UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Addressing the Healthy Donor Effect while Studying the Impact of Blood Donation on Donors' Long-Term Health Outcomes.

by Md Morshadur Rahman

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of
Dr Andrew Hayen & Dr Surendra Karki

University of Technology Sydney
School of Public Health
Faculty of Health

August 2023

Certificate of Original Authorship

I, Md Morshadur Rahman, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Public Health at the Faculty of Health at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: 11/08/2023

Abstract

Background

The healthy donor effect (HDE) is a methodological problem that arises in donor health research when comparing donor versus non-donor or regular donor versus casual donor. This may distort the true causal relationship between blood donation and long-term health outcomes.

Aims and Objectives

The aim of this thesis was to identify and apply the method/s to provide an unbiased estimate of the impact of blood donation on blood donors' long-term health outcomes (all-cause mortality and some cancers).

Methods

I conducted a systematic review to identify and summarised the methods that were used to account for the HDE in published studies. I then applied a causal inference method called 'target trial emulation' and another less biased method called 'exposure window' method to adjust for the HDE using the Sax Institute's 45 and Up Study data, linked with blood donor data and other health data sets. For all-cause mortality, I used both the target trial and 5-year exposure window methods, along with adjustments from g-methods (inverse probability weighting, targeted minimum loss-based estimator, and sequentially doubly robust estimator). I also investigated the association of regular blood donation with gastrointestinal, colorectal, and haematological cancer using the 5-year exposure window method.

Results

The results from the systematic review suggested that most of the existing methods used to mitigate the HDE were inadequate to effectively address this bias. A lack of use of appropriate causal inference techniques was also observed. In my analyses, the target trial emulation technique did not show any statistically significant association between the initiation of blood donation and the mortality risk. The use of the 5-year exposure window method also did not reveal any significant relationship between higher-frequency blood donation and all-cause mortality when compared to lower-frequency donors. For gastrointestinal and haematological

cancers, I also found no statistically significant difference in the risk of cancers for the higher-frequency blood donors compared to the lower-frequency donors.

Conclusion

Through a systematic review, I found that the methods used in blood donor research to mitigate the impact of HDE are inadequate. Using some of the less biased methods, I found no significant association between regular, higher-frequency blood donation and long-term health outcomes such as mortality and some cancers after adjusting the HDE. The findings from this study can provide crucial insights for the Australian Red Cross Lifeblood's strategic planning and directing future research.

Acknowledgements

I would like to express my deepest gratitude to everyone who supported and encouraged me to complete my PhD thesis.

First and foremost, I would like to express my heartfelt gratefulness to my supervisors, Professor Andrew Hayen, and Dr Surendra Karki, for their excellent mentoring, insightful advice, encouragement, and constructive feedback. Andrew allowed me to explore my academic interests and offered his utmost assistance. His continuous support in my research motivated me and boosted my confidence. Surendra was always ready to help me when I needed anything, and his passion for research has always inspired me to work harder. I learnt a lot from him as he always provided insightful remarks and constructive criticism on my research ideas and writings.

I would like to thank the University of Technology Sydney for providing scholarships to complete my PhD. Their financial support has been instrumental in enabling me to focus on my research. I am also very grateful to Dr David Irving, Director, Research & Development, Australian Red Cross Lifeblood, and Surendra for allowing me to access their facilities at Lifeblood. The Australian government funds the Australian Red Cross Lifeblood for the provision of blood, blood products, and services to the Australian community. Interacting with the dedicated team at Lifeblood and being able to utilise their resources enriched my knowledge and understanding of donor-health research.

I would also like to express my gratitude to my colleagues and staff members in the School of Public Health, Lin, Krishna, Priya, and Michelle, for their help and support during my PhD journey. I am grateful to all my friends who always supported and encouraged me to pursue my PhD and dream big.

Finally, I would like to express my deepest gratitude to my parents. Their unconditional love, support and kindness have kept me going and helped me reach my goal. I am also thankful to my wife, Sinthia Pervez, who accompanied me during the whole period of my PhD. It was never easy to come and start my PhD in Australia during the Covid-19 pandemic. Her love, support and unwavering faith in my ability have made it possible to complete my Ph.D.

List of Publications

Journal Publications

1. **Rahman M M,** Karki S and Hayen A. A Methods Review of the 'Healthy Donor Effect' in Studies of Long-Term Health Outcomes in Blood Donors. *Transfusion* 62(3), 698-712 (2022).
2. **Rahman M M,** Karki S and Hayen A. A target trial emulation to estimate the impact of blood donation on mortality in blood donors in Australia. (Submitted to journal Transfusion).
3. **Rahman M M,** Karki S and Hayen A. High-frequency whole blood donation and its impact on mortality: evidence from a data linkage study in Australia. (Submitted to Sax Institute for technical review).
4. **Rahman M M,** Karki S, Cust A E, Olynyk J K and Hayen A. Regular whole blood donation and gastrointestinal, and haematological cancer risk among older Australian blood donors. (Ready to submit to Sax Institute for technical review).

Conference Presentations

1. **Rahman M M,** Karki S and Hayen A. A target trial emulation to estimate the impact of blood donation on mortality risk in older blood donors in Australia. *STAToSPHERE Annual Scientific Meeting 2022 Sydney Australia*.
2. **Rahman M M,** Karki S and Hayen A. Regular blood donation and the risk of mortality: minimizing the impact of healthy donor effect. *5th European Conference on Donor Health and Management 2023 Austria*.
3. Karki S, **Rahman M M**, Hayen A and Irving D O. Frequent apheresis donation does not increase the risk of bone fractures in donors. *5th European Conference on Donor Health and Management 2023 Austria*.

Table of Contents

List of Tables

List of Figures

Abbreviation

BCAs	Blood Collection Agencies
ROS	Reactive Oxygen Species
LDL	Low-Density Lipoprotein
CVD	Cardiovascular Disease
CHD	Coronary Heart Disease
BMD	bone mineral density
HDE	Healthy Donor Effect
HRE	Healthy Registration Effect
HDSE	Healthy Donor Survivor Effect
HDCE	Healthy Donor Career Effect
IgG	Immunoglobulin G
ANOVA	Analysis of Variance
BMI	Body Mass Index
RCT	Randomised Control Trial
ITT	Intention to treat
IPW	Inverse Probability Weights /Weighting/Weighted
HR	Hazard Ratio
TMLE	Targeted Minimum Loss-Based Estimator
SDR	Sequentially Doubly Robust
NBMS	National Blood Management System
APDC	Admitted Patient Data Collection
NSWCCR	NSW Central Cancer Registry
EDDC	Emergency Department Data Collection
NCIMS	Notifiable Conditions Information Management System
RBDM	Registry of Birth, Deaths, and Marriages- Deaths Registrations
MBS	Medicare Benefits Schedule
PBS	Pharmaceutical Benefits Scheme
CHeReL	Centre for Health Record Linkage
SURE	Secure Unified Research Environment
PHRN	Population Health Research Network

CI	Confidence Interval
HIV	Human Immunodeficiency Virus
HCV	Hepatitis C Virus
HTLV	Human T-cell Lymphotropic Virus
CCI	Charlson Co-morbidity Index
GP	General Practitioner
ICD-10	International Classification of Diseases 10th Revision
ATC	Anatomical Therapeutic Chemical
HREC	Human Research Ethics Committee
RR	Risk Ratio/ Relative Risk
RD	Risk Difference
WHO	World Health Organization
ICD-9	International Classification of Diseases 9th Revision

Chapter 1 Introduction

1.1 Background

1.1.1 Blood donation and long-term health outcomes

Blood donation is a life-saving act for patients in need. In 2017, the global demand for blood and blood products was more than 304 million units [1]. The majority of this demand is met by donations from voluntary donors. While blood and plasma donation are generally safe for donors, few may develop immediate adverse events such as fainting, and a small proportion may also develop iron-deficiency or anaemia after regular whole blood donations [2].

Blood collection agencies (BCAs) employ a number of strategies to guarantee the donors' continuous health and safety as well as the transfusion-safety of donated blood [3]. The primary approach involves choosing the healthiest portion of the population who can tolerate the physiological alterations caused by donation, ensuring the donated blood is safe for those receiving it [4]. Also, BCAs conduct research to understand the impact of donation on health and continually refine their donor selection criteria to ensure the ongoing health of donors [3].

In Australia, one-third of the population will require blood and blood products at some point in their lives, and more than 29,000 donations per week are required to meet demand [5]. Additionally, roughly 3% of Australians donate blood annually, with many donors making multiple donations [5]. With the aging of Australia's population, including blood donors, it is anticipated that health issues related to aging will increase [6, 7]. For instance, older patients who suffer from acute myocardial infarction and have a low haematocrit level upon admission often require blood transfusions. These transfusions have been associated with a lower short-term mortality rate in such patients [8]. This will consequently lead to a greater need for blood and blood products in the country. If the current pool of donors does not grow, it could result in depending on existing donors to donate more frequently to meet supply needs. However, the

long-term health effects of regular blood donation remain unclear. As such, it's appropriate to conduct a comprehensive study on the long-term health outcomes of regular blood donors to assess if donation leads to beneficial or harmful effects.

The majority of current donor health research is focused on the immediate and mid-term effects of blood donation, such as hematomas, vasovagal reactions, and iron-deficiency [9]. The long-term impact of blood donation on the health of a donor is a relatively neglected area of research. Only a few studies has investigated the consequences of regular whole blood donation and plasma/platelet donation, including potential associations to cancer due to reduced body iron levels and the potential harmful effects of citrate exposure on bone health [10-12]. However, the findings from these studies are inconclusive which is described extensively in the later chapters.

Whole blood donation increases the risk of iron-deficiency or anaemia. After each whole blood donation, donors typically lose 200 to 250 mg of iron and subsequently are exposed to a lower level of body stores of iron for a considerable period of time [13]. Similarly, while donating plasma and platelets by apheresis (apheresis is a medical technology where the blood of a donor is passed through an apparatus that separates a particular component of blood for donation, and the rest is returned back to the donor's circulation), donors are exposed to a small quantity of citrate anticoagulant which causes acute disturbance in calcium and phosphate metabolism in the body [10]. In addition, high volume or regular plasma donation also leads to a transient reduction in serum proteins, including immunoglobulins [14]. In general, blood donors are healthier than the rest of the population and usually practice healthier lifestyles [15]. They tend to smoke less, exercise more, have better self rated health and suffer less from chronic diseases than general population [16]. However, repeated and chronic exposure to changes in the normal physiologic level of minerals and proteins in the body may have acute as well as long-term impacts on health.

The lowered stored iron in the body due to whole blood donation has been considered as something that can impact the health of donors. Frequent blood donors are exposed to low stores of body iron for a considerable time compared to if they had not donated blood [13]. A lower store of iron in the human body is hypothesised to decrease the risk of cardiovascular diseases [17]. Although the precise mechanism of lowering such risk is unknown, one of the highly debated potential mechanisms is lower iron stores leading to less oxidative stress (due to less amount of reactive oxygen species (ROS) production). A lower iron level reduces the amount of ROS production, which in turn reduces the vascular reactivity and decreases the peroxidation of low-density lipoprotein (LDL)-cholesterol. This process may have some preventive effect on atherosclerosis, subsequently decreasing the risk of cardiovascular diseases [18]. Another proposed mechanism is that low iron store reduces the risk of development of type 2 diabetes which is associated with an increased risk of cardiovascular diseases [19, 20].

Only a small number of studies have examined the risk of CVD or biomarkers of CVD in blood donors in relation to donation intensity, and the results are inconsistent. A study from the Netherlands reported that whole blood donation intensity is not associated with a lower prevalence of metabolic syndromes, which are recognised as risk factors for CVD [21]. Another two studies did not find any association between high-intensity donation and subclinical atherosclerosis and high-intensity donation and myocardial infarction [22, 23]. However, a few other studies have reported a decreased risk of coronary heart disease in frequent whole blood donors. A sub-study under the Nebraska Diet Heart Survey has suggested a possible protective effect of blood donation on CVD in non-smoking men [12]. Another study among males has suggested that endothelial dysfunction, which increases the risk of CHD and occurs secondary to systemic inflammation and oxidative stress, may be reduced by regular blood donation [24].

Another health outcome that is hypothesised to be associated with iron is cancer. Animal models and human studies suggest that high levels of iron in the body can increase the risk of cancer development [25, 26]. Thus lowering body iron stores may lead to a lower risk of cancer. On the other hand, low levels and less activity of natural killer cells after each blood donation is also hypothesised to increase the risk of cancers, particularly non-Hodgkin's lymphoma [27, 28]. Results from studies examining the effect of a high level of iron in the body on cancer risk are not consistent. A systematic review suggested that there is a positive association between increasing heme iron uptake (such as through red meat consumption) and cancer risk [11]. A large study of Swedish and Danish blood donors did not find any overall association of increased or decreased cancer risk in repeat blood donors but reported that frequent plasma donors were at higher risk of non-Hodgkin's lymphoma [29]. The study also reported that among men, when considering 3-7 years of latency, there was a decreasing trend in the risk of liver, lung, colon, oesophagus, and stomach cancer as the amount of iron loss increased due to donation [29]. In another study from the US, the authors did not find any association between regular blood donation and colorectal cancer risk when adjusted for many lifestyles and dietary factors [30]. Similarly, another analysis from the same cohort also did not find any association between blood donation frequency and non-Hodgkin's lymphoma risk [28]. The detail of the study designs and methodologies used in the above studies are discussed in the literature review chapter where I conducted a systematic methods review of these studies.

Apheresis donation is also believed to be associated with long-term bone health, but previous studies showed inconsistent results. Amrein et al. reported lower bone mineral density (BMD) at the lumbar spine in apheresis donors than in non-donor controls, but another study by Boot et al. did not find any difference in BMD of the lumbar spine in apheresis donors compared to whole blood donors [10, 31]. The first and the only study examining the long-term risk of bone fractures and high-frequency plasma donation in Swedish donors did not find any association

[32]. Regular plasma donation also exposes the donor to chronic low levels of serum protein and immunoglobulins [33, 34]. Immunoglobulins play an essential role in preventing infection in the early phase of exposure to antigens. A review of available evidence until 2007 concluded that prolonged and deficient immunoglobulins levels increase the risk of infection [35]. A detailed description of the methods used in the above studies is in Chapter 2.

1.1.2 Addressing the Healthy Donor Effect (HDE)

The research examining the impact of blood donation on donors' long-term health outcomes has shown inconsistent results. Some studies found a positive association of blood donation with reduced risk of certain diseases, while other studies found no association of blood donation with those health outcomes [12, 23, 28-32, 36-48]. A few studies even found a negative association of blood donation with long-term health outcomes [10, 39, 49-51].

This may have been caused due to the fact that blood donors are selected on health criteria for their eligibility to donate and may differ systematically in relation to the status of health and utilisation of healthcare services compared to the general population. This donor selection procedure gives rise to a bias in donor health research known as the "Healthy Donor Effect" (HDE), which, if not adequately adjusted, indicates a better health condition and lower disease morbidity in blood donors when compared to the general population. In the field of health research among blood donors, this issue is considered to be a methodological problem, and it is a combination of selection bias and confounding [15].

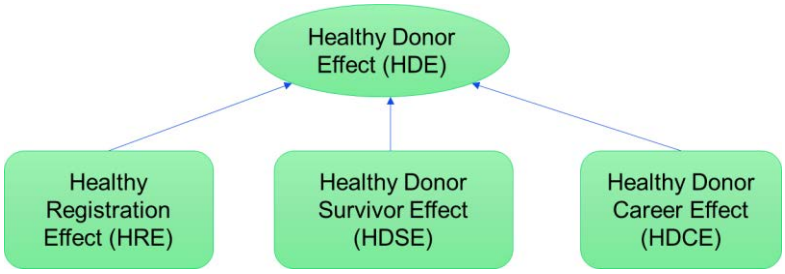


Figure 1.1 Different types of Healthy Donor effects.

HDE arises due to a combination of selection biases known as the Healthy Registration Effect (HRE), the Healthy Donor Survivor Effect (HDSE), and the Healthy Donor Career Effect (HDCE) [15]. Healthy individuals are more capable of and interested in donating blood. When a person wishes to donate blood, they must meet specific health and lifestyle requirements. As a result, when compared with the general population, blood donors are healthier, and this gives rise to HRE [15, 42, 52]. Further, less healthy donors usually stop donating sooner. Thus, it is more likely that active donors who regularly donate are, on average healthier than inactive/lapsed donors (donors who have stopped donating for some period of time). This gives rise to a bias known as HDSE [15, 42, 52]. HDCE affects analyses that compare health outcomes within the active donor population [15, 42, 52]. As blood donors must undergo repeated medical check-ups throughout their donation careers, donors who regularly donate for a longer time period are healthier than those donors who have been making donations for a relatively short period of time. Because of these selection effects, it is very challenging in donor health studies to make causal inferences about whether the health outcomes measured occurred due to the effect of blood donation or were caused by HDE. Therefore, it is crucial to acknowledge and appropriately adjust the HDE when assessing the health effects of blood donation.

1.1.3 Current blood donation practice in Australia

In Australia, blood donation is a voluntary act. Individuals have the option to either donate whole blood or choose to donate plasma/platelets through the process of apheresis. Typically, if they satisfy all other eligibility requirements, individuals can make up to 4 whole blood donations and 26 plasma donations annually [53]. Individuals between the ages of 18 and 75 can initiate whole blood, platelets, or plasma donations, while ongoing donors who have a record of donation within the country can continue donating at any age, provided they meet other qualifying conditions . Prior to every whole blood donation, a thorough assessment of the donor's eligibility is carried out, which includes a series of questions about their overall health, medical history, and recent travel. This process is aimed at ensuring the donor's safety during the donation and reducing the risk of transmitting diseases via transfusion to recipients.

Before each donation, donors are tested for haemoglobin levels. If a potential issue with low haemoglobin is detected, further tests for serum ferritin are conducted. Donors diagnosed with anaemia are deemed ineligible for donation and are advised to consult with their doctor [53]. They can return to donate blood after a six-month period, assuming their iron deficiency has been resolved. For those donating plasma for the first time, an initial screening is performed to check their complete blood count, total protein, albumin, and immunoglobulin G (IgG) levels. Regular plasma donors also have their total protein, albumin, and IgG levels measured annually [54]. Donors may be temporarily or permanently deferred from donating based on factors such as age, health status, or certain diseases, including anaemia.

Table 1.1 Interval between blood donation types in Australia

Last donation type	**Next donation type**	**Minimum interval between donations**
Whole blood	Whole blood	84 days (12 weeks)
Whole blood	Apheresis plasma	28 days (4 weeks)
Whole blood	Apheresis platelets (single or double)	28 days (4 weeks)

Any type of apheresis donation	Any type of donation	14 days (2 weeks)

1.2 Aims and objectives

The main aim of this thesis is to identify and apply the method/s that adequately adjust the HDE and provide an unbiased estimate of the impact of blood donation on blood donors' long-term health outcomes. The specific objectives of this thesis are:

1. To summarise the methods that have been used to adjust the HDE and identify any additional/new approaches that may adjust the HDE adequately.
2. To apply the appropriate method/s to The Sax Institute's 45 and Up Study data which is linked with other administrative health data sets to examine the association between donation and various long-term health outcomes such as mortality and cancer.

1.4 Thesis chapters organisation

This thesis has five main chapters with a publication-style structure. There are four articles in this thesis, among which one is already published (Chapter 2), one is submitted to a journal (Chapter 4, Section 1), one is under technical review at Sax Institute (Chapter 4, Section 2), and last one is ready to be submitted to Sax Institute for technical review (Chapter 4, Section 3).

Chapter 1 provides the background of the study, the aims and objectives of the study, a brief description of the study design and methods, and the organization of the study.

Chapter 2 provides a review of existing approaches/methods that were used to adjust the HDE.

Chapter 3 provides a description of the appropriate method/s and data to be used in the analysis.

Chapter 4 provides the application of the HDE adjustment methods to the Sax Institute's 45 and up study data to find the association between blood donation and various health outcomes, i.e., all-cause mortality and cancers.

Chapter 5 discusses the key findings from all the previous chapters, future research direction and recommendation and conclusion.

Chapter 2 Literature Review

A Methods Review of the 'Healthy Donor Effect' in Studies of Long-Term Health Outcomes in Blood Donors

Authors

Rahman MM[1,2,3], Karki S[3,4], Hayen A[1]

Affiliations

[1]School of Public Health, University of Technology Sydney, Sydney, Australia

[2]Department of Statistics, University of Dhaka, Bangladesh

[3]Research and Development, Australian Red Cross Lifeblood, Sydney, Australia

[4]School of Population Health, UNSW, Sydney, Australia

Journal: Transfusion

Type of Publication: Research paper

Stage of Publication: Published online on 06 January 2022

URL: https://doi.org/10.1111/trf.16791

2.1.1 Abstract

Background: The impact of blood donation on donors' long-term health outcomes shows inconsistent results. This may have been caused by the 'healthy donor effect' (HDE). In this study, we aimed to determine the extent to which studies examining the relationship between blood donation and long-term health outcomes acknowledge and adjust for the HDE.

Study Design and Methods: We conducted a systematic literature search examining the relationship between blood or plasma donation and long-term health outcomes. Then, we extracted data on several important study characteristics and information on how authors acknowledged and adjusted for the HDE.

Results: We identified 8784 articles, out of which 27 were included in this review. Among all, 19 (74%) studies mentioned potential bias resulting from HDE while the rest of the studies did not. Of those 19 studies that did mention the bias due to HDE, 13 studies reported the HDE as an important limitation. The most common method used to adjust for the effect of HDE was regression methods. Many studies also used comparison within blood donor population and few used qualification/exposure window techniques and restriction of the analysis to healthier subjects as a means to minimise HDE.

Conclusions: We provide a summary of how previous studies acknowledged and adjusted the HDE. Causal inference methods may be more appropriate and useful when selection bias is present in observational studies. Researchers should consider collecting information on relevant confounders in the design phase of the study to be able to apply causal methods in observational study settings.

Keywords: Healthy donor effect, HDE, blood donors, health outcomes

2.1.2 Background

Blood donation is a life-saving act for patients in need of blood products. In 2017, the global demand for blood and blood products was more than 304 million units [1]. The majority of this demand is met by donations from voluntary donors. While donating blood and plasma is generally safe for donors, few may develop immediate adverse events such as fainting, and some donors may also develop iron-deficiency or anaemia after regular whole blood donations [2]. Blood Collection Agencies (BCAs) employ several strategies to ensure the ongoing health and safety of donors, and the donated blood is safe for transfusion to patients. To ensure donor health, a key strategy is to start with the selection of the healthiest segment of the population that can withstand the physiological changes brought upon by the donation. Also, BCAs conduct research to understand the impact of donation on health and continually refine their donor selection criteria and policies to ensure the ongoing health of donors.

The majority of donor health research until recently was focused on the immediate effects of blood donation. However, recently several researchers have examined the relationship between donating blood or plasma and the health outcomes of donors in the longer-term. After Sullivan suggested [17] that the high level of stored iron in the body increases risk of cardiovascular diseases, many studies have examined whether lowering iron through blood donation can reduce cardiovascular disease risk. Some studies reported that blood donation has a protective effect on the incidence of cardiovascular diseases as well as cancer, mortality, and hospitalisation [12, 55, 56]. However, the observed protective effect of blood donation on donor's health may be potentially due to donors being healthier than the general population as they face rigorous donor selection procedures throughout their donation career.

These selection procedures also give rise to a bias while studying the effect of donation on health outcomes which is described as "Healthy Donor Effect" (HDE). The HDE arises when blood donors are compared to the general population, as well as when comparing donors of

different levels of donation frequency or donation career. This happens, as there are inherent health differences between the groups compared. The HDE is regarded as a methodological problem in health research in blood donors, and it is a combination of selection bias and confounding [15].

HDE may arise in a combination of selection biases termed as the Healthy Registration Effect (HRE), the Healthy Donor Survivor Effect (HDSE), and the Healthy Donor Career Effect (HDCE) [15]. Healthy people are usually more capable and interested to become blood donors. Also, when people want to donate blood, they must pass a number of health and lifestyle criteria set by the blood collection agencies. As a result, when comparing the health outcomes with general population, blood donors are healthier, and this effect is called HRE. Further, less healthier donors usually stop donating sooner, as a result it is more likely that active donors are on average healthier than inactive/lapsed donors. This bias is known as HDSE. HDCE arises in analyses that compare health outcomes within active donor population. Since blood donors face repeated medical check-ups throughout their donation career, donors who donate at a high frequency and higher number of donations over the lifetime are healthier than low frequency donors [15]. Because of these selection biases, it is very challenging in donor health studies to make causal inference whether the health outcome studied occurred due to blood donation or due to HDE. Therefore, acknowledging and adequately adjusting the HDE is extremely important when comparing health effects of blood donation.

In this methods review, we assessed the methods used to acknowledge and adjust the HDE in studies examining the relationship between blood/plasma donation and long-term health outcomes. This review summarizes the extent of this issue and currently used approaches to mitigate the impact of HDE on the relationship between donation and health outcomes.

2.1.3 Materials and Methods

Data sources and searches

We searched 4 electronic databases PubMed, Scopus, EMBASE, CINAHL to identify relevant peer-reviewed articles, and ProQuest for the inclusion of dissertation and theses available up to November 2020. We did not specify the starting publication date in the search. We used a combination of keywords and index terms for all available studies published in the English language. We also searched the reference list of included papers and conducted a forward citation search. The search strategy was based on the PubMed database and was replicated in other databases after necessary modifications. A detailed search strategy is provided in supplementary materials (Appendix A).

Study Selection

We exported all the articles found from search results to EndNote 9 for data management. We screened titles and abstracts in Covidence after importing from the EndNote. Detailed inclusion-exclusion criteria are given in table 1. At first, duplicate titles were removed, and then the title and abstract screening were done by two independent reviewers, MMR and AH. The reviewers resolved any disagreement by discussing first and if necessary, these were resolved by SK. Screening of full-text articles was carried out by MMR and SK independently, and AH resolved disagreements following discussion.

Table 2.1 Inclusion and Exclusion criteria for the study selection.

Inclusion Criteria	Exclusion Criteria
• Studies examining the association between blood donation and long-term health outcomes. • Adults study population. • Randomised controlled trial, Cohort, Case control and Cross-sectional studies. • Health outcomes such as cardiovascular/heart diseases, cancer, diabetes, bone density or fracture, all-cause mortality, morbidity and infections.	• Health outcomes other than cardiovascular diseases, cancer, bone density or fracture, all-cause mortality, morbidity and infections. • Studies where blood donation is not studied as an intervention/exposure. • Studies conducted on blood/blood product recipients. • Short term blood donation-related reactions and adverse events after blood donation such as localised swelling, vasovagal reactions, iron deficiency, anaemia. • Studies including donors donating blood for therapeutic reasons. (e.g. haemochromatosis, polycythaemia rubra vera). • Studies where study population are autologous donors.

Data Extraction

We extracted data on several study characteristics such as publication year, country, study hypothesis, study population, intervention, study design, data collection method for both exposure and outcome, sample size, follow-up duration, donation type, main statistical method used, primary outcome, and results relating to the primary outcome. We also extracted information on acknowledgement of the HDE as a potential limitation in their analysis, and whether they attempted to adjust for it.

Evidence synthesis/Analysis

To assess the acknowledgement and adjustment of the HDE, we calculated the percentage of studies acknowledging that the HDE may be a potential bias in their analysis and those attempting to adjust this effect. We considered studies to have acknowledged the HDE if they named the term HDE or the inclusion of healthy subject on the results of the study anywhere in their article. We also defined 'HDE acknowledged as a limitation' if the studies acknowledged that the result of their studies could be impacted by the HDE. We further explored the methods used in the studies described to mitigate the HDE and provided a qualitative summary of these methods. We also highlighted the likely residual HDE which could still impact the study results after the author's described their adjustment methods. We did not perform a risk of bias assessment of the included studies as our focus was to provide a descriptive assessment of the methods used to adjust the HDE and not on the outcome of the studies.

2.1.4 Results

Search Results

Figure 2.1 shows the flow diagram of the study selection process. We identified 8784 studies by searching 5 electronic databases. After removing duplicates, 4976 papers were left for screening. The title/abstract screening further excluded 4896 papers and left 80 papers for full-text screening. Of the 80 articles, 57 were excluded with various reasons, and 23 were found eligible for the review. We also included 4 more articles through a forward citation search, resulting 27 articles to be included in the review.

Study characteristics

Table 2.2 shows important study characteristics such as publication year, country of publication, intervention, study design, follow-up duration, donation type, and main statistical

method used in the studies. All the studies were published between 1983 and 2020. Of the 27 studies, 30% were from the USA, and 26% were conducted in either Sweden or Denmark.

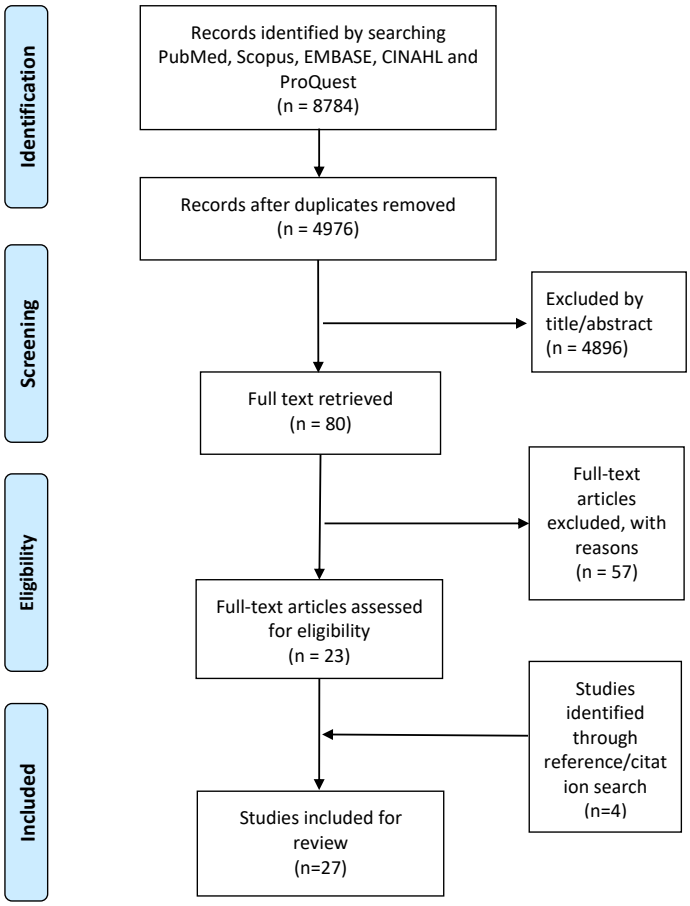


Figure 2.1 Flow diagram of the study selection process.

In terms of comparator group, 14, studies (52%) made comparisons within donor population, whereas 10 (37%) studies made a comparison between the donor and non-donor population.

The study design of 19(70%) studies was cohort study, whereas 4 (15%) reported cross-sectional, and 3 (11%) reported case-control study design. We found only one randomised controlled trial in the study design. Twelve studies (44 %) did not mention the types of blood donors in their studies (such as whole blood donor only or mixed donors), while 8 (29%) studies

reported that they studied only whole blood donors, 6 (22%) studies reported that they studied apheresis donors, and only 1 study (4%) reported that it studied mixed donors.

In terms of analytic methods, 8 (29%) articles used only Cox proportional hazard regression as a main statistical model in their analyses, and 7 (26%) studies used only logistic regression in their analyses. The rest of the studies used Analysis of Variance (ANOVA), Poisson

Table 2.2. Characteristics of the included studies.

