3,147.775
IPW, met	1.118	1.304	1.496	1.909	1.882	339.257
IPW, sulf	2.197	4.849	6.447	17.905	11.788	3,147.775
IPW, $\geq$2000, overall	1.018	1.091	1.165	2.765	1.394	2,282.572
IPW, $\geq$2000, met	1.018	1.085	1.143	1.247	1.274	28.539
IPW, $\geq$2000, sulf	1.295	2.899	4.496	14.854	9.419	2,282.572

Table C.38: ATE weights summary for bowel cancer incidence risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	1.003	1.052	1.163	2.082	1.578	218.960
Fewer, met	1.003	1.037	1.099	1.479	1.294	84.099
Fewer, sulf	1.011	1.206	1.610	3.540	2.652	218.960
Fewer, ≥2000, overall	1.002	1.032	1.100	2.171	1.380	304.239
Fewer, ≥2000, met	1.002	1.027	1.072	1.169	1.197	7.775
Fewer, ≥2000, sulf	1.157	2.141	3.197	7.955	6.259	304.239
Complete, overall	1.000	1.017	1.066	2.483	1.291	2,533.461
Complete, met	1.000	1.013	1.045	1.175	1.147	48.014
Complete, sulf	1.008	1.506	2.271	9.716	4.837	2,533.461
Complete, ≥2000, overall	1.000	1.015	1.052	2.387	1.207	2,096.414
Complete, ≥2000, met	1.000	1.013	1.040	1.121	1.123	22.850
Complete, ≥2000, sulf	1.059	2.067	3.347	12.466	7.429	2,096.414
Separate, overall	1.002	1.040	1.128	2.056	1.508	310.178
Separate, met	1.002	1.027	1.076	1.485	1.256	96.764
Separate, sulf	1.006	1.152	1.490	3.437	2.432	310.178
Separate, ≥2000, overall	1.001	1.023	1.073	2.103	1.323	450.934
Separate, ≥2000, met	1.001	1.019	1.052	1.171	1.156	17.458
Separate, ≥2000, sulf	1.050	1.739	2.685	7.484	5.668	450.934
IPW, overall	1.118	1.334	1.598	4.360	2.507	3,147.775
IPW, met	1.118	1.304	1.496	1.909	1.882	339.257
IPW, sulf	2.197	4.849	6.447	17.905	11.788	3,147.775
IPW, ≥2000, overall	1.018	1.091	1.165	2.765	1.394	2,282.572
IPW, ≥2000, met	1.018	1.085	1.143	1.247	1.274	28.539
IPW, ≥2000, sulf	1.295	2.899	4.496	14.854	9.419	2,282.572

Table C.39: ATT weights summary for bowel cancer incidence risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.003	0.056	0.238	0.632	1	83.099
Fewer, met	0.003	0.037	0.099	0.479	0.294	83.099
Fewer, sulf	1	1	1	1	1	1
Fewer, ≥2000, overall	0.002	0.032	0.100	0.292	0.385	6.775
Fewer, ≥2000, met	0.002	0.027	0.072	0.169	0.197	6.775
Fewer, ≥2000, sulf	1	1	1	1	1	1
Complete, overall	0.0001	0.017	0.067	0.301	0.355	47.014
Complete, met	0.0001	0.013	0.045	0.175	0.147	47.014
Complete, sulf	1	1	1	1	1	1
Complete, ≥2000, overall	0.0001	0.015	0.052	0.219	0.208	21.850
Complete, ≥2000, met	0.0001	0.013	0.040	0.121	0.123	21.850
Complete, ≥2000, sulf	1	1	1	1	1	1
Separate, overall	0.002	0.042	0.201	0.636	1	95.764
Separate, met	0.002	0.027	0.076	0.485	0.256	95.764
Separate, sulf	1	1	1	1	1	1
Separate, ≥2000, overall	0.001	0.023	0.073	0.293	0.354	16.458
Separate, ≥2000, met	0.001	0.019	0.052	0.171	0.156	16.458
Separate, ≥2000, sulf	1	1	1	1	1	1
IPW, overall	0.0001	0.023	0.099	0.780	0.625	328.551
IPW, met	0.0001	0.018	0.064	0.388	0.235	328.551
IPW, sulf	1.197	2.010	2.621	2.949	3.386	26.426
IPW, ≥2000, overall	0.0001	0.016	0.058	0.274	0.238	27.290
IPW, ≥2000, met	0.0001	0.014	0.044	0.141	0.138	27.290
IPW, ≥2000, sulf	1.027	1.191	1.292	1.336	1.429	3.226

Table C.40: ATT weights summary for bowel cancer incidence risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.011	1	1	1.451	1	217.960
Fewer, met	1	1	1	1	1	1
Fewer, sulf	0.011	0.206	0.610	2.540	1.652	217.960
Fewer, ≥2000, overall	0.157	1	1	1.879	1	303.239
Fewer, ≥2000, met	1	1	1	1	1	1
Fewer, ≥2000, sulf	0.157	1.141	2.197	6.955	5.259	303.239
Complete, overall	0.008	1	1	2.182	1	2,532.461
Complete, met	1	1	1	1	1	1
Complete, sulf	0.008	0.506	1.271	8.716	3.837	2,532.461
Complete, ≥2000, overall	0.059	1	1	2.168	1	2,095.414
Complete, ≥2000, met	1	1	1	1	1	1
Complete, ≥2000, sulf	0.059	1.067	2.347	11.466	6.429	2,095.414
Separate, overall	0.006	1	1	1.421	1	309.178
Separate, met	1	1	1	1	1	1
Separate, sulf	0.006	0.152	0.490	2.437	1.432	309.178
Separate, ≥2000, overall	0.050	1	1	1.810	1	449.934
Separate, ≥2000, met	1	1	1	1	1	1
Separate, ≥2000, sulf	0.050	0.739	1.685	6.484	4.668	449.934
IPW, overall	0.053	1.273	1.440	3.580	1.756	3,146.533
IPW, met	1.114	1.269	1.418	1.521	1.623	14.095
IPW, sulf	0.053	1.595	3.412	14.956	8.721	3,146.533
IPW, ≥2000, overall	0.095	1.063	1.095	2.491	1.153	2,281.483
IPW, ≥2000, met	1.015	1.062	1.090	1.106	1.132	2.146
IPW, ≥2000, sulf	0.095	1.483	3.138	13.518	8.119	2,281.483

Table C.41: ATU weights summary for bowel cancer incidence risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.011	1	1	1.451	1	217.960
Fewer, met	1	1	1	1	1	1
Fewer, sulf	0.011	0.206	0.610	2.540	1.652	217.960
Fewer, ≥2000, overall	0.157	1	1	1.879	1	303.239
Fewer, ≥2000, met	1	1	1	1	1	1
Fewer, ≥2000, sulf	0.157	1.141	2.197	6.955	5.259	303.239
Complete, overall	0.008	1	1	2.182	1	2,532.461
Complete, met	1	1	1	1	1	1
Complete, sulf	0.008	0.506	1.271	8.716	3.837	2,532.461
Complete, ≥2000, overall	0.059	1	1	2.168	1	2,095.414
Complete, ≥2000, met	1	1	1	1	1	1
Complete, ≥2000, sulf	0.059	1.067	2.347	11.466	6.429	2,095.414
Separate, overall	0.006	1	1	1.421	1	309.178
Separate, met	1	1	1	1	1	1
Separate, sulf	0.006	0.152	0.490	2.437	1.432	309.178
Separate, ≥2000, overall	0.050	1	1	1.810	1	449.934
Separate, ≥2000, met	1	1	1	1	1	1
Separate, ≥2000, sulf	0.050	0.739	1.685	6.484	4.668	449.934
IPW, overall	0.053	1.273	1.440	3.580	1.756	3,146.533
IPW, met	1.114	1.269	1.418	1.521	1.623	14.095
IPW, sulf	0.053	1.595	3.412	14.956	8.721	3,146.533
IPW, ≥2000, overall	0.095	1.063	1.095	2.491	1.153	2,281.483
IPW, ≥2000, met	1.015	1.062	1.090	1.106	1.132	2.146
IPW, ≥2000, sulf	0.095	1.483	3.138	13.518	8.119	2,281.483