First Author	Year Published	Country/Region	Intervention (exposure & control)	Study Design	Sample Size	Duration of follow-up	Donation type	Main statistical method
Casale[55]	1983	Italy	Blood donors vs Non donors	Cross Sectional	731	N/A	Not mentioned	Chi square test and correlation
Merk[56]	1990	Sweden	Comparison between observed and expected no. of cases among blood donors	Cohort	37795	1970 to 1986	Not mentioned	Observed over expected ratio
Lasek[43]	1994	Poland	Comparison between observed and expected no. of cases among blood donors	Cohort	3126	1969 to 1991	Whole blood	Observed over expected ratio
Meyers[12]	1997	USA	Comparison among blood donors and non-donors	Cohort	3855	From Oct 1992 to Nov 1993 to the last 10 years.	Not mentioned	ANOVA, Logistic regression
Tuomainen[36]	1997	Finland	Blood donors vs Non donors	Cohort	2682	1984 to 1992	Not mentioned	Cox regression
Salonen[37]	1998	Finland	Blood donors vs Non donors	Cohort	2682	1984 to 1995	Not mentioned	Cox regression
Ascherio[23]	2001	USA	Comparison across 0/1-4/5-9/10-19/20-29/>= 30 donations	Cohort	38244 (male only)	Jan 1992 to Jan 1996	Not reported	Multiple logistic regression
Meyers[38]	2002	USA	Frequent (more than 1 donation in 3 year period) vs. casual donors (1 donation in 3 year period).	Retrospective cohort study	3016	1990 to 2000	Whole blood	ANOVA, Logistic regression, Cox regression
Jiang[57]	2004	USA	Comparison across 0, 1-5, 6-9, 10-19, 20-	Cohort	38394	the date of returning the 1992 questionnair	Not reported	Cox regression

First Author	Year Published	Country/Region	Intervention (exposure & control)	Study Design	Sample Size	Duration of follow-up	Donation type	Main statistical method
			29 and >= 30 donations.			e (in the analysis of the association between blood donations and diabetes incidence) to the date of the first diagnosis of type 2 diabetes, death, or 1 June 1998,		
Edgren[29]	2008	Sweden and Denmark	Comparisons across 1-8, 9-16, 17-25, > 25 numbers of donations in 3-12 years before diagnosis and comparisons across 0-4, 5-8, 9-12, > 12 numbers of donations in 3-7 and 8-12 years before diagnosis	Nested case control study	1110212 (cases 10866, control 107140)	January 1, 1968, to December 31, 2002	Whole blood and Plasma	Conditional logistic regression
Amrein[10]	2010	Austria	Apheresis donors vs. matched non-donors	Cross-sectional	204(102 donors and 102 controls)	N/A	Apheresis	ANOVA
Zhang[30]	2012	USA	Comparison across), 1-5, 6-9, 10-19, 20-29 and >= 30 donations	Cohort	35121	1992 to the event (death, cancer) or censored on Jan 1, 2008	No reported	Cox regression
Germain[44]	2013	Canada	Comparison among eligible vs. disqualified donors	Retrospective cohort study (quasi-random experiment)	12357 disqualified donors vs. 50889 eligible donors	June 1990 to March 2007	Whole blood	Cox regression
Vahidnia[39]	2013	USA	Comparison across donors with cancer and non-donors with cancer	Cohort	56058 and 36672	from first repository donation to cancer diagnosis, death, or the end of study follow-up (December 31, 2009)	Not mentioned	Poisson regression and Cox regression
Germain[40, 44, 58]	2013	Canada	Comparison between allowed donor	Retrospective Cohort	Deferred: 6076,	1 year	Not mentioned	Logistic regression

First Author	Year Published	Country/Region	Intervention (exposure & control)	Study Design	Sample Size	Duration of follow-up	Donation type	Main statistical method
			with an atypical pulse and deferred donor with atypical pulse		Active: 10671			
Gallerani[40]	2014	Italy	Comparison across non-donors and several blood donor categories.	Retrospective cohort study	55000 (11862 donors, 43138 non-donors)	January 2005 to December 2010	Not mentioned	Logistic regression
Boot[31]	2015	Netherlands	Apheresis donation (treatment) vs whole blood donation (control)	Cross Sectional (pilot study)	40	15 years	Apheresis	ANOVA
Ullum[41]	2015	Sweden and Denmark	Comparison across donation rates (0.01-0.50, 0.51-1.50, 1.51-2.50, 2.51-3.50, 3.51-4.50)	Cohort	1182495	1982 to 2012	Whole blood	Poisson Regression
Ishii[28]	2016	USA	>20 donations vs 0 donations	Cohort	36576	Jan 1992 to Jan 2010	Not reported	Cox regression
Edgren[45]	2016	Sweden and Denmark	Comparisons across 1-8, 9-20, 21-32, ≥ 33 numbers of donations, in different time windows	Nested case control study	1,435,968	from date of first whole blood donation in 1980 or later, until the date of the first diagnosis of PV[a], death, other cancer (including other MPNs[b]), emigration, or end of follow-up (December 31, 2012).	Whole blood	Conditional logistic regression
Grau[32]	2017	Sweden	Comparison across 1-8, 9-24, 25-49, 50-99 and ≥ 100 apheresis donors,	Retrospective cohort	140289	From 1990 to fracture event, censored at December 31, 2012	Apheresis	Poisson Regression
Haron[47]	2018	Malaysia	Comparison across less than 20 vs. more than 50 plateletpheresis	Cross-sectional study	50	N/A	Apheresis (plateletpheresis)	Wilcoxon signed-rank test, Mann-Whitney test, Chi-Square test.

First Author	Year Published	Country/Region	Intervention (exposure & control)	Study Design	Sample Size	Duration of follow-up	Donation type	Main statistical method
Hendig[49]	2018	Germany	Comparison across whole blood, low-frequency plasma, high-frequency plasma, and frequently donate plasma for 5+ years	Cohort	243	N/A	ALL	Multivariate variant analysis
Peffer[42]	2019	Netherlands	Comparison across low/medium/high frequency donors based on tertiles; sex-specific analyses given	Cohort	159934	Time to CVD event; censored Dec 31 2010	Whole blood	Cox regression
Bialkowski[48]	2019	USA	Apheresis blood donations (treatment) vs zero or whole blood donations (control)	Randomised controlled trial	58	1 year	Apheresis	Multiple logistic regression
Zhao[46]	2020	Sweden	Comparison across 1-5, 6-10, 11-25, 26-50 and >50 donations.	Cohort and nested case-control	1021433	from the date of their first whole blood donation until the date of the first diagnosis of any haematological malignancy, death, emigration, or December 31, 2017,	Whole blood	Conditional logistic regression
Zhao[50]	2020	Sweden	Comparison with apheresis donation with LRS[c] and without LRS[c]	Cohort	74408	1996 to 2017 (date of the first apheresis to first infection event, censored at December 31, 2017)	Platelet and plasmapheresis	Cox regression

[a] Polycythemia vera
[b] Myeloproliferative neoplasms
[c] Leukoreduction system

regression, standardised incidence ratios, and a combination of ANOVA or logistic regression or Cox regression in their analysis.

Strategies to address the 'Healthy Donor Effect'

Figure 2.2 shows information on the proportion of studies acknowledging the HDE and the methods used to mitigate the impact of the HDE. We found that 19 studies (70%) acknowledged about HDE bias while the rest of the studies did not mention anything about the HDE. Of these 19 studies that did acknowledge HDE, 13 studies (68%) reported the HDE as a limitation of their studies.

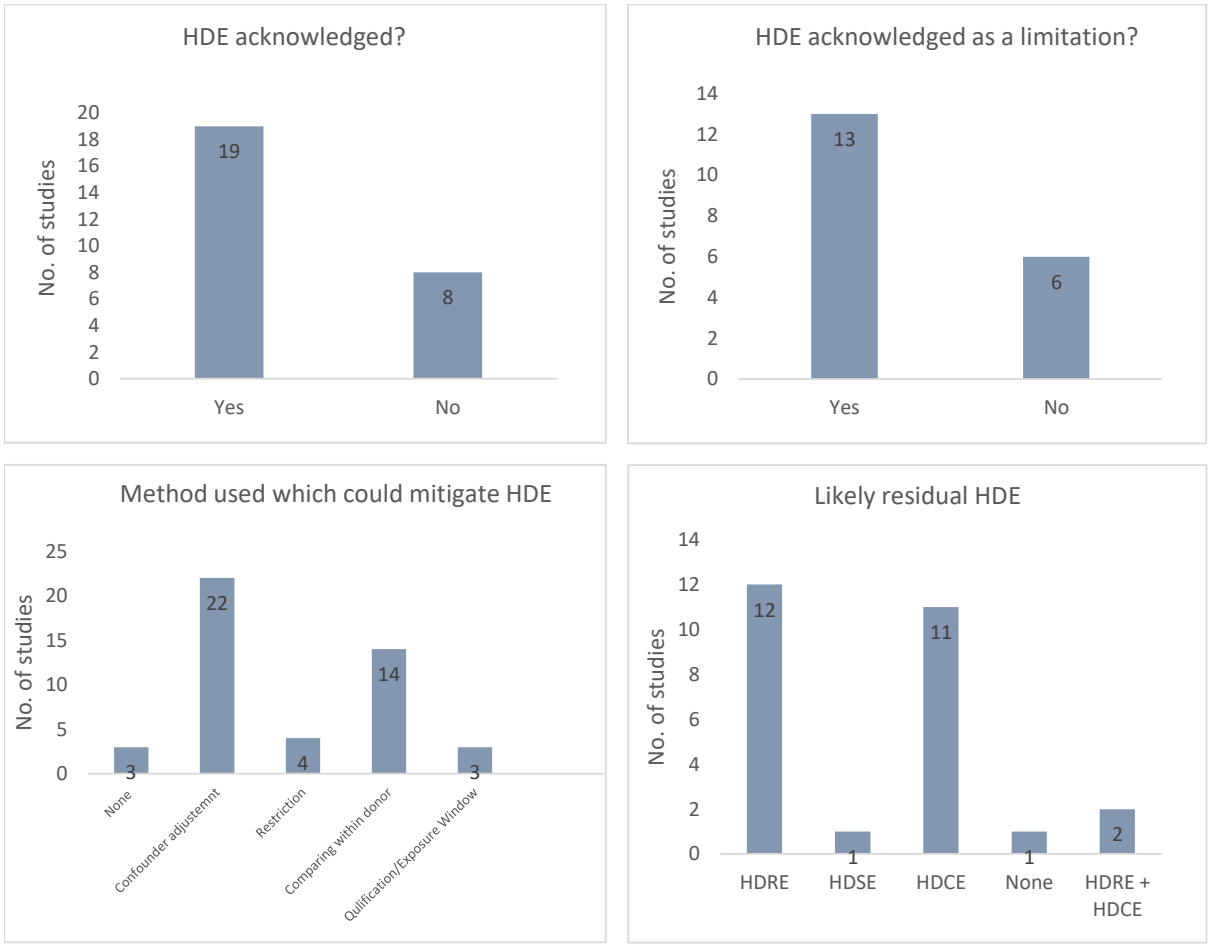


Figure 2.2 Bar charts representing the acknowledgement of the HDE, its adjustment methods, and likely residual HDE information. (HDE= Healthy donor effect).

Of the 27 studies, 3 (11%) did not use any method that could mitigate the HDE while the rest of the studies used one or more methods that could potentially mitigate the HDE. However, 22 (81%) studies used some form of regression analysis or ANOVA to adjust the confounding factors with or without using some other methods which could mitigate the HDE issue. We found 5 (18%) out of the 27 studies restricted the analysis to a healthier group of people along with other methods as a means to adjust the HDE. Another method to adjust for HDE that was used by 14 (52%) out of 27 studies was to compare the difference in outcomes within the donor population. We also found 3 (11%) studies which used qualification period/exposure window method which could potentially mitigate the HDE.

Peffer et al. first proposed the 10-year qualification period method (i.e. all donors must have donated for at least 10 years to be qualified for the study) as a means to adjust for HDE [21, 42]. In this method, the authors adjusted HDE by introducing 10-years qualification period where donors must remain an active donor for at least 10 years. During this period of 10 years, the number of donations was calculated, and comparison was done across the group formed based on donation tertiles. The actual follow-up period started after the qualification period.

Ullum et al. also used a 5-year exposure window preceding the last donation as a means to adjust the HDE [41]. The authors adjusted HDE by only considering outcome among donors with a last donation at ages 64.5-65.0 years and who survived at least 2 years after their most recent donation (these donors were called “retired”). They calculated average number of donations per year in the preceding 5-year window and an indicator for ongoing donation (i.e. not “retired” yet) and used Poisson regression model to estimate the HDE adjusted effect of donation rate on mortality [41]. Similar to Ullum et al., Zhao et al. used 10-year exposure window to mitigate the HDE while examining the effect of platelets donation on lymphopenia and risk of infections [50]. They defined the exposure as the cumulative number of apheresis

donations in a 10-year “exposure window”. For further adjustment of HDE, the 10-year exposure window was delayed by one year where donors might modify their donation habits in the months leading up to an outcome event.

We also made inference about the possibility of residual HDE in the studies included in the review. Of the 27 studies, 12 (44%) studies were likely to be affected by the HDRE, whereas 11 (41%) were likely to be affected by the HDCE. 2 (7%) studies were likely to be impacted both by the HDRE and HDCE, while 1 (4%) study may have been impacted by HDSE. Moreover, we believe that one study was likely not affected by the residual HDE because they used randomised controlled trial as their study design.

Main outcomes reported and confounding factors considered

The study hypothesis, primary outcome, and main results reported in the included studies are shown in a table provided in Appendix II. The included articles studied the association of blood donation with cancers, cardiovascular diseases, bone mineral density, and all-cause mortality. Age and sex were the most common confounding factors followed by body mass index (BMI) and smoking status which were adjusted in the analysis. In general, the observed association between blood donation and long-term health outcomes across the studies were inconsistent. Several studies showed blood donation has a protective effect on CVD, cancer, and death [12, 36-42, 55, 56], while other studies showed blood donation had no association with CVD, cancer, and death [23, 28-30, 40, 43-46]. Some studies also found that apheresis donation was associated with reduction in bone mineral density (BMD) and increased risk of fracture [10, 49], while other studies found no association between blood donation and BMD markers [31, 32, 47, 48]. One of the study reported that apheresis donors had an increased risk of bacterial infection [50].

2.1.5 Discussion

The healthy donor effect is an important methodological problem when studying the long-term health impact of blood donation. We found that about 70% of the studies included in this review acknowledged the HDE in their studies and most of them acknowledged it as a study limitation. We also found while authors used a variety of methods to mitigate against HDE, only a few used methods that may mitigate the HDE substantially. Almost all of these studies were observational studies and have a possibility of residual HDE in their results.

The most common method used by the studies which could mitigate the HDE to some extent was adjusting confounding factors by some form of regression or ANOVA. Many studies used the methods to adjust the bias while some studies used regression for prediction purposes. Atsma et al. defined the HDE as a combination of selection bias and confounding [15]. Adjusting confounding factors can reduce the HDE significantly if all the variables responsible for the HDE are considered and are available in the data. However, it is almost impossible to determine and measure all potential confounding factors. In addition to this, the factors responsible for the HDE are also intermediary determinants of exposure to donation and the health outcomes which means that they are influenced by exposure, and they also affect the probability of subsequent exposure [21]. Thus, adjusting for confounding in regression models is usually not enough to address the HDE adequately.

Another method used by many of the studies was to compare or restrict the study within the active donor population. This approach helps to mitigate the HDRE. Limiting the study within the donor population can considerably diminish HDE, however it is unlikely to eliminate it entirely. The length of the donation career (e.g., short versus long) can still lead to selection bias (for example, donors who are in better health and can withstand all the physiological changes brought upon by donation are likely to have a longer donation career) [59]. As a result,

the existence of residual HDE cannot be ruled out from most of the observational studies even though they restrict the analysis to the active donor population.

Few studies used 'qualification period' or 'exposure window' techniques within the active donor population to adjust the HDE, particularly the HDCE. These methods may be better at mitigating the HDCE than other methods as they ensure that only people who have donated for a long period of time and only active donor at the end of the qualification (exposure) period are included in the analysis. In the qualification period technique described by Peffer [42, 60] they separated the exposure period and the follow-up period and as a result exposure status cannot influence the survival probability which is a major advantage of this technique. However, they admitted that it is impossible to completely rule out the presence of residual HDE in observational study settings [42].

Overall, the methods used in previous studies appear not adequate to fully adjust the HDE when they are used alone or in combination. If the methods are used in combination, then there is a possibility that the HDE might be further minimized. Interestingly, many of the previous studies used these methods together but still they could not rule out the existence of the residual HDE. Thus, it is useful to investigate the use of other methods such as causal inference methods as a means of reducing the HDE in observational studies examining health outcomes in blood donors. In observational study settings, causal inference methods such as, target trial emulation, regression discontinuity, G-estimation etc. can be used to measure the true causal relationship between exposure and outcome [61-63]. These methods could be more effective to adjust for the effect of selection bias. By using these causal inference methods and existing methods together, the HDE may be further minimized.

It is important to note that the extent of bias due to the HDE for the same health outcome may differ between studies based on the donor population, which may vary by blood collection

agency. Different blood collection agencies may have different health-related eligibility criteria and due to this, the selected blood donor population may have some differences in their level of health status. These differences may lead to variation in the magnitude of HDE when the donor population is compared with non-donor population to study the incidence of certain disease outcomes. Most high-income countries such as the USA, Canada, Australia and countries in Europe have strict and similar criteria to screen their donor population to protect the health of donors and transfusion recipients, and so it is less relevant in such context. Also, some donor populations are described to be less healthier [64] and this may also impact the level of HDE in studies examining health outcomes. Further, there is an increasing trend of using blood samples collected from blood donors to measure the sero-prevalence of SARS-CoV-2 [65, 66] and such studies should carefully examine how the health and demographic differences between the background population and blood donor population may impact the result of their studies.

To the best of our knowledge, this is the first review that assessed the HDE adjustment methods thoroughly in studies that examined the long-term health outcomes among blood donors. However, few limitations should be considered while interpreting the findings from this review. We only searched five databases, and we did not include any articles published in languages other than English. We also did not do any risk of bias assessment of the included studies as our focus was to provide a descriptive assessment of the methods used to adjust the HDE and not on the outcome of the studies. In this review, we only included studies with specific outcomes that are more common when comparing donor with non-donor populations in their analyses. Paid and remunerated blood donors could have different health status and studies involving one or other of these population may have different impacts on bias from the HDE. However, we did not differentiate studies between these donors as for some studies this information was not available.

2.1.6 Conclusions

The ‘healthy donor effect’ is an important methodological issue in studies examining the effect of blood donation on long-term health outcomes. Most of the studies studying such relationship acknowledged the impact of HDE and mentioned it as their study limitation. These studies used several forms of adjustment methods to address the HDE, however none of the studies attempted to use any robust causal inference methods. In absence of a feasible/ethical randomised controlled trial, use of causal methods using observational data may be helpful to further mitigate the HDE. However, such studies will need access to a comprehensive set of variables related to the health of donors at baseline and during the follow-up to effectively use them in causal methods, and collection of this information should be taken into consideration in the design phase of studies.

2.1.7 Acknowledgement

The authors are thankful to University of Technology Sydney (UTS) as first author (Rahman) is supported by PhD scholarship for this project. They would also like to thank the Research and Development program of the Australian Red Cross Lifeblood for their support in this project.

Chapter 3 General Methodology

This chapter provides a general overview of the appropriate causal inference methods to be used to adjust for the healthy donor effect while assessing the relationship between regular blood donation and long-term health outcomes. Detailed descriptions of the methods and data analysis are provided in each consecutive chapters.

3.1 Target Trial Emulation

RCTs are considered the gold standard for determining the effectiveness and safety of interventions. However, there are instances where RCTs are impractical or unattainable. In such cases, observational studies can serve as an alternative. This approach is known as "target trial emulation" using observational data. The challenge with observational studies lies in the potential for confounding bias and self-inflicted biases due to study design flaws. Target trial emulation addresses these challenges by applying the principles of randomised trials to observational studies [61, 62, 67]. If successful, this approach can produce the same results as the intended trial would have [61].

To design a target trial, a protocol of the hypothetical randomised trial is specified, considering the constraints of the available observational data. This involves defining the eligibility criteria, treatment strategies, assignment procedures, outcomes, follow-up, causal contrasts of interest, and statistical analysis plan [61, 67]. In this thesis, we will specify a hypothetical blood donor trial and emulate it by our observational data. The detailed description will be given in the next chapter for the separate outcome of interest, mortality and cancers.

3.2 Analysis Plan from Trial

3.2.1 Intention-to-treat (ITT) effect

In a randomised controlled trial, the ITT effect is estimated by comparing the incidence rate of the events in the treated compared to the control group. This effect is commonly estimated by fitting a Cox proportional hazard regression, referred to as the intention-to-treat hazard ratio. One can also approximate this hazard ratio by fitting a pooled logistic regression model, which includes a flexible time function (polynomials or splines) as a time since the start of follow-up [68].

In our study, we will estimate the hazard ratios by fitting the above-described pooled logistic regression model by expanding the data set so that each observation represents a particular individual's 1-month follow-up observation in a person-trial. For example, if someone is eligible for trial one and their follow-up time is 60 months, that individual will contribute to 60 observations in that trial.

The model for the intention to treat analysis is:

$$logit[\Pr(Y_{m+t+1} = 1 \mid \bar{Y}_{m+t} = 0, A_m, L_0, L_m)] = \alpha_{0,m+t} + \alpha_1 A_m + \alpha_2^T L_0 + \alpha_3^T L_m,$$

where Y_{m+t+1} is 1 if someone dies at month t+1 of trial m and 0 otherwise, $m = 0,1,\ldots,59$, $t = 0,1,\ldots.59$, $\alpha_{0,m+t}$ is a time-varying intercept estimated as constant plus linear and quadratic terms of both month m and time t, $\bar{Y}_{m+t}$ is the event history up to time t of trial m, A_m is the indicator for donation initiation (1: Donor, 0: Non-donor), L_0 is the vector of the potential confounding factors at the start of the follow-up when someone becomes eligible and L_m is the vector of potential confounding factors at the start of the trial m. As many individuals participated in more than one trial, we used a robust variance estimator to calculate conservative 95% confidence intervals.

3.2.2 Per-protocol effect

If participants are not fully adherent to the treatment they are given at the baseline, the ITT effect can move towards the null. In a randomised trial, the standard method to deal with this imperfect adherence is to censor the person-months when the participants deviate from their original assigned treatment, which is known as per-protocol analysis. There are mainly two approaches to analyse the per-protocol effect:

- Restricting the analysis to the subjects who continued their baseline treatment for the duration of the follow-up period
- Censoring the individual person-time if/when they deviate from their initial baseline treatment

In our target trial, we will use the second approach to estimate the effect of continuous blood donation, where we compare the risk of all-cause mortality in donors if all participants had donated continuously to the risk in non-donors if all the participants did not donate during the follow-up time. We stop following a donor person-trial if they stop blood donation, and we stop following a non-donor person-trial if they start blood donation.

The outcome model described for the ITT analysis can be applied to this artificially censored population to obtain the per-protocol effect. However, the above model can only adjust baseline confounders. The adherence of donors and non-donors to their baseline donation status can be influenced by post-baseline factors and vice versa. Adjusting for these factors can introduce selection bias. To adjust these time-varying confounding factors, we will use an inverse probability weighted marginal structural model. In general, the IP weight model's denominator is the probability that a subject received his or her observed treatment given their prior treatments and confounder histories, and the numerator, which serves as a stabilising factor, is the probability that a subject received their observed treatments given their prior treatments

and baseline confounders. In our analysis, we will use the following stabilized IP weight model for each patient i at each time $m + t$ as:

$$sw_{m+t}^{A} = \prod_{j=m}^{m+t} \frac{f_N(A_j|\bar{A}_{j-1}, L_0, \bar{Y}_{j-1} = 0)}{f_D(A_j|\bar{A}_{j-1}, L_{0,}\bar{L}_j, \bar{Y}_{j-1} = 0)}$$

Where overbar denotes a variable's history since the start of the trial m. We estimated the numerator by fitting the following logistic model

$$logit\left[\Pr\left(A_j = 1 \,\middle|A_{j-1} = a, L_0, \bar{Y}_{j-1} = 0\right)\right] = \gamma_0 + \gamma_1^T L_0$$

and the denominator by $logit\left[\Pr\left(A_j = 1 \,\middle|A_{j-1} = a, L_0, \bar{L}_j, \bar{Y}_{j-1} = 0\right)\right] = \delta_0 + \delta_1^T L_0 + \delta_2^T L_j$

In the denominator model, the baseline value L_0 and the most recent value L_j of the confounding factors were used to summarize the history $\bar{L}_j$. We will fit two sets of logistic regression models for $a = 1$ (who were donors in the previous month) and $a = 0$ (who were non-donors in the previous month) to calculate the stabilized IP weights. These stabilized IP weights create a 'pseudo population' where donation status is independent of confounding factors. When someone started donating blood, they were treated as continuous donors for the next six months. As a result, we will exclude these observations from our IP weight models and their weights for the pooled logistic regression model were set to 1. To prevent the influence of outliers with the large weights, we will truncate our weights maximum to 10 for the per protocol analysis. Finally, we will fit a weighted marginal structural model by fitting the same logistic regression model used for ITT analysis to our artificially censored population as:

$$logit[\Pr(Y_{m+t+1} = 1 \,|\bar{Y}_{m+t-1} = 0, A_m, L_0, L_m, \bar{C}_{m+t+1} = 0)]$$
$$= \beta_{0,m+t} + \beta_1 A_m + \beta_2^T L_0 + \beta_3^T L_m$$

where, $\bar{C}_{m+t+1}$ is the artificial censoring indicator where someone discontinues their initial donation status at time $m+1$ of trial m.

3.2.3 Standardised survival curve

The average Hazard Ratio (HR) might not provide useful information due to the possibility of HRs changing over specific periods and the presence of inherent selection bias causing these changes [69]. These issues can be resolved by summarising the study findings as appropriately adjusted survival curves which can be termed standardised survival curves. To calculate the standardised survival curves, we will fit a pooled logistic regression model by including product terms between donation indicator and linear and quadratic terms of follow-up time in our previously described intention to treat the model as:

$$logit[\Pr(Y_{m+t+1} = 1 \,|\bar{Y}_{m+t} = 0, A_m, L_0, L_m)] \\ = \alpha_{0,m+t} + \alpha_1 A_m + \alpha_2 A_m t + \alpha_3 A_m t^2 + \alpha_2^T L_0 + \alpha_3^T L_m$$

We used this model to estimate the predicted survival probability at time $m+t$ for individual i under each treatment indicator conditional on the baseline confounders:

$$\hat{s}^a_{i,\,m+t} = \prod_{j=m}^{m+t}[1 - \frac{\exp\,(\alpha_{0,t} + \alpha_1 a_m + \alpha_2 a_m j + \alpha_3 a_m j^2 + \alpha_4^T L_{i,0} + \alpha_5^T L_{i,m}\,)}{1 + \exp\,(\alpha_{0,t} + \alpha_1 a_m + \alpha_2 a_m j + \alpha_3 a_m j^2 + \alpha_4^T L_{i,0} + \alpha_5^T L_{i,m}\,)}]$$

We then calculated standardised survival probabilities at each time point by using the observed distribution of baseline confounding factors in the entire study population:

$$\hat{s}^a{}_{m+t} = \frac{1}{n}\sum_{i=1}^{n} \hat{s}^a_{i,\,m+t}$$

Where n represents the total number of individuals who participated in multiple monthly trials.

3.3 Exposure Window Technique

Peffer et al. suggested a 10-year qualification period inspired by a clinical trial setting [70]. They divided this timeframe, separating the phase where exposure is assessed from when disease events occur. This duration serves as a fix exposure period, during which donors must demonstrate their eligibility. During this timeframe, the exposure, i.e., the number of donations, was determined. The actual follow-up for cardiovascular events among donors begins only after this qualification period. By distinctly setting the exposure determination period, there's an assurance that the survival probability isn't influenced by the exposure status, as they occur in separate phases.

3.3.1 Analysis Plan

3.3.1.1 Inverse Probability Weighting of Marginal Structural Model

We estimated the risk of the outcome among exposure group, risk difference and risk ratio by inverse probability weighted marginal structural model. We fitted a pooled logistic regression model by adding a constant plus linear and quadratic terms of time and also linear and quadratic product terms of donation status and time. The baseline covariates were adjusted by inverse probability weighting and outcome regression. The IPW was truncated at the 99th percentile to remove any extreme weights from outliers. Finally, we used non-parametric bootstrapping with 500 samples to calculate all the 95% CIs. Unweighted and inverse probability weighted survival curves were also plotted.

3.3.1.2 Doubly Robust Methods

Besides using the IP weighted model, I also used alternative g-computation methods, targeted minimum loss-based estimator (TMLE) and sequentially doubly robust estimators (SDR) [71, 72]. These estimators, including IPW, depend on two mathematical models: Treatment and outcome models, which are the function of the confounders. IPW is a singly robust method, as

its correctness depends on specifying the treatment model correctly. However, TMLE and SDR are doubly robust estimators as the estimation from these models remain correct if any of the treatment or outcome model is mis specified. In addition to this, an inverse probability weighted marginal structural model can suffer from positivity assumption violation. In contrast, doubly robust estimators often produce less biased results than the IPW method, even if the positivity assumption is extremely violated [73, 74]. Furthermore, doubly robust estimators like TMLE and SDR can employ machine learning algorithms to fit the treatment and outcome models, which may capture complex associations that are not possible with simple regression-based approaches [72, 75]. Moreover, I also adjusted time-varying exposure and confounders by using TMLE and SDR, as blood donation behaviour was assumed to be time-varying in nature. I used R packages “SuperLearner” and “lmtp” to implement the analysis with doubly robust estimators. [76].

3.4 Data Description

This study did not require primary data collection. We used the Sax Institute's 45 and Up Study [77] dataset linked with several health administrative data to apply the methods developed to examine whether blood donation has any association with donors' long-term health outcomes.

The Sax Institute's 45 and Up Study is the largest prospective health-related cohort study ever conducted in Australia. During 2006-2009, a total of 267,000 NSW residents aged 45 years and over were recruited to the study and provided detailed information on their health, lifestyle, and demographic characteristics. The participants also consented to link their data with other health administration databases and be contacted for additional sub-studies.

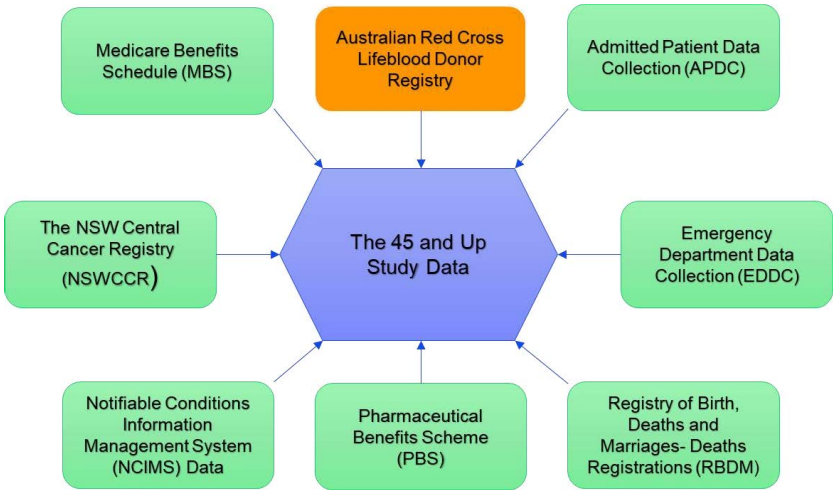


Figure 3.1 Data sources linked with the 45 and Up Study Data.

The cohort is followed-up every five years through surveys, and researchers can ask for periodic data-linkage. This data is already linked to the Australian Red Cross Blood Service Donor Registry, Admitted Patient Data Collection (APDC), The NSW Central Cancer Registry (NSWCCR), Emergency Department Data Collection (EDDC), Notifiable Conditions Information Management System (NCIMS) Data, Registry of Birth, Deaths and Marriages-Deaths Registrations (RBDM), Medicare Benefits Schedule (MBS), Pharmaceutical Benefits Scheme (PBS).

Regarding studies of the health impact of blood donation in previous studies, the most common limitation was the unavailability of important confounding factors to adjust in their studies. The linked 45 and Up study data overcame this problem as it had access to a range of sociodemographic, behavioural, and health-related variables which were not accessible in previous studies. It is important to note that 45 and Up Study data are self-reported. However, data linkage enabled us to validate and cross-reference the 45 and Up Study data with other administrative health data sets which mitigated the concerns about data accuracy and reliability to a significant extent.

3.4.1 Data Linkage Method

The data linkage was conducted by the NSW Centre for Health Record Linkage (CHeReL). CHeReL also deidentified the dataset so that no unwarranted identification can be made. The data linkage methodology adopted by CHeReL is the procedure outlined in Kelman, Bass and Holman [78]. Briefly, the method revolves around a separation principle, where data linkage is conducted separate from the data custodians, and all identifiers (including unique institutional identifiers or full patient identifiers) are withheld from the researchers analysing the linked data. The linkage methodology adopted by CHeReL is depicted in Appendix. The linked data is stored in the Secure Unified Research Environment (SURE) located at the Sax Institute. SURE is a high-powered computing environment that helps to bring researchers together from all over the world to collaborate on large-scale projects. It was established with funding from the Australian Government as a part of the Population Health Research Network (PHRN).