Table C.42: ATU weights summary for bowel cancer incidence risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.003	0.056	0.238	0.632	1	83.099
Fewer, met	0.003	0.037	0.099	0.479	0.294	83.099
Fewer, sulf	1	1	1	1	1	1
Fewer, ≥2000, overall	0.002	0.032	0.100	0.292	0.385	6.775
Fewer, ≥2000, met	0.002	0.027	0.072	0.169	0.197	6.775
Fewer, ≥2000, sulf	1	1	1	1	1	1
Complete, overall	0.0001	0.017	0.067	0.301	0.355	47.014
Complete, met	0.0001	0.013	0.045	0.175	0.147	47.014
Complete, sulf	1	1	1	1	1	1
Complete, ≥2000, overall	0.0001	0.015	0.052	0.219	0.208	21.850
Complete, ≥2000, met	0.0001	0.013	0.040	0.121	0.123	21.850
Complete, ≥2000, sulf	1	1	1	1	1	1
Separate, overall	0.002	0.042	0.201	0.636	1	95.764
Separate, met	0.002	0.027	0.076	0.485	0.256	95.764
Separate, sulf	1	1	1	1	1	1
Separate, ≥2000, overall	0.001	0.023	0.073	0.293	0.354	16.458
Separate, ≥2000, met	0.001	0.019	0.052	0.171	0.156	16.458
Separate, ≥2000, sulf	1	1	1	1	1	1
IPW, overall	0.0001	0.023	0.099	0.780	0.625	328.551
IPW, met	0.0001	0.018	0.064	0.388	0.235	328.551
IPW, sulf	1.197	2.010	2.621	2.949	3.386	26.426
IPW, ≥2000, overall	0.0001	0.016	0.058	0.274	0.238	27.290
IPW, ≥2000, met	1.027	1.191	1.292	1.336	1.429	3.226
IPW, ≥2000, sulf	0.0001	0.014	0.044	0.141	0.138	27.290

C.8 Weights summary for bowel cancer mortality risks

Table C.43: ATE weights summary for bowel cancer mortality risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	1.016	1.116	1.284	2.063	1.784	43.323
Fewer, met	1.016	1.094	1.199	1.492	1.462	35.186
Fewer, sulf	1.025	1.285	1.711	3.196	3.113	43.323
Fewer, ≥2000, overall	1.018	1.116	1.282	2.065	1.794	49.376
Fewer, ≥2000, met	1.018	1.093	1.186	1.304	1.381	3.545
Fewer, ≥2000, sulf	1.184	1.909	2.701	4.471	5.141	49.376
Complete, overall	1.001	1.061	1.182	2.533	1.570	359.774
Complete, met	1.001	1.050	1.119	1.258	1.274	5.144
Complete, sulf	1.023	1.493	2.088	6.984	4.935	359.774
Complete, ≥2000, overall	1.001	1.060	1.165	2.372	1.525	264.460
Complete, ≥2000, met	1.001	1.048	1.115	1.223	1.257	5.415
Complete, ≥2000, sulf	1.141	1.883	2.892	7.351	5.695	264.460
Separate, overall	1.008	1.081	1.227	2.063	1.674	47.262
Separate, met	1.009	1.070	1.147	1.542	1.385	47.262
Separate, sulf	1.008	1.207	1.555	3.097	2.715	37.021
Separate, ≥2000, overall	1.009	1.079	1.217	2.010	1.661	45.419
Separate, ≥2000, met	1.009	1.062	1.136	1.315	1.326	7.911
Separate, ≥2000, sulf	1.072	1.579	2.287	4.208	4.488	45.419
IPW, overall	1.125	1.392	1.758	4.141	3.007	522.220
IPW, met	1.125	1.348	1.540	1.905	2.005	24.291
IPW, sulf	2.040	3.761	5.481	11.950	9.029	522.220
IPW, ≥2000, overall	1.022	1.138	1.283	2.690	1.769	286.092
IPW, ≥2000, met	1.022	1.119	1.219	1.352	1.417	6.226
IPW, ≥2000, sulf	1.577	2.447	3.694	8.485	6.898	286.092

Table C.44: ATE weights summary for bowel cancer mortality risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	1.016	1.116	1.284	2.063	1.784	43.323
Fewer, met	1.016	1.094	1.199	1.492	1.462	35.186
Fewer, sulf	1.025	1.285	1.711	3.196	3.113	43.323
Fewer, ≥2000, overall	1.018	1.116	1.282	2.065	1.794	49.376
Fewer, ≥2000, met	1.018	1.093	1.186	1.304	1.381	3.545
Fewer, ≥2000, sulf	1.184	1.909	2.701	4.471	5.141	49.376
Complete, overall	1.001	1.061	1.182	2.533	1.570	359.774
Complete, met	1.001	1.050	1.119	1.258	1.274	5.144
Complete, sulf	1.023	1.493	2.088	6.984	4.935	359.774
Complete, ≥2000, overall	1.001	1.060	1.165	2.372	1.525	264.460
Complete, ≥2000, met	1.001	1.048	1.115	1.223	1.257	5.415
Complete, ≥2000, sulf	1.141	1.883	2.892	7.351	5.695	264.460
Separate, overall	1.008	1.081	1.227	2.063	1.674	47.262
Separate, met	1.009	1.070	1.147	1.542	1.385	47.262
Separate, sulf	1.008	1.207	1.555	3.097	2.715	37.021
Separate, ≥2000, overall	1.009	1.079	1.217	2.010	1.661	45.419
Separate, ≥2000, met	1.009	1.062	1.136	1.315	1.326	7.911
Separate, ≥2000, sulf	1.072	1.579	2.287	4.208	4.488	45.419
IPW, overall	1.125	1.392	1.758	4.141	3.007	522.220
IPW, met	1.125	1.348	1.540	1.905	2.005	24.291
IPW, sulf	2.040	3.761	5.481	11.950	9.029	522.220
IPW, ≥2000, overall	1.022	1.138	1.283	2.690	1.769	286.092
IPW, ≥2000, met	1.022	1.119	1.219	1.352	1.417	6.226
IPW, ≥2000, sulf	1.577	2.447	3.694	8.485	6.898	286.092

Table C.45: ATT weights summary for bowel cancer mortality risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.016	0.137	0.463	0.662	1	34.186
Fewer, met	0.016	0.094	0.199	0.492	0.462	34.186
Fewer, sulf	1	1	1	1	1	1
Fewer, ≥2000, overall	0.018	0.116	0.284	0.471	1	2.545
Fewer, ≥2000, met	0.018	0.093	0.186	0.304	0.381	2.545
Fewer, ≥2000, sulf	1	1	1	1	1	1
Complete, overall	0.001	0.062	0.186	0.423	0.994	4.144
Complete, met	0.001	0.050	0.119	0.258	0.274	4.144
Complete, sulf	1	1	1	1	1	1
Complete, ≥2000, overall	0.001	0.060	0.166	0.369	0.584	4.415
Complete, ≥2000, met	0.001	0.048	0.115	0.223	0.257	4.415
Complete, ≥2000, sulf	1	1	1	1	1	1
Separate, overall	0.009	0.100	0.389	0.696	1	46.262
Separate, met	0.009	0.070	0.147	0.542	0.385	46.262
Separate, sulf	1	1	1	1	1	1
Separate, ≥2000, overall	0.009	0.080	0.231	0.479	1	6.911
Separate, ≥2000, met	0.009	0.062	0.136	0.315	0.326	6.911
Separate, ≥2000, sulf	1	1	1	1	1	1
IPW, overall	0.001	0.083	0.264	0.898	1.429	19.441
IPW, met	0.001	0.065	0.165	0.454	0.439	19.441
IPW, sulf	1.242	1.639	1.969	2.448	2.600	15.404
IPW, ≥2000, overall	0.001	0.064	0.183	0.440	0.656	5.076
IPW, ≥2000, met	0.001	0.052	0.127	0.252	0.288	5.076
IPW, ≥2000, sulf	1.025	1.115	1.205	1.251	1.319	1.996

Table C.46: ATT weights summary for bowel cancer mortality risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.025	1	1	1.400	1	42.323
Fewer, met	1	1	1	1	1	1
Fewer, sulf	0.025	0.285	0.711	2.196	2.113	42.323
Fewer, ≥2000, overall	0.184	1	1	1.593	1	48.376
Fewer, ≥2000, met	1	1	1	1	1	1
Fewer, ≥2000, sulf	0.184	0.909	1.701	3.471	4.141	48.376
Complete, overall	0.023	1	1	2.110	1	358.774
Complete, met	1	1	1	1	1	1
Complete, sulf	0.023	0.493	1.088	5.984	3.935	358.774
Complete, ≥2000, overall	0.141	1	1	2.003	1	263.460
Complete, ≥2000, met	1	1	1	1	1	1
Complete, ≥2000, sulf	0.141	0.883	1.892	6.351	4.695	263.460
Separate, overall	0.008	1	1	1.367	1	36.021
Separate, met	1	1	1	1	1	1
Separate, sulf	0.008	0.207	0.555	2.097	1.715	36.021
Separate, ≥2000, overall	0.072	1	1	1.530	1	44.419
Separate, ≥2000, met	1	1	1	1	1	1
Separate, ≥2000, sulf	0.072	0.579	1.287	3.208	3.488	44.419
IPW, overall	0.246	1.265	1.409	3.243	1.695	520.768
IPW, met	1.098	1.264	1.371	1.450	1.535	4.850
IPW, sulf	0.246	1.369	2.694	9.503	6.654	520.768
IPW, ≥2000, overall	0.194	1.046	1.082	2.250	1.170	285.010
IPW, ≥2000, met	1.009	1.046	1.075	1.100	1.122	1.550
IPW, ≥2000, sulf	0.194	1.111	2.398	7.234	5.622	285.010

Table C.47: ATU weights summary for bowel cancer mortality risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.025	1	1	1.400	1	42.323
Fewer, met	1	1	1	1	1	1
Fewer, sulf	0.025	0.285	0.711	2.196	2.113	42.323
Fewer, ≥2000, overall	0.184	1	1	1.593	1	48.376
Fewer, ≥2000, met	1	1	1	1	1	1
Fewer, ≥2000, sulf	0.184	0.909	1.701	3.471	4.141	48.376
Complete, overall	0.023	1	1	2.110	1	358.774
Complete, met	1	1	1	1	1	1
Complete, sulf	0.023	0.493	1.088	5.984	3.935	358.774
Complete, ≥2000, overall	0.141	1	1	2.003	1	263.460
Complete, ≥2000, met	1	1	1	1	1	1
Complete, ≥2000, sulf	0.141	0.883	1.892	6.351	4.695	263.460
Separate, overall	0.008	1	1	1.367	1	36.021
Separate, met	1	1	1	1	1	1
Separate, sulf	0.008	0.207	0.555	2.097	1.715	36.021
Separate, ≥2000, overall	0.072	1	1	1.530	1	44.419
Separate, ≥2000, met	1	1	1	1	1	1
Separate, ≥2000, sulf	0.072	0.579	1.287	3.208	3.488	44.419
IPW, overall	0.246	1.265	1.409	3.243	1.695	520.768
IPW, met	1.098	1.264	1.371	1.450	1.535	4.850
IPW, sulf	0.246	1.369	2.694	9.503	6.654	520.768
IPW, ≥2000, overall	0.194	1.046	1.082	2.250	1.170	285.010
IPW, ≥2000, met	1.009	1.046	1.075	1.100	1.122	1.550
IPW, ≥2000, sulf	0.194	1.111	2.398	7.234	5.622	285.010

Table C.48: ATU weights summary for bowel cancer mortality risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.016	0.137	0.463	0.662	1	34.186
Fewer, met	0.016	0.094	0.199	0.492	0.462	34.186
Fewer, sulf	1	1	1	1	1	1
Fewer, ≥2000, overall	0.018	0.116	0.284	0.471	1	2.545
Fewer, ≥2000, met	0.018	0.093	0.186	0.304	0.381	2.545
Fewer, ≥2000, sulf	1	1	1	1	1	1
Complete, overall	0.001	0.062	0.186	0.423	0.994	4.144
Complete, met	0.001	0.050	0.119	0.258	0.274	4.144
Complete, sulf	1	1	1	1	1	1
Complete, ≥2000, overall	0.001	0.060	0.166	0.369	0.584	4.415
Complete, ≥2000, met	0.001	0.048	0.115	0.223	0.257	4.415
Complete, ≥2000, sulf	1	1	1	1	1	1
Separate, overall	0.009	0.100	0.389	0.696	1	46.262
Separate, met	0.009	0.070	0.147	0.542	0.385	46.262
Separate, sulf	1	1	1	1	1	1
Separate, ≥2000, overall	0.009	0.080	0.231	0.479	1	6.911
Separate, ≥2000, met	0.009	0.062	0.136	0.315	0.326	6.911
Separate, ≥2000, sulf	1	1	1	1	1	1
IPW, overall	0.001	0.083	0.264	0.898	1.429	19.441
IPW, met	0.001	0.065	0.165	0.454	0.439	19.441
IPW, sulf	1.242	1.639	1.969	2.448	2.600	15.404
IPW, ≥2000, overall	0.001	0.064	0.183	0.440	0.656	5.076
IPW, ≥2000, met	0.001	0.052	0.127	0.252	0.288	5.076
IPW, ≥2000, sulf	1.025	1.115	1.205	1.251	1.319	1.996

C.9 Weights summary for lung cancer incidence risks

Table C.49: ATE weights summary for lung cancer incidence risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	1.004	1.053	1.163	2.083	1.579	218.193
Fewer, met	1.004	1.037	1.099	1.479	1.295	84.336
Fewer, sulf	1.011	1.206	1.610	3.540	2.652	218.193
Fewer, $\geq$2000, overall	1.002	1.033	1.101	2.170	1.383	305.727
Fewer, $\geq$2000, met	1.002	1.027	1.072	1.170	1.197	7.838
Fewer, $\geq$2000, sulf	1.156	2.139	3.193	7.919	6.254	305.727
Complete, overall	1.000	1.017	1.067	2.500	1.293	2,464.718
Complete, met	1.000	1.014	1.045	1.175	1.148	47.864
Complete, sulf	1.008	1.506	2.272	9.790	4.849	2,464.718
Complete, $\geq$2000, overall	1.000	1.015	1.053	2.408	1.209	2,564.806
Complete, $\geq$2000, met	1.000	1.013	1.041	1.122	1.124	22.749
Complete, $\geq$2000, sulf	1.059	2.065	3.343	12.592	7.409	2,564.806
Separate, overall	1.002	1.040	1.128	2.058	1.509	305.362
Separate, met	1.002	1.027	1.077	1.485	1.257	96.847
Separate, sulf	1.006	1.152	1.491	3.441	2.433	305.362
Separate, $\geq$2000, overall	1.001	1.023	1.074	2.104	1.325	447.830
Separate, $\geq$2000, met	1.001	1.019	1.052	1.172	1.157	17.464
Separate, $\geq$2000, sulf	1.050	1.738	2.682	7.462	5.652	447.830
IPW, overall	1.118	1.334	1.599	4.381	2.513	3,073.533
IPW, met	1.118	1.304	1.496	1.908	1.883	336.721
IPW, sulf	2.195	4.842	6.439	17.990	11.751	3,073.533
IPW, $\geq$2000, overall	1.018	1.091	1.166	2.789	1.396	2,844.448
IPW, $\geq$2000, met	1.018	1.085	1.144	1.248	1.275	28.413
IPW, $\geq$2000, sulf	1.293	2.897	4.493	14.984	9.380	2,844.448