A brief description of the datasets that were used in the proposed studies is shown in Table 3.1. The detailed variable name, their derivation and roles in the analysis are described in the Chapter 4 and tables are given in the Appendix G, Appendix K, and Appendix O.

Table 3.1 Brief description of the data sets linked with 45 and Up Study data.

Name	Database Description
Australian Red Cross Lifeblood Donor Registry	List of Australian blood donors and a history of blood donations. It keeps a centralised, national digital database of all donors and data pertaining to donations. Data-linkage was performed using a donation dataset that only contained donors who made donations on or after June 1, 2002.

Name	Database Description
Admitted Patient Data Collection (APDC)	All admitted patient services are provided by New South Wales Public Hospitals, Public Psychiatric Hospitals, Public Multi-Purpose Services, Private Hospitals, and Private Day Procedures Centres.
The NSW Central Cancer Registry (NSWCCR)	The NSWCCR keeps track of all cancer patients in NSW. Information is gathered from pathology labs, hospitals, radiotherapy, the department of medical oncology, age care facilities, and the registry of births, deaths, and marriages-deaths registration.
Emergency Department Data Collection (EDDC)	The EDDC data records emergency department visits by residents of NSW at public hospitals, which account for the majority of the population.
Notifiable Conditions Information Management System (NCIMS) Data	According to the NSW Public Health Act 2010, labs must report confirmed cases of influenza. NSW Health then aggregates these reports into the NCIMS database. The NCIMS data consists of the specified condition, estimated onset date, laboratory confirmation details, and specimen type.
Registry of Birth, Deaths and Marriages- Deaths Registrations (RBDM)	The RBDM records all deaths records in NSW residents.
Medicare Benefits Schedule (MBS)	Medicare Australia collects MBS claims data which includes all attendances to GPs under the Medicare act 1973. These data are then regularly provided to the Australian Government Department of Health.
Pharmaceutical Benefits Scheme (PBS)	The PBS dataset includes details on prescription drugs that are eligible for benefits as per the National Health Act 1953 and for which a claim has been submitted.

Chapter 4 Application of the HDE adjustment methods

Section 1: A target trial emulation to estimate the impact of blood donation on mortality in blood donors in Australia.

Md Morshadur Rahman[1,2], Surendra Karki[2], Andrew Hayen[1]

[1] School of Public Health, University of Technology Sydney, Sydney, Australia

[2]Research and Development, Australian Red Cross Lifeblood, Sydney, Australia

Journal: Transfusion

Type of Publication: Research paper

Stage of Publication: Submitted

Abstract

Background: The healthy donor effect (HDE) is a bias evident when comparing the effect of blood donation on health, often leading to conclusions such as donors with a high frequency of donation have better health outcomes than non-donors or donors with a lower frequency of donation. To overcome this bias, we proposed a target trial emulation method and investigated the relationship between blood donation and mortality in blood donors in Australia.

Methods: We emulated 60 target trials from July 2006 to June 2011 using the Sax Institute's 45 and Up Study data, which was linked with other electronic health databases, including blood donation data in NSW, Australia. We conducted observational analogues of intention-to-treat (ITT) analyses comparing donors with non-donors, adjusting our analyses for variables that impact mortality. Hazard ratios were approximated by the pooled logistic regression model.

Results: The 60 trials generated 263300 person-trials with 121967 unique participants included in the analysis. The unadjusted ITT mortality hazard ratio (95% Confidence Interval (CI)) between donors and non-donors was 0.57 (95% CI 0.42, 0.77). The HR was 0.70 (95% CI 0.52, 0.95) and 1.0 (95% CI 0.73, 1.35) after adjustment for age and sex, and then all the baseline covariates, respectively.

Conclusion: The ITT effect did not show any statistically significant association between mortality and blood donation, although blood donors tended to have a lower crude mortality rate than non-donors.

Background

The majority of studies on the effects of blood donation have considered short or medium-term safety outcomes, such as bruises, fainting, and iron deficiency [9]. However, recent advances in data linkage and access to large data sets now allow for research on long-term safety outcomes and potential health benefits, such as the impact on cardiovascular disease and cancer. Examining these long-term effects is essential for blood collection agencies to fulfil their ethical responsibility of ensuring blood donation safety and uncovering potential benefits.

Studies of blood donors have claimed that regular blood donors have a lower incidence of death, cancer, and cardiovascular disease compared to the general population [41, 42, 51, 55, 56]. This suggests that there may be a beneficial health effect associated with blood donation. Also, several studies have indicated that there is no association between those health outcomes and blood donation [23, 28, 29, 40, 43, 46]. It is therefore important to note that beneficial health effects may be because blood donors are typically healthier than the general population. Individuals must meet specific health and lifestyle criteria to be eligible for initiating and continuing blood donation [41].

Blood donors tend to have better self-reported health than the general population, which is a strong determinant of blood donation activity [15, 52]. This phenomenon is known as the "healthy donor effect," which is a combination of selection bias and confounding [15]. The healthy donor effect may act as reverse causation, meaning that it shows a slower disease rate in donors than non-donors and in current donors than in previous donors. This can be misinterpreted as a beneficial effect of blood donation when in reality, it is due to the healthier characteristics of blood donors [41].

In this study, we used a target trial emulation technique to examine the relationship between whole blood donation and mortality risk in Australian blood donors at least 45 years of age.

We used data from the Sax Institute's 45 and Up Study with linked records from Lifeblood's blood donation data sets and other external health data sets [77, 79]. The availability of multiple self-reported and verified health-related variables in our data sets and a target trial emulation offered a chance to account for the "healthy donor effect comprehensively

Methods

Target Trial Emulation

Conducting a randomised controlled trial can be difficult or impossible due to feasibility, ethical considerations, cost, and time limitations. In such cases, researchers may attempt to simulate a trial by using existing large observational databases. This approach is known as "target trial emulation" using observational data. If successful, this approach can produce the same results as the intended trial would have [61].

Hypothetical Blood donor Target Trial

Eligibility

The blood donor target trial will enrol participants who are at least 18 years but younger than 70 years of age, with no prior history of circulatory diseases, any type of cancer, or infectious diseases (e.g., human immunodeficiency virus (HIV), hepatitis, human T-lymphotropic virus (HTLV)), and chronic diseases (e.g., chronic kidney disease, autoimmune hepatitis). Additionally, trial participants cannot have made any whole blood donations within two years before enrolment. The duration of the enrolment period will be determined based on the required sample size and the pace of recruitment.

Treatment and Control Groups

The participants would be randomly assigned to the treatment and control groups if eligible to be blood donors. In the treatment group, each participant will start donating blood and continue to do so for the duration of the trial. Each participant must donate at least once every six months

to be considered a continuous donor. The control group participants cannot donate blood for the duration of the trial. Each participant will be followed until they die, are lost to follow-up, or the administrative end of the study, whichever occurs first.

Emulating the trial using observational data

Data source

We emulate the above hypothetical randomised trial by using the Sax Institute's 45 and Up Study data, which is linked to other electronic health databases such as the Australian Red Cross Lifeblood Donor Registry, Admitted Patient Data Collection (APDC), Registry of Birth, Deaths, and Marriages- Deaths Registrations (RBDM), Medicare claims, and Pharmaceutical Benefits Scheme (PBS).

The Sax Institute conducted the 45 and Up Study, enrolling 267,357 participants aged 45 or older in New South Wales, Australia, from 2005 to 2009. Prospective participants were randomly selected from the Services Australia Medicare enrolment database and had a 19.2% participation rate [79]. Residents living in rural and remote areas and people 80+ years of age were oversampled. The participants completed an initial questionnaire covering socio-demographics, health, lifestyle, and behaviour and consented to long-term follow-up by linking their data to various administrative databases.

The Australian Red Cross Lifeblood manages the entire blood collection, processing, and distribution process and maintains a centralized National Blood Management System (NBMS) for donor records. Before 2007, Lifeblood had varying methods for storing donor data. After the 2007 national merger, all donor information was consolidated within the NBMS. However, complete data for donations in New South Wales (NSW) was only available from June 1st, 2002, onwards. Consequently, data linkage utilised the dataset of blood donations made between June 1st, 2002, and December 31st, 2018.

The Admitted Patient Data Collection (APDC) database in New South Wales (NSW) contains extensive records of all inpatient admissions in the state, including admission and discharge dates, primary diagnosis, and up to 49 secondary diagnoses that may impact treatment or length of stay. This data was used to compute the Charlson Co-morbidity Index (CCI), a tool that evaluates the impact of multiple chronic conditions on patient outcomes. This data is complete up to June 2018.

The Medicare claims data compiles information on medical services and procedures subsidized by the Australian government under Medicare. It includes extensive details on the types and frequency of services offered by medical practitioners, such as consultations, diagnostic tests, surgical procedures, and allied health services like physiotherapy and psychology. This data was utilised to determine the annual total number of GP visits and the number of specialist consultations and pathology test referrals. MBS data is complete up to December 2017.

The Australian Pharmaceutical Benefits Scheme (PBS) data contains information on subsidized medicines under the program, including details about the type, quantity, and cost of each medicine. Using PBS data, Rx-Risk co-morbidity index was calculated to measure the overall burden of illness and co-existing medical conditions in a patient population. The dataset is complete until December 2017.

The NSW Registry of Births, Deaths, and Marriages (RBDM) database holds records of residents' birth, death, and marriage dates. In our study, we used the RBDM death data to establish the date of death and all-cause mortality. It is important to note that the RBDM dataset was completely updated up to December 2018.

According to the Public Health Act 2010, it is mandatory for laboratories, hospitals, medical professionals, educational institutions, and childcare facilities to inform either NSW Health or their local public health department about specific infectious diseases and any adverse events

that occur after immunisation. These reports are gathered and organized within the Notifiable Conditions Information Management System (NCIMS), overseen by the Communicable Diseases Branch of Health Protection NSW. This data was used to exclude the people who had any type of Hepatitis or HIV diseases 2 years before being included in the trial.

The NSW Centre for Health Record Linkage (CHeReL) conducted the process of connecting the 45 and Up Study data with NBMS, APDC, and RBDM, utilising a probabilistic matching approach. Past quality assurance assessments of CHeReL's linkages using the master linkage key have indicated an estimated false positive rate of 0.5%. The Sax Institute supplied Services Australia with distinct identifier information for the 45 and Up Study participants, and Services Australia provided the corresponding Medicare claim and PBS data based on these unique identifiers. The Sax Institute then linked the Medicare claim and PBS data to the 45 and Up Study data by the deterministic matching procedure.

Eligibility

We identified all individuals in the linked dataset who had enrolled in the 45 and Up Study before July 2006 and were less than 70 years old as of July 2006. We then excluded individuals who had donated whole blood at least once within the two years before July 2006. Furthermore, we excluded individuals with a history of cancer, cardiovascular disease, infectious disease, or a chronic disease before July 2006. By implementing the specified criteria, we identified a total of 20,507 individuals that met all requirements for the trial of July 2006.

Treatment and Follow-up

The individuals enrolled in the July 2006 trial were classified into two groups: the donor (treatment) group who donated blood in the month of July 2006 and the non-donor (control) group who did not donate during that month. We considered a fixed follow-up period for everyone. Everyone was followed until their death or until five years after the start of their

follow-up time, whichever occurred first. Of the 20,507 participants from the July 2006 trial, 30 were donors, and 299 died during the follow-up. No one from the donor group died during the follow-up period.

Emulating a sequence of trials

Our trial for July 2006 only resulted in a small number of donors, and none of them experienced an event. As a result, it was impossible to analyse the data from this trial only. To increase the number of donors and events, we applied the eligibility criteria described above to every month between July 2006 and June 2011, emulating 60 randomised trials. Each trial had a 1-month enrolment period, and we followed every participant until their death or five years after each trial's start, whichever occurred first. A total of 153,393 individuals met the eligibility criteria for at least one of these 60 trials. Many participants were included in more than one trial as they met the eligibility criteria for that trial month. For example, many eligible non-donors from the July 2006 trial still met the eligibility criteria for the August 2006 trial and were thus enrolled in that trial as well. However, donors from the July 2006 trial did not meet the eligibility criteria for the August 2006 trial as they fell into the donor category who donated within the last two years for that trial month. After pooling all the trial data, there were 5,180,763 eligible person-trials with 153,393 unique people. For computational efficiency, we randomly chose 5% of the non-donors from each trial, which generated 263,300 person-trials, of which 5,837 were donors, and 3425 died (44 among donors). The average duration of follow-up time was 58.9 months for donors and 58.7 months for non-donors. Figure 4.1 illustrates the person-trials enrolled in the study.

Confounding factors

Unlike randomised control trials, trial emulation from observational data requires adjustment for confounding as the treatment and control arm are not randomly assigned. We selected several baseline variables from the 45 and Up Study data as potential confounding factors: age,

sex(male, female), geographical location (major city, regional/remote), education (no formal education, school to diploma, university), gross annual household income (<20, 20-39, 40-69, 70+ thousand), BMI (0-18.4, 18.5-24.9, 25-29.9, 30+ kg/m^2), self-reported health (excellent, very good, good, fair/poor), smoking status (never, former, regular), daily alcohol intake (none, ≤1/day, >1/day), weekly physical activity (<1/week, ≥1/week), and daily fruits or raw vegetable consumption (0-2, 3-4, 5+). We also adjusted for time-varying variables such as the number of general practice (GP) visits in the last three months (0,1, 2-4, 5+), the number of specialist consultations and pathology test referrals for the last three months (0, 1, 2-4, 5+), Charlson's comorbidity index for the previous one year (0, ≥1), and the previous year's Rx-Risk index (-6 to -1, 0 to 2, 3+); these are described below. The number of GP visits and referral numbers were calculated from the Medicare claim data. Charlson and Rx-Risk comorbidity index was calculated from Admitted Patient Data Collection (APDC) and Pharmaceutical Benefits Scheme (PBS) data.

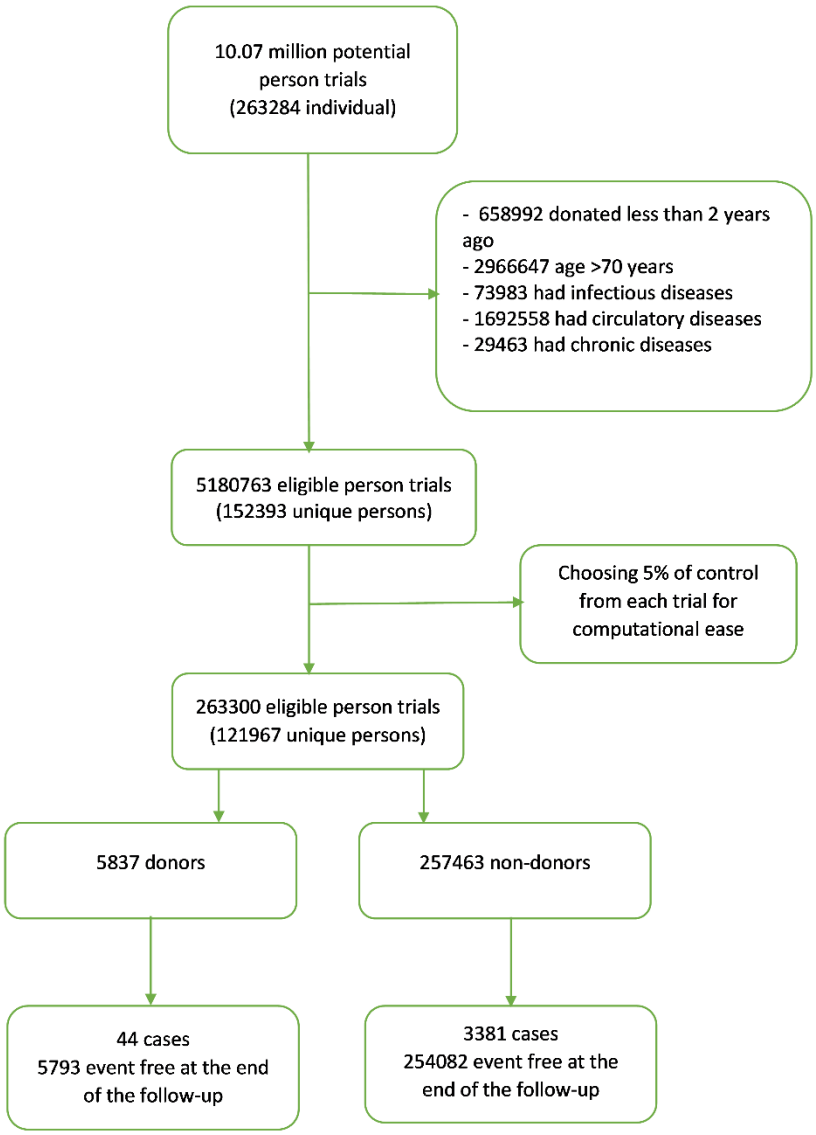


Figure 4.1 Flowchart of the selection of person-trials in the target trial emulation process

The Charlson Comorbidity Index (CCI) predicts one-year mortality rates for patients with chronic conditions [80]. This tool assigns weights to 19 medical conditions based on their association with mortality, with larger weights given to more severe or life-threatening conditions, ranging from 1 to 6. A patient's CCI score is calculated by adding up the weights of all their comorbidities. We used the updated CCI index calculated from hospital discharge

data using the ICD-10 coding algorithm [81]. The Rx-Risk Comorbidity Index uses PBS data to identify comorbidities and assign weights to 45 medical conditions based on their association with increased healthcare use [82]. These weights range from 1 to 3, with higher weights given to more severe or complex conditions. A patient's Rx-Risk score is determined by adding up the weights of all their comorbidities. In this analysis, we used the Australian version of the Rx-Risk index, which is calculated using medicines mapped to the Anatomical Therapeutic Chemical (ATC) classification system [83]. The detailed method of calculation of these two co-morbidity indexes can be found elsewhere [80-84]. Comorbidity scores are frequently utilised in observational research to minimise confounding risks. The primary benefit of these summary scores is that they simplify the process of combining individual covariates related to each comorbidity into one comprehensive score.

The distribution of confounding factors for donors and non-donors included in the trial is shown in Table 4.1.

Table 4.1 Characteristics of the study participants at the start of the trial's follow-up

Characteristics	Non-donor 257463 person-trials	Donor 5837 person-trials
Male %	41.7	39.0
Age in years, mean (standard deviation)	58.0 (6.3)	56.1 (5.8)
Body mass index kg/m^2, mean (sd)	27.0 (5.0)	27.0 (4.6)
Smoking Status %		
Never	57.4	61.5
Former	32.7	32.9
Regular	9.4	5.2
Unknown	0.5	0.3
Self-rated health %		
Excellent	18.1	25.5
Very good	38.8	44.5
Good	30.3	24.1
Fair/Poor	10.0	4.0
Unknown	2.8	2.0
Alcohol consumption/day, mean (sd)	1.0 (1.4)	1.1 (1.3)

Characteristics	Non-donor 257463 person-trials	Donor 5837 person-trials
Education level %		
No formal education	8.9	5.1
School to Diploma	62.5	63.2
University	27.6	30.9
Unknown	1.1	0.8
Annual household income %		
<20k	12.7	6.9
20k-39k	15.8	13.7
40k-69k	21.0	23.6
70k+	31.5	40.6
Unknown	19.0	15.2
Physical activity/week, mean (sd)	1.9 (4.9)	2.1 (2.8)
Daily fruits/vegetables consumed %		
0-2	6.6	5.2
3-4	24.6	24.7
5+	53.9	57.6
Unknown	14.9	12.5
Location %		
Major city	51.6	46.6
Regional/Remote	46.4	51.0
Unknown	2.0	2.5
No. of GP visits in the past 3 months, mean (sd)	1.3 (1.7)	1.1 (1.4)
No. of referrals in the past 3 months, mean (sd)	0.9 (1.3)	0.6 (1.0)
Charlson co-morbidity index %		
0	98.7	99.6
>= 1	1.3	0.4
Rx-Risk index %		
None	43.7	55.5
-6 to -1	27.8	23.0
0 to 2	22.6	18.3
3+	5.9	3.3

Analysis

Intention-to-treat (ITT) effect

In a randomised controlled trial, the ITT effect is estimated by comparing the incidence rate of the events in the treated compared to the control group. This effect is commonly estimated by fitting a Cox proportional hazard regression, referred to as the intention-to-treat hazard ratio. One can also approximate this hazard ratio by fitting a pooled logistic regression model, which

includes a flexible time function (polynomials or splines) as a time since the start of follow-up [68].

In our study, we estimated the hazard ratios by fitting the above-described pooled logistic regression model by expanding the data set so that each observation represents a particular individual's 1-month follow-up observation in a person-trial. For example, if someone is eligible for trial one and their follow-up time is 60 months, that individual will contribute to 60 observations in that trial.

The model for the intention to treat analysis is:

$$logit[\Pr(Y_{m+t+1} = 1 \mid \bar{Y}_{m+t} = 0, A_m, L_0, L_m)] = \alpha_{0,m+t} + \alpha_1 A_m + \alpha_2^T L_0 + \alpha_3^T L_m,$$

where Y_{m+t+1} is 1 if someone dies at month t+1 of trial m and 0 otherwise, $m = 0,1, \ldots,59$, $t = 0, 1, \ldots.59$, $\alpha_{0,m+t}$ is a time-varying intercept estimated as constant plus linear and quadratic terms of both month m and time t, $\bar{Y}_{m+t}$ is the event history up to time t of trial m, A_m is the indicator for donation initiation (1: Donor, 0: Non-donor), L_0 is the vector of the potential confounding factors at the start of the follow-up when someone becomes eligible and L_m is the vector of potential confounding factors at the start of the trial m.

Standardised survival curve

In the above intention-to-treat model, we fitted a pooled logistic regression model by including product terms between donation indicator and linear and quadratic terms of follow-up time:

$$logit[\Pr(Y_{m+t+1} = 1 \mid \bar{Y}_{m+t} = 0, A_m, L_0, L_m)]$$
$$= \alpha_{0,m+t} + \alpha_1 A_m + \alpha_2 A_m t + \alpha_3 A_m t^2 + \alpha_2^T L_0 + \alpha_3^T L_m$$

We used this model to estimate the predicted survival probability at time $m + t$ for individual i under each treatment indicator conditional on the baseline confounders:

$$\hat{s}^{a}_{i,\ m+t} = \prod_{j=m}^{m+t}[1 - \frac{\exp\left(\alpha_{0,t} + \alpha_1 a_m + \alpha_2 a_m j + \alpha_3 a_m j^2 + \alpha_4^T L_{i,0} + \alpha_5^T L_{i,m}\right)}{1 + \exp\left(\alpha_{0,t} + \alpha_1 a_m + \alpha_2 a_m j + \alpha_3 a_m j^2 + \alpha_4^T L_{i,0} + \alpha_5^T L_{i,m}\right)}]$$

We then calculated standardised survival probabilities at each time point by using the observed distribution of baseline confounding factors in the entire study population:

$$\hat{s}^{a}{}_{m+t} = \frac{1}{n}\sum_{i=1}^{n} \hat{s}^{a}_{i,\ m+t}$$

Where *n* represents the total number of individuals who participated in multiple monthly trials.

Sensitivity analyses

For the purpose of the sensitivity analyses, we examined several scenarios. We created a target trial where we emulated 60 trials. Each individual was followed until they died or until an administrative end of the follow-up (June 2016), whichever occurred first. We also created another target trial where we emulated 120 trials by applying the same eligibility criteria described earlier and followed each individual until their death or five years after the start of their follow-up time, or until July 2016, whichever occurred first. We also did a negative control analysis using injury-related hospitalization as a negative control outcome. For this, we emulated 60 and 120 trials. We followed the individual until the injury-related hospitalization occurred, or until their death or five years after the start of their follow-up time, or until July 2016, whichever occurred first.

Ethics approval

The 45 and Up Study received approval from the Human Research Ethics Committee (HREC) at the University of New South Wales. Additionally, the study was approved by the NSW Population Health Research Ethics Committee (HREC) under the reference number 2016/02/633 and the Lifeblood HREC with the reference number 2015#13.

Results

ITT effect and Standardised survival curves

Table 4.2 shows the adjusted and unadjusted hazard ratio of all-cause mortality and 95% confidence intervals for our trial emulation. The total number of person trials was 263,300, with 121,967 unique individuals. We found a total of 3,425 deaths, where 1,560 were unique deaths. The unadjusted ITT hazard ratios were 0.57 (95% CI 0.42, 0.77) and 0.70 (95% CI 0.52, 0.95) after adjusting for age and sex. When adjusted for all the baseline confounding factors, the hazard ratio was 1.00 (95% CI 0.73, 1.35).

Table 4.2 Intention to treat (ITT) hazard ratio for 60 trials with 95% confidence intervals.

	Donor vs. non-donor
Unique Persons	121967
Cases	3425
Unique cases	1560
Person trials	263300
Unadjusted[a]	0.57 (0.42, 0.77)
Age-sex adjusted[a]	0.70 (0.52, 0.95)
Adjusted for all baseline covariates[ab]	1.00 (0.73, 1.35)

[a]Confidence intervals are calculated using a robust variance estimator as many individuals participated in more than one trial.
[b]Baseline variables in Table 4.1 were included as covariates.

Figure 4.2 shows the standardised survival curve for the 60-trial emulation. When the hazard ratio is allowed to vary over time, it also depicted very insignificant differences between the donor and non-donor survival rates (Figure 4.2).

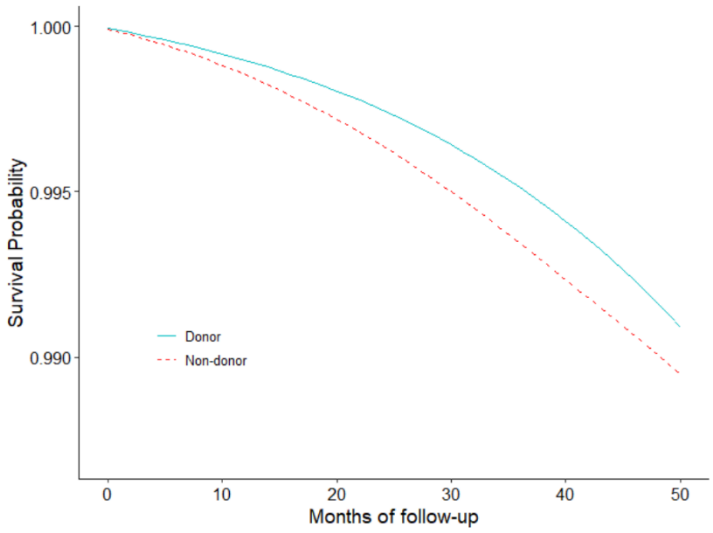


Figure 4.2 Standardised survival curves for donor and non-donor group for 60 trials.

Sensitivity analyses

The fully adjusted hazard ratio for 60 trials where each of the individuals was followed until they died or until an administrative end of the follow-up (June 2016) was 0.87 (95% CI, 0.68-1.10), whereas for 120 trials where everyone was followed until their death or until five years after the start of their follow-up time, or until July 2016 whichever occurred first was 0.83 (95% CI, 0.62-1.11). The fully adjusted hazard ratio for 60 trials for the negative control was 0.94 (95% CI, 0.81-1.10), and it was 0.98 (95% CI, 0.86-1.12) for 120 trials. Detailed results of the sensitivity analysis are given in the appendices.

Discussion

We examined whether blood donation is associated with reduced mortality risk by applying target trial emulation technique to minimise the HDE. Our adjusted mortality hazard ratio and standardised survival curves did not show a significant association between blood donation and reduced mortality. We consistently implemented the qualification criteria for participation in

60 consecutive trials to reduce the influence of the HDE. This was achieved by evaluating individuals for eligibility for blood donation at the start of each month for each trial and only including those who met the criteria. Additionally, we adjusted the analyses for a diverse array of potential confounding factors, including self-reported health indicators, to mitigate further against the potential effects of the HDE. Furthermore, we conducted sensitivity analyses, including emulation of 120 trials and variations in the conditions of our primary 60 trials. In both analyses, we did not find a significant association between blood donation and mortality. Additionally, our negative control analysis found no significant association between blood donation and hospitalization due to injuries, further supporting the fact that our analysis may have adequately adjusted the HDE.

Many previous studies have reported lower mortality among blood donors compared to the general population and among high-frequency donors compared to low-frequency donors [85]. Most of these studies could not rule out the existence of residual healthy donor effect and did not use any causal inference methods which could adjust the selection biases [85]. Casale et al. found longer life expectancy in blood donors than non-donors; however, the findings may have been impacted by the HDE as they did not use any causal inference methods or adjust any confounding factors [55]. Edgren et al. also found lower mortality among donors, but they did not conclude the results were due to the impact of blood donation [51]. Other studies also found lower mortality rates among high-frequency blood donors compared to low-frequency blood donors but suggested they could not rule out the existence of residual HDE [38, 41, 52].

Our study has several strengths. We used linked data with a range of important potential confounding factors, which helped us to emulate the target trial and also used for the statistical adjustments [15]. The target trial method enabled us to create less biased comparison groups to compare the outcome between donors and non-donors as it can mimic an RCT, and a successful trial emulation can produce the same results as the intended trial would have [61].

In the target trial method, we generated a sequence of trials and assessed the eligibility criteria at each trial month, which can reduce the HDE to some extent. Further statistical adjustment of overall baseline confounding factors and trial-specific baseline confounders further reduced the HDE. It is important to note that the 45 and Up Study data is self-reported. However, data linkage enabled us to validate and cross-reference the 45 and Up Study data with other administrative health data sets which mitigated the concerns about data accuracy and reliability to a significant extent.

There are also some limitations in our study. In an RCT, participants may not fully be adherent to the treatment they were given at the baseline. The common method to deal with this imperfect adherence is to censor the person-months when the participants deviate from their originally assigned treatment which is known as Per-Protocol analysis. In this study, using a marginal structural model to adjust for the time-varying selection bias could further reduce the healthy donor effect in the per-protocol analysis. However, we only had six events in the donor group after artificially censoring the non-adherent participants. In addition, 55.8% of donors switched to the non-donor group within one year of the follow-up. Due to this low number of events and the large number of non-adherences within the donor group, we did not conduct a per-protocol analysis as it could yield misleading findings. Another limitation of our research is that our study population comprised older Australian donors, which may not be generalizable to the younger donor population. Furthermore, we evaluated individuals for eligibility to donate at the beginning of each trial using some available data, but in actual blood donor assessment, a large number of factors are assessed and considered to determine whether someone is eligible to donate or not.

In summary, the emulation of the target trial from linked data sets and further adjustment of several potential confounders helped us to reduce the HDE while examining the effect of blood donation on all-cause mortality. Our analyses did not observe a statistically significant

association between whole blood donation and all-cause mortality in the ITT analysis. Further studies are needed to understand the effect of strict regular blood donation versus no donation on mortality. This can give more insight into whether regular blood donation has any impact on mortality.

Acknowledgements

This research was completed using data collected through the 45 and Up Study (www.saxinstitute.org.au). The 45 and Up Study is managed by the Sax Institute in collaboration with major partner Cancer Council NSW; and partners: the Heart Foundation and the NSW Ministry of Health. We thank the many thousands of people participating in the 45 and Up Study and also those who donated blood to save the lives of others. We express our gratitude to Services Australia for providing the Medicare claim and PBS data. The linked data was managed and accessed via the Sax Institute's Secure Unified Research Environment (SURE), which provides secure data access. We thank CHeReL for supplying the linked data (www.cherel.org.au). The Australian government funds the Australian Red Cross Lifeblood for the provision of blood, blood products, and services to the Australian community.

Section 2: High-frequency whole blood donation and its impact on mortality: evidence from a data linkage study in Australia.

Md Morshadur Rahman[1,2], Surendra Karki[2], Andrew Hayen[1]

[1] School of Public Health, University of Technology Sydney, Sydney, Australia

[2]Research and Development, Australian Red Cross Lifeblood, Sydney, Australia

Journal: To be determined

Type of Publication: Research paper

Stage of Publication: Submitted to the Sax Institute for technical review.

Abstract

Background

Previous studies examining the health of blood donors have consistently suggested a decreased mortality risk among blood donors. However, this observed lower mortality risk among blood donors may, to some extent, be attributed to an inadequate adjustment for the "healthy donor effect" (HDE).