Table C.50: ATE weights summary for lung cancer incidence risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	1.004	1.053	1.163	2.083	1.579	218.193
Fewer, met	1.004	1.037	1.099	1.479	1.295	84.336
Fewer, sulf	1.011	1.206	1.610	3.540	2.652	218.193
Fewer, $\geq$2000, overall	1.002	1.033	1.101	2.170	1.383	305.727
Fewer, $\geq$2000, met	1.002	1.027	1.072	1.170	1.197	7.838
Fewer, $\geq$2000, sulf	1.156	2.139	3.193	7.919	6.254	305.727
Complete, overall	1.000	1.017	1.067	2.500	1.293	2,464.718
Complete, met	1.000	1.014	1.045	1.175	1.148	47.864
Complete, sulf	1.008	1.506	2.272	9.790	4.849	2,464.718
Complete, $\geq$2000, overall	1.000	1.015	1.053	2.408	1.209	2,564.806
Complete, $\geq$2000, met	1.000	1.013	1.041	1.122	1.124	22.749
Complete, $\geq$2000, sulf	1.059	2.065	3.343	12.592	7.409	2,564.806
Separate, overall	1.002	1.040	1.128	2.058	1.509	305.362
Separate, met	1.002	1.027	1.077	1.485	1.257	96.847
Separate, sulf	1.006	1.152	1.491	3.441	2.433	305.362
Separate, $\geq$2000, overall	1.001	1.023	1.074	2.104	1.325	447.830
Separate, $\geq$2000, met	1.001	1.019	1.052	1.172	1.157	17.464
Separate, $\geq$2000, sulf	1.050	1.738	2.682	7.462	5.652	447.830
IPW, overall	1.118	1.334	1.599	4.381	2.513	3,073.533
IPW, met	1.118	1.304	1.496	1.908	1.883	336.721
IPW, sulf	2.195	4.842	6.439	17.990	11.751	3,073.533
IPW, $\geq$2000, overall	1.018	1.091	1.166	2.789	1.396	2,844.448
IPW, $\geq$2000, met	1.018	1.085	1.144	1.248	1.275	28.413
IPW, $\geq$2000, sulf	1.293	2.897	4.493	14.984	9.380	2,844.448

Table C.51: ATT weights summary for lung cancer incidence risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.004	0.057	0.239	0.632	1	83.336
Fewer, met	0.004	0.037	0.099	0.479	0.295	83.336
Fewer, sulf	1	1	1	1	1	1
Fewer, $\geq$2000, overall	0.002	0.033	0.101	0.293	0.389	6.838
Fewer, $\geq$2000, met	0.002	0.027	0.072	0.170	0.197	6.838
Fewer, $\geq$2000, sulf	1	1	1	1	1	1
Complete, overall	0.0001	0.017	0.068	0.302	0.358	46.864
Complete, met	0.0001	0.014	0.045	0.175	0.148	46.864
Complete, sulf	1	1	1	1	1	1
Complete, $\geq$2000, overall	0.0001	0.015	0.053	0.220	0.210	21.749
Complete, $\geq$2000, met	0.0001	0.013	0.041	0.122	0.124	21.749
Complete, $\geq$2000, sulf	1	1	1	1	1	1
Separate, overall	0.002	0.042	0.202	0.636	1	95.847
Separate, met	0.002	0.027	0.077	0.485	0.257	95.847
Separate, sulf	1	1	1	1	1	1
Separate, $\geq$2000, overall	0.001	0.023	0.074	0.294	0.357	16.464
Separate, $\geq$2000, met	0.001	0.019	0.052	0.172	0.157	16.464
Separate, $\geq$2000, sulf	1	1	1	1	1	1
IPW, overall	0.0001	0.023	0.100	0.781	0.632	326.015
IPW, met	0.0001	0.018	0.064	0.388	0.237	326.015
IPW, sulf	1.197	2.001	2.607	2.944	3.382	26.426
IPW, $\geq$2000, overall	0.0001	0.016	0.059	0.275	0.240	27.164
IPW, $\geq$2000, met	0.0001	0.014	0.045	0.142	0.139	27.164
IPW, $\geq$2000, sulf	1.027	1.191	1.291	1.335	1.428	3.226

208

Table C.52: ATT weights summary for lung cancer incidence risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.011	1	1	1.451	1	217.193
Fewer, met	1	1	1	1	1	1
Fewer, sulf	0.011	0.206	0.610	2.540	1.652	217.193
Fewer, ≥2000, overall	0.011	0.206	0.610	2.540	1.652	217.193
Fewer, ≥2000, met	1	1	1	1	1	1
Fewer, ≥2000, sulf	0.156	1	1	1.877	1	304.727
Complete, overall	0.008	1	1	2.198	1	2,463.718
Complete, met	1	1	1	1	1	1
Complete, sulf	0.008	0.506	1.272	8.790	3.849	2,463.718
Complete, ≥2000, overall	0.059	1	1	2.188	1	2,563.806
Complete, ≥2000, met	1	1	1	1	1	1
Complete, ≥2000, sulf	0.059	1.065	2.343	11.592	6.409	2,563.806
Separate, overall	0.006	1	1	1.422	1	304.362
Separate, met	1	1	1	1	1	1
Separate, sulf	0.006	0.152	0.491	2.441	1.433	304.362
Separate, ≥2000, overall	0.050	1	1	1.810	1	446.830
Separate, ≥2000, met	1	1	1	1	1	1
Separate, ≥2000, sulf	0.050	0.738	1.682	6.462	4.652	446.830
IPW, overall	0.054	1.272	1.440	3.600	1.756	3,072.028
IPW, met	1.114	1.269	1.417	1.521	1.622	14.095
IPW, sulf	0.054	1.593	3.414	15.046	8.697	3,072.028
IPW, ≥2000, overall	0.096	1.063	1.095	2.513	1.154	2,843.339
IPW, ≥2000, met	1.015	1.062	1.089	1.106	1.132	2.146
IPW, ≥2000, sulf	0.096	1.485	3.139	13.649	8.102	2,843.339

Table C.53: ATU weights summary for lung cancer incidence risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.011	1	1	1.451	1	217.193
Fewer, met	1	1	1	1	1	1
Fewer, sulf	0.011	0.206	0.610	2.540	1.652	217.193
Fewer, $\geq$2000, overall	0.156	1	1	1.877	1	304.727
Fewer, $\geq$2000, met	1	1	1	1	1	1
Fewer, $\geq$2000, sulf	0.156	1.139	2.193	6.919	5.254	304.727
Complete, overall	0.008	1	1	2.198	1	2,463.718
Complete, met	1	1	1	1	1	1
Complete, sulf	0.008	0.506	1.272	8.790	3.849	2,463.718
Complete, $\geq$2000, overall	0.059	1	1	2.188	1	2,563.806
Complete, $\geq$2000, met	1	1	1	1	1	1
Complete, $\geq$2000, sulf	0.059	1.065	2.343	11.592	6.409	2,563.806
Separate, overall	0.006	1	1	1.422	1	304.362
Separate, met	1	1	1	1	1	1
Separate, sulf	0.006	0.152	0.491	2.441	1.433	304.362
Separate, $\geq$2000, overall	0.050	1	1	1.810	1	446.830
Separate, $\geq$2000, met	1	1	1	1	1	1
Separate, $\geq$2000, sulf	0.050	0.738	1.682	6.462	4.652	446.830
IPW, overall	0.054	1.272	1.440	3.600	1.756	3,072.028
IPW, met	1.114	1.269	1.417	1.521	1.622	14.095
IPW, sulf	0.054	1.593	3.414	15.046	8.697	3,072.028
IPW, $\geq$2000, overall	0.096	1.063	1.095	2.513	1.154	2,843.339
IPW, $\geq$2000, met	1.015	1.062	1.089	1.106	1.132	2.146
IPW, $\geq$2000, sulf	0.096	1.485	3.139	13.649	8.102	2,843.339

Table C.54: ATU weights summary for lung cancer incidence risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.004	0.057	0.239	0.632	1	83.336
Fewer, met	0.004	0.037	0.099	0.479	0.295	83.336
Fewer, sulf	1	1	1	1	1	1
Fewer, ≥2000, overall	0.002	0.033	0.101	0.293	0.389	6.838
Fewer, ≥2000, met	0.002	0.027	0.072	0.170	0.197	6.838
Fewer, ≥2000, sulf	1	1	1	1	1	1
Complete, overall	0.0001	0.017	0.068	0.302	0.358	46.864
Complete, met	0.0001	0.014	0.045	0.175	0.148	46.864
Complete, sulf	1	1	1	1	1	1
Complete, ≥2000, overall	0.0001	0.015	0.053	0.220	0.210	21.749
Complete, ≥2000, met	0.0001	0.013	0.041	0.122	0.124	21.749
Complete, ≥2000, sulf	1	1	1	1	1	1
Separate, overall	0.002	0.042	0.202	0.636	1	95.847
Separate, met	0.002	0.027	0.077	0.485	0.257	95.847
Separate, sulf	1	1	1	1	1	1
Separate, ≥2000, overall	0.001	0.023	0.074	0.294	0.357	16.464
Separate, ≥2000, met	0.001	0.019	0.052	0.172	0.157	16.464
Separate, ≥2000, sulf	1	1	1	1	1	1
IPW, overall	0.0001	0.023	0.100	0.781	0.632	326.015
IPW, met	0.0001	0.018	0.064	0.388	0.237	326.015
IPW, sulf	1.197	2.001	2.607	2.944	3.382	26.426
IPW, ≥2000, overall	0.0001	0.016	0.059	0.275	0.240	27.164
IPW, ≥2000, met	0.0001	0.014	0.045	0.142	0.139	27.164
IPW, ≥2000, sulf	1.027	1.191	1.291	1.335	1.428	3.226