Methods

We utilised the Sax Institutes' 45 and Up Study data, along with databases containing blood donation history and other health-related information, to investigate the association between regular, frequent whole blood (WB) donation and the risk of mortality. To address the potential impact of the healthy donor effect (HDE), we implemented a "5-year exposure window" criterion and compared the mortality outcome between individuals classified as regular, high-frequency WB donors (with at least two WB donations in each exposure year) and others (low-frequency donors). We employed statistical methods such as the inverse probability weighted (IPW) marginal structural model and doubly robust Targeted Minimum Loss-Based Estimator (TMLE), which incorporated machine learning algorithms and time-varying analyses.

Results

A total of 4750 (64.7%) low-frequency and 2588 (35.3%) high-frequency donors were identified from the 5-year exposure window, and 69 (1.5%) from the low-frequency and 45 (1.7%) from the high-frequency group died during the 7-year follow-up period. We did not find any significant association between high-frequency blood donors and mortality when compared to low-frequency donors (IPW RR = 0.98 95% CI 0.68, 1.28). TMLE model also showed similar results to IPW (RR = 0.97 95% CI 0.80, 1.16). Time-varying TMLE did not find any significant association between high-frequency donation and all-cause mortality either (RR = 0.98 95% CI 0.74, 1.29).

Conclusions

This study did not observe any significant association between high-frequency WB donation and all-cause mortality when compared to low-frequency blood donation.

Background

Blood donation saves millions of lives worldwide each year. According to the World Health Organization (WHO), approximately 118.4 million blood donations are collected globally every year, with 47% of donations collected in high-income countries [86]. In Australia, 1 in 3 people need blood and blood products in their lifetime, with more than 29,000 donations required every week to meet demand [5].

There is a growing interest in understanding the potential effects of blood donation on donors. Most studies on the effects of blood donation have focused on short-term or medium-term safety outcomes, such as bruising, fainting, and iron deficiency [9]. Recent advances in data linkage and access to large data sets have also made it possible to study long-term safety outcomes and potential health effects, such as the impact of blood donation on cardiovascular disease, cancer or mortality. It is important for blood collection agencies to examine these long-term effects to fulfil their ethical responsibility of ensuring blood donation safety and uncovering potential benefits.

Several studies have found that frequent blood donors have a lower risk of death, cancer, and cardiovascular disease when compared to the general population [41, 42, 51, 55, 56]. These findings suggest that donating blood may have a positive effect on health. However, other studies have contradicted this claim, suggesting that there is no clear relationship between blood donation and these health outcomes [23, 28, 29, 40, 43, 46]. It is crucial to consider that these positive health effects may be due to the fact that blood donors tend to be healthier than the general population, as individuals must meet specific health and lifestyle requirements to be eligible for blood donation [41].

Blood donors often report better overall health than the general population, which is a significant factor influencing blood donation rates [15, 52]. The healthier lifestyle and the better

self-reported health of blood donors is known as the "healthy donor effect," which arises from a combination of selection bias and confounding effects [15]. This effect can create a biased association indicating a lower disease rate among donors compared to non-donors and among current donors compared to previous donors, leading to a misinterpretation of the beneficial effects of blood donation. In reality, the observed positive health effects may be attributed to the generally healthier characteristics of blood donors rather than to the act of blood donation itself [41].

In this study, we aim to examine the relationship between regular high-frequency blood donation and mortality risk in Australian blood donors aged 45 years and older. To adjust for the "healthy donor effect", we used an "exposure window" technique similar to Peffer et al. [42] and further statistical adjustment was done by employing models that use machine learning algorithms to adjust for the time-varying nature of the blood donation exposure and confounders. We used data from the Sax Institute's 45 and Up Study with linked records from Lifeblood's blood donation data sets and other external administrative health data sets [77, 79].

Methods

Data

We used data from the Sax Institute's 45 and Up Study, which is linked to several other electronic health databases. These include the Australian Red Cross Lifeblood Donor Registry, the Admitted Patient Data Collection (APDC), the NSW Central Cancer Registry (NSWCCR), the Registry of Births, Deaths, and Marriages- Deaths Registrations (RBDM), the Medicare claims data, and the Pharmaceutical Benefits Scheme (PBS).

The Sax Institute's 45 and Up Study is a large-scale longitudinal study that recruited 267,357 participants aged 45 years and above who were living in New South Wales, Australia, between 2005 and 2009 [77]. The participants were randomly selected from the Services Australia Medicare enrolment database and invited to participate, and the participation rate was 19.2%

[79]. All the participants completed a comprehensive baseline questionnaire that collected information on various aspects of their lives, including socio-demographics, health, lifestyle, and behaviours. Additionally, the participants provided their consent for their data to be linked to various administrative databases [77].

Australian Red Cross Lifeblood is solely responsible for managing the collection, processing, and distribution of blood and its components in Australia. They have established a centralized national digital record of donors and their related information using the National Blood Management System (NBMS). Every attempted donation date is recorded in the NBMS. Before 2007, Lifeblood functioned as state-based centres, leading to different methods of storing donor records across the states. However, after the 2007 national merger, donor data was merged and updated in the central NBMS. Nonetheless, complete data for donations made in New South Wales (NSW) was only uploaded to the NBMS for donations made on or after June 1st, 2002. Therefore, only the dataset of blood donations made by donors on or after June 1st, 2002, until December 31, 2018, was used for data linkage.

The Admitted Patient Data Collection (APDC) database in New South Wales (NSW) maintains a comprehensive record of every inpatient admission in the state. This includes details such as the admission and discharge dates, the primary diagnosis for admission, and up to 49 secondary diagnoses that may affect the treatment or length of stay of the patient. We used this data to calculate the Charlson Co-morbidity index (CCI), which is used to assess the burden of multiple chronic conditions on patient outcomes. This data is available up to June 2018.

The Medicare claims data is a collection of information on the medical services and procedures that are subsidised by the Australian government under the Medicare system. It contains detailed information on the types and frequency of medical services provided by medical practitioners, such as consultations, diagnostic tests, and surgical procedures, as well as allied

health services, such as physiotherapy and psychology. We used this data to calculate the yearly total number of GP visits and a number of diagnostic test referrals up to December 2017.

The Australian Pharmaceutical Benefits Scheme (PBS) data contains information on medicines that are subsidized by the Australian Government, including details on the type, quantity, and cost of each medicine. We used PBS data to calculate the one-year co-morbidity index, which is a measure of the overall burden of illness and co-existing medical conditions in a patient population until December 2017.

The NSW Registry of Births, Deaths, and Marriages (RBDM) database maintains records of the dates of births, deaths, and marriages of residents in NSW. For our study, we used the RBDM deaths data to determine the date of death and all-cause mortality. The RBDM dataset was fully updated until December 2018.

The NSW Centre for Health Record Linkage (CHeReL) employed a probabilistic matching technique to connect the data from the 45 and Up Study with NBMS, APDC, and RBDM. CHeReL has previously conducted quality assurance evaluations using the master linkage key, which indicated a false positive rate of approximately 0.5%. The Sax Institute provided Services Australia with distinct identifier information for the participants in the 45 and Up Study. Services Australia then supplied the corresponding Medicare claims and PBS data based on these unique identifiers. Finally, the Sax Institute used a deterministic matching procedure to link the Medicare claims and PBS data to the 45 and Up Study data.

Exposure Window

Our study used a similar methodology as Peffer et al. by employing a five-year "exposure window/qualification period" to select the participants for our analyses [42]. The exposure window refers to the duration in which the donor is required to donate blood actively while meeting other criteria for eligibility. In our research, this exposure window spans three years

before the recruitment into the 45 and Up study and two years after. To qualify for inclusion in our study, donors must have made at least one whole blood (WB) donation during the first and fifth years of the exposure window and remained alive during the whole 5-year period. Therefore, our study population consists of individuals who had an active donation career of five years during the exposure window period.

Exposure and outcome variable

To assess the frequency and regularity of blood donations made by participants during each year of the exposure period, we divided the exposure variable into two categories: (i) individuals who donated a minimum of two whole blood units during each year of the exposure window and (ii) those who donated fewer than two whole blood units at least one year of the exposure window. We made this categorization based on the knowledge that most whole blood donors need up to six months to restore their pre-donation levels of stored iron. Therefore, donating blood twice within a year would result in consistently lower levels of stored iron than before donating. The primary outcome variable of this study is mortality from any cause. To determine the outcome of death, we linked the RBDM database with the 45 and Up Study and other electronic datasets.

Follow-up period

We commenced the follow-up period from the last day of the fifth year of the exposure window and ended it at the conclusion of either seven years from the commencement of follow-up or the death year, whichever occurred first. We also conducted several sensitivity analyses where we used a three-year exposure window with a fixed seven-year follow-up and a fixed five-year follow-up for the seven-year exposure window.

Confounder Variables

We selected several demographic and health-related variables from the 45 and Up study data as potential confounding factors, such as age, sex, geographical location, education, gross annual household income, BMI, self-reported health, smoking status, daily alcohol intake, weekly physical activity, daily fruits or raw vegetable consumption. We also adjusted for time-varying variables such as the number of general practitioner (GP) visits in the one last year, the number of specialist consultations and pathology test referrals in the last one year, the Charlson's co-morbidity index for the previous year, and the previous year's Rx-Risk index. The detailed categorization of the variables can be found in Appendix K.

The Charlson Comorbidity Index (CCI) is a tool used to predict one-year mortality rates for patients with chronic illnesses [80]. It assigns weights to 19 different medical conditions based on their severity and association with mortality, with higher weights given to more severe conditions. Patients' CCI scores are calculated by adding up the weights of all their comorbidities. We utilised an updated version of the CCI, which was calculated using hospital discharge data and the ICD-10 coding algorithm [81]. The Rx-Risk Comorbidity Index is another tool that identifies comorbidities using pharmacy claims data and assigns weights to 45 medical conditions based on their association with increased healthcare utilization [82]. These weights range from 1 to 3, with higher weights given to more severe or complex conditions. Patients' Rx-Risk scores are determined by adding up the weights of all their comorbidities. In our analysis, we used the Australian version of the Rx-Risk index, which is calculated using medicines mapped to the Anatomical Therapeutic Chemical (ATC) classification system [83]. The detailed method of calculation of these two co-morbidity indexes can be found elsewhere [80-84]. Both these comorbidity scores are frequently used in observational studies to minimize confounding. The main benefit of these scores is that they simplify the combination of individual covariates for each comorbidity into a single score.

Ethics approval

The 45 and Up Study received approval from the Human Research Ethics Committee (HREC) at the University of New South Wales. Additionally, the study was approved by the NSW Population Health Research Ethics Committee (HREC) under the reference number 2016/02/633 and the Lifeblood HREC with the reference number 2015#13.

Statistical Methods

We estimated the 7-year mortality risk, risk difference and risk ratio using an inverse probability weighting (IPW) of a marginal structural model. We fitted a pooled logistic regression model by adding a constant plus linear and quadratic terms of time and also linear and quadratic product terms of donation status and time. The baseline covariates were adjusted by inverse probability weighting. The weights were truncated at the 99th percentile to remove any extreme weights. Finally, we used non-parametric bootstrapping with 500 samples to calculate all the 95% Confidence Intervals. Unweighted and inverse probability weighted survival curves were also plotted.

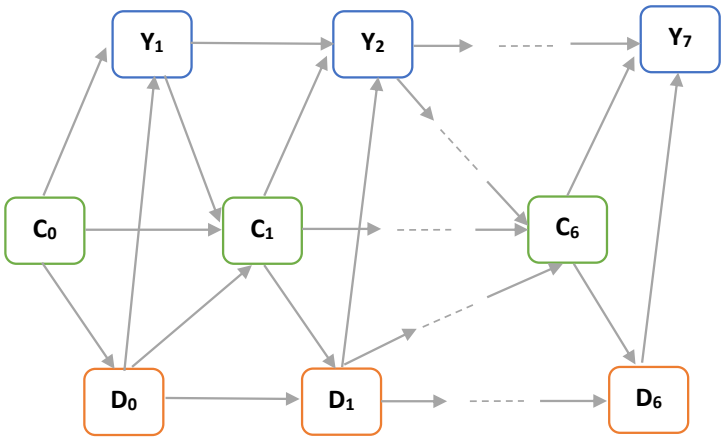


Figure 4.3 Directed acyclic graph of the relationship between blood donation (D_t), confounders (C_t) and mortality (Y_t) for the 7-year follow-up period.

Blood donation behaviour is assumed to be time-varying in nature. It means that current blood donation behaviour impacts subsequent donation. In this process, the factors that determine a donor's donation behaviour also lie in the causal pathway between donation status and mortality. Figure 4.3 illustrates the assumed directed acyclic graph of the relationship between blood donation (D_t), confounders (C_t) and mortality (Y_t) for the 7-year follow-up period. If traditional regression methods adjust C_t, it will produce a biased estimate as adjusting the post-baseline confounders; for example, C_1 will open a path from D_0 to Y_1 even if D_0 has no direct effect on Y_1. There is also treatment-confounder feedback exists in this relationship, which makes traditional regression methods unsuitable analysis techniques [62]. G methods such as inverse probability weighting, parametric g-formula, and doubly robust methods such as Targeted minimum loss-based estimator are suitable alternatives to adjust for time-varying confounding factors in such scenarios [62].

In this study, besides adjusting baseline confounders, we also adjusted for time-varying confounders and donation status by using the targeted minimum loss-based estimator (TMLE) [72]. TMLE depends on two kinds of mathematical models: Treatment and outcome models, which are the function of the confounders. IPW is a singly robust method, as its correctness depends on specifying the treatment model correctly. However, TMLE is a doubly robust estimator as it remains correct if any of the treatment or outcome models are misspecified. In addition to this, the inverse probability weighted marginal structural models can produce biased estimates if the positivity assumption is violated. In contrast, doubly robust estimators often produce less biased results than the IPW method, even if the positivity assumption is extremely violated [73, 74]. Furthermore, doubly robust estimators like TMLE can employ machine learning algorithms to fit the treatment and outcome models, which may capture complex associations that are not possible with simple regression-based approaches [72, 75].

In addition to the IPW model, we also utilised TMLE to estimate both single time point and time-varying 7-year mortality risks, risk differences and risk ratios. The treatment and outcome regressions were estimated through an ensemble of machine learning models, which leveraged super learner algorithms with 5-fold cross-validation. The candidate libraries used include generalized linear models and multivariate adaptive regression splines [87]. Additionally, a 5-fold cross-fitting component was incorporated into the process. We used the R package "SuperLearner" and "lmtp" to implement this analysis [76].

A few variables had missing values, with a maximum percentage of approximately 16%. Despite assuming that the data were missing at random, we opted for multiple imputations to estimate the missing values due to the limited number of cases. Removing participants with missing values would have further reduced the available cases. The imputation process employed a fully conditional specification approach, utilising classification and regression trees, and was implemented using the R package 'mice' [88].

We considered several sensitivity analyses by considering – i) a complete case analysis by removing all the missing values with a follow-up period was seven years, ii) 3 years exposure window which ended at the recruitment date of the 45 and Up Study data for each participant with follow-up was seven years, iii) 7-years exposure window which spans three years before the recruitment into the 45 and Up study and four years after with a follow-up for five years.

We used R version 4.2.2 to conduct all the statistical analyses.

Results

Table 4.3 shows the distribution of characteristics of 7338 donors who were selected by exposure window methods. Of these, 4750 (64.7%) were low-frequency donors, whereas 2588 (35.3%) were high-frequency donors. Among 7338 donors, 69 (1.5%) experienced death from

low-frequency donor groups while 45 (1.7%) experienced death from high-frequency donor groups during the 7-year follow-up period for each whole blood donor.

Table 4.3 Characteristics of the study participants among high-frequency and low-frequency donors at the start of the follow-up

Characteristics	Low-frequency	High-frequency
Total participant number (%)	4750 (64.7)	2588 (35.3)
Sex, n (%)		
Male	2189 (46.0)	1473 (56.7)
Female	2575 (54.1)	1123 (43.3)
Age in years, mean (standard deviation)	58.8 (6.6)	60.3 (6.9)
Body mass index kg/m^2, n (%)		
Underweight	16 (0.3)	6 (0.2)
Normal	1566 (32.9)	747 (28.8)
Overweight	1924 (40.4)	1122 (43.2)
Obese	971 (20.4)	577 (22.2)
Unknown	287 (6.0)	144 (5.6)
Smoking Status, n (%)		
Never	2966 (62.3)	1669 (64.3)
Former	1589 (33.4)	830 (32.0)
Regular	193 (4.1)	87 (3.4)
Unknown	16 (0.3)	10 (0.4)
Self-rated health, n (%)		
Excellent	1331 (27.9)	733 (28.2)
Very good	2189 (46.0)	1229 (47.3)
Good	1004 (21.1)	546 (21.0)
Fair/Poor	152 (3.2)	54 (2.1)
Unknown	88 (1.9)	34 (1.3)
Alcohol consumption/day, n (%)		
None	1034 (21.7)	567 (21.8)
<=1/day	1855 (38.9)	1005 (38.7)
>1/day	1838 (38.6)	1008 (38.8)
Unknown	37 (0.8)	16 (0.6)
Education level, n (%)		
No formal education	265 (5.6)	167 (6.4)
School to Diploma	3010 (63.2)	1804 (69.5)
University	1454 (30.5)	611 (23.5)
Unknown	35 (0.7)	14 (0.5)
Annual household income, n (%)		
<20k	367 (7.7)	245 (9.4)
20k-39k	637 (13.4)	463 (17.8)
40k-69k	1191 (25.0)	713 (27.5)
70k+	1813 (38.1)	778 (30.0)
Unknown	756 (15.9)	397 (15.3)
Physical activity/week, n (%)		

Characteristics	Low-frequency	High-frequency
<1/week	1714 (36.0)	874 (33.7)
>=1/week	2499 (52.5)	1404 (54.1)
Unknown	551 (11.6)	318 (12.3)
Daily fruits/vegetables consumed, n (%)		
0-2	275 (5.8)	149 (5.7)
3-4	1131 (23.7)	631 (24.3)
5+	2784 (58.4)	1507 (58.1)
Unknown	574 (12.1)	309 (11.9)
Location, n (%)		
Major city	2307 (48.4)	982 (37.8)
Regional/Remote	2347 (49.3)	1560 (60.1)
Unknown	110 (2.3)	54 (2.1)
No. of GP visits in the past 1 year, mean (sd)	4.72 (4.16)	4.17 (3.42)
No. of referrals in the past 1 year, mean (sd)	2.9 (3.08)	2.51 (2.38)
Charlson co-morbidity index, n (%)		
0	4707 (98.8)	2569 (99.0)
>= 1	57 (1.2)	27 (1.0)
Rx-Risk index, n (%)		
None	2461 (51.7)	1327 (51.1)
-6 to -1	1245 (26.1)	722 (27.8)
0 to 2	913 (19.2)	483 (18.6)
3+	145 (3.0)	64 (2.5)
Outcome, n (%)	69 (1.5%)	45 (1.7%)

Figure 4.4 shows the unweighted and inverse probability weighted survival curves for 7-year mortality of less than two and at least two donor groups. Unweighted survival curves showed a small amount of elevated risk of mortality in at least two donor groups compared to less than two donor groups after 5th year of follow-up. However, the survival curves almost overlapped in the IP-weighted model for the 7-year follow-up period.

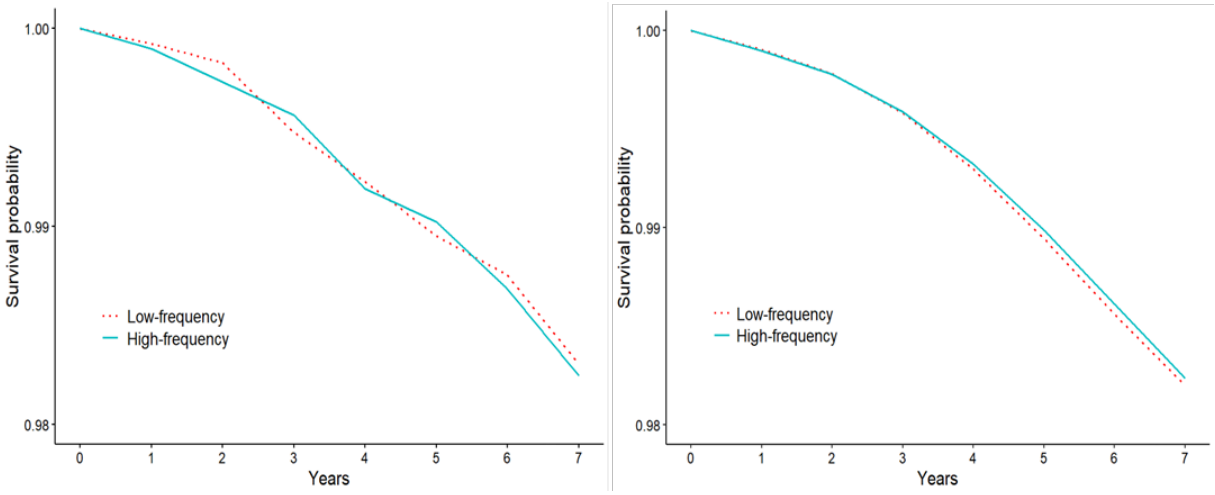


Figure 4.4 Unweighted and Inverse probability weighted survival curves for seven years follow-up period.

Table 4.4 presents the estimated 7-year mortality risks, risk differences and risk ratios calculated by inverse probability weighting, targeted minimum loss-based estimator (TMLE) and time-varying TMLE. The inverse probability weighted risk of mortality was 1.8% (95% CI 1.5%, 2.1%) for the low-frequency donor group and 1.8% (95% CI 1.4%, 2.1%) for the high-frequency donor group, which resulted in the risk difference of 0% (95% CI -0.5%, 0.5%) and risk ratio of 0.98 (95% CI 0.68, 1.28). TMLE estimated the risk of mortality as 1.8% (95% CI 1.5%, 2.0%) for the low-frequency donor group and 1.7% (95% CI 1.5%, 1.9%) for the high-frequency donor group resulted in the risk difference of -0.1% (95% CI -0.4%, 0.3%) and risk ratio of 0.97 (95% CI 0.80, 1.16). The time-varying TMLE produced almost similar results to IPW and single time point TMLE; estimated mortality risk was 1.9% (95% CI 1.5%, 2.3%) in low-frequency donor group and 1.9% (95% CI 1.5%, 2.2%) in high-frequency donor group which produced the risk difference of 0% (95% CI -0.6%, 0.5%) and risk ratio of 0.98 (95% CI 0.74, 1.29).

Table 4.4 Estimated 7-year mortality risk, risk difference and risk ratios for high and low-frequency donors.

Models	Risk, % (95% CI)		Risk Difference (RD), % (95% CI)	Risk Ratio (RR) (95% CI)
	Low-Frequency	High-Frequency		
Inverse probability weighted[a]	1.8 (1.5, 2.1)	1.8 (1.4, 2.1)	0.0 (-0.5, 0.5)	0.98 (0.68, 1.28)
Targeted minimum loss-based estimator (TMLE)[a]	1.8 (1.5, 2.0)	1.7 (1.5, 1.9)	-0.1 (-0.4, 0.3)	0.97 (0.80, 1.16)
TMLE (time-varying)[ab]	1.9 (1.5, 2.3)	1.9 (1.5, 2.2)	0.0 (-0.6, 0.5)	0.98 (0.74, 1.29)

[a]adjusted for sex, age, BMI, smoking status, self-rated health, alcohol consumption, education, annual income, physical activity, daily consumption of fruits and vegetables, location, no. of GP visits in the past 1 year, no. of referrals in the past 1 year, Charlson co-morbidity index, Rx-Risk index.

[b]time-varying TMLE included yearly exposure status, yearly Charlson co-morbidity index, yearly Rx-Risk index, yearly GP visits and yearly referral information.

Sensitivity Analyses

In the complete case analysis, we found a bit lower risk of mortality for the high-frequency donors compared to the low-frequency donor [RD: -0.3% (95% CI -0.9%, 0.2%), which resulted in the risk ratio of 0.83 (95% CI 0.47, 1.19). Single time point TMLE model also produced almost identical results to the IPW model [RR: 0.81 (95% CI 0.65, 1.01)], while time-varying TMLE produced null risk difference and null risk ratio [RR: 1.00 (95% CI 0.60, 1.66)]. For the three-year exposure window analysis, both IPW and time-fixed TMLE produced similar results to our main analysis. However, time-varying TMLE found a significantly lower mortality risk among high-frequency blood donors compared to low-frequency blood donors [RR: 0.63 (95% CI 0.46, 0.86)] although the risk difference was still less than 1% [RD: -0.8% 95% CI (-1.3, -0.2)]. The 7-year exposure window produced a similar pattern to the main analysis for all the models, and the results were not statistically significant.

Discussion

This study investigated the association between regular high-frequency whole blood donation and mortality risk among Australian blood donors. We used the exposure window technique and applied IPW and TMLE models, where later one involved machine learning algorithms while accounting for the healthy donor effect. We adjusted baseline confounders in the time-fixed models and both baseline and time-varying confounders in the time-varying model to account for the lagged effect of blood donation on mortality and did not find any significant impact of high-frequency blood donation on mortality risk when compared to low-frequency blood donation.

We used 5-years exposure window technique to select the donors for our analysis and categorized them into exposure group (high-frequency donor) and control group (low-frequency donor) based on frequency and regularity of donation, which is similar to the qualification period used by Peffer et al. where they measured the relationship between high-frequency blood donation and CVD outcomes [42]. However, there are some dissimilarities between the analysis conducted by Peffer et al. and ours. They employed a three-category exposure variable based on the tertiles of WB donations made in the 10-year qualification period. In contrast, we used a distinct categorization of the exposure variable based on the frequency and consistency of the donation pattern.

As described in Figure 4.3, the HDE poses a challenge because the factors responsible for the HDE can simultaneously confound and be a part of the causal pathway between exposure and outcome. There was also treatment-confounder feedback exists between exposure and outcome. G-methods are suitable in these situations [62]. However, Peffer et al. did not use these methods as one of the essential assumptions of these methods (positivity assumption) could be violated because the strict nature of donor exclusion criteria may produce a probability of zero donation. In our analysis, we used TMLE estimator as both time-fixed and time-varying

models, which are doubly robust and can still have unbiased estimates if the positivity assumption is extremely violated.

As we used doubly robust TMLE to adjust for time-varying confounders, we could have used shorter exposure window and then adjusted for all the time-varying confounders. Moreover, shorter exposure periods also decrease the size of the healthy donor effect [60]. Nonetheless, the impact of blood donation is assumed to be lagged. Donating blood for shorter periods can hardly be considered to have causal effects on mortality. Thus, from a biological perspective, a longer exposure window is required to determine the causal effect, if there is one. For this reason, using a 5-year exposure window is justified in our primary analysis.

In our sensitivity analysis, the time-varying effect of a 3-year exposure period showed stronger protective effects of regular high-frequency WB donation than the main analysis, and also the results were statistically significant. This significant effect of the 3-year exposure window could be non-causal, as described earlier from the biological perspective of lagged effect of donation, and risk differences are also less than 1%. Moreover, results from other sensitivity analyses for time-varying models did not find any significant association, which ruled out any association between high-frequency blood donation and mortality risk.

Few other studies have studied the association between whole blood donation and all-cause mortality. Casale et al. found longer life expectancy in blood donors than non-donors [55]. However, they did not adjust for the potential set of confounders and compared between donor and general population, which may have biased the result with the HDE. Edgren et al. also found 30% lower mortality among donors, but they did not use any causal method and did not conclude that the results were due to the impact of blood donation [51]. Ullum et al. tried to calculate the internal HDE by considering mortality among donors who retired due to age criteria [41]. The study used a Poisson regression model to analyse the impact of the donation

rate and ongoing donation on non-retired donors. They found an HDE-adjusted effect on donation rate by including an interaction between these variables. Although they found a 7.5% decreased mortality risk at each additional annual donation, they could not confirm that this effect was unbiased as the adjustment factor was estimated among elderly donors. These previous studies found reduced mortality risk between donor versus non-donor or high-frequency donor versus casual donors, and most of them could not confirm it as conclusive evidence of a beneficial effect of blood donation.

Our study has several strengths. Firstly, the exposure window reduced the HDE by comparing the mortality among active donor populations who had long donation careers and were likely to differ less in their health status. Secondly, our data linkage enabled us to adjust a range of potential confounders that most of the previous studies lacked. Furthermore, we adjusted co-morbidity indexes such as Charlson and Rx-Risk in both time-fixed and time-varying analyses. These indexes are powerful independent predictors of mortality. Finally, we used a doubly robust TMLE estimator for both time-fixed and time-varying models, which can still give unbiased estimates if any of the treatment or outcome models are misspecified.

There are also some limitations in this study. Our study population comprises mainly older donors, who probably began donating blood well before the exposure window. As our records of donations are only available on or after June 2002, we could not consider the time since the first donation or the overall impact of the entire donation history in our analyses. Moreover, compared to a few previous studies, our sample size is slightly smaller, and the follow-up time is shorter (maximum of 7 years), resulting in fewer events. A longer follow-up time could have estimated the effect more precisely.

In conclusion, this study found no significant impact of high-frequency blood donation on mortality risk among Australian donors after accounting for the healthy donor effect using

robust estimation models and adjusting for numerous potential confounders. Despite the methodological strengths, further investigation with a longer follow-up period may provide a more precise estimate of the impact of regular blood donation on mortality risk.

Acknowledgments

The research utilised data gathered from the 45 and Up Study, which is administered by the Sax Institute in collaboration with Cancer Council NSW, the Heart Foundation, and the NSW Ministry of Health. We extend our appreciation to the numerous individuals who participated in the 45 and Up Study, as well as those who generously donated blood to save lives. We are grateful to Services Australia for granting access to the Medicare claims and PBS data. The linked data was securely managed and accessed through the Sax Institute's Secure Unified Research Environment (SURE). We would like to acknowledge CHeReL for providing the linked data. The Australian government funds Australian Red Cross Lifeblood to provide blood, blood products, and services to the Australian population.

Section 3: Regular whole blood donation and gastrointestinal and haematological cancer risk among older Australian blood donors

Md Morshadur Rahman[1,2], Surendra Karki[2], Anne E. Cust[3,4], John K. Olynyk[5], David O Irving[1,2], Andrew Hayen[1]

[1]School of Public Health, University of Technology Sydney, Sydney, Australia

[2]Research and Development, Australian Red Cross Lifeblood, Sydney, Australia

[3]Sydney School of Public Health, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia

[4]The Daffodil Centre, The University of Sydney, a Joint Venture with Cancer Council NSW, Sydney, Australia

[5]Curtin Medical School, Curtin University, Bentley, and Fiona Stanley Hospital, Murdoch, Western Australia, Australia

Journal: To be determined

Type of Publication: Research paper

Stage of Publication: Ready for submission

Abstract

Background

Some studies suggest that blood donors are less likely to suffer from gastrointestinal cancers, whilst others indicate an increased risk of haematological cancers among blood donors due to enhanced cell proliferation. We aimed to clarify whether the reported discrepancies in cancer association with blood donation may be due to the potential impact of the 'healthy donor effect' (HDE).

Method

To examine the relationship between regular high-frequency blood donation and gastrointestinal and haematological cancer risk, we implemented a 5-year exposure window to determine the exposure status (high or low-frequency donors) and also adjust for the HDE. We used the Sax Institute's 45 and Up Study data, combining it with databases containing information about blood donation history and other health-related factors, for the purpose of the analyses. We calculated 5-year cancer risks, risk differences and risk ratios by utilising inverse probability weighting of a marginal structural model, along with other advanced doubly robust g-methods which can utilise ensemble machine learning algorithms.

Results

In our study, we identified a total of 3888 (57.6%) donors as low-frequency and 2867 (42.4%) as high-frequency donors within a 5-year exposure window. The inverse probability weighted 5-year risk difference between high and low-frequency donors for gastrointestinal cancer was 0.2% (95% CI -0.1%, 0.5%) with a risk ratio of 1.25 (95% CI 0.83, 1.68). Regarding haematological cancer, the risk difference was 0% (95% CI -0.3%, 0.5%) with a risk ratio of 0.97 (95% CI 0.55, 1.40). Our doubly robust estimators targeted minimum loss-based estimator (TMLE) and sequentially doubly robust (SDR) estimator yielded similar results, but none of our findings were statistically significant.

Conclusion

After accounting for the healthy donor effect (HDE), our study did not identify any statistically significant differences in the risk of gastrointestinal or haematological cancer between high and low-frequency blood donors.