C.10 Weights summary for lung cancer mortality risks

Table C.55: ATE weights summary for lung cancer mortality risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	1.046	1.205	1.434	2.026	2.056	16.218
Fewer, met	1.046	1.165	1.301	1.603	1.636	15.682
Fewer, sulf	1.059	1.445	1.994	2.761	3.034	16.218
Fewer, ≥2000, overall	1.041	1.208	1.413	2.029	2.047	16.516
Fewer, ≥2000, met	1.041	1.150	1.276	1.414	1.525	3.936
Fewer, ≥2000, sulf	1.251	1.903	2.451	3.500	3.854	16.516
Complete, overall	1	1.037	1.166	1.900	1.536	30.227
Complete, met	1	1.026	1.100	1.258	1.272	3.950
Complete, sulf	1.058	1.400	2.111	4.119	4.396	30.227
Complete, ≥2000, overall	1	1.029	1.129	2.031	1.456	50.971
Complete, ≥2000, met	1	1.023	1.089	1.240	1.228	3.921
Complete, ≥2000, sulf	1.071	1.396	2.362	5.010	3.700	50.971
Separate, overall	1.007	1.096	1.274	1.994	1.721	17.756
Separate, met	1.007	1.066	1.179	1.592	1.549	17.756
Separate, sulf	1.015	1.173	1.547	2.692	2.224	17.067
Separate, ≥2000, overall	1.010	1.089	1.235	2.039	1.768	38.147
Separate, ≥2000, met	1.010	1.067	1.149	1.348	1.435	4.736
Separate, ≥2000, sulf	1.045	1.268	1.835	3.690	3.347	38.147
IPW, overall	1.150	1.485	1.918	3.697	3.331	58.937
IPW, met	1.150	1.409	1.684	2.018	2.228	9.018
IPW, sulf	2.031	4.154	5.844	9.502	11.141	58.937
IPW, ≥2000, overall	1.020	1.119	1.270	2.328	1.729	56.418
IPW, ≥2000, met	1.020	1.100	1.196	1.374	1.437	4.482
IPW, ≥2000, sulf	1.280	1.848	2.896	5.924	4.773	56.418

Table C.56: ATE weights summary for lung cancer mortality risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	1.046	1.205	1.434	2.026	2.056	16.218
Fewer, met	1.046	1.165	1.301	1.603	1.636	15.682
Fewer, sulf	1.059	1.445	1.994	2.761	3.034	16.218
Fewer, ≥2000, overall	1.041	1.208	1.413	2.029	2.047	16.516
Fewer, ≥2000, met	1.041	1.150	1.276	1.414	1.525	3.936
Fewer, ≥2000, sulf	1.251	1.903	2.451	3.500	3.854	16.516
Complete, overall	1	1.037	1.166	1.900	1.536	30.227
Complete, met	1	1.026	1.100	1.258	1.272	3.950
Complete, sulf	1.058	1.400	2.111	4.119	4.396	30.227
Complete, ≥2000, overall	1	1.029	1.129	2.031	1.456	50.971
Complete, ≥2000, met	1	1.023	1.089	1.240	1.228	3.921
Complete, ≥2000, sulf	1.071	1.396	2.362	5.010	3.700	50.971
Separate, overall	1.007	1.096	1.274	1.994	1.721	17.756
Separate, met	1.007	1.066	1.179	1.592	1.549	17.756
Separate, sulf	1.015	1.173	1.547	2.692	2.224	17.067
Separate, ≥2000, overall	1.010	1.089	1.235	2.039	1.768	38.147
Separate, ≥2000, met	1.010	1.067	1.149	1.348	1.435	4.736
Separate, ≥2000, sulf	1.045	1.268	1.835	3.690	3.347	38.147
IPW, overall	1.150	1.485	1.918	3.697	3.331	58.937
IPW, met	1.150	1.409	1.684	2.018	2.228	9.018
IPW, sulf	2.031	4.154	5.844	9.502	11.141	58.937
IPW, ≥2000, overall	1.020	1.119	1.270	2.328	1.729	56.418
IPW, ≥2000, met	1.020	1.100	1.196	1.374	1.437	4.482
IPW, ≥2000, sulf	1.280	1.848	2.896	5.924	4.773	56.418

Table C.57: ATT weights summary for lung cancer mortality risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.046	0.232	0.731	0.748	1	14.682
Fewer, met	0.046	0.165	0.301	0.603	0.636	14.682
Fewer, sulf	1	1	1	1	1	1
Fewer, ≥2000, overall	0.041	0.208	0.445	0.587	1	2.936
Fewer, ≥2000, met	0.041	0.150	0.276	0.414	0.525	2.936
Fewer, ≥2000, sulf	1	1	1	1	1	1
Complete, overall	0	0.037	0.177	0.424	1	2.950
Complete, met	0	0.026	0.100	0.258	0.272	2.950
Complete, sulf	1	1	1	1	1	1
Complete, ≥2000, overall	0	0.029	0.138	0.399	1	2.921
Complete, ≥2000, met	0	0.023	0.089	0.240	0.228	2.921
Complete, ≥2000, sulf	1	1	1	1	1	1
Separate, overall	0.007	0.120	0.624	0.741	1	16.756
Separate, met	0.007	0.066	0.179	0.592	0.549	16.756
Separate, sulf	1	1	1	1	1	1
Separate, ≥2000, overall	0.010	0.096	0.352	0.540	1	3.736
Separate, ≥2000, met	0.010	0.067	0.149	0.348	0.435	3.736
Separate, ≥2000, sulf	1	1	1	1	1	1
IPW, overall	0	0.050	0.272	0.970	1.601	8.217
IPW, met	0	0.037	0.158	0.436	0.459	6.735
IPW, sulf	1.487	1.974	2.277	2.815	3.160	8.217
IPW, ≥2000, overall	0	0.033	0.160	0.478	1.049	3.339
IPW, ≥2000, met	0	0.024	0.100	0.269	0.265	3.339
IPW, ≥2000, sulf	1.042	1.096	1.146	1.263	1.373	2.394

Table C.58: ATT weights summary for lung cancer mortality risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.059	1	1	1.278	1	15.218
Fewer, met	1	1	1	1	1	1
Fewer, sulf	0.059	0.445	0.994	1.761	2.034	15.218
Fewer, $\geq$2000, overall	0.251	1	1	1.443	1	15.516
Fewer, $\geq$2000, met	1	1	1	1	1	1
Fewer, $\geq$2000, sulf	0.251	0.903	1.451	2.500	2.854	15.516
Complete, overall	0.058	1	1	1.475	1	29.227
Complete, met	1	1	1	1	1	1
Complete, sulf	0.058	0.400	1.111	3.119	3.396	29.227
Complete, $\geq$2000, overall	0.071	1	1	1.631	1	49.971
Complete, $\geq$2000, met	1	1	1	1	1	1
Complete, $\geq$2000, sulf	0.071	0.396	1.362	4.010	2.700	49.971
Separate, overall	0.015	1	1	1.253	1	16.067
Separate, met	1	1	1	1	1	1
Separate, sulf	0.015	0.173	0.547	1.692	1.224	16.067
Separate, $\geq$2000, overall	0.045	1	1	1.499	1	37.147
Separate, $\geq$2000, met	1	1	1	1	1	1
Separate, $\geq$2000, sulf	0.045	0.268	0.835	2.690	2.347	37.147
IPW, overall	0.218	1.298	1.512	2.727	1.927	56.987
IPW, met	1.148	1.298	1.478	1.582	1.670	4.899
IPW, sulf	0.218	1.306	2.997	6.687	5.380	56.987
IPW, $\geq$2000, overall	0.099	1.047	1.081	1.851	1.176	55.311
IPW, $\geq$2000, met	1.010	1.050	1.078	1.104	1.134	1.511
IPW, $\geq$2000, sulf	0.099	0.468	1.624	4.661	3.688	55.311