Introduction

Iron, an essential element for human life, serves as the main component of heme proteins and ferritin, an iron storage protein [89]. However, iron may also have a potential detrimental role as a contributor to carcinogenesis through iron-induced oxidative stress, involving the formation of reactive oxygen species (ROS) and free radicals [89, 90]. Evidence from various studies demonstrates that higher levels of iron, even within the upper end of the accepted reference range, might play a role in different types of cancer, including the liver, colon, stomach, and oesophagus [29, 91-94]. Since a blood donor loses approximately 200mg of iron after each whole blood donation, studying the relationship between regular blood donation and iron-related cancers has gained interest in recent years.

Epidemiological studies that have been conducted to evaluate the risk of cancers in connection to iron intake or indicators of body iron storage have shown inconclusive findings [30, 93, 95-99]. One of the major limitations of these studies was to use non-specific measures of body iron stores which could be mitigated by the use of blood donation history as a marker of iron levels [89]. A few studies have evaluated cancer risk among blood donors, but the findings were still inconsistent. A US cohort study conducted on male blood donors hypothesised that frequent blood donation, which reduces body iron stores, might decrease the risk of colorectal cancer, but the findings were insignificant [30]. Two studies found decreased overall cancer risk among blood donors compared to the general population, which may mainly reflect the healthy donor effect [51, 56]. In another study, Edgren et al. reported decreasing the risk for cancers of the liver, lung, colon, stomach, and oesophagus with increased estimated iron loss, but only among male blood donors [29]. Finally, Zacharski et al. demonstrated that phlebotomy therapy of older men to reduce serum ferritin levels from a median of 120 to 80μg/L was associated with a 35% reduction in the risk of visceral malignancies over a 5-year period [100].

Moreover, observing a higher incidence of cancers among blood donors compared to the general population was also not uncommon. A US study indicated higher cancer risk among blood donors, potentially due to demographic differences, education level, and access to primary care [39]. Although Merk et al. found lower overall cancer risk among blood donors, haematological malignancies were higher among donors compared to non-donors [56]. Frequent plasma donors were also at higher risk of non-Hodgkin's lymphoma [29]. Similar results for haematological malignancies were also found in other studies [39, 51]. There is also a hypothesis that blood donors may be more likely to develop haematological malignancies due to enhanced cell proliferation, which is itself thought to be a risk factor for cancer and is sparked by the frequent removal of blood cells [29]. Nonetheless, a recent study in Sweden did not find any risk of haematological malignancy among blood donors [46].

Considering the contradictory findings in the literature, the ongoing uncertainty surrounding the role of iron in cancer risk and also the biological plausibility of haematological malignancy among blood donors, the aim of this study was to investigate the possible association between regular blood donation and the risk of gastrointestinal, and haematological cancers among blood donors in Australia. To mitigate the 'healthy donor effect', a methodological pitfall in donor health studies, we utilised a 5-year exposure period method, similar to Peffer's 'qualification period' [42]. To conduct the analysis, we utilised the Sax Institute's 45 and Up Study data [77], which was linked with records from Lifeblood's blood donation data sets and various other external administrative health data sets.

Methods

Data sources and linkage

To explore the relationship between whole blood donation and gastrointestinal, colorectal and haematological malignancies, we utilised the Sax Institute's 45 and Up Study data, which is

linked to other electronic health databases such as the Australian Red Cross Lifeblood Donor Registry, Registry of Birth, Deaths, and Marriages- Deaths Registrations (RBDM), NSW Central Cancer Registry (NSWCCR), and Medicare claims.

The 45 and Up Study, conducted by the Sax Institute, involved the enrolment of 267,357 individuals aged 45 years or above in New South Wales, Australia, between 2005 and 2009. The study recruited prospective participants through random selection from the Services Australia Medicare enrolment database, resulting in a participation rate of 19.2% [79]. In order to ensure representation, the study oversampled residents residing in rural and remote areas, as well as individuals aged 80 years and above. Participants completed an initial questionnaire that covered a wide range of topics, including socio-demographic information, health status, lifestyle choices, and behaviours. Additionally, they provided consent for their data to be linked with various administrative databases, allowing for long-term follow-up analysis.

Australian Red Cross Lifeblood is responsible for collecting, processing, and distributing blood and blood products in Australia. It also keeps track of donor data in a central system called the National Blood Management System (NBMS). Prior to 2007, the methods used by Lifeblood to store donor data varied. However, after a national merger in 2007 of what was to that time separate, state-based sets of donor data, all donor information was consolidated within the NBMS. In terms of data availability, complete records for blood donations in New South Wales (NSW) were only accessible starting from June 1st, 2002. Therefore, for the purpose of data linkage, the dataset used included blood donation information spanning from June 1st, 2002, to December 31st, 2018.

The Medicare claims data consolidates information regarding healthcare services and procedures that are subsidized by the Australian government through Medicare. This dataset provides comprehensive information on the various types and frequency of services delivered

by medical practitioners, including consultations, diagnostic tests, surgical procedures, and allied health services such as physiotherapy and psychology. We utilised this data to calculate the overall annual count of visits to general practitioners (GPs) and referrals for diagnostic tests. It is important to note that the Medicare Benefits Schedule (MBS) data is complete up until December 2017.

The NSW Cancer Registry (NSWCR) is responsible for keeping track of individuals diagnosed with cancer in NSW. Since 1972, the NSWCR has maintained comprehensive records that include demographic information, incidence data, and death details for individuals who have been diagnosed with cancer. In our study, we used this dataset to ascertain the date of cancer diagnosis. The data is complete up to December 2015.

The NSW Registry of Births, Deaths, and Marriages (RBDM) database holds records of residents' birth, death, and marriage dates. In our study, we used the RBDM death data to establish the date of death and all-cause mortality. It's important to note that the RBDM dataset was completely updated up to December 2018.

The NSW Centre for Health Record Linkage (CHeReL) employed a probabilistic matching method to connect the data from the 45 and Up Study with NBMS, APDC, and RBDM. Previous quality assurance assessments of CHeReL's linkages using the master linkage key indicated an estimated false positive rate of 0.5%. The Sax Institute provided Services Australia with specific identifier information for the participants in the 45 and Up Study. In turn, Services Australia supplied the corresponding Medicare claim and PBS data based on these unique identifiers. The Sax Institute then linked the Medicare claim and PBS data to the 45 and Up Study data using a deterministic matching procedure.

Study population, exposure window

We employed a 5-year window to determine the participants and exposure status inspired by the method used by Peffer and colleagues [60] depicted in Figure 4.5. The exposure window refers to the time in which the donor is needed to actively give blood while satisfying other requirements for eligibility. In our research, this exposure window includes three years before the enrolment into the 45 and Up trial and two years thereafter. To qualify for participation in our research, donors must have made at least one whole blood (WB) donation between the first and fifth years of the exposure window and stayed alive and cancer free for the full 5-year period. We excluded donors who performed any plasma or platelet donation during the 5-year period to keep only WB donors for the analysis. Therefore, our research cohort consisted of persons who had an active WB donation career of five years within the exposure window period and who did not develop any kind of cancer during that time or in the time leading up to the exposure window.

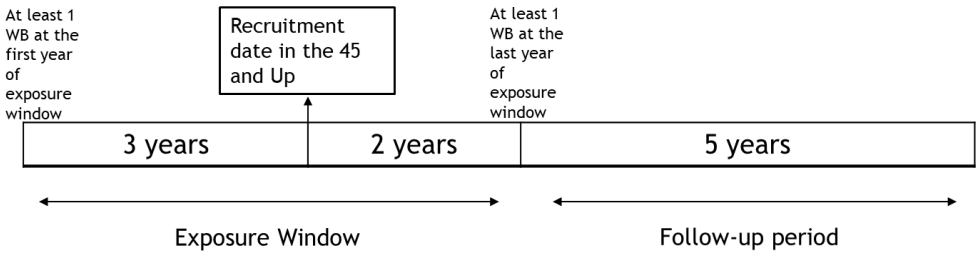


Figure 4.5 5-year exposure window and follow-up period for cancer analysis

Exposure variable

We have considered several exposure scenarios to measure the frequency and regularity of blood donations made by participants during each year of exposure window. We considered: i) at least one WB donation during each year of exposure window vs. less than one donation

during each year of exposure but still meeting the eligibility criteria ii) at least two WB donations during each year of exposure window vs. less than two donations during each year of exposure window iii) at least three WB donation during each year of exposure window vs. less than three donations during each year of exposure window.

Ascertainment of WB donation

Utilising linked Lifeblood donation history data, whole blood donation was determined. If a person successfully donated a unit of WB, the individual was regarded as a WB donor.

Ascertainment of Cancer

The primary outcomes of this study were gastrointestinal, and haematological cancers. All the cancer information was ascertained from the linked NSW cancer registry data. By using the revision 9 codes of the International Classification of Diseases, an individual was confirmed to have experienced either gastrointestinal or colorectal cancer if the cancer diagnosis codes were C15 (oesophageal) or C16 (stomach) or C17 (small intestinal) or C22 (liver) or C23-C24 (gallbladder) or C25 (pancreatic) or C18 (colon) or C19-C21 (rectal). Moreover, an individual was confirmed to have experienced haematological malignancy if the diagnosis codes were C920 (acute myeloid leukaemia) or C910 (acute lymphoblastic leukaemia) or C81 (Hodgkin lymphoma) or C8890 (multiple myeloma) or C82 (non-Hodgkin lymphoma) or C919 (other lymphoid leukaemia) or C929 (other myeloid leukaemia) or C94 (other specified leukaemia). We only considered the first diagnosed cancer for this analysis if a person had multiple malignancies over the follow-up period.

Follow-up period

The follow-up period commenced from the last day of the exposure window and ended at the conclusion of either five years from the start of the follow-up or the death date or the cancer diagnosis date, whichever occurred first. For the purpose of the sensitivity analysis, we also

considered an administrative end date for the follow-up so that the study started from the last day of the exposure period and ended on December 30, 2015, or the death date or cancer diagnosis death, whichever occurred first.

Potential confounding factors

A number of demographic/socioeconomic, health status and blood donation-related variables were considered as potential confounding factors. The demographic/socioeconomic variables were age, sex, geographical location (major city, regional/remote), education (no formal education, school to diploma, university) and gross annual household income (<20, 20-39, 40-69, 70+ thousand). The health status-related variables were BMI (0-18.4, 18.5-24.9, 25-29.9, 30+ kg/m2), self-reported health (excellent, very good, good, fair/poor), smoking status (never, former, regular), daily alcohol intake (none, ≤1/day, >1/day), weekly physical activity (<1/week, ≥1/week), and daily fruits or raw vegetable consumption (0-2, 3-4, 5+), intake of multivitamins and minerals (no, yes), number of general practice (GP) visits in the last one year, and number of specialist consultations and pathology test referrals in the last one year, family history of cancers (no, yes), any cancer screening (no, yes). Blood donation-related variables were average blood pressure levels during the exposure period, average haemoglobin level during the exposure period, and blood group (O, non-O). The detailed derivation of the variables is given in the appendix.

Statistical Methods

We calculated 5-year cancer risk, risk difference and risk ratio by inverse probability weighting (IPW) of a marginal structural model for gastrointestinal and haematological cancers separately. We fitted a pooled logistic regression model by adding a constant plus linear and quadratic terms of time and also linear and quadratic product terms of donation status and time. The baseline covariates were adjusted by calculating the inverse probability weights and then using the weights in the outcome regression model. The IPW was truncated at the 99th

percentile to remove any extreme weights from outliers. Finally, we used non-parametric bootstrapping with 500 samples to calculate all the 95% CIs. Inverse probability weighted Kaplan Meier survival curves were also plotted for the two cancer outcomes with three different exposure definitions.

We also utilised two alternative g-methods, namely the targeted minimum loss-based estimator (TMLE) and the sequentially doubly robust estimators (SDR), to compute 5-year cancer risk, risk difference, and risk ratios [71, 72]. These estimators, including inverse probability weighting (IPW), rely on two mathematical models: the treatment model and the outcome model, both of which are functions of the confounding variables. The IPW is a singly robust estimator, as its accuracy depends on correctly specifying the treatment model. On the other hand, TMLE and SDR are doubly robust estimators, meaning that their estimates remain unbiased even if one of the treatment or outcome models is misspecified.

Additionally, the inverse probability weighted marginal structural models can produce a biased estimate if affected by violations of the positivity assumption. In contrast, doubly robust estimators often produce less biased results than IPW estimators, even if the positivity assumption is extremely violated [73, 74]. Moreover, these doubly robust estimators have the advantage of being able to utilise machine learning algorithms to fit the treatment and outcome models, allowing them to capture complex associations that may not be possible with simple regression-based approaches [72, 75]. As blood donation behaviour is assumed to be time-varying in nature, we also estimated time-varying TMLE and SDR estimators in one of the sensitivity analyses. We used the R package “SuperLearner” and “lmtp” to implement this analysis [76].

A few variables had missing values (maximum of approximately 16%). Although we assumed that the data were missing at random, we still did multiple imputations to calculate missing

values as we had a lower number of cases and removing participants with missing values could further lower the number of cases. The imputation was a fully conditional specification that used classification and regression trees and was implemented by the R package 'mice' [88].

We used R version 4.2.2 to conduct all the statistical analyses.

Results

Table 4.5 depicts the characteristics distribution of 6755 whole blood donors, of whom 2667 (42.4%) donated at least two whole blood units in each year of the exposure period, whereas 3888 (57.6%) did not donate at least two whole blood units during each year of the exposure period. High-frequency donors were mostly male (55.3%) and also slightly older (average age 60.3 years) than low-frequency donors. Among 6755 donors, 25 (0.6%) experienced gastrointestinal cancer from low-frequency blood donor groups during five years of follow-up, while 27 (0.9%) gastrointestinal cancer in the high-frequency donor group. For haematological cancer, we found 23 (0.6%) cases from the low-frequency donor group and 20 (0.7%) from the high-frequency donor group during the 5-year follow-up period.

Table 4.5 Characteristics of the study participants who were donating or not donating at least 2 WB donations in each year of the exposure period.

Characteristics	At least 2 whole blood donations in each year of the exposure period	
	No	Yes
No. (%)	3888 (57.6)	2867 (42.4)
Sex, n (%)		
Male	1717 (44.2)	1585 (55.3)
Female	2171 (55.8)	1282 (44.7)
Age at baseline, mean (sd)	57.72 (6.68)	60.3 (6.9)
Haemoglobin, g/dl, mean (sd)	140.99 (10.36)	143.38 (9.86)
Systolic blood pressure, mean (sd)	127.39 (12.05)	128.66 (11.17)
Diastolic blood pressure, mean (sd)	76.95 (6.84)	77.26 (6.25)
Total no. of WB donation in exp period, mean(sd)	9.78 (3.56)	16.85 (2.65)
Blood group, n (%)		
Non-O	1976 (50.8)	1401 (48.9)
O	1912 (49.2)	1466 (51.1)

Characteristics	At least 2 whole blood donations in each year of the exposure period	
	No	Yes
Body mass index kg/m^2, n (%)		
Underweight	10 (0.3)	8 (0.3)
Normal	1306 (33.6)	897 (31.3)
Overweight	1527 (39.3)	1231 (42.9)
Obese	793 (20.4)	577 (20.1)
Missing	252 (6.5)	154 (5.4)
Smoking Status, n (%)		
Never	2435 (62.6)	1884 (64.3)
Former	1282 (33.0)	921 (32.1)
Regular	157 (4.0)	90 (3.1)
Missing	14 (0.4)	12 (0.4)
Self-rated health, n (%)		
Excellent	1040 (26.8)	850 (29.7)
Very good	1791 (46.1)	1361 (47.5)
Good	854 (22.0)	564 (19.7)
Fair/Poor	130 (3.3)	57 (2.0)
Missing	73 (1.9)	35 (1.2)
Alcohol consumption/day, n (%)		
None	877 (22.6)	600 (20.9)
<=1/day	1521 (39.1)	1100 (38.4)
>1/day	1461 (37.6)	1148 (40.0)
Missing	29 (0.8)	19 (0.7)
Education level, n (%)		
No formal education	215 (5.5)	175 (6.1)
School to Diploma	2432 (62.6)	1927 (67.2)
University	1213 (31.2)	747 (26.1)
Missing	28 (0.7)	18 (0.6)
Annual household income, n (%)		
<20k	313 (8.1)	257 (9.0)
20k-39k	521 (13.4)	503 (17.5)
40k-69k	954 (25.5)	762 (26.6)
70k+	1484 (38.2)	901 (31.4)
Missing	616 (15.8)	444 (15.5)
Location, n (%)		
Major city	1909 (49.1)	1161 (40.5)
Regional/Remote	1888 (48.6)	1646 (57.4)
Missing	91 (2.3)	60 (2.1)
Daily fruits/vegetable consumed, n (%)		
0-2	229 (5.9)	160 (5.6)
3-4	928 (23.9)	688 (24.0)
5+	2259 (58.1)	1685 (58.8)
Missing	472 (12.1)	334 (11.7)
Taking any vitamin or mineral, n (%)		
No	2975 (76.5)	2236 (78.0)
Yes	912 (23.5)	631 (22.0)
Missing	1 (0.0)	0 (0.0)

Characteristics	At least 2 whole blood donations in each year of the exposure period	
	No	Yes
Consumption of red meat, n (%)		
<5/week	2954 (76.0)	2134 (74.4)
>=5/week	865 (22.3)	697 (24.3)
Missing	68 (1.8)	36 (1.3)
Consumption of processed meat, n (%)		
<3/week	2869 (73.8)	2097 (73.1)
>=3/week	577 (14.8)	459 (16.0)
Missing	442 (11.4)	311 (10.9)
Family history of cancer, n (%)		
No	2058 (52.9)	1517 (52.9)
Yes	1830 (47.1)	1350 (47.1)
Cancer screening, n (%)		
No	421 (10.8)	301 (10.5)
Yes	3428 (88.2)	2545 (88.8)
Missing	39 (1.0)	21 (0.7)
No. of GP visits in the past 1 year, mean (sd)	4.68 (4.15)	4.15 (3.41)
No. of referrals in the past 1 year, mean (sd)	2.84 (2.69)	2.51 (2.35)
Outcomes		
Gastrointestinal, n (%)	25 (0.6)	27 (0.9)
Haematological, n (%)	23 (0.6)	20 (0.7)

Table 4.6 presents the estimated 5-year cancer risk for gastrointestinal and haematological cancer, their risks, risk differences and risk ratios calculated by IPW, TMLE and SDR estimators. The IPW risk of gastrointestinal cancer was 0.7% (95% CI 0.5%, 0.9%) for low-frequency donors and 0.9% (95% CI 0.6%, 1.2%) for high-frequency donors resulted in the risk difference of 0.2% (95% CI -0.1%, 0.5%) and risk ratio of 1.25 (95% CI 0.83, 1.68). We found almost identical results from TMLE; the risk for low-frequency donors was 0.7% (95% CI 0.5, 0.9), and the risk for high-frequency donors was 0.9% (95% CI 0.7%, 1,1%), which resulted in risk difference of 0.2% (95% CI -0.1%, 0.5%) and risk ratio of 1.25 (95% CI 0.86, 1.81). The SDR estimator produced almost similar results (Table 3) to IPW and TMLE. Moreover, the IPW risk of haematological cancer was 0.6% (95% CI 0.5%, 0.8%) for low-frequency donors and 0.6% (95% CI 0.4%, 0.8%) for high-frequency donors, which produced a risk difference of 0.0% (95% CI -0.3%, 0.2%) and risk ratio of 0.97 (95% CI 0.55, 1.40). The

TMLE produced almost similar results; risk of 0.6% (95% CI 0.5%, 0.8%) for low-frequency donors, risk of 0.6% (95% CI 0.5%, 0.8%) for high-frequency donors, risk difference of 0.0% (95% CI -0.3%, 0.2%), and risk ratio of 0.96 (0.66, 1.40). The SDR estimator produced similar results to IPW and TMLE, except the risk ratio was slightly higher than both estimators [RR = 1.01 (95% CI 0.71, 1.43)]. None of the results for both gastrointestinal and haematological cancer were statistically significant, indicating no increased/ decreased risk of gastrointestinal/colorectal and haematological cancers among blood donors.

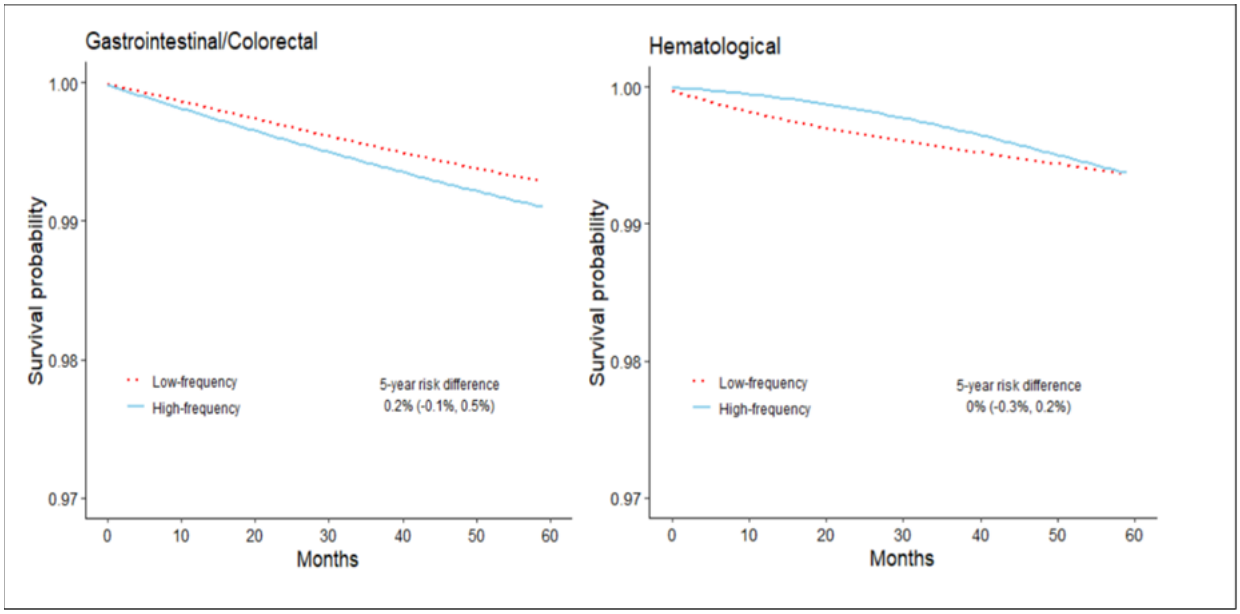
Figure 4.6 shows the inverse probability weighted Kaplan Meyer survival curves for gastrointestinal and haematological cancers for 5-year follow-up, and both curves show very insignificant risk differences between low and high-frequency donors.

Table 4.6 Estimated 5-year cancer risk, risk difference and risk ratios for high and low-frequency donors*.

Outcomes	Models	Risk, % (95% CI)		Risk Difference, % (95% CI)	Risk Ratio (95% CI)
		Low-frequency	High-frequency		
Gastrointestinal	IPTW	0.7 (0.5, 0.9)	0.9 (0.6, 1.2)	0.2 (-0.1, 0.5)	1.25 (0.83, 1.68)
	TMLE	0.7 (0.5, 0.9)	0.9 (0.7, 1.1)	0.2 (-0.1, 0.5)	1.25 (0.86, 1.81)
	SDR	0.8 (0.6, 0.9)	1.0 (0.7, 1.2)	0.2 (-0.1, 0.5)	1.27 (0.89, 1.80)
Haematological	IPTW	0.6 (0.5, 0.8)	0.6 (0.4, 0.8)	0.0 (-0.3, 0.2)	0.97 (0.55, 1.40)
	TMLE	0.6 (0.5, 0.8)	0.6 (0.5, 0.8)	0.0 (-0.3, 0.2)	0.96 (0.66, 1.40)
	SDR	0.7 (0.5, 0.9)	0.7 (0.6, 0.9)	0.0 (-0.2, 0.3)	1.01 (0.71, 1.43)

*Adjusted for sex, age, haemoglobin, systolic blood pressure, diastolic blood pressure, blood group, BMI, smoking status, self-rated health, alcohol consumption, education, annual income, physical activity, daily consumption of fruits and vegetables, vitamin/mineral intake, red meat consumption, processed meat consumption, family history of cancer, cancer screening, location, no. of GP visits in the past 1 year, no. of referrals in the past 1 year.

Figure 4.6 Weighted survival curves for a 5-year follow-up period for gastrointestinal and haematological cancers.



Sensitivity analysis

We found similar results to our main analysis when we ended the follow-up on 31 December 2015 instead of a fixed 5-year follow-up for each participant. The IPW risk ratio for this analysis was 1.27 (95% CI 0.74, 1.80) for gastrointestinal cancer and 0.92 (95% CI 0.53, 1.30) for haematological cancer. When we changed the definition of exposure to at least 1 WB donation during each year of exposure period as high-frequency donors and other donation frequency as low-frequency donors, we found a slightly decreased gastrointestinal cancer risk ratio of 1.11 (0.44, 1.78) with almost zero risk differences of 0.1% (95% CI -0.3%, 0.5%). For haematological cancer, we found 0.4% (95% CI -0.8%, 0.0%) decreased risk among high-frequency donors and a risk ratio of 0.57 (95% CI 0.01, 1.12) when we used this exposure definition. For the exposure definition of at least 3 WB donations in each year of the exposure period vs other frequencies of donation in each year of the exposure period, we found zero risk difference [0.0% (95% CI -0.4, 0.4)] and risk ratio close to null value [1.02 (95% CI 0.41,

1.64)] when considered the gastrointestinal cancer incidence. Haematological cancer showed a slightly elevated risk ratio of 1.44 (95% CI 0.86, 2.01), but still with very small risk differences [0.3% (-0.1, 0.7)] between high and low-frequency donors. We found zero risk differences [0.0% (95% CI -0.5%, 0.4%)] from time-varying TMLE estimator among high and low-frequency donors with a risk ratio of 0.97 (95% CI 0.56, 1.66)] for gastrointestinal cancer. The time-varying SDR estimator produced a risk ratio of 0.79 (95% CI 0.40, 1.56) with a slightly lower risk for high-frequency donors compared to low-frequency donors. For haematological cancer, time-varying TMLE and SDR estimators produced a risk ratio of 0.80 (95% CI 0.27, 2.36) and 1.07 (95% CI 0.42, 2.76), respectively, where the risk differences were almost zero for both of the estimators.

Discussion

In this study, we investigated the association between frequent whole blood donation and the risk of gastrointestinal and haematological malignancies among Australian blood donors. We used the exposure window method to determine the exposure status and fitted inverse probability-weighted and doubly robust statistical models while adjusting several potential demographics, socio-economic, health status-related, and blood donation-related variables. Our research did not discover any evidence of a statistically significant relationship between frequent whole blood donations and an increased or decreased risk of developing cancer. There was a 0.2% increased risk of gastrointestinal cancer among high-frequency blood donors. However, the results were not statistically significant, and time-varying models used in sensitivity analysis revealed that there was no difference in risk between high-frequency and low-frequency blood donors.

We used the 5-year exposure window technique to ascertain the exposure (high-frequency donor) and control (low-frequency donor), which is comparable to the qualification period used by Peffer et al. to examine the association between high-frequency blood donation and CVD

outcomes [42]. The healthy donor effect has a substantial impact on the studies that used the lifetime number of donations to determine exposure status. In contrast to previous studies, Peffer et al. separated the exposure status determination period from the follow-up period, which significantly reduced the "healthy donor effect" [42]. Our 5-year exposure window had a comparable effect on lowering the HDE. However, there are a few differences between our analysis and that of Peffer et al. They used a three-category exposure variable based on the tertiles of donations made during the 10-year qualification period. In contrast, we categorised the exposure variable that was based on the frequency and consistency of the donation pattern.

One recent study in Australia also used a 5-year qualification period to assess the relationship between high-frequency, regular blood donation and the risk of cardiovascular diseases [101] and found no statistically significant association between blood donation and CVD. This study also used the same data linkage we used in our analysis, and their categorisation of the exposure variable is also similar to ours. However, their exposure period started from the date of recruitment into the 45 and Up Study data, whereas we started our exposure three years before the recruitment and ended two years after the recruitment into the 45 and Up Study data, as we have cancer data available until December 2015, and starting three years before the recruitment into the 45 and Up Study data provided extra follow-up time and cases.

In our primary analysis, we used a 5-year exposure window to distinguish between high- and low-frequency donors. In addition, it has been suggested that a shortened exposure interval can further reduce the healthy donor effect, given that donors are chosen prior to each donation based on health parameters that could simultaneously predict future cancer events, leaving space for a residual HDE. If the duration of exposure had been shortened, donors would have been selected and categorised based on HDE-causing health criteria less frequently. However, it is believed that blood donation has a lagged effect on health [42]. Donating blood for a shortened duration is unlikely to have causal effects on malignancy. Consequently, from a

biological standpoint, an extended exposure interval is required to ascertain the possibility of a causal effect. This justifies the use of a 5-year exposure window in our primary analysis.

Blood donation is a good indicator of serum ferritin, so numerous studies have examined the incidence of cancer among blood donors. Several of these studies have reported a lower risk of cancer occurrence and mortality among blood donors [43, 51, 56]. However, these findings may be influenced by the healthy behaviours and overall well-being of the donor population. In one particular Swedish study, researchers utilised a nested case-control design to investigate the impact of iron depletion through blood donation on Swedish and Danish donors [29]. The study found a trend towards a reduced risk of liver, lung, colon, stomach, and oesophageal cancers in males with a latency period of 3 to 7 years, comparing the lowest to highest estimated iron loss from donation (OR = 0.70, 95% CI: 0.58, 0.84). Nevertheless, the authors acknowledged their inability to account for several important confounding factors, such as smoking, alcohol consumption, nutrition, physical activity, anthropometric measures, and occupational exposures, which might have influenced the observed results. Another cohort study conducted in the United States examined the risk of colorectal cancer in regular blood donors compared to non-donors among male participants [30]. Although their results could be a consequence of the healthy donor effect, our findings of gastrointestinal and colon cancer are consistent with their findings.

It is not uncommon in donor health studies to find increased cancer risk among high-frequency blood donors. There is also a hypothesis that blood donors may be more likely to develop haematological malignancies due to enhanced cell proliferation, which is itself thought to be a risk factor for cancer and is sparked by the frequent removal of blood cells [29]. Mark et al. reported increased haematological cancer among blood donors, although their findings are believed to be impacted by the HDE [56]. Edgreen et al. found an increased risk of non-Hodgkin lymphoma among frequent plasma donors, although they did not adjust some

important confounding factors [29]. Recently, one Swedish study used a standardised incidence ratio and nested case-control study and did not find an increased risk of haematological malignancy among frequent blood donors, which is consistent with our findings for haematological malignancies [46]. Moreover, they found a slightly higher risk of chronic lymphocytic leukemia (CLL) (SIR, 1.07; 95% CI 1.01, 1.15). However, as there was no evidence of a dose-response relationship between no. of donations and the risk of these cancers, they concluded this risk as non-causal.

Although none of our findings was statistically significant, our point estimates for gastrointestinal/colorectal cancer in the main analysis were higher than the null value [IPW RR: 1.25 (95% CI 0.83, 1.68)]. In our sensitivity analysis, we define the high-frequency exposure group with at least 1 and 3 donations each year of the exposure period and found no dose-response relationship. It ruled out any possibility of increased risk, which could not be detected by our sample. In addition to that, time-varying TMLE and SDR estimators also found almost zero risk differences among high and low-frequency donors. Moreover, because of blood donors' continuous screening during their donation career and comparatively higher health consciousness, it is not uncommon to have more cancer detection among frequent blood donors compared to casual donors [39, 46].

This is the first study to use the exposure window method to evaluate gastrointestinal and haematological cancer outcomes among blood donors. Our study has several strengths. First, the use of an exposure window decreased the HDE by comparing cancer outcomes among active donor populations with a lengthy donation career and presumably less variance in health status. Second, our data linkage allowed us to control for a variety of potential confounding variables, something that was lacking in the majority of previous studies. In addition, we utilised doubly robust models, such as TMLE and SDR, which incorporated machine learning algorithms to determine the risk estimates. Since the findings of our IPTW model and our

doubly robust models are nearly identical, our treatment and outcome models are not misspecified.

This study has some limitations as well. The majority of our research population consists of elderly donors who likely began donating blood well before the beginning of the exposure window. Due to the fact that our donation records are only available on or after June 2002, we were unable to analyse the duration since the first donation or the cumulative impact of the entire donation history. Moreover, compared to some previous studies, our sample size is somewhat small, and our follow-up period is somewhat shorter (a maximum of 5 years), resulting in a smaller number of events. This may have indicated that our study lacked the statistical power to detect clinically important small effect sizes. Because of the smaller number of events, we also did not do a sex-stratified analysis, which would have been more appropriate while assessing iron-induced outcomes.