Table C.59: ATU weights summary for lung cancer mortality risks between metformin and sulfonylureas

Reference (metformin)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.059	1	1	1.278	1	15.218
Fewer, met	1	1	1	1	1	1
Fewer, sulf	0.059	0.445	0.994	1.761	2.034	15.218
Fewer, ≥2000, overall	0.251	1	1	1.443	1	15.516
Fewer, ≥2000, met	1	1	1	1	1	1
Fewer, ≥2000, sulf	0.251	0.903	1.451	2.500	2.854	15.516
Complete, overall	0.058	1	1	1.475	1	29.227
Complete, met	1	1	1	1	1	1
Complete, sulf	0.058	0.400	1.111	3.119	3.396	29.227
Complete, ≥2000, overall	0.071	1	1	1.631	1	49.971
Complete, ≥2000, met	1	1	1	1	1	1
Complete, ≥2000, sulf	0.071	0.396	1.362	4.010	2.700	49.971
Separate, overall	0.015	1	1	1.253	1	16.067
Separate, met	1	1	1	1	1	1
Separate, sulf	0.015	0.173	0.547	1.692	1.224	16.067
Separate, ≥2000, overall	0.045	1	1	1.499	1	37.147
Separate, ≥2000, met	1	1	1	1	1	1
Separate, ≥2000, sulf	0.045	0.268	0.835	2.690	2.347	37.147
IPW, overall	0.218	1.298	1.512	2.727	1.927	56.987
IPW, met	1.148	1.298	1.478	1.582	1.670	4.899
IPW, sulf	0.218	1.306	2.997	6.687	5.380	56.987
IPW, ≥2000, overall	0.099	1.047	1.081	1.851	1.176	55.311
IPW, ≥2000, met	1.010	1.050	1.078	1.104	1.134	1.511
IPW, ≥2000, sulf	0.099	0.468	1.624	4.661	3.688	55.311

216

Table C.60: ATU weights summary for lung cancer mortality risks between metformin and sulfonylureas, continued

Reference (sulfonylureas)	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Fewer, overall	0.046	0.232	0.731	0.748	1	14.682
Fewer, met	0.046	0.165	0.301	0.603	0.636	14.682
Fewer, sulf	1	1	1	1	1	1
Fewer, ≥2000, overall	0.041	0.208	0.445	0.587	1	2.936
Fewer, ≥2000, met	0.041	0.150	0.276	0.414	0.525	2.936
Fewer, ≥2000, sulf	1	1	1	1	1	1
Complete, overall	0	0.037	0.177	0.424	1	2.950
Complete, met	0	0.026	0.100	0.258	0.272	2.950
Complete, sulf	1	1	1	1	1	1
Complete, ≥2000, overall	0	0.029	0.138	0.399	1	2.921
Complete, ≥2000, met	0	0.023	0.089	0.240	0.228	2.921
Complete, ≥2000, sulf	1	1	1	1	1	1
Separate, overall	0.007	0.120	0.624	0.741	1	16.756
Separate, met	0.007	0.066	0.179	0.592	0.549	16.756
Separate, sulf	1	1	1	1	1	1
Separate, ≥2000, overall	0.010	0.096	0.352	0.540	1	3.736
Separate, ≥2000, met	0.010	0.067	0.149	0.348	0.435	3.736
Separate, ≥2000, sulf	1	1	1	1	1	1
IPW, overall	0	0.050	0.272	0.970	1.601	8.217
IPW, met	0	0.037	0.158	0.436	0.459	6.735
IPW, sulf	1.487	1.974	2.277	2.815	3.160	8.217
IPW, ≥2000, overall	0	0.033	0.160	0.478	1.049	3.339
IPW, ≥2000, met	0	0.024	0.100	0.269	0.265	3.339
IPW, ≥2000, sulf	1.042	1.096	1.146	1.263	1.373	2.394

Bibliography

Abraham, N.S., Young, J.M. and Solomon, M.J. A systematic review of reasons for nonentry of eligible patients into surgical randomized controlled trials. *Surgery*, 139(4):469–483, 2006.

Adams, C.P. and Brantner, V.V. Estimating the cost of new drug development: is it really $802 million? *Health Affairs*, 25(2):420–428, 2006.

Algra, A.M. and Rothwell, P.M. Effects of regular aspirin on long-term cancer incidence and metastasis: a systematic comparison of evidence from observational studies versus randomized trials. *The Lancet Oncology*, 13(5):518–527, 2012.

Aronoff, D.M. and Neilson, E.G. Antipyretics: mechanisms of action and clinical use in fever suppression. *The American Journal of Medicine*, 111(4):304–315, 2001.

Ateren, K.O. Metformin associated with lower cancer mortality in type 2 diabetes: ZODIAC-16. *Diabetes Care*, 33(2):322–326, 2010.

Baggs, J.G., Schmitt, M.H., Mushlin, A.I., Mitchell, P.H. et al. Association between nurse-physician collaboration and patient outcomes in three intensive care units. *Critical Care Medicine*, 27(9):1991–1998, 1999.

Bai, R.Y., Staedtke, V., Aprhys, C.M., Gallia, G.L. et al. Antiparasitic mebendazole shows survival benefit in 2 preclinical models of glioblastoma multiforme. *Neuro-Oncology*, 13(9):974–982, 2011.

Benjamin, E.J., Virani, S.S., Callaway, C., Chamberlain, A. et al. Heart disease and stroke statistics-2018 update: a report from the american heart association. *Circulation*, 137(12):e67–e492, 2018.

Berk, R.A. An introduction to ensemble methods for data analysis. *Sociological Methods & Research*, 34(3):263–295, 2006.

Bielawski, K., Winnicha, K. and Bielawska, A. Inhibition of DNA topoisomerases I and II, and growth inhibition of breast cancer MCF-7 cells by ouabain, digoxin and proscillaridin A. *Biological and Pharmaceutical Bulletin*, 29(7):1493–1497, 2006.

Bodmer, M. Long-term metformin use is associated with decreased risk of breast cancer. *Diabetes Care*, 33(6):1304–1308, 2010.

Borgquist, S., Tamimi, R.M., Chen, W.Y., Garber, J.E. et al. Statin use and breast cancer risk in the nurses' health study. *Cancer Epidemiology, Biomarkers & Prevention*, 25(1):201–206, 2016.

Bowker. Increased cancer-related mortality for patients with type 2 diabetes who use sulfonylureas or insulin. *Diabetes Care*, 29(2):254–258, 2006.

Breslow, N.E. Discussion of the paper by DR Cox. *Journal of the Royal Statistical Society*, 34:216–217, 1972.

Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J. et al. Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12): 1149–1156, 2006.

Buer, J.K. Origins and impact of the term 'NSAID'. *Inflammopharmacology*, 22(5): 263–267, 2014.

Chen, T.M., Lin, C.C., Huang, P.T. and Wen, C.F. Metformin associated with lower mortality in diabetic patients with early stage hepatocellular carcinoma after radiofrequency ablation. *Journal of Gastroenterology and Hepatology*, 26(5):858–865, 2011.

Cox, D.R. Regression models and life tables. *Journal of the Royal Statistical Society*, 34(2):187–220, 1972.

Cox, D.R. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

Currie, C.J., Poole, C.D. and Gale, E.A.M. The influence of glucose-lowering therapies on cancer risk in type 2 diabetes. *Diabetologia*, 52(9):1766–1777, 2009.

Cuzick, J., Otto, F., Baron, J.A., Brown, P.H. et al. Aspirin and non-steroidal anti-inflammatory drugs for cancer prevention: an international consensus statement. *The Lancet Oncology*, 10(5):501–507, 2009.

Dalton, S.O., Johansen, C., Poulsen, A.H., Nørgaard, M. et al. Cancer risk among users of neuroleptic medication: a population-based cohort study. *British Journal of Cancer*, 95(7):934–939, 2006.

Diamanti-Kandarakis, E., Christakou, C.D., Kandaraki, E. and Economou, F.N. Metformin: an old medication of new fashion: evolving new molecular mechanisms and clinical implications in polycystic ovary syndrome. *European Journal of Endocrinology*, 162(2):193–212, 2010.

Doudican, N., Rodriguez, A., Osman, I. and Orlow, S.J. Mebendazole induces apoptosis via Bcl-2 inactivation in chemoresistant melanoma cells. *Molecular Cancer Research*, 6(8):1308–1315, 2008.

Doudican, N.A., Byron, S.A., Pollock, P.M. and Orlow, S.J. XIAP downregulation accompanies mebendazole growth inhibition in melanoma xenografts. *Anti-Cancer Drugs*, 24(2):181–188, 2013.

Dunn, O.J. and Clark, V.A. *Basic Statistics: A Primer for the Biomedical Sciences.* John Wiley & Sons, 4th edition, 2009.

Elbaz, H.A., Stueckle, T.A., Tse, W., Rojanasakul, Y. et al. Digitoxin and its analogs as novel cancer therapeutics. *Experimental Hematology & Oncology*, 1(4), 2012.

Elder, D.J.E., Hague, A., Hicks, D.J. and Paraskeva, C. Differential growth inhibition by the aspirin metabolite salicylate in human colorectal tumor cell lines: enhanced apoptosis in carcinoma and in vitro-transformed adenoma relative to adenoma cell lines. *Cancer Research*, 56(10):2273–2276, 1996.

Evans. Metformin and reduced risk of cancer in diabetic patients. *BMJ*, 330(7503): 1304–1305, 2005.

Fife, R.S., Sledge Jr., G., Roth, B.J. and Proctor, C. Effects of doxycycline on human prostate cancer cells in vitro. *Cancer Letters*, 127(1-2):37–41, 1998.

Franciosi, M., Lucisano, G., Lapice, E., Strippoli, G.F.M. et al. Metformin therapy and risk of cancer in patients with type 2 diabetes: systematic review. *PLOS One*, 8(8), 2013.

Fung, E.K. and Lore, J.M. Randomized controlled trials for evaluating surgical questions. *JAMA Otolaryngology - Head & Neck Surgery*, 128(6):631–634, 2002.

Garrett, C.R., Hassabo, H.M., Bhadkamkar, N.A., Wen, S. et al. Survival advantage observed with the use of metformin in patients with type II diabetes and colorectal cancer. *British Journal of Cancer*, 106(8):1374–1378, 2012.

Hall. Diabetes and the risk of lung cancer. *Diabetes Care*, 28(3):590–594, 2005.

Harder, V.S., Stuart, E.A. and Anthony, J.C. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, 15(3):234–249, 2010.

Hay, M., Thomas, D.W., Craighead, J.L., Economides, C. et al. Clinical development success rates for investigational drugs. *Nature Biotechnology*, 32(1):40–51, 2014.

He, X., Tu, S.M., Lee, M. and Yeung, S.J. Thiazolidinediones and metformin associated with improved survival of diabetic prostate cancer patients. *Annals of Oncology*, 22(12):2640–2645, 2011.

He, X., Esteva, F.J., Ensor, J., Hortobagyi, G.N. et al. Metformin and thiazolidinediones are associated with improved breast cancer-specific survival of diabetic women with HER2+ breast cancer. *Annals of Oncology*, 23(7):1771–1780, 2012.

Hernán, M.A. With great data comes great responsibility: publishing comparative effectiveness research in epidemiology. *Epidemiology*, 22(3):290–291, 2011.

Hernán, M.A. and Robins, J.M. Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8):758–764, 2016.

Ho, D.E., Imai, K., Kingand, G. and Stuart, E.A. Matching as nonparametric pre-processing for reducing model dependence in parametric causal inference. *Political Analysis*, 15:199–236, 2007.

Imbens, G.W. and Rubin, D.B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, 2015.

Kelley, T.C. *Iterative Methods for Optimization.* Society for Industrial and Applied Mathematics, 1999.

Kim, M.H. and Chung, J. Synergistic cell death by EGCG and ibuprofen in DU-145 prostate cancer cell line. *Anticancer Research*, 27(6B):3947–3956, 2007.

Klein, J.P. and Moeschberger, M.L. *Survival Analysis: Techniques for Censored and Truncated data.* Springer, 2003.

Kola, I. and Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery*, 3(8):711–715, 2004.

Kordes, S., Pollak, M.N., Zwinderman, A.H., Mathôt, R.A. et al. Metformin in patients with advanced pancreatic cancer: a double-blind, randomized, placebo-controlled phase 2 trial. *The Lancet Oncology*, 16(7):839–847, 2015.

Kumar, S., Bryant, C.S., Chamala, S., Qazi, A. et al. Ritonavir blocks AKT signaling, activates apoptosis and inhibits migration and invasion in ovarian cancer cells. *Molecular Cancer*, 8(26), 2009.

Kunz, R., Vist, G.E. and Oxman, A.D. Randomization to protect against selection bias in healthcare trials. *The Cochrane Database of Systematic Reviews*, 18(2): MR000012, 2007.

Lee, E.T. and Wang, J.W. *Statistical Methods for Survival Data Analysis.* John Wiley & Sons, 3rd edition, 2003.

Leite, W. *Practical Propensity Score Methods Using R.* SAGE Publications, Inc, 2017.

Leonard, I., Beales, P., Hensley, A., Loke, Y. et al. Reduced esophageal cancer incidence in statin users, particularly with cyclo-oxygenase inhibition. *World Journal of Gastrointestinal Pharmacology and Therapeutics*, 4(3):69–79, 2013.

Li, F., Dou, J., Wei, L., Li, S. et al. The selective estrogen receptor modulators in breast cancer prevention Italian study. *Cancer Chemotherapy and Pharmacology*, 77(5):895–903, 2016.

Lilford, R.J., Thornton, J.G. and Braunholtz, D. Clinical trials and rare diseases: a way out of a conundrum. *BMJ*, 311(7020):1621–1625, 1995.

Little, R.J. and Rubin, D.B. *Statistical Analysis with Missing Data.* John Wiley & Sons, 2nd edition, 2002.

Lord, J.M., Flight, I.H.K. and Norman, R.J. Metformin in polycystic ovary syndrome: systematic review and meta-analysis. *BMJ*, 327(951), 2003.

Lytras, T., Nikolopoulos, G. and Bonovas, S. Statins and the risk of colorectal cancer: an updated systematic review and meta-analysis of 40 studies. *World Journal of Gastroenterology*, 20(7):1858–1870, 2014.

Maeng, Y.S., Lee, R., Lee, B., Choi, S.i. et al. Lithium inhibits tumor lymphangio-genesis and metastasis through the inhibition of TGFBIp expression in cancer cells. *Scientific Reports*, 6(20739), 2016.

Marubini, E. and Valsecchi, G.M. *Analyzing Survival Data from Clinical Trials and Observational Studies.* John Wiley & Sons, 2004.

Mayers, W.R. Handling missing data in clinical trials: an overview. *Drug Information Journal*, 34(2):525–533, 2000.

Mcconkey, D.J., Lin, Y., Nutt, L.K., Ozel, H.Z. et al. Cardiac glycosides stimulate Ca^{2+} increases and apoptosis in androgenindependent, metastatic human prostate adenocarcinoma cells. *Cancer Research*, 60(14):3807–3812, 2000.

Mcquay, H.J. and Moore, R.A. Dose-response in direct comparisons of different doses of aspirin, ibuprofen and paracetamol (acetaminophen) in analgesic studies. *British Journal of Clinical Pharmacology*, 63(3):271–278, 2007.

Miner, J. The discovery of aspirin's antithrombotic effects. *The Texas Heart Institute Journal*, 34(2):179–186, 2007.

Molenberghs, G., Fitzmaurice, G., Kenward, M.G., Tsiatis, A. et al. *Handbook of Missing Data Methodology.* Chapman and Hall/CRC, 2014.

Mortensen, P.B. Neuroleptic medication and reduced risk of prostate cancer in schizophrenic patients. *Acta Psychiatrica Scandinavica*, 85(5):390–393, 1992.

Mouratidis, P.X., Colston, K.W. and Dalgleish, A.G. Doxycycline induces caspase-dependent apoptosis in human pancreatic cancer cells. *International Journal of Cancer*, 120(4):743–52, 2007.

Mullard, A. 2016 FDA drug approvals. *Nature Publishing Group*, 16(2):73–76, 2017.

Nakai, M. and Ke, W. Review of the methods for handling missing data in longitudinal data analysis. *International Journal of Mathematical Analysis*, 5(1):1–13, 2011.