In conclusion, we did not find any convincing evidence of an increased or decreased risk of gastrointestinal and haematological malignancy among older Australian blood donors. Further exploration is needed in an Australian cohort with a longer follow-up time to better understand the relationship between these cancer outcomes and regular whole blood donation.

Acknowledgements

This study utilised data collected from the 45 and Up Study, which is administered by the Sax Institute in collaboration with Cancer Council NSW, the Heart Foundation, and the NSW Ministry of Health. We express our gratitude to the numerous participants of the 45 and Up Study, as well as those who selflessly donated blood to save lives. We acknowledge Services Australia for granting us access to the Medicare claims and PBS data, which were securely managed and accessed through the Sax Institute's Secure Unified Research Environment (SURE). We also extend our thanks to CHeReL for providing the linked data

(www.cherel.org.au). It is important to note that the Australian government funds Australian Red Cross Lifeblood to ensure the provision of blood, blood products, and services to the Australian population.

Chapter 5 General Discussion

5.1 Summary of the main findings

The Healthy Donor Effect (HDE) is a methodological pitfall in donor health research. The aim of this thesis was to summarise the methods that have been used to adjust for the Healthy Donor Effect (HDE) and identify any additional or new methods to address it. The methods were applied to Sax Institute's 45 and Up Study data, which was linked with other electronic health databases, including the blood donor data from Australian Red Cross Lifeblood, to explore the association of regular blood donation with long-term health outcomes (mortality and cancer). To achieve this, I conducted a systematic review of methods previously used in donor health research that examined the association between blood donation and long-term health outcomes. I assessed if the studies used methods to address the HDE, and if yes, then assessed the suitability of those methods. I then applied some appropriate methods to minimise the HDE using the linked 45 and Up Study data. For this exploration, I chose two long-term health outcomes: all-cause mortality and cancer (specifically gastrointestinal/colorectal, haematological). To assess the outcome of mortality, I used the 'target trial emulation' technique and the 'exposure window' method to evaluate both the initiation of blood donation and the effect of donating in the long term. For the cancer outcome, I selected iron-related gastrointestinal and colorectal cancers, as well as haematological cancers as primary outcomes and applied the 5-year exposure window technique to mitigate the HDE. In this section, I will discuss the main findings obtained from the analyses, organised according to the thesis objectives.

Objective 1: To summarise the methods that have been used to adjust the HDE and identify any additional/new approaches that may adjust the HDE adequately.

To address the first objective, I conducted a systematic methods review that included studies assessing the impact of blood donation on donors' long-term health outcomes, such as cardiovascular/heart diseases, cancer, diabetes, bone density or fracture, all-cause mortality, morbidity, and infections.

The HDE presents a significant methodological challenge when examining the long-term health impact of blood donation. In this review, approximately 70% of the included studies recognised the HDE, and 68% of these identified it as a limitation in their study. However, the methods used to address the HDE varied, and only a few seemed to effectively mitigate its impact. Most studies were observational and likely contained residual HDE in their findings. Common mitigation methods included adjusting for confounders using regression or ANOVA. While these methods can decrease the HDE if all relevant variables are considered, it is often challenging to identify and measure all potential confounding factors. Additionally, some of these factors can also act as intermediary determinants of exposure to donation and health outcomes, making simple adjustments in regression models insufficient to address the HDE adequately.

Some studies have restricted their analysis to an active donor population, which can help reduce HDE, though not eliminate it entirely. For instance, the duration of a donation career within an active donor population can still introduce selection bias [59]. Therefore, even studies that limit their analysis to active donors cannot completely rule out the presence of residual HDE. A few studies have used 'qualification period' or 'exposure window' techniques within the active donor population to adjust HDE, specifically the Healthy Donor Career Effect (HDCE) [42, 60]. These techniques, which include only long-term, active donors at the end of the qualification period, may offer better mitigation of the HDCE than other methods. Although these methods, too, cannot fully exclude the possibility of residual HDE, they seem to be superior to other adjusting methods used in studies included in the review.

Overall, the methods used in previous studies do not seem to have adequately adjusted for HDE. If these methods are used in combination, there is a possibility that the HDE might be further minimised. Interestingly, some of the previous studies employed these methods together, yet they still could not rule out the existence of residual HDE [29, 38, 41, 46]. Therefore, investigating the use of other causal methods to further reduce the HDE in observational studies examining health outcomes in blood donors was a worthwhile choice. As blood donation behaviour is assumed to be time-varying in nature, current blood donation behaviour impacts subsequent donations. In this process, the factors that determine a donor's donation behaviour also lie in the causal pathway between the donation status and long-term health outcome. In such scenarios, modern causal inference methods such as target trial emulation and g-methods could be used to measure the true causal relationship between exposure and outcome [62]. These methods might be more effective in adjusting for selection bias and, when used in conjunction with other existing methods, could further minimise HDE. Thus, for my thesis, I chose to use the target trial emulation in conjunction with other g-methods and the 'Exposure Window' technique identified from this review to mitigate the HDE while assessing the long-term health outcomes in blood donors.

Objective 2: To apply the appropriate method/s to the Sax Institute's 45 and Up Study data which is linked with other administrative health data sets to examine the association between donation and various long-term health outcomes such as mortality and cancer.

To address the second objective, I chose a relatively new causal inference framework known as Target Trial Emulation along with the 'Exposure Window' technique from the methods review, used in conjunction with other G-methods such as Inverse Probability Weighting (IPW) of the Marginal Structural Model, Targeted Minimum Loss Based Estimator (TMLE), and Sequentially Doubly Robust Estimator (SDR). I focused on long-term health outcomes such as all-cause mortality and cancer incidence (gastrointestinal, colorectal, and haematological). I

attempted to answer three research questions while adjusting for the HDE. The first question was whether initiating whole blood donation could reduce the risk of mortality (Chapter 4: Section 1), followed by whether regular high-frequency whole blood donation could reduce the mortality risk compared to low-frequency blood donation (Chapter 4: Section 2). The third question I sought to answer was whether regular high-frequency whole blood donation could increase or decrease the risk of gastrointestinal, colorectal, or haematological cancer compared to low-frequency blood donation (Chapter 4: Section 3).

While examining the relationship between initiating a blood donation and mortality using the target trial emulation method, the adjusted hazard ratio [HR: 1.0 (95% CI 0.74, 1.35)] and standardised survival curves, as shown in Chapter 4: Section 1 did not indicate any statistically significant results. This means that initiating a blood donation has no effect on mortality. I found similar results when applying the 5-year 'Exposure Window' technique to assess the impact of high-frequency blood donation on mortality in Chapter 4: Section 2. The IPW model did not show any significant association of high-frequency blood donation with mortality when compared to low-frequency blood donation [RR: 0.98 (95% CI 0.68, 1.28)]. Both time-fixed and time-varying TMLE models also produced results similar to the IPW model. Overall, both the initiation of blood donation and regular high-frequency blood donation showed no significant association with mortality risk.

There are a few other studies that have explored the association between whole blood donation and all-cause mortality. Casale et al. discovered that blood donors generally have a longer lifespan than non-donors [55]. However, their results are likely biased by the HDE as they did not account for potential confounders and compared donors with the general population. Edgren et al. noted a 30% reduction in mortality among donors but did not employ any causal inference methods or attribute the results explicitly to blood donation [51]. Further, Ullum et al. sought to quantify the internal HDE, focusing on mortality among donors who retired due

to age-related criteria [41]. Using a Poisson regression model, they examined the influence of donation rate and continued donation on active donors, identifying an HDE-adjusted impact on the donation rate. Despite finding a 7.5% decrease in mortality risk with each additional annual donation, they couldn't verify if this effect was unbiased, as the adjustment was based on data from older donors. They also noted that causality could not be inferred as the calculated mortality rate ratios were defined by comparing donors with different donation rates. Although these prior studies showed reduced mortality risk among donors, particularly high-frequency donors, compared to non-donors or occasional donors, they largely couldn't definitively confirm this as conclusive evidence of a beneficial effect of blood donation.

While I compared donors and non-donors in the target trial emulation method that I employed, this method differs from previous studies comparing donors and non-donors. In my analyses, the eligibility criteria to participate in each trial were consistently implemented in 60 consecutive trials to reduce the influence of the HDE. A wide range of potential baseline confounding factors, including self-reported health indicators, were also adjusted. Sensitivity analyses, including emulation of 120 consecutive trials, also found no significant association between blood donation and mortality. Additionally, a negative control analysis found no significant association between blood donation and hospitalisation due to injuries, further providing evidence that this analysis has adequately adjusted the impact of HDE.

It is worth noting that, in the target trial analyses, there was imperfect adherence to treatment (donation status). This occurs when participants deviate from their originally assigned treatment after the start of the trial. Both in a randomised controlled trial and target trial emulation, the common method to deal with this imperfect adherence is to censor the person-months when the participants deviate from their originally assigned treatment, a process known as per-protocol analysis. However, I did not report per protocol mortality hazard ratios in this study as 55.8% of donors switched to the non-donor group within one year of the follow-up,

which resulted in only six events in the donor group after censoring the non-adherent participants. Thus, in this analysis, I only reported the effect of starting/initiating blood donation on all-cause mortality, not the effect of continuous blood donation effect.

The use of a 5-year 'Exposure Window' in Chapter 4: Section 2 enabled me to examine the effect of high-frequency regular whole blood donation compared to low-frequency donation. Although Ullum et al. used a similar exposure window approach to determine the donation rate beforehand [41], it was first used comprehensively by Peffer and colleagues when they studied the association between regular blood donation and cardiovascular risk [42, 60]. They used a three-category exposure variable based on the tertiles of donations made in 10 years. However, in this thesis, I classified donors into high-frequency (exposure group) and low-frequency donors (control group), focusing on the frequency and consistency of donation patterns. Unlike Peffer's study, I adjusted for time-varying exposure and time-varying confounding factors during the follow-up period using the doubly robust TMLE estimator, which can still produce unbiased estimates even if the positivity assumption is violated. As I adjusted for time-varying exposure and confounding factors, one could argue that using a shorter exposure window and adjusting for time-varying confounders would have been more appropriate in this analysis. However, the impact of blood donation is assumed to be lagged, and donating blood for shorter periods can hardly be considered to have a causal effect on mortality. Thus, from a biological perspective, a longer exposure window is required to determine the causal effect of blood donation, if there is one.

In Chapter 4: Section 3, I again used the 5-year ‘Exposure Window’ to investigate whether regular high-frequency whole blood donation could increase or decrease the risk of gastrointestinal, colorectal, or haematological cancers compared to low-frequency blood donation. The IPW risk ratio for gastrointestinal/colorectal cancers was 1.25 (95% CI 0.83, 1.68), whereas it was 0.97 (95% CI 0.55, 1.40) for haematological cancers. In addition to

TMLE, I used the SDR estimator in this analysis, and they also found an almost similar pattern of results, none of which were statistically significant. This means there were no increased or decreased risks of gastrointestinal/colorectal or haematological cancers among high-frequency blood donors when compared to low-frequency donors.

Numerous studies have suggested a relationship between blood donation and cancer incidence, indicating a lower cancer risk among blood donors, possibly due to their healthier lifestyle choices [43, 51, 56]. However, none of these previous studies that examined the impact of blood donation on cancer risk used the 'Exposure Window' or 'Qualification Period' technique. A Swedish research study on iron depletion in blood donors showed a trend of reduced risk for several types of cancer among male donors who experienced significant iron loss from donation [29]. However, it acknowledged the potential influence of confounding factors like smoking, alcohol consumption, diet, exercise, body measurements, and work-related exposures. Similarly, a US cohort study found a lower colorectal cancer risk in regular male blood donors, possibly due to HDE [30]. Contrarily, some research has also suggested that high-frequency donors might have an increased risk of haematological malignancies, potentially due to enhanced cell proliferation triggered by frequent blood cell removal [29, 56]. Yet, a recent Swedish study found no heightened risk of such cancers in frequent donors, dismissing the higher risk of chronic lymphocytic leukemia as non-causal due to the lack of a dose-response relationship [46].

5.2 Strengths and Limitations

Through a systematic review of methods, followed by their application to the linked Sax Institute's 45 and Up Study data, this thesis found no statistically significant association between regular blood donation and long-term health outcomes (mortality and some types of cancers). The thesis has several notable strengths. First, I adjusted a comprehensive set of potential confounding factors in all the analyses, which most of the previous studies lacked.

Secondly, the use of 'Target Trial Emulation' and 'Exposure Window' methods helped to create less biased comparison groups, which many of the previous studies struggled to do. In the Target Trial, the continuous application of eligibility criteria and adjusting trial-specific baseline confounders largely mitigated the overall HDE. In the 'Exposure Window' method, the restriction of the comparison groups to the active donor population reduced the Healthy Registration Effect (HRE) and Healthy Donor Survivor Effect (HDSE). Additionally, the separation of the exposure period and follow-up period mitigated the Healthy Donor Career Effect (HDCE). Finally, the utilisation of doubly robust models with machine learning algorithms helped to produce more robust results and provided protection against model misspecification.

This thesis has some limitations as well. The majority of the research population in this study consists of older Australian donors, which may restrict the generalisability of the findings to younger donors and other donor populations. People who donate blood for therapeutic reasons such as patient with hereditary hemochromatosis and polycythaemia rubra vera were also excluded from the analysis. In the target trial analysis, an important issue was non-adherence to assigned treatment, resulting in a significant number of donors switching to the non-donor group. This was exacerbated by a low number of events, leading to the decision against conducting a per-protocol analysis. Another notable limitation in the 'Exposure Window' analysis was the unavailability of donation records prior to June 2002, which made it impossible to evaluate the impact of the entire donation history or the time since the first donation. Moreover, a relatively smaller sample size and shorter follow-up time compared to some of the previous studies might have impacted the precision of the effect estimation.

It is important to note that the variables I used in the analysis from 45 and Up Study data are self-reported. However, data linkage enabled us to validate and cross-reference the 45 and Up

Study data with other administrative health data sets which mitigated the concerns about data accuracy and reliability to a significant extent.

5.3 Recommendations for Future Study

The healthy donor effect is a significant methodological pitfall that can bias the studies investigating the association between blood donation and long-term health outcomes. To minimise this bias and generate robust effect estimates, future studies can utilise a few strategies. Firstly, a blend of study design and analysis plan should be implemented. Only statistical modelling techniques without appropriate study design and data collection will keep room for the residual HDE. The use of causal inference methods (Target trial, G-methods) should be considered.

Further, as also recommended by Atsma and colleagues, the analysis restricted only to the donor population will reduce the healthy donor effect significantly [15]. However, one can incorporate the target trial emulation technique if they want to compare between donor and non-donor. This approach, which is a type of observational study designed to mimic the characteristics of a randomised controlled trial, can help to produce more reliable and valid results.

As blood donation can be thought of as a time-varying exposure, Inverse Probability Weighting of Marginal Structural Models can be used to adjust for time-varying exposure and confounding factors. More advanced alternative g methods, such as Targeted Minimum Loss Based Estimator (TMLE) and the Sequentially Doubly Robust (SDR) models, can also be valuable. These tools are very useful for both time-fixed and time-varying analyses as they can employ machine learning algorithms, which can significantly improve the accuracy of the findings. Nonetheless, the implementation of these methods can be computationally intensive and time-consuming.

Lastly, to effectively apply these causal methods, researchers must have access to a comprehensive set of variables related to the health of donors both at baseline and during the follow-up period. The collection of this data is critical and should be considered during the design phase of studies to ensure that they offer the most robust and insightful results possible.

5.4 Conclusion

This thesis aimed to address the Healthy Donor Effect in examining the relationship between regular whole blood donation and long-term health outcomes, specifically mortality and gastrointestinal, and haematological cancers. Through a systematic review of methods, followed by their application to the linked Sax Institute's 45 and Up Study data, this thesis found no significant association between regular blood donation and these health outcomes. These findings can provide crucial insights for the Australian Red Cross Lifeblood's strategic planning, aiding in setting priorities and directing future research. In summary, while blood donation does not offer any clear protective benefits against mortality and cancers, it does not pose any harm too, thus improving our understanding of its long-term health impacts.

References

1. Roberts, N., et al., *The global need and availability of blood products: a modelling study.* The Lancet Haematology, 2019. **6**(12): p. e606-e615.
2. Amrein, K., et al., *Adverse events and safety issues in blood donation—a comprehensive review.* Blood reviews, 2012. **26**(1): p. 33-42.
3. Masser, B., G. Smith, and L.A. Williams, *Donor research in Australia: challenges and promise.* Transfusion Medicine and Hemotherapy, 2014. **41**(4): p. 296-301.
4. Eder, A.F., *Improving safety for young blood donors.* Transfusion medicine reviews, 2012. **26**(1): p. 14-26.
5. Australian Government Department of Health and Aged Care. *Blood and blood products in Australia*. 2020 March 1, 2023]; Available from: https://www.health.gov.au/topics/blood-and-blood-products/blood-and-blood-products-in-australia.
6. Australian Institute of Health and Welfare & Australasian Association of Cancer Registries, *Cancer in Australia: an overview, 2012*, in *Cancer series no. 74. Cat. no. CAN 70. Canberra: AIHW.* 2012.
7. AIHW: Mathur, S., *Epidemic of coronary heart disease and its treatment in Australia.*, in *Cardiovascular Disease Series No. 20. AIHW Cat. No. CVD 21. Canberra: Australian Institute of Health and Welfare.* 2002.
8. Wu, W.-C., et al., *Blood transfusion in elderly patients with acute myocardial infarction.* New England Journal of Medicine, 2001. **345**(17): p. 1230-1236.
9. Crocco, A. and D. D'Elia, *Adverse reactions during voluntary donation of blood and/or blood components. A statistical-epidemiological study.* Blood transfusion = Trasfusione del sangue, 2007. **5**(3): p. 143-152.
10. Amrein, K., et al., *Apheresis affects bone and mineral metabolism.* Bone, 2010. **46**(3): p. 789-95.
11. Fonseca-Nunes, A., P. Jakszyn, and A. Agudo, *Iron and cancer risk--a systematic review and meta-analysis of the epidemiological evidence.* Cancer Epidemiol Biomarkers Prev, 2014. **23**(1): p. 12-31.
12. Meyers, D.G., et al., *Possible association of a reduction in cardiovascular events with blood donation.* Heart, 1997. **78**(2): p. 188-193.

13. Kiss, J.E., et al., *Oral iron supplementation after blood donation: a randomized clinical trial.* Jama, 2015. **313**(6): p. 575-583.
14. Laub, R., et al., *Specific protein content of pools of plasma for fractionation from different sources: impact of frequency of donations.* Vox Sang, 2010. **99**(3): p. 220-31.
15. Atsma, F., et al., *The healthy donor effect: a matter of selection bias and confounding.* Transfusion Medicine and Hemotherapy, 2011. **51**(9): p. 1883-1885.
16. Gemelli, C.N., et al., *Demographic and health profile of older Australian blood donors: results from the Extended Donor Vigilance data linkage study (EDV: Link).* ISBT Science Series, 2018. **13**(4): p. 412-420.
17. Sullivan, J., *Iron and the sex difference in heart disease risk.* The Lancet Haematology, 1981. **317**(8233): p. 1293-1294.
18. Fernández-Real, J.M. and M. Manco, *Effects of iron overload on chronic metabolic diseases.* Lancet Diabetes Endocrinol, 2014. **2**(6): p. 513-26.
19. Sarwar, N., et al., *Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies.* Lancet, 2010. **375**(9733): p. 2215-22.
20. Datz, C., et al., *Iron homeostasis in the metabolic syndrome.* Eur J Clin Invest, 2013. **43**(2): p. 215-24.
21. Peffer, K., et al., *Donation intensity and metabolic syndrome in active whole-blood donors.* Vox Sang, 2015. **109**(1): p. 25-34.
22. Peffer, K., et al., *The effect of frequent whole blood donation on ferritin, hepcidin, and subclinical atherosclerosis.* Transfusion, 2013. **53**(7): p. 1468-74.
23. Ascherio, A., et al., *Blood donations and risk of coronary heart disease in men.* Circulation, 2001. **103**(1): p. 52-57.
24. Yücel, H., et al., *Regular blood donation improves endothelial function in adult males.* Anatol J Cardiol, 2016. **16**(3): p. 154-8.
25. MacKinnon, A. and J. Bancewicz, *Sarcoma after injection of intramuscular iron.* Br Med j, 1973. **2**(5861): p. 277-279.
26. Richmond, H., *Induction of Sarcoma in the Rat by Iron—Dextran Complex.* British medical journal, 1959. **1**(5127): p. 947.
27. Marquet, R.L., et al., *Blood donation leads to a decrease in natural killer cell activity: a study in normal blood donors and cancer patients.* Transfusion, 1993. **33**(5): p. 368-73.

28. Ishii, K., et al., *A prospective analysis of blood donation history and risk of non-Hodgkin lymphoma.* Leukemia & lymphoma, 2016. **57**(6): p. 1423-1428.
29. Edgren, G., et al., *Donation frequency, iron loss, and risk of cancer among blood donors.* Journal of the National Cancer Institute, 2008. **100**(8): p. 572-579.
30. Zhang, X., et al., *Blood donation and colorectal cancer incidence and mortality in men.* PLoS One, 2012. **7**(6): p. e39319.
31. Boot, C., et al., *Bone density in apheresis donors and whole blood donors.* Vox Sanguinis, 2015. **109**(4): p. 410-413.
32. Grau, K., et al., *No association between frequent apheresis donation and risk of fractures: a retrospective cohort analysis from S weden.* Transfusion, 2017. **57**(2): p. 390-396.
33. Lewis, S.L., et al., *Plasma proteins and lymphocyte phenotypes in long-term plasma donors.* Transfusion, 1994. **34**(7): p. 578-85.
34. Tran-Mi, B., et al., *The impact of different intensities of regular donor plasmapheresis on humoral and cellular immunity, red cell and iron metabolism, and cardiovascular risk markers.* Vox Sang, 2004. **86**(3): p. 189-97.
35. Furst, D.E., *Serum immunoglobulins and risk of infection: how low can you go?* Semin Arthritis Rheum, 2009. **39**(1): p. 18-29.
36. Tuomainen, T.-P., et al., *Cohort study of relation between donating blood and risk of myocardial infarction in 2682 men in eastern Finland.* Bmj, 1997. **314**(7083): p. 793.
37. Salonen, J.T., et al., *Donation of blood is associated with reduced risk of myocardial infarction: the Kuopio Ischaemic Heart Disease Risk Factor Study.* American journal of epidemiology, 1998. **148**(5): p. 445-451.
38. Meyers, D.G., K.C. Jensen, and J.E. Menitove, *A historical cohort study of the effect of lowering body iron through blood donation on incident cardiac events.* Transfusion, 2002. **42**(9): p. 1135-1139.
39. Vahidnia, F., et al., *Cancer incidence and mortality in a cohort of US blood donors: a 20-year study.* Journal of cancer epidemiology, 2013. **2013**.
40. Gallerani, M., et al., *Risk of illness, hospitalization and death in a cohort of blood donors in Italy.* Current medical research opinion, 2014. **30**(9): p. 1803-1812.
41. Ullum, H., et al., *Blood donation and blood donor mortality after adjustment for a healthy donor effect.* Transfusion, 2015. **55**(10): p. 2479-2485.
42. Peffer, K., et al., *Cardiovascular risk in 159 934 frequent blood donors while addressing the healthy donor effect.* Heart, 2019. **105**(16): p. 1260-1265.

43. Lasek, W., M. Jakobisiak, and T. Stoklosa, *Decreased natural killer cell activity in whole-blood donors does not seem to result in increased cancer incidence.* Transfusion, 1994. **34**(4): p. 359-360.
44. Germain, M., et al., *Iron and cardiac ischemia: a natural, quasi-random experiment comparing eligible with disqualified blood donors (CME).* Transfusion, 2013. **53**(6): p. 1271-1279.
45. Edgren, G., et al., *Blood donation and risk of polycythemia vera.* Transfusion, 2016. **56**(6pt2): p. 1622-1627.
46. Zhao, J., et al., *Risk of hematological malignancy in blood donors: A nationwide cohort study.* Transfusion, 2020. **60**(11): p. 2591-2596.
47. Haron, N.H., et al., *Donors' Calcium Level and Bone Density after Frequent and Regular Plateletpheresis Blood Donation.* Malaysian Journal of Medicine and Health Sciences, 2018. **14**(Supp 1): p. 7-10.
48. Bialkowski, W., et al., *Impact of frequent apheresis blood donation on bone density: A prospective, longitudinal, randomized, controlled trial.* Bone Reports, 2019. **10**: p. 100188.
49. Hendig, D., et al., *Donor safety in haemapheresis-influence of frequent plasma donations on parameters of calcium-phosphate and bone metabolism.* Transfusion Medicine and Hemotherapy, 2018. **45**(Supplement 1): p. 16-17.
50. Zhao, J., et al., *Frequent platelet donation is associated with lymphopenia and risk of infections: A nationwide cohort study.* Transfusion, 2020: p. 1-10.
51. Edgren, G., et al., *Improving health profile of blood donors as a consequence of transfusion safety efforts.* Transfusion, 2007. **47**(11): p. 2017-2024.
52. Atsma, F., et al., *Healthy donor effect: its magnitude in health research among blood donors.* Transfusion, 2011. **51**(8): p. 1820-1828.
53. Australian Red Cross Blood Service, *Guidelines for the selection of blood donors*. 2017: Australian Red Cross Blood Service.
54. Pink, J., et al., *Safe and sustainable plasmapheresis.* ISBT Science Series, 2017. **12**(4): p. 471-482.
55. Casale, G., M. Bignamini, and P. De Nicola, *Does blood donation prolong life expectancy?* Vox sanguinis, 1983. **45**(5): p. 398-399.
56. MERK, K., et al., *The incidence of cancer among blood donors.* International journal of epidemiology, 1990. **19**(3): p. 505-509.

57. Jiang, R., et al., *Dietary iron intake and blood donations in relation to risk of type 2 diabetes in men: a prospective cohort study.* The American journal of clinical nutrition, 2004. **79**(1): p. 70-75.
58. Germain, M., et al., *Donation by donors with an atypical pulse rate does not increase the risk of cardiac ischaemic events.* Vox sanguinis, 2013. **104**(4): p. 309-316.
59. van den Hurk, K., et al., *Associations of health status with subsequent blood donor behavior—An alternative perspective on the Healthy Donor Effect from Donor InSight.* PLoS One, 2017. **12**(10): p. e0186662.
60. Peffer, K., *Blood donation and cardiovascular disease. Addressing the healthy donor effect*. 2015, (Doctoral dissertation, [Sl: sn]).
61. Hernán, M.A. and J.M. Robins, *Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available.* Am J Epidemiol, 2016. **183**(8): p. 758-64.
62. Hernán, M.A. and J.M. Robins, *Causal Inference: What If.* 2020: Boca Raton: Chapman & Hall/CRC.
63. Bor, J., et al., *Regression discontinuity designs in epidemiology: causal inference without randomized trials.* Epidemiology, 2014. **25**(5): p. 729-37.
64. Hughes, J.A., et al., *Characterization of health issues in young first-time blood donors.* Vox Sanguinis, 2021. **116**(3): p. 288-295.
65. Jones, J.M., et al., *Estimated US infection-and vaccine-induced SARS-CoV-2 seroprevalence based on blood donations, July 2020-May 2021.* J Jama, 2021. **326**(14): p. 1400-1409.
66. Uyoga, S., et al., *Prevalence of SARS-CoV-2 antibodies from a national serosurveillance of Kenyan blood donors, January-March 2021.* J Jama, 2021. **326**(14): p. 1436-1438.
67. Matthews, A.A., et al., *Target trial emulation: applying principles of randomised trials to observational studies.* bmj, 2022. **378**.
68. Danaei, G., et al., *Observational data for comparative effectiveness research: an emulation of randomised trials of statins and primary prevention of coronary heart disease.* Statistical methods in medical research, 2013. **22**(1): p. 70-96.
69. Hernán, M.A., *The hazards of hazard ratios.* Epidemiology, 2010. **21**(1): p. 13.
70. Knipschild, P., P. Leffers, and A.R. Feinstein, *The qualification period.* Journal of clinical epidemiology, 1991. **44**(6): p. 461-464.
71. Luedtke, A.R., et al., *Sequential double robustness in right-censored longitudinal models.* J arXiv preprint 2017.

72. van der Laan, M.J. and S. Rose, *Targeted learning in data science: causal inference for complex longitudinal studies*. 2018: Springer.
73. Brooks, J.C., et al., *Targeted minimum loss-based estimation of causal effects in right-censored survival data with time-dependent covariates: Warfarin, stroke, and death in atrial fibrillation.* Journal of Causal Inference, 2013. **1**(2): p. 235-254.
74. Tran, L., et al., *Double robust efficient estimators of longitudinal treatment effects: comparative performance in simulations and a case study.* The international journal of biostatistics, 2019. **15**(2): p. 20170054.
75. Chernozhukov, V., et al., *Double/debiased machine learning for treatment and structural parameters*. 2018, Oxford University Press Oxford, UK.
76. Hoffman, K.L., et al., *Introducing longitudinal modified treatment policies: a unified framework for studying complex exposures.* arXiv preprint arXiv:2304.09460, 2023.
77. Collaborators, a.U.S., et al., *Cohort profile: the 45 and up study*, in *International journal of epidemiology*. 2008. p. 941-947.
78. Kelman, C.W., A.J. Bass, and C.D.J. Holman, *Research use of linked health data—a best practice protocol.* Australian and New Zealand journal of public health, 2002. **26**(3): p. 251-255.
79. Bleicher, K., et al., *Cohort Profile Update: The 45 and Up Study.* International Journal of Epidemiology, 2022.
80. Charlson, M.E., et al., *A new method of classifying prognostic comorbidity in longitudinal studies: development and validation.* Journal of chronic diseases, 1987. **40**(5): p. 373-383.
81. Quan, H., et al., *Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries.* American journal of epidemiology, 2011. **173**(6): p. 676-682.
82. Clark, D.O., et al., *A chronic disease score with empirically derived weights.* Medical care, 1995. **33**(8): p. 783-795.
83. Pratt, N.L., et al., *The validity of the Rx-Risk comorbidity index using medicines mapped to the anatomical therapeutic chemical (ATC) classification system.* BMJ open, 2018. **8**(4): p. e021122.
84. Deyo, R.A., D.C. Cherkin, and M.A. Ciol, *Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases.* Journal of clinical epidemiology, 1992. **45**(6): p. 613-619.

85. Rahman, M.M., S. Karki, and A. Hayen, *A methods review of the "healthy donor effect" in studies of long-term health outcomes in blood donors.* Transfusion, 2022. **62**(3): p. 698-712.
86. World Health Organization. *Blood Safety and Availability*. 2022 March 1, 2023]; Available from: https://www.who.int/news-room/fact-sheets/detail/blood-safety-and-availability.
87. Milborrow, S., et al. *earth: Multivariate adaptive regression splines*. 2021.
88. Van Buuren, S. and K. Groothuis-Oudshoorn, *mice: Multivariate imputation by chained equations in R.* Journal of statistical software, 2011. **45**: p. 1-67.
89. Huang, X., *Iron overload and its association with cancer risk in humans: evidence for iron as a carcinogenic metal.* Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 2003. **533**(1-2): p. 153-171.
90. Halliwell, B. and J.M.C. Gutteridge, *Oxygen free radicals and iron in relation to biology and medicine: some problems and concepts.* Archives of biochemistry and biophysics, 1986. **246**(2): p. 501-514.
91. Tiniakos, G. and R. Williams, *Cirrhotic process, liver cell carcinoma and extrahepatic malignant tumors in idiopathic haemochromatosis. Study of 71 patients treated with venesection therapy.* Applied Pathology, 1988. **6**(2): p. 128-138.
92. Selby, J.V. and G.D. Friedman, *Epidemiologic evidence of an association between body iron stores and risk of cancer.* International journal of cancer, 1988. **41**(5): p. 677-682.
93. Knekt, P., et al., *Body iron stores and risk of cancer.* International journal of cancer, 1994. **56**(3): p. 379-382.
94. Nelson, R.L., *Iron and colorectal cancer risk: human studies.* Nutrition reviews, 2001. **59**(5): p. 140-148.
95. Cross, A.J., et al., *Iron and colorectal cancer risk in the α-tocopherol, β-carotene cancer prevention study.* International journal of cancer, 2006. **118**(12): p. 3147-3152.
96. Kabat, G.C., et al., *A cohort study of dietary iron and heme iron intake and risk of colorectal cancer in women.* British journal of cancer, 2007. **97**(1): p. 118-122.
97. Kato, I., et al., *Iron intake, body iron stores and colorectal cancer risk in women: a nested case-control study.* International journal of cancer, 1999. **80**(5): p. 693-698.
98. Larsson, S.C., et al., *Re: Heme iron, zinc, alcohol consumption, and risk of colon cancer.* Journal of the National Cancer Institute, 2005. **97**(3): p. 232-233.

99. Wurzelmann, J.I., et al., *Iron intake and the risk of colorectal cancer.* Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 1996. **5**(7): p. 503-507.
100. Zacharski, L.R., et al., *Decreased cancer risk after iron reduction in patients with peripheral arterial disease: results from a randomized trial.* Journal of the National Cancer Institute, 2008. **100**(14): p. 996-1002.
101. Karki, S., et al., *Regular high-frequency whole blood donation and risk of cardiovascular disease in middle-aged and older blood donors in Australia.* Transfusion, 2023.

Appendices

A. Search Strategy for PubMed Database.

("Blood Donors"[mh] OR "blood donor*"[tiab] OR "blood donation"[tiab] OR "donating blood"[tiab] OR "plasma donor*"[tiab] OR "plasma donation*"[tiab] OR "donating plasma"[tiab] OR "platelet donor*"[tiab] OR "platelet donation*"[tiab] OR "donating platelet"[tiab] OR "apheresis donor*"[tiab] OR "apheresis donation*"[tiab])

AND

("Coronary Disease"[mh] OR "coronary disease*"[tiab] OR "coronary heart disease*"[tiab] OR "Myocardial Infarction"[mh] OR "myocardial infarction*"[tiab] OR "heart attack*"[tiab] OR "Heart Diseases"[mh] OR "heart disease*"[tiab] OR "heart failure"[tiab] OR "cardiac disease*"[tiab] OR cardiovasc*[tiab] OR cvd[tiab] OR Neoplasms[mh] OR cancer[tiab] OR "Bone Density"[mh] OR "Fractures, Bone*"[mh] OR fracture*[tiab] OR "bone mineral density"[tiab] OR "bone density"[tiab] OR "Life Expectancy"[mh] OR "life expectancy"[tiab] OR "Mortality"[mh] OR "mortality"[tiab] OR "morbidity"[tiab] OR "all-cause mortality"[tiab] OR infection[tiab])

NOT

("Blood Transfusion"[mh] OR "blood transfusion*"[tiab] OR "Blood Transfusion, Autologous"[mh] OR "autologous blood transfusion"[tiab] OR "autotransfusion"[tiab] OR "autologous blood"[tiab] OR vasovagal[tiab] OR fainting[tiab] OR haematoma[tiab] OR "arterio-venous fistula"[tiab] OR "compartment syndrom*"[tiab] OR hypovolemic [tiab] OR bruising[tiab] OR hypocalcemia[tiab] OR "Transfusion Reaction"[mh] OR "adverse reaction*"[tiab] OR hemolytic[tiab] OR hypotensive[tiab] OR "delayed serologic"[tiab] OR "allergic reaction"[tiab] OR "circulatory overload*"[tiab] OR dyspnea[tiab] OR "posttransfusion purpura"[tiab] OR Transplantation[mh] OR transplantation[tiab] OR HIV[mh] OR HIV[tiab] OR "human immunodeficiency virus"[tiab] OR "Acquired Immunodeficiency Syndrome"[mh] OR aids[tiab] OR htlv[tiab] OR "human t cell leukemia"[tiab] OR "human t lymphotropic"[tiab] OR "human t cell lymphotropic"[tiab] OR "Hepatitis B"[mh] OR "Hepatitis C"[mh] OR hepatitis[tiab] OR "Syphilis"[mh] OR syphilis[tiab] OR tuberculosis[tiab] OR "blood group*"[tiab] OR DNA[mh] OR dna[tiab] OR RNA[mh] OR rna[tiab] OR Genetics[mh] OR "genetic*"[tiab] OR Genes[mh] OR "genes"[tiab] OR "stem cell*"[tiab] OR "Blood Safety"[mh] OR "blood supply safety"[tiab] OR hemovigilance[tiab] OR "blood product*"[tiab] OR "blood recipient*"[tiab] OR Child[mh] OR child[tiab] OR Infant[mh] OR infant*[tiab] OR newborn[tiab])

B. Study hypothesis, outcomes, and association with blood donation from methods review.

Table 0.1 Study hypothesis, outcomes, and association with blood donation.

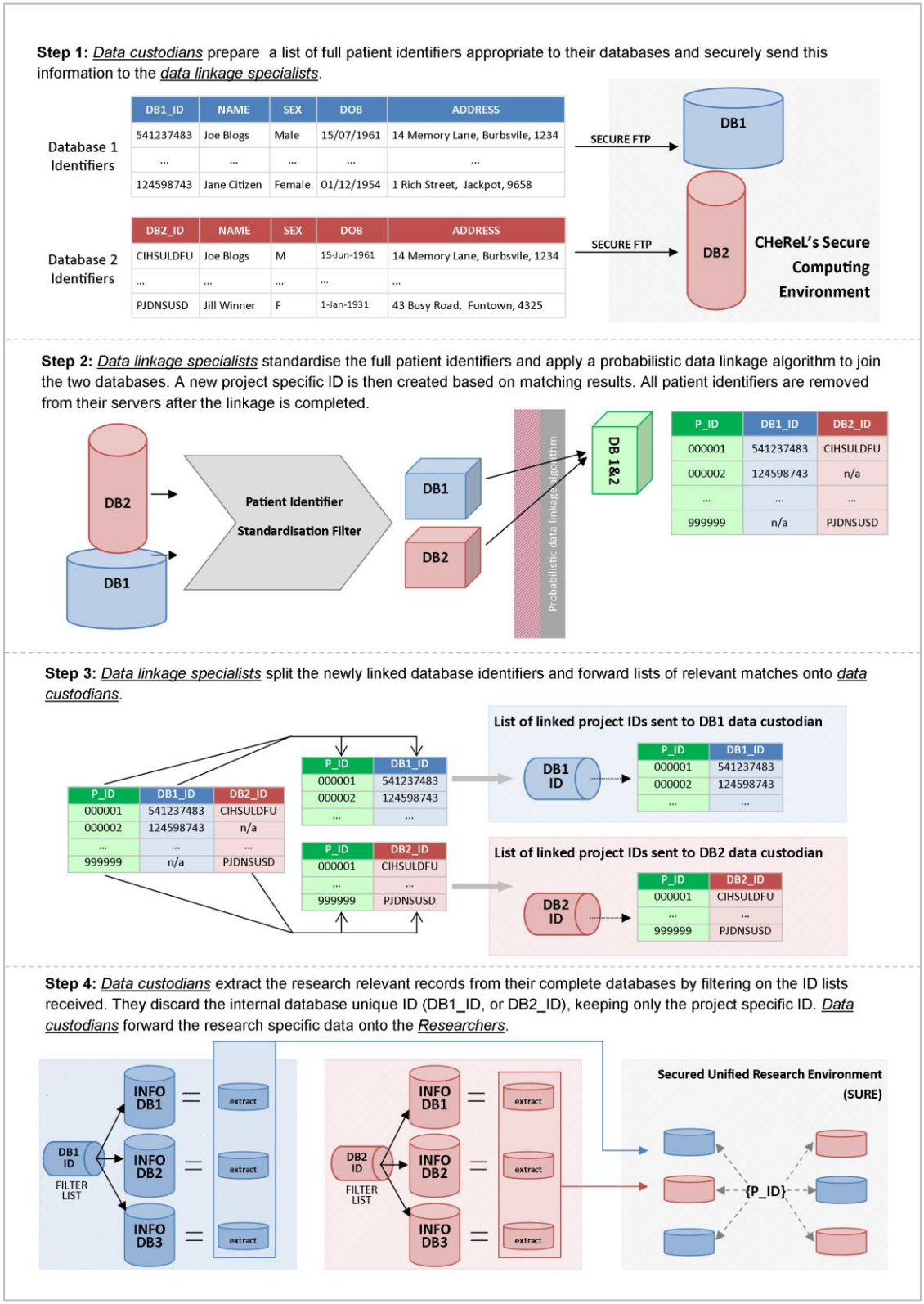
First author and year	Hypothesis/argument	Primary Outcome	Factors controlled (adjusted)	Observed association
Casale 1983	Longer life expectancy in blood donors than non-donors	Death	N/A	Blood donors had a lower death rate than non-donor
Merk 1990	Blood donation might be associated with cancer development.	Cancer	Age	Blood donors had a lower cancer risk
Lasek 1994	Investigated the cancer incidence among blood donors in relation to NK[a] cells	Cancer	N/A	No association of blood donation and cancer incidence in relation with NK[a] cells
Meyers 1997	Depletion of iron through blood donation may reduce CHD[b] among donors.	Cardiovascular events	Education level, physical activity, lipid disorders, hypertension, and diabetes mellitus	Non-smoker male blood donors had a lower CHD[b] risk
Tuomainen 1997	Investigated the association of donating blood with the risk of acute myocardial infarction	Myocardial Infarction	Age, diseases and family history, smoking, blood pressure, apolipoprotein B	Blood donor had a lower myocardial infarction rate
Salonen 1998	Voluntary blood donation is associated with a reduced risk of acute myocardial infarction	Myocardial Infarction	Age, diseases and family history, biologic risk factors, behavioral risk factors, psychological risk factors	Blood donor had a lower myocardial infarction rate
Ascherio 2001	Regular blood donation reduces the risk of myocardial infarction	Fatal CHD[b] and nonfatal myocardial infarction.	Age, BMI[c], smoking, physical activity, alcohol, vit E, history of myocardial infarction, diabetes, hypertension, high blood cholesterol	No association between blood donation and myocardial infarction
Meyers 2002	Whole blood donation might be associated with a reduced risk of cardiovascular events.	Cardiovascular diseases:	Age, BMI[c], smoking, diabetes drug, Antihypertensive drugs, Lipid-modifying drugs, family history, prior blood donations	blood donor had a lower myocardial infarction rate
Jiang 2004	Examined dietary iron intake and history of blood donations in relation to the incidence of type 2 diabetes	Type 2 diabetes	Age, (BMI; in kg/m2), family history of diabetes, physical activity, smoking, alcohol consumption, and dietary variables	No association between blood donation and type 2 diabetes
Edgren 2008	Repeated blood donation is associated with cancer incidence.	Cancer	Age, sex, country of residence.	No association between blood donation and cancer overall. Although the risk of non-Hodgkin lymphoma was

First author and year	Hypothesis/argument	Primary Outcome	Factors controlled (adjusted)	Observed association
				higher among frequent plasma donors
Amrein 2010	Investigated the effects of apheresis on acid-base balance, bone, and mineral metabolism and compared BMD[d] at the lumbar spine and hip of donors to matched control subjects	BMD[d]	BMI[c], physical activity, daily calcium intake	Apheresis donor had a lower BMD[d] rate
Zhang 2012	Frequent blood donation is associated with a lower risk of colorectal cancer	Colorectal Cancer	Age, smoking, colorectal cancer history in the family, colonoscopy history, BMI[c], aspirin use, physical activity,	No association between colorectal cancer and blood donation
Germain 2013	The rate of CHD[b] would be lower among donors who remained eligible, that is, potentially exposed to bloodletting, compared to disqualified (unexposed) donors.	CHD[b] related hospitalization and death	Age, sex, region, number of previous donations, year of entry in the study	No association between blood donation and CHD[b]
Vahidnia 2013	Investigated cancer incidence among blood donors, also investigated HDE[e]	All-cause mortality and cancer	Cancer site, adjusted for age at diagnosis, sex, race, socio-economic status, tumor stage, and grade at diagnosis	The donor had a lower death rate (among cancer subjects)
Germain 2012	Temporarily deferring prospective donors who have an atypical pulse rate can prevent the triggering of CHD[b] events.	Death/hospitalization due to CHD[b]	Age, sex, residence, previous donations,	No association between atypical pulse rate and CHD[b]
Gallerani 2014	Blood donors exhibit significant differences in an increased risk of illness, hospitalization, and death when compared with non-blood donors.	Diseases (malignancy, leukemias, lymphomas, myeloma), hospitalization, death	Age, sex, number of hospitalizations	Blood donor had a lower death rate, No association with malignancy
Boot 2015	BMD[d] is lower in postmenopausal apheresis donors compared to postmenopausal whole blood donors of similar age.	BMD[d]	Age, BMI[c], postmenopausal years	Apheresis donors did not have lower BMD[d]
Ullum 2015	Investigated the relation between blood donation frequency and mortality.	All-cause mortality	Age, sex, country, calendar period, Haemoglobin (hb)	Frequent blood donors had a lower mortality rate
Ishii 2016	Frequent blood donation increases the risk of Non-Hodgkin's lymphoma	Non-Hodgkin's lymphoma (including subtypes)	Age, BMI[c], alcohol, smoking, height, physical activity, race	No association between blood donation and NHL
Edgren 2016	Frequent donors have a higher PV[f] risk than less frequent donors.	PV[f]	Effect modification (sex, country, age), haemoglobin concentration	No association between blood donation and PV[f]

First author and year	Hypothesis/argument	Primary Outcome	Factors controlled (adjusted)	Observed association
Grau 2017	Frequent apheresis donation is associated with fracture risk	All fractures, especially osteoporosis-related fractures	Age (in 1-year categories), sex, calendar period of observation (in 1-year categories),	No association between apheresis blood donation and fracture
Haron 2018	Long-term regular plateletpheresis donation affects calcium and magnesium levels and BMD[d]	Calcium magnesium and BMD[d]	N/A	No association between apheresis blood donation and BMD[d]
Hendig 2018	Investigated the impact of frequent plasma donations on parameters of calcium-phosphate and bone metabolism	Bone metabolism marker	N/A	Apheresis donor has increased risk of bone turnover
Peffer 2019	Higher frequency blood donation decreases cardiovascular risk	Cardiovascular disease events (including cardiovascular death)	Age, donation career, SBP[g], DBP[h], BMI[c], blood type	Women donors had decreased cardiovascular disease risk; Man donors had no association with cardiovascular diseases
Bialkowski 2019	High-frequency apheresis blood donation causes a decline in BMD[d]	BMD[d]	Age, risk factors, BMI[c], baseline BMD[d], health condition, diet	Apheresis donation had no association with BMD[d] for male donors
Zhao 2020	Blood donations increase the risk of developing haematological malignancies, specifically ALL[i], AML[j], , CLL[k], CML[l], Hodgkin lymphoma, and myeloma, as well other non-Hodgkin lymphomas.	ALL[i], AML[j], CLL[k], CML[l], Hodgkin lymphoma, multiple myeloma, and other non-Hodgkin lymphomas	Age, sex, age at first donation, country of birth	No association between blood donation and malignancy
Zhao 2020	Investigated the risk of infection in Plateletpheresis donors using an LRS chamber.	Infection (common bacterial infections, immunosuppression-related infections)	Age, sex, the interaction of age and sex, region, and calendar year	Apheresis donors had an increased risk of infection

[a] NK, Natural killer [b] CHD, Coronary heart disease [c] BMI, Body mass index [d] BMD, Bone mineral density [e] HDE, Healthy donor effect
[f] PV, Polycythemia vera [g] SBP, Systolic blood pressure [h] DBP, Diastolic blood pressure [i] ALL, Acute lymphoblastic leukemia
[j] AML, Acute myeloid leukemia [k] CLL, Chronic lymphocytic leukemia [l] CML, Chronic myeloid leukemia

C. Overview of the separation principle for a general database linkage methodology as described by Kelman, Bass and Holman [78]



D. Intention to treat (ITT) mortality hazard ratios for 60 trials with 95% CI.

Table 0.2 Intention to treat (ITT) mortality hazard ratios for 60 trials with 95% CI (Administrative end June 2016).

	Donor vs non-donor
Unique Persons	121967
Cases	5605
Unique cases	2800
Person trials	263300
Unadjusted[a]	0.52 (0.41, 0.65)
Age-sex adjusted[a]	0.64 (0.50, 0.81)
Adjusted for all baseline covariates[b]	0.87 (0.68, 1.10)

[a]Confidence intervals are calculated using a robust variance estimator as many individuals participated in more than one trial.
[b]Baseline variables in Table 4.1 were included as covariates.

E. Intention to treat (ITT) mortality hazard ratios for 120 trials with 95% CI.

Table 0.3 Intention to treat (ITT) mortality hazard ratios for 120 trials with 95% CI (5 years follow-up or administrative end July 2016).

	Donor vs non-donor
Unique Persons	142476
Cases	5504
Unique cases	1748
Person trials	590800
Unadjusted[a]	0.49 (0.36, 0.65)
Age-sex adjusted[a]	0.59 (0.44, 0.79)
Adjusted for all baseline covariates[b]	0.83 (0.62, 1.11)

[a]Confidence intervals are calculated using a robust variance estimator as many individuals participated in more than one trial.
[b]Baseline variables in Table 4.1 were included as covariates.

F. Intention to treat (ITT) injury-hospitalization hazard ratios for 60 and 120 trials with 95% CI.

Table 0.4 Intention to treat (ITT) injury-hospitalization hazard ratios for 60 and 120 trials with 95% CI.

	60 trials Donor vs non-donor	120 trials Donor vs non-donor
Unadjusted[a]	0.88 (0.76, 1.01)	0.92 (0.81, 1.04)
Age-sex adjusted[a]	0.89 (0.76, 1.03)	0.93 (0.81, 1.05)
Adjusted for all baseline covariates[b]	0.94 (0.81, 1.10)	0.98 (0.86, 1.12)

[a]Confidence intervals are calculated using a robust variance estimator as many individuals participated in more than one trial.
[b]Baseline variables in Table 4.1 were included as covariates.

G. Categorisation and derivation of variables used in the target trial study.

Table 0.5 Categorisation and derivation of variables used in this study from 45 and Up Study data, APDC, MBS and PBS data set.

Variable names and categories in this study	**45 and Up Study baseline questions or other data set's variables name**	**Recategorization/Derivation**
Age at baseline (continuous variable)	#What is your date of birth? #What is today's date?	(today's date-date of birth)/365.25 + 730
Body mass index at recruitment (kg/m^2)	#How tall are you without shoes? #About how much do you weigh?	Body mass index is calculated as-weight in kilogram/(height in meter)^2. Unknown category represents when information is lacking to calculate BMI or the calculated BMI is invalid (< 9 or 50)
<18.5		
18.5 – 24.9		
25-29.9		
30+		
Unknown		
Smoking Status	#Have you ever been a regular smoker? Yes ▼ No If No – please go to question x How old were you when you started smoking regularly? □ years old Are you a regular smoker now? Yes, No If No – how old were you when you stopped smoking regularly? □ years old	Never= 'Have you ever been a regular smoker- No', Current= 'Are you a regular smoker now?-Yes' Past= 'Have you ever been a regular smoker- Yes' but 'Are you a regular smoker now?-No' Unknown when not enough information is available to categorise in to above three categories
Never		
Former		
Regular		
Unknown		
Self-rated health at recruitment	#In general, how would you rate your: overall health? -Excellent -Very good -Good -Fair -Poor	fair and poor categories were combined as one category fair/poor. Unknown= No information provided
Excellent		
Very good		
Good		
Fair/Poor		
Unknown		
Alcohol consumption/day	#About how many alcoholic drinks do you have each week? one drink = a glass of wine, middy of beer or nip of spirits (put "0" if you do not drink, or have less than one drink each week) □ □ number of alcoholic drinks each week	None=0 or <1 drink each week, ≤1/day= number of alcoholic drinks each week is ≤7 >1/day= number of alcoholic drinks each week is >7 Unknown= No information provided
None		
≤1/day		
>1/day		
Education level	#What is the highest qualification you have completed? (please put a cross in the most appropriate box) -no school certificate or other qualifications	No formal education ='no school certificate or other qualifications' School to diploma= 'school or intermediate certificate (or equivalent)', 'higher school or
No formal education		
School to Diploma		
University		
Unknown		

Variable names and categories in this study	45 and Up Study baseline questions or other data set's variables name	Recategorization/Derivation
	-school or intermediate certificate (or equivalent) -higher school or leaving certificate (or equivalent) -trade/apprenticeship (e.g. hairdresser, chef) -certificate/diploma (e.g. child care, technician) -university degree or higher	leaving certificate (or equivalent)', 'trade/apprenticeship (e.g. hairdresser, chef)', 'certificate/diploma (e.g. child care, technician)' University='university degree or higher' Unknown= no information provided
Annual household income <20k 20k-39k 40k-69k 70k+ Unknown	#What is your usual yearly HOUSEHOLD income before tax, from all sources? (please include benefits, pensions, superannuation, etc) -less than $5,000 per year -$5,000-$9,999 per year -$10,000-$19,999 per year -$20,000-$29,999 per year -$30,000-$39,999 per year -$40,000-$49,999 per year -$50,000-$69,999 per year -$70,000 or more per year -I would rather not answer this question	<20k= 'less than $5,000 per year', '$5,000-$9,999 per year' , '$10,000-$19,999 per year' 20k-39k= '$20,000-$29,999 per year', '$30,000-$39,999 per year' 40k-69k= '$40,000-$49,999 per year', '$50,000-$69,999 per year' 70k+= '$70,000 or more per year' Unknown = 'I would rather not answer this question' or when no information available
Physical activity/week <1/week ≥1/week Unknown	#How many TIMES did you do each of these activities LAST WEEK? (put "0" if you did not do this activity) Vigorous physical activity (that made you breathe harder or puff and pant, like jogging, cycling, aerobics, competitive tennis, but not household chores or gardening) □ □ times in the last week	<1/week= No vigorous physical activity in the last week ≥1/week= 1 or more times in the last week Unknown= No information is available
Daily fruits/vegetable consumed 0-2 3-4 5+ Unknown	#About how many serves of fruit or glasses of fruit juice do you usually have each day? A serve is 1 medium piece or 2 small pieces or 1 cup of diced or canned fruit pieces (put "0" if you eat less than one serve a day) □ □ number of serves of fruit each day □ □ number of glasses of fruit juice each day □ I don't eat fruit # About how many serves of vegetables do you usually eat each day? A serve is half a cup of cooked vegetables or one cup of	Absolute number of serves of fruit and vegetable consumed in a day was combined. The combined number was used for categorisation. 0-2= when 'number of serves of fruit each day'+ 'number of serves of cooked vegetables each day'+ 'number of serves of raw vegetables each day (e.g. salad)'≤2 3-4= when 'number of serves of fruit each day'+ 'number of serves of cooked vegetables each day'+ 'number of serves of raw vegetables each day (e.g. salad)'3 to 4

Variable names and categories in this study	45 and Up Study baseline questions or other data set's variables name	Recategorization/Derivation
	salad (please include potatoes and put "0" if less than one a day) □ □ number of serves of cooked vegetables each day □ □ number of serves of raw vegetables each day (e.g. salad) □ I don't eat vegetables	5+= when 'number of serves of fruit each day'+ 'number of serves of cooked vegetables each day'+ 'number of serves of raw vegetables each day (e.g. salad)'>=5 Unknown= No information available
Location Major city Regional/Remote Unknown	# RA_NAME_2011	Geographical location of the participant was categorised as — major city, inner regional or outer regional/remote, according to the Accessibility/Remoteness Index of Australia [ARIA+] derived from postcode at recruitment. Major city, when RA_NAME_2011= 'Major Cities of Australia' Regional and remote, when RA_NAME_2011= 'Inner Regional Australia' or 'Outer Regional Australia' or 'Remote Australia' or 'Very Remote Australia' Unknown= No information available
No. of gp visits in the past 3 months 0-1 2-4 5+	# date of service in Medicare claims data	Counting number of services taken in the past 3 months for an individual. Three categories were created for this No. of Gp visits variable.
No. of referrals in the past 3 months 0 1 2-4 5+	# date of referral in Medicare claims data	Counting number of diagnostic tests referrals given in the past 3 months for an individual. Four categories were created for this No. of referrals variable.
Charlson co-morbidity index 0 ≥ 1	# Diagnoses for the episode of care from APDC data. Principal diagnosis has 'P' suffix. diagnosis_codeP, diagnosis_code1 -diagnosis_code50 # The diagnosis code used ICD-10 algorithm.	The weighted Charlson Co-morbidity index were calculated by using the following reference: Quan, H., et al., *Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries.* American journal of epidemiology, 2011. **173**(6): p. 676-682.
Rx-Risk index None -6 to -1 0 to 2	# Anatomical Therapeutic Chemical (ATC) classification code (atc_code) and date on which	The weighted Rx-Risk index was calculated by using the following reference:

Variable names and categories in this study	45 and Up Study baseline questions or other data set's variables name	Recategorization/Derivation
3+	the PBS item was supplied (date_of_supply)	Pratt, N.L., et al., *The validity of the Rx-Risk comorbidity index using medicines mapped to the anatomical therapeutic chemical (ATC) classification system.* BMJ open, 2018. **8**(4): p. e021122.

H. Estimated 7-year mortality risk, risk difference and risk ratios for high and low-frequency donors in a complete case analysis.

Table 0.6 Estimated 7-year mortality risk, risk difference and risk ratios for high and low-frequency donors (Complete case analysis).

Models	Risk, % (95% CI)		Risk Difference, % (95% CI)	Risk Ratio (95% CI)
	Low Frequency	High Frequency		
Inverse probability weighted[a]	1.8 (1.4, 2.2)	1.5 (1.1, 1.9)	-0.3 (-0.9, 0.2)	0.83 (0.47, 1.19)
Targeted minimum loss based estimator (TMLE)[a]	1.8 (1.5, 2.1)	1.5 (1.2, 1.7)	-0.3 (-0.7, 0.1)	0.81 (0.65, 1.01)
TMLE (time-varying)[b]	1.6 (1.2, 2.1)	1.6 (1.0, 2.3)	0.0 (-0.8, 0.8)	1.00 (0.60, 1.66)

[a]adjusted for sex, age, BMI, smoking status, self-rated health, alcohol consumption, education, annual income, physical activity, daily consumption of fruits and vegetables, location, no. of GP visits in the past 1 year, no. of referrals in the past 1 year, Charlson co-morbidity index, Rx-Risk index.

[b]time-varying TMLE included yearly exposure status, yearly Charlson co-morbidity index, yearly Rx-Risk index, yearly GP visits and yearly referral information.

I. Estimated 7-year mortality risk, risk difference and risk ratios for high and low-frequency donors with a 3-year exposure period.

Table 0.7 Estimated 7-year mortality risk, risk difference and risk ratios for high and low-frequency donors with a 3-year exposure period.

Models	Risk, % (95% CI)		Risk Difference, % (95% CI)	Risk Ratio (95% CI)
	Low Frequency	High Frequency		
Inverse probability weighted	1.7 (1.4, 2.0)	1.6 (1.4, 1.9)	-0.1 (-0.5, 0.3)	0.94 (0.68, 1.21)
Targeted minimum loss based estimator (TMLE)	1.7 (1.5, 1.9)	1.6 (1.4, 1.8)	-0.1 (-0.4, 0.2)	0.94 (0.79, 1.11)
TMLE (time-varying)*	2.0 (1.6, 2.5)	1.3 (1.0, 1.6)	-0.8 (-1.3, -0.2)	0.60 (0.44, 0.83)

[a]Adjusted for sex, age, BMI, smoking status, self-rated health, alcohol consumption, education, annual income, physical activity, daily consumption of fruits and vegetables, location, no. of GP visits in the past 1 year, no. of referrals in the past 1 year, Charlson co-morbidity index, Rx-Risk index.

[b]time-varying TMLE included yearly exposure status, yearly Charlson co-morbidity index, yearly Rx-Risk index, yearly GP visits and yearly referral information.

J. Estimated 5-year mortality risk, risk difference and risk ratios for high and low-frequency donors with a 7-year exposure period.

Table 0.8 Estimated 5-year mortality risk, risk difference and risk ratios for high and low-frequency donors with a 7-year exposure period.

Model	Risk, % (95% CI)		Risk Difference, % (95% CI)	Risk Ratio (95% CI)
	Low Frequency	High Frequency		
Inverse probability weighted	1.2 (0.9, 1.5)	1.0 (0.6, 1.4)	-0.2 (-0.6, 0.3)	0.86 (0.38, 1.34)
Targeted minimum likelihood estimator (TMLE)	1.2 (0.9, 1.4)	0.9 (0.7, 1.2)	-0.2 (-0.6, 0.1)	0.80 (0.57, 1.11)
TMLE (time-varying)*	1.8 (1.1, 2.5)	1.5 (1.0, 1.9)	-0.3 (-1.2, 0.5)	0.81 (0.49, 1.33)

[a]adjusted for sex, age, BMI, smoking status, self-rated health, alcohol consumption, education, annual income, physical activity, daily consumption of fruits and vegetables, location, no. of GP visits in the past 1 year, no. of referrals in the past 1 year, Charlson co-morbidity index, Rx-Risk index.

[b]time-varying TMLE included yearly exposure status, yearly Charlson co-morbidity index, yearly Rx-Risk index, yearly GP visits and yearly referral information.

K. Categorisation and derivation of variables used in the all-cause mortality exposure window study.

Table 0.9 Categorisation and derivation of variables used in the exposure window all-cause mortality from 45 and Up Study data, APDC, Medicare claims and PBS data set.