Newson, R. Confidence intervals for rank statistics: Somers' D and extensions. *The Stata Journal*, 6(3):309–334, 2006.

Newson, R. Attributable and unattributable risks and fractions and other scenario comparisons. *The Stata Journal*, 13(4):672–698, 2013.

Newson, R.B. Interpretation of Somers' D under four simple models. Technical report, Imperial College London, 2014.

Noto, H., Goto, A., Tsujimoto, T. and Noda, M. Cancer risk in diabetic patients treated with metformin: a systematic review and meta-analysis. *PLOS One*, 7(3), 2012.

Nygren, P., Fryknäs, M., Ågerup, B. and Larsson, R. Repositioning of the anthelmintic drug mebendazole for the treatment for colon cancer. *Journal of Cancer Research and Clinical Oncology*, 139(12):2133–2140, 2013.

of Medicine, I. *Drug Repurposing and Repositioning: Workshop Summary*. Washington, DC: The National Academies Press, 2014.

Onnelly, O. New users of metformin are at low risk of incident cancer: a cohort study among people with type 2 diabetes. *Diabetes Care*, 32(9):1620–1625, 2009.

Pantziarka, P., Sukhatme, V., Bouche, G., Meheus, L. et al. Repurposing drugs in oncology (ReDO)-diclofenac as an anti-cancer agent. *Ecancermedicalscience*, 10 (610), 2016.

Paul, S.M., Mytelka, D.S., Dunwiddie, C.T., Persinger, C.C. et al. How to improve R&D productivity- the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9(3):203–214, 2010.

Phillips, I., Langley, R., Gilbert, D. and Ring, A. Aspirin as a treatment for cancer. *Clinical Oncology*, 25(6):333–335, 2013.

Pollak, M. Overcoming drug development bottlenecks with repurposing: repurposing biguanides to target energy metabolism for cancer treatment. *Nature Medicine*, 20 (6):591–593, 2014.

Powell, J.L. Estimation of Semiparametric Models. *Handbook of Econometrics*, 4: 2443–2521, 1994.

Robins, J. A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

Robins, J.M., Hernán, M.A. and Brumback, B. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

Romero, I.L., Mccormick, A., Mcewen, K.A. and Park, S. Relationship of type II diabetes and metformin use to ovarian cancer progression, survival, and chemosensitivity. *Obstetrics & Gynecology*, 119(1):61–67, 2012.

Rosenbaum, P.R. Design sensitivity and efficiency in observational studies. *Journal of the American Statistical Association*, 105(490):692–702, 2010.

Rosenbaum, P.R. and Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Rubin, D.B. Matching to remove bias in observational studies. *Biometrics*, 29(1): 159–183, 1973.

Rubins, J.B., Charboneau, D., Alter, M.D., Bitterman, P.B. et al. Inhibition of mesothelioma cell growth in vitro by doxycycline. *Journal of Laboratory and Clinical Medicine*, 138(2):101–106, 2001.

Sacks, H., Chalmers, T.C. and Smith, H.J. Randomized versus historical controls for clinical trials. *The American Journal of Medicine*, 72(2):233–240, 1982.

Sadeghi, N., Abbruzzese, J.L., Yeung, S.c.J., Hassan, M. et al. Metformin use is associated with better survival of diabetic patients with pancreatic cancer. *Clinical Cancer Research*, 18(10):2905–2912, 2012.

Sasaki, J.i., Ramesh, R., Chada, S., Gomyo, Y. et al. The anthelmintic drug mebendazole induces mitotic arrest and apoptosis by depolymerizing tubulin in non-small cell lung cancer cells. *Molecular Cancer Therapeutics*, 1(13):1201–1209, 2002.

Schulz, K.F. and Grimes, D.A. Sample size calculations in randomized trials: mandatory and mystical. *The Lancet*, 365(9467):1348–1353, 2005.

Schulz, K.F., Altman, D.G. and Moher, D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomized trials. *BMJ*, 8(18), 2010.

Setoguchi, S., Schneeweiss, S., Brookhart, M.A., Glynn, R.J. et al. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety*, 17(6):546–555, 2008.

Skinner, H.D., Crane, C.H., Garrett, C.R., Eng, C. et al. Metformin use and improved response to therapy in rectal cancer. *Cancer Medicine*, 2(1):99–107, 2013a.

Skinner, H.D., Mccurdy, M.R., Echeverria, A.E., Steven, H. et al. Metformin use and improved response to therapy in esophageal adenocarcinoma. *Acta Oncologica*, 52 (5):1002–1009, 2013b.

Sleire, L., Elise, H., Anne, I., Leiss, L. et al. Drug repurposing in cancer. *Pharmacological Research*, 124:74–91, 2017.

Solomon, M.J. and McLeod, R.S. Should we be performing more randomized controlled trials evaluating surgical operations? *Surgery*, 118(3):459–467, 1995.

Song, H., Fares, M., Maguire, K.R., Siden, A. et al. Cytotoxic effects of tetracycline analogues (doxycycline, minocycline and COL-3) in acute myeloid leukemia HL-60 cells. *PLOS One*, 9(12), 2014.

Srirangam, A., Mitra, R., Wang, M., Gorski, J.C. et al. Effects of HIV protease inhibitor ritonavir on Akt-regulated cell proliferation in breast cancer. *Clinical Cancer Research*, 12(6):1883–1896, 2006.

Sturmer, T., Schneeweiss, S., Avorn, J. and J, G.R. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American Journal of Epidemiology*, 162(3):279–289, 2005.

Su, B., Zheng, B., Somai, M., Bassil, D. et al. Harnessing diverse bioinformatics approaches to repurpose drugs for alzheimer's disease. Technical document, Imperial College London, 2018.

Sun, A., Shanmugam, I., Song, J., Terranova, P.F. et al. Lithium suppresses cell proliferation by interrupting E2F-DNA interaction and subsequently reducing S-phase gene expression in prostate cancer. *The Prostate*, 67(9):976–988, 2007.

Tamargo, R.J., Bok, R.A. and Brem, H. Angiogenesis inhibition by minocycline. *Cancer Research*, 51(1):672–675, 1991.

Tsilidis, K.K., Capothanassi, D., Allen, N.E., Rizos, E.C. et al. Metformin does not affect cancer risk: a cohort study in the U.K. Clinical Practice Research Datalink analyzed like an intention-to-treat trial. *Diabetes Care*, 37(9):2522–2532, 2014a.

Tsilidis, K.K., Kasimis, J.C., Lopez, D.S., Ntzani, E.E. et al. Type 2 diabetes and cancer: umbrella review of meta-analyses of observational studies. *BJM*, 350(g7607), 2014b.

Tutton, P.J.M. and Barkla, D.H. Influence of inhibitors of serotonin uptake on intestinal epithelium and colorectal carcinomas. *British Journal of Cancer*, 46(2): 260–265, 1982.

van Buuren, S. and Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 2011.

Walker, E., Hernandez, A.V. and Kattan, M.W. Meta-analysis: its strengths and limitations. *Cleveland Clinic Journal of Medicine*, 75(6):431–439, 2008.

Wang, P., Li, Y. and Reddy, C.K. Machine learning for survival analysis: a survey. 2017.

Weber, G.M., Adams, W.G., Bernstam, E.V., Bickel, J.P. et al. Biases introduced by filtering electronic health records for patients with "complete data". *Journal of the American Medical Informatics Association*, 24(6):1134–1141, 2017.

Westreich, D., Lessler, J. and Funk, M.J. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63(8):826–833, 2010.

Wu, D., Hu, D., Chen, H., Shi, G. et al. Glucose-regulated phosphorylation of TET2 by AMPK reveals a pathway linking diabetes to cancer. *Nature*, 559(7715):637–641, 2018.

Wu, X., Zeng, K., Xue, F., Chen, J. et al. Statins are associated with reduced risk of gastric cancer: a systematic review and meta-analysis. *Annals of Oncology*, 24(7): 1721–1730, 2013.

Yeung. Antidiabetic therapies affect risk of pancreatic cancer. *Gastroenterology*, 137 (2):482–488, 2009.

Zhong, G.C., Liu, Y., Ye, Y.Y., Hao, F.b. et al. Meta-analysis of studies using statins as a reducer for primary liver cancer risk. *Scientific Reports*, 6(26256), 2016.