Variable names and categories in this study	**45 and Up Study baseline questions or other data set's variables name**	**Recategorization/Derivation**
Age at baseline (continuous variable)	#What is your date of birth? #What is today's date?	(today's date-date of birth)/365.25 + 730
Body mass index at recruitment (kg/m^2) <18.5 18.5 – 24.9 25-29.9 30+ Unknown	#How tall are you without shoes? #About how much do you weigh?	Body mass index is calculated as-weight in kilogram/(height in meter)^2. Unknown category represents when information is lacking to calculate BMI or the calculated BMI is invalid (< 9 or 50)
Smoking Status Never Former Regular Unknown	#Have you ever been a regular smoker? Yes ▼ No If No – please go to question x How old were you when you started smoking regularly? □ years old Are you a regular smoker now? Yes, No If No – how old were you when you stopped smoking regularly? □ years old	Never= 'Have you ever been a regular smoker- No', Current= 'Are you a regular smoker now?-Yes' Past= 'Have you ever been a regular smoker- Yes' but 'Are you a regular smoker now?-No' Unknown when not enough information is available to categorise in to above three categories
Self-rated health at recruitment Excellent Very good Good Fair/Poor Unknown	#In general, how would you rate your: overall health? -Excellent -Very good -Good -Fair -Poor	fair and poor categories were combined as one category fair/poor. Unknown= No information provided
Alcohol consumption/day None ≤1/day >1/day Unknown	#About how many alcoholic drinks do you have each week? one drink = a glass of wine, middy of beer or nip of spirits (put "0" if you do not drink, or have less than one drink each week) □ □ number of alcoholic drinks each week	None=0 or <1 drink each week, ≤1/day= number of alcoholic drinks each week is ≤7 >1/day= number of alcoholic drinks each week is >7 Unknown= No information provided
Education level No formal education School to Diploma University	#What is the highest qualification you have completed? (please put a cross in the most appropriate box)	No formal education ='no school certificate or other qualifications' School to diploma= 'school or intermediate certificate (or

Variable names and categories in this study	45 and Up Study baseline questions or other data set's variables name	Recategorization/Derivation
Unknown	-no school certificate or other qualifications -school or intermediate certificate (or equivalent) -higher school or leaving certificate (or equivalent) -trade/apprenticeship (e.g. hairdresser, chef) -certificate/diploma (e.g. child care, technician) -university degree or higher	equivalent)', 'higher school or leaving certificate (or equivalent)', 'trade/apprenticeship (e.g. hairdresser, chef)', 'certificate/diploma (e.g. child care, technician)' University='university degree or higher' Unknown= no information provided
Annual household income <20k 20k-39k 40k-69k 70k+ Unknown	#What is your usual yearly HOUSEHOLD income before tax, from all sources? (please include benefits, pensions, superannuation, etc) -less than $5,000 per year -$5,000-$9,999 per year -$10,000-$19,999 per year -$20,000-$29,999 per year -$30,000-$39,999 per year -$40,000-$49,999 per year -$50,000-$69,999 per year -$70,000 or more per year -I would rather not answer this question	<20k= 'less than $5,000 per year', '$5,000-$9,999 per year' , '$10,000-$19,999 per year' 20k-39k= '$20,000-$29,999 per year', '$30,000-$39,999 per year' 40k-69k= '$40,000-$49,999 per year', '$50,000-$69,999 per year' 70k+= '$70,000 or more per year' Unknown = 'I would rather not answer this question' or when no information available
Physical activity/week <1/week ≥1/week Unknown	#How many TIMES did you do each of these activities LAST WEEK? (put "0" if you did not do this activity) Vigorous physical activity (that made you breathe harder or puff and pant, like jogging, cycling, aerobics, competitive tennis, but not household chores or gardening) □ □ times in the last week	<1/week= No vigorous physical activity in the last week ≥1/week= 1 or more times in the last week Unknown= No information is available
Daily serves of fruits/vegetables 0-2 3-4 5+ Unknown	#About how many serves of fruit or glasses of fruit juice do you usually have each day? A serve is 1 medium piece or 2 small pieces or 1 cup of diced or canned fruit pieces (put "0" if you eat less than one serve a day) □ □ number of serves of fruit each day □ □ number of glasses of fruit juice each day □ I don't eat fruit # About how many serves of vegetables do you usually eat each day? A serve is half a cup of	Absolute number of serves of fruit and vegetable consumed in a day was combined. The combined number was used for categorisation. 0-2= when 'number of serves of fruit each day'+ 'number of serves of cooked vegetables each day'+ 'number of serves of raw vegetables each day (e.g. salad)'≤2 3-4= when 'number of serves of fruit each day'+ 'number of serves of cooked vegetables each day'+ 'number of serves of raw

Variable names and categories in this study	45 and Up Study baseline questions or other data set's variables name	Recategorization/Derivation
	cooked vegetables or one cup of salad (please include potatoes and put "0" if less than one a day) □ □ number of serves of cooked vegetables each day □ □ number of serves of raw vegetables each day (e.g. salad) □ I don't eat vegetables	vegetables each day (e.g. salad)'3 to 4 5+= when 'number of serves of fruit each day'+ 'number of serves of cooked vegetables each day'+ 'number of serves of raw vegetables each day (e.g. salad)'>=5 Unknown= No information available
Location Major city Regional/Remote Unknown	# RA_NAME_2011	Geographical location of the participant was categorised as — major city, inner regional or outer regional/remote, according to the Accessibility/Remoteness Index of Australia [ARIA+] derived from postcode at recruitment. Major city, when RA_NAME_2011= 'Major Cities of Australia' Regional and remote, when RA_NAME_2011= 'Inner Regional Australia' or 'Outer Regional Australia' or 'Remote Australia' or 'Very Remote Australia' Unknown= No information available
No. of GP visits in the past 1 year	# date of service in Medicare claims data with the following MBS item numbers 3-4, 23-24, 36-37, 44, 47, 193, 195, 197, 199, 585, 594, 599, 2497-2559, 5000-5067 and 90020-90051.	Counting number of services taken in the past 1 year for an individual.
No. of referrals in the past 1 year	# date of referral in Medicare claims data	Counting number of specialist consultations and pathology tests in the past 1 year for an individual.
Charlson co-morbidity index 0 ≥ 1	# Diagnoses for the episode of care from APDC data. Principal diagnosis has 'P' suffix. diagnosis_codeP, diagnosis_code1 -diagnosis_code50 # The diagnosis code used ICD-10 algorithm.	The weighted Charlson Co-morbidity index were calculated by using the following reference: Quan, H., et al., *Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries.* American journal of epidemiology, 2011. **173**(6): p. 676-682.

Variable names and categories in this study	45 and Up Study baseline questions or other data set's variables name	Recategorization/Derivation
Rx-Risk index	# Anatomical Therapeutic Chemical (ATC) classification code (atc_code) and date on which the PBS item was supplied (date_of_supply)	The weighted Rx-Risk index was calculated by using the following reference: Pratt, N.L., et al., *The validity of the Rx-Risk comorbidity index using medicines mapped to the anatomical therapeutic chemical (ATC) classification system.* BMJ open, 2018. **8**(4): p. e021122.

L. Estimated 5-year cancer risk, risk difference and risk ratios for high and low-frequency donors for time-varying analysis.

Table 0.10 Estimated 5-year cancer risk, risk difference and risk ratios for high and low-frequency donors for time-varying analysis*.

Outcomes	Models	Risk, % (95% CI)		Risk Difference, % (95% CI)	Risk Ratio (95% CI)
		Low frequency	High frequency		
Gastrointestinal/ Colorectal	TMLE	0.8 (0.5, 1.2)	0.8 (0.5, 1.1)	0.0 (-0.5, 0.4)	0.97 (0.56, 1.66)
	SDR	0.8 (0.4, 1.1)	0.6 (0.3, 0.9)	-0.2 (-0.6, 0.3)	0.79 (0.40, 1.56)
Haematological	TMLE	0.6 (0.3, 0.9)	0.5 (0.0, 0.9)	-0.1 (-0.6, 0.4)	0.80 (0.27, 2.36)
	SDR	0.6 (0.2, 0.9)	0.6 (0.1, 1.0)	0.0 (-0.5, 0.6)	1.07 (0.42, 2.76)

*Adjusted at baseline for sex, age, haemoglobin, systolic blood pressure, diastolic blood pressure, blood group, BMI, smoking status, self-rated health, alcohol consumption, education, annual income, physical activity, daily consumption of fruits and vegetables, vitamin/mineral intake, red meat consumption, processed meat consumption, family history of cancer, cancer screening, location, no. of GP visits in the past 1 year, no. of referrals in the past 1 year as well as time-varying exposure, time-varying GP visits and time-varying referral.

M. Estimated 5-year cancer risk, risk difference and risk ratios for high and low-frequency donors for the follow-up ending on 31 December 2015.

Table 0.11 Estimated 5-year cancer risk, risk difference and risk ratios for high and low-frequency donors for the follow-up ending on 31 December 2015*.

Outcomes	Models	Risk, % (95% CI)		Risk Difference, % (95% CI)	Risk Ratio (95% CI)
		Low frequency	High frequency		
Gastrointestinal/ Colorectal	IPTW	0.7 (0.5, 0.9)	0.9 (0.6, 1.2)	0.2 (-0.2, 0.5)	1.27 (0.74, 1.80)
Haematological	IPTW	0.6 (0.5, 0.8)	0.6 (0.4, 0.8)	-0.1 (-0.3, 0.2)	0.92 (0.53, 1.30)

*Adjusted at baseline for sex, age, haemoglobin, systolic blood pressure, diastolic blood pressure, blood group, BMI, smoking status, self-rated health, alcohol consumption, education, annual income, physical activity, daily consumption of fruits and vegetables, vitamin/mineral intake, red meat consumption, processed meat consumption, family history of cancer, cancer screening, location, no. of GP visits in the past 1 year, no. of referrals in the past 1 year.

N. Estimated 5-year cancer risk, risk difference and risk ratios for high and low-frequency donors for different exposure settings.

Table 0.12 Estimated 5-year cancer risk, risk difference and risk ratios for high and low-frequency donors for different exposure settings*.

Exposure definition	Outcomes	Risk, % (95% CI)		Risk Difference, % (95% CI)	Risk Ratio (95% CI)
		Low-frequency	High-frequency		
At least 1 per every exposure year vs other	Gastrointestinal/ Colorectal	0.7 (0.4, 1.1)	0.8 (0.6, 1.0)	0.1 (-0.3, 0.5)	1.11 (0.44, 1.78)
	Haematological	1.0 (0.6, 1.4)	0.5 (0.4, 0.7)	-0.4 (-0.8, 0.0)	0.57 (0.01, 1.12)
At least 3 per every exposure year vs other	Gastrointestinal/ Colorectal	0.8 (0.6, 0.9)	0.8 (0.4, 1.2)	0.0 (-0.4, 0.4)	1.02 (0.41, 1.64)
	Haematological	0.6 (0.5, 0.7)	0.9 (0.5, 1.3)	0.3 (-0.1, 0.7)	1.44 (0.86, 2.01)

*Adjusted at baseline for sex, age, haemoglobin, systolic blood pressure, diastolic blood pressure, blood group, BMI, smoking status, self-rated health, alcohol consumption, education, annual income, physical activity, daily consumption of fruits and vegetables, vitamin/mineral intake, red meat consumption, processed meat consumption, family history of cancer, cancer screening, location, no. of GP visits in the past 1 year, no. of referrals in the past 1 year.

O. Categorisation and derivation of variables used in the cancer exposure window study.

Table 0.13 Categorisation and derivation of variables used in the gastrointestinal/colorectal and haematological cancer study from 45 and Up Study data and Medicare claims data set.

Variable names and categories in this study	**45 and Up Study baseline questions or other data set's variables name**	**Recategorization/Derivation**
Age at baseline (continuous variable)	#What is your date of birth? #What is today's date?	(today's date-date of birth)/365.25 + 730
Body mass index at recruitment (kg/m^2) <18.5 18.5 – 24.9 25-29.9 30+ Missing	#How tall are you without shoes? #About how much do you weigh?	Body mass index is calculated as-weight in kilogram/(height in meter)^2. Unknown category represents when information is lacking to calculate BMI or the calculated BMI is invalid (< 9 or 50)
Smoking Status Never Former Regular Missing	#Have you ever been a regular smoker? Yes ▼ No If No – please go to question x How old were you when you started smoking regularly? □ years old Are you a regular smoker now? Yes, No If No – how old were you when you stopped smoking regularly? □ years old	Never= 'Have you ever been a regular smoker- No', Current= 'Are you a regular smoker now?-Yes' Past= 'Have you ever been a regular smoker- Yes' but 'Are you a regular smoker now?-No' Unknown when not enough information is available to categorise in to above three categories
Self-rated health at recruitment Excellent Very good Good Fair/Poor Missing	#In general, how would you rate your: overall health? -Excellent -Very good -Good -Fair -Poor	fair and poor categories were combined as one category fair/poor. Unknown= No information provided
Alcohol consumption/day None ≤1/day >1/day Missing	#About how many alcoholic drinks do you have each week? one drink = a glass of wine, middy of beer or nip of spirits (put "0" if you do not drink, or have less than one drink each week) □ □ number of alcoholic drinks each week	None=0 or <1 drink each week, ≤1/day= number of alcoholic drinks each week is ≤7 >1/day= number of alcoholic drinks each week is >7 Unknown= No information provided
Education level No formal education School to Diploma	#What is the highest qualification you have completed? (please put a cross in the most appropriate box)	No formal education ='no school certificate or other qualifications'

Variable names and categories in this study	45 and Up Study baseline questions or other data set's variables name	Recategorization/Derivation
University Missing	-no school certificate or other qualifications -school or intermediate certificate (or equivalent) -higher school or leaving certificate (or equivalent) -trade/apprenticeship (e.g. hairdresser, chef) -certificate/diploma (e.g. child care, technician) -university degree or higher	School to diploma= 'school or intermediate certificate (or equivalent)', 'higher school or leaving certificate (or equivalent)', 'trade/apprenticeship (e.g. hairdresser, chef)', 'certificate/diploma (e.g. child care, technician)' University='university degree or higher' Unknown= no information provided
Annual household income <20k 20k-39k 40k-69k 70k+ Missing	#What is your usual yearly HOUSEHOLD income before tax, from all sources? (please include benefits, pensions, superannuation, etc) -less than $5,000 per year -$5,000-$9,999 per year -$10,000-$19,999 per year -$20,000-$29,999 per year -$30,000-$39,999 per year -$40,000-$49,999 per year -$50,000-$69,999 per year -$70,000 or more per year -I would rather not answer this question	<20k= 'less than $5,000 per year', '$5,000-$9,999 per year' , '$10,000-$19,999 per year' 20k-39k= '$20,000-$29,999 per year', '$30,000-$39,999 per year' 40k-69k= '$40,000-$49,999 per year', '$50,000-$69,999 per year' 70k+= '$70,000 or more per year' Unknown = 'I would rather not answer this question' or when no information available
Physical activity/week <1/week ≥1/week Missing	#How many TIMES did you do each of these activities LAST WEEK? (put "0" if you did not do this activity) Vigorous physical activity (that made you breathe harder or puff and pant, like jogging, cycling, aerobics, competitive tennis, but not household chores or gardening) □ □ times in the last week	<1/week= No vigorous physical activity in the last week ≥1/week= 1 or more times in the last week Unknown= No information is available
Daily serves of fruits/vegetables 0-2 3-4 5+ Missing	#About how many serves of fruit or glasses of fruit juice do you usually have each day? A serve is 1 medium piece or 2 small pieces or 1 cup of diced or canned fruit pieces (put "0" if you eat less than one serve a day) □ □ number of serves of fruit each day □ □ number of glasses of fruit juice each day □ I don't eat fruit # About how many serves of vegetables do you usually eat each	Absolute number of serves of fruit and vegetable consumed in a day was combined. The combined number was used for categorisation. 0-2= when 'number of serves of fruit each day'+ 'number of serves of cooked vegetables each day'+ 'number of serves of raw vegetables each day (e.g. salad)'≤2 3-4= when 'number of serves of fruit each day'+ 'number of serves of cooked vegetables each day'+ 'number of serves of raw

Variable names and categories in this study	45 and Up Study baseline questions or other data set's variables name	Recategorization/Derivation
	day? A serve is half a cup of cooked vegetables or one cup of salad (please include potatoes and put "0" if less than one a day) □ □ number of serves of cooked vegetables each day □ □ number of serves of raw vegetables each day (e.g. salad) □ I don't eat vegetables	vegetables each day (e.g. salad)'3 to 4 5+= when 'number of serves of fruit each day'+ 'number of serves of cooked vegetables each day'+ 'number of serves of raw vegetables each day (e.g. salad)'>=5 Missing= No information available
Taking any vitamin & mineral? No Yes Missing	# Have you taken any medications, vitamins or supplements for most of the last 4 weeks: □ multivitamins+minerals	No = Didn't take multivitamins and minerals Yes = Took multivitamins and minerals Missing = No information available
Consumption of red meat <5/week >= 5/week Missing	# About how many times each week do you eat beef, lamb or pork? □ □ number of times eaten each week.	<5/week = No. of time eaten beef, lamb or pork <5 per week >= 5/week = No. of time eaten beef, lamb or pork >= 5 per week Missing = No information available
Consumption of processed meat <3/week >= 3/week Missing	# About how many times each week do you eat processed meat (include bacon, sausages, salami, devon, burgers, etc) □ □ number of times eaten each week.	<3/week = No. of time eaten processed eat were <3 per week >= 3/week = No. of time eaten processed meat were >= 3 per week Missing = No information available
Family History of cancer No Yes	#Have your mother, father, brother(s) or sister(s) ever had: "Breast cancer, bowel cancer, lung cancer, melanoma, prostate cancer, ovarian cancer" (blood relatives only)	No= No history of any cancer in mother or father or brother or sister Yes= History of any type of cancer in mother or father or brother or sister
Cancer screening No Yes Missing	# Have you ever had a blood test ordered by your doctor to check for prostate disease? (PSA test) (for men) □ Yes □ No # Have you ever been for a breast screening mammogram? □ Yes □ No	For men, No = No PSA test or bowel cancer screening Yes = Either PSA test or bowel cancer screening For women, No = No mammogram or bowel cancer screening

Variable names and categories in this study	45 and Up Study baseline questions or other data set's variables name	Recategorization/Derivation
	# Have you ever been screened for colorectal (bowel) cancer? □ Yes □ No	Yes = Either mammogram or bowel cancer screening Missing = No information available
Location Major city Regional/Remote Unknown	# RA_NAME_2011	Geographical location of the participant was categorised as — major city, inner regional or outer regional/remote, according to the Accessibility/Remoteness Index of Australia [ARIA+] derived from postcode at recruitment. Major city, when RA_NAME_2011= 'Major Cities of Australia' Regional and remote, when RA_NAME_2011= 'Inner Regional Australia' or 'Outer Regional Australia' or 'Remote Australia' or 'Very Remote Australia' Unknown= No information available
No. of GP visits in the past 1 year	# date of service in Medicare claims data	Counting number of services taken in the past 1 year for an individual.
No. of referrals in the past 1 year	# date of referral in Medicare claims data	Counting number of diagnostic tests referrals given in the past 1 year for an individual.

P. SAS and R codes used in the study.

/*SAS codes for calculating Charlson and Rx-Risk co-morbidity index*/

```
libname apdc 'G:\edvlink_2019_original_extracted_data\NSW APDC' access=readonly;
libname rbdm 'G:\edvlink_2019_original_extracted_data\NSW RBDM Deaths' access=readonly;
libname mylib 'G:/morshad_sas';

options nofmterr;
* this adds on the Charlson CCI for individual episodes of care;
* and the elixhauser index;


data _NULL_;
        if 0 then set apdc.pr2015422_apdc_sensitive_health nobs=n;
        call symputx('nrows',n);
run;

%put nobs=&nrows;

data charlson;
        set apdc.pr2015422_apdc_sensitive_health
/*      (obs= 11000)*/
;
        if _N_ = 1 then put "Record is:";
        if mod(_N_,10000) = 0 then do;
                x = round(_N_/&nrows*100,0.1);
                put x '%';

        end;

        array dx(*) diagnosis_codeP diagnosis_code1-diagnosis_code50;
        array cc_grp(17) cc_grp_1-cc_grp_17;
        array elix(31) elix_grp_1-elix_grp_31;
        do i = 1 to 17;
                cc_grp(i)=0;
        end;
        do i = 1 to 31;
                elix(i) = 0;
        end;

        do i = 1 to dim(dx);
                if dx(i) in: ('I21','I22','I25.2') then cc_grp_1=1;
                LABEL cc_grp_1 = 'Acute myocardial infarction';

                if dx(i) in:
('I43','I50','I09.9','I11.0','I13.0','I13.2','I25.5','I42.0','I42.5','I42.6','I42.7','I42.8','I42.9','P29.0') then cc_grp_2=1;
                LABEL cc_grp_2 = 'Congestive heart failure';

                if dx(i) in:
('I71','I72','I73.1','I73.8','I73.9','I77.1','I79.0','I79.2','K55.1','K55.8','K55.9','Z95.8','Z95.9') then cc_grp_3=1;
                LABEL cc_grp_3 = 'Peripheral vascular disease';


                if dx(i) in: ('G45','G46','I6','H34.0') then cc_grp_4=1;
                LABEL cc_grp_4 = 'Cerebrovascular disease';
```

```
                    if dx(i) in: ('F00','F01','F02','F03','G30','F05.1','G31.1') then cc_grp_5=1;
                    LABEL cc_grp_5 = 'Dementia';

                                        if ('J40' <=: dx(i) <=: 'J47') or ('J60' <=: dx(i) <=: 'J67') or dx(i) in:
('I27.8','I27.9','J68.4','J70.1','J70.3')  then cc_grp_6=1;
                    LABEL cc_grp_6 = 'Chronic Pulmonary Disease';

                    if ('M05' <=: dx(i) <=: 'M06')  or ('M32' <=: dx(i) <=: 'M34') or dx(i) in:
('M31.5','M35.1','M35.3','M36.0') then cc_grp_7=1;
                    LABEL cc_grp_7 = 'Connective Tissue Disease';

                    if ('K25' <=: dx(i) <=: 'K28')   then cc_grp_8=1;
                    LABEL cc_grp_8 = 'Peptic Ulcer Disease';

                    if dx(i)="B18" or ('K73' <=: dx(i) <=: 'K74') or ('K70.0' <=: dx(i) <=: 'K70.3') or dx(i)='K70.3'
                              or dx(i)=: 'K70.9' or dx(i)=: 'K71.7' or ('K71.3' <=: dx(i) <=: 'K71.5') or dx(i)='K76.0'
                              or  ('K76.2' <=: dx(i) <=: 'K76.4') or  ('K76.8' <=: dx(i) <=: 'K76.9') or dx(i) =: 'Z94.4' then
cc_grp_9=1;
                    LABEL cc_grp_9 = 'Mild Liver  Disease';

                    if dx(i) in: ('E10.0','E10.1','E10.6','E10.8','E10.9','E11.0', 'E11.1',
                                                  'E11.6','E11.8','E11.9','E12.0','E12.1','E12.6','E12.8',
                                                  'E12.9','E13.0','E13.1','E13.6','E12.8','E13.9','E14.0',
                                                  'E14.1','E14.6','E14.8','E14.9') then cc_grp_10 = 1;
                    LABEL cc_grp_10 = 'Diabetes without complications';



                    if dx(i) in: ('E10.2','E10.3','E10.4','E10.5','E10.6','E11.2', 'E11.3',
                                                  'E11.4','E11.5','E11.7','E12.2','E12.3','E12.4','E12.5',
                                                  'E12.7','E13.2','E13.3','E13.4','E13.5','E13.7','E14.2',
                                                  'E14.3','E14.4','E14.5','E14.7') then cc_grp_11 = 1;
                    LABEL cc_grp_11 = 'Diabetes with complications';

                    if ('G81' <=: dx(i) <= 'G82') or dx(i) in: ('G04.1','G11.4','G80.1',
                              'G80.2','G83.0','G83.1','G83.2','G83.3','G83.4','G83.9')
                              then cc_grp_12 = 1;
                              LABEL CC_GRP_12 = 'Paraplegia and Hemiplegia';


                    if ('N18' <=: dx(i) <=: 'N19') or ('N05.2' <=: dx(i) <=: 'N05.7') or
                              dx(i) in: ('N25.0','I21.0','I31.','N03.2','N03.3','N03.4',
                                                            'N03.5','N03.6','N03.7','Z49.0','Z49.1','Z49.2',
                                                            'Z94.0','Z99.2') then  cc_grp_13=1;
                    LABEL cc_grp_13 = 'Renal Disease';

                    if ('C00' <=: dx(i) <=: 'C26') or ('C30' <=: dx(i) <=: 'C34') or ('C37' <=: dx(i) <=: 'C41') or dx(i)=:
'C43'
                              or  ('C45' <=: dx(i) <=: 'C58')           or  ('C60' <=: dx(i) <=: 'C76') or ('C81' <=: dx(i) <=:
'C85 ') or dx(i)=: 'C88'
                              or ('C90' <=: dx(i) <=: 'C97') then cc_grp_14=1;
                    LABEL cc_grp_14 = 'Cancer';

                    if dx(i) in: ('K70.4','K71.1','K72.1','K72.9','K76.5','K76.6',
                                                            'K76.7','I85.0','I85.9','I86.4','I98.2')
                                                            then cc_grp_15=1;
```

```
                    LABEL cc_grp_15 = 'Moderate or Severe Liver  Disease';

                    if ('C77' <=: dx(i) <=: 'C80')   then cc_grp_16=1;
                    LABEL cc_grp_16 = 'Metastatic carcinoma';

                    if ('B20' <=: dx(i) <=: 'B22') or dx(i)=:'B24'   then cc_grp_17=1;
                    LABEL cc_grp_17 = 'HIV/AIDS';

          *elixhauser;

                    if dx(i) in:
('I09.9','I11.0','I13.0','I13.2','I25.5','I42.0','I42.5','I42.6','I42.7','I42.8','I42.9','I43','I50','P29.0') then elix_grp_1=1;
                    LABEL elix_grp_1 = 'Congestive heart failure';

                    if dx(i) in: ('I44.1','I44.2','I44.3','I45.6','I45.9','I47',
                                                'I48','I49','R00.0','R00.1','R00.8','T82.1',
                                                'Z45.0','Z95.0') then elix_grp_2 = 1;
                    LABEL elix_grp_2 = 'Cardiac aryhtmia';
                    if dx(i) in: ('A52.0','I05','I06','I07','I08','I09.1',
                                                'I09.8','I34','I37','I38','I39','Q23.0',
                                                'Q23.1','Q23.2','Q23.3','Z95.2','Z95.3','Z95.4') then elix_grp_3 = 1;
                    LABEL elix_grp_3 = 'Valvular disease';

                    if dx(i) in: ('I26','I27','I28.0','I28.8','I28.9') then elix_grp_4 = 1;
                    LABEL elix_grp_4 = 'Pulmonary circulation disorders';

                    if dx(i) in: ('I70','I71','I73.1','I73.8','I73.9','I77.1',
                                                'I79.0','I79.2','K55.1','K55.8','K55.9','Z95.8','Z95.9')
                              then elix_grp_5 = 1;
                    LABEL elix_grp_5 = 'peripheral vascular disorders';

                    if dx(i) in: ('I10') then elix_grp_6 = 1;
                     LABEL elix_grp_6 = 'Hypertension uncomplicated';

                    if ('I11' <=:  dx(i)<=: ('I13')) or  dx(i) in: ('I15') then elix_grp_7 = 1;
                    LABEL elix_grp_7 = 'Hypertension complicated';


                    if ('G81' <=: dx(i) <= 'G82') or dx(i) in: ('G04.1','G11.4','G80.1',
                              'G80.2','G83.0','G83.1','G83.2','G83.3','G83.4','G83.9')
                              then elix_grp_8 = 1;
                              LABEL elix_GRP_8 = 'Paralysis';


                    if ('G10' <=: dx(i) <= 'G13') or  ('G20' <=: dx(i) <= 'G22') or dx(i) in:
('G25.4','G25.5','G31.2','G31.8','G31.9','G32')
                                        or ('G35' <=: dx(i) <= 'G37') or dx(i) in:
('G40','G41','G93.1','G93.4','R47.0','R56') then elix_grp_9=1;
                    LABEL elix_grp_9 = "Other neurological disorders";



                              if ('J40' <=: dx(i) <=: 'J47') or ('J60' <=: dx(i) <=: 'J67') or dx(i) in:
('I27.8','I27.9','J68.4','J70.1','J70.3')  then ELIX_grp_10=1;
                    LABEL ELIX_grp_10 = 'Chronic Pulmonary Disease';
```

```
if dx(i) in: ('E10.0','E10.1','E10.9','E11.0', 'E11.1',
                                    'E11.9','E12.0','E12.1',
                                    'E12.9','E13.0','E13.1','E13.9','E14.0',
                                    'E14.1','E14.9') then elix_grp_11 = 1;
LABEL elix_grp_11 = 'Diabetes without complications';



if dx(i) in: ('E10.2','E10.3','E10.4','E10.5','E10.6','E10.7','E10.8',
          'E11.2', 'E11.3', 'E11.4','E11.5','E11.6','E11.7','E11.7','E11.8',
                    'E12.2','E12.3','E12.4','E12.5','E12.6', 'E12.7','E12.8',
                    'E13.2','E13.3','E13.4','E13.5','E13.6','E13.7','E13.8',
                    'E14.2','E14.3','E14.4','E14.5','E14.6','E14.7','E14.8') then elix_grp_12 = 1;
LABEL elix_grp_12 = 'Diabetes with complications';

if ('E00' <=: dx(i) <=: 'E04') or dx(i) in: ('E89.0') then elix_grp_13 = 1;
LABEL elix_grp_13 = 'Hyperthyroidism';

if dx(i) in: ('I12.0','I13.1','N18','N19','N25.0','Z49.0','Z49.1',
                                    'Z49.2','Z94.0','Z99.2') THEN elix_grp_14 =1;
LABEL elix_grp_14 = 'Renal failure';


if dx(i) in: ('B18','I85','I86.4','I98.2','K70','K71.1','K71.3','K71.4',
                                    'K71.5','K71.7') OR ('K72' <=: DX(i) <=: 'K74')
                                OR ('K76.2' <=: DX(i) <=: 'K76.9') or dx(i) IN: ('Z94.4')
                                 THEN elix_grp_15 =1;
LABEL elix_grp_15 = 'Liver disease';



if dx(i) in ('K25.7','K25.9','K26.7','K26.9','K27.7','K27.9',
                                    'K28.7','K28.9')   then elix_grp_16=1;
LABEL elix_grp_16 = 'Peptic Ulcer Disease w/o bleeding';

if ('B20' <=: dx(i) <=: 'B22') or dx(i)=:'B24'   then elix_grp_17=1;
LABEL elix_grp_17 = 'HIV/AIDS';


if ('C81' <=:  dx(i)<=: ('C85')) or  dx(i) in: ('C88','C96','C90.0','C90.2') then elix_grp_18 = 1;
LABEL elix_grp_18 = 'Lymphoma';

if ('C77' <=: dx(i) <=: 'C80')   then elix_grp_19=1;
LABEL elix_grp_19 = 'Metastatic carcinoma';

          if ('C00' <=: dx(i) <=: 'C26') or ('C30' <=: dx(i) <=: 'C34') or ('C37' <=: dx(i) <=: 'C41') or
dx(i)=: 'C43'
          or  ('C45' <=: dx(i) <=: 'C58')          or  ('C60' <=: dx(i) <=: 'C76')
          or ('C90' <=: dx(i) <=: 'C95') or dx(i) in: ('C97') then elix_grp_20=1;
LABEL elix_grp_20 = 'Solid tumour without metastasis';

if dx(i) in: ('L94.0','L94.1','L94.3') or ('M05' <=: dx(i) <=: 'M06')  or dx(i) in: ('M08')
          or dx(i) in: ('M12.0','M12.3','M30','M31.0','M31.1','M31.2','M31.3')
          or ('M32' <=: dx(i) <=: 'M35') or dx(i) in: ('M45','M46.1','M46.8','M46.9') then
elix_grp_21=1;
LABEL elix_grp_21 = 'Rheumatoid arthritis/collagen';
```

```
                    if ('D65' <=: dx(i) <=: 'D68') or  dx(i) in: ('D69.1','D69.3','D69.4','D69.5','D69.6') then
elix_grp_22 = 1;
                    LABEL elix_grp_22 = 'Coagulopathy';


                    if dx(i) in: ('E66') then elix_grp_23 = 1;
                    LABEL elix_grp_23 = 'Obesity';

                    if ('E40' <=: dx(i) <=: 'E46') or  dx(i) in: ('R63.4','R64') then elix_grp_24 = 1;
                    LABEL elix_grp_24 = 'Weight loss';


                    if dx(i) in : ('E22.2','E86','E87') then elix_grp_25 = 1;
                    LABEL elix_grp_25 = 'Fluid/electrolyte disorders';

                    if dx(i) in: ('D50.0') then elix_grp_26 = 1;
                    LABEL elix_grp_26 = 'Blood loss anaemia';

                    if dx(i) in: ('D50.8','D50.9') or  ('D51' <=:  dx(i)<=: ('D53'))then elix_grp_27 = 1;
                    LABEL elix_grp_27 = 'Deficiency anaemia';


                    if dx(i) in: ('F10','E52','G62.1','I42.6','K29.2','K70.0','K70.3','K70.1','T51','Z50.2','Z71.4','Z72.1')
then elix_grp_28 = 1;
                    LABEL elix_grp_28 = 'Alochol abuse';

                    if ('F11' <=: dx(i) <=: 'F16') or ('F18' <=: dx(i) <=: 'F19') or dx(i) in: ('Z71.5','Z72.2') then
elix_grp_29 = 1;
                    LABEL elix_grp_29 = 'Drug abuse';

                    if dx(i) in: ('F20') or ('F22' <=: dx(i) <=: 'F29') or dx(i) in: ('F30.2','F31.2','F31.5') then
elix_grp_30 = 1;
                    LABEL elix_grp_30 = 'Psychoses';

                    if dx(i) in: ('F20.4') or ('F31.3' <=: dx(i) <=: 'F31.5') or dx(i) in: ('F32','F33','F34.1','F41.2','F43.2')
then elix_grp_31 = 1;
                    LABEL elix_grp_31 = 'Depression';
          end;

          cci= sum(of cc_grp_1-cc_grp_17);
          wgt_cci = sum(of cc_grp_1-cc_grp_10)+cc_grp_11*2+cc_grp_12*2
                                    +cc_grp_13*2+cc_grp_14*2+cc_grp_15*2+cc_grp_16*6+cc_grp_17*6;
          elix_sum = sum(of elix_grp_1-elix_grp_31);
          year_admitted = year(episode_start_date);
          drop i x;

run;

proc freq data = charlson;
          tables cc_grp_1-cc_grp_17  elix_grp_1- elix_grp_31 cc_grp_14*cc_grp_16 cci wgt_cci year_admitted
elix_sum elix_sum*cci;
run;

proc print data = charlson;
where elix_grp_1 ne cc_grp_2;
run;
```

```
proc sort;
	by ppn episode_start_date;
run;

data charlson_lon;
set mylib.charlson;
keep ppn episode_end_date cc_grp_1-cc_grp_17 elix_grp_1-elix_grp_31;
run;


%macro loopy_dates(start,end);

dm 'clear log';
dm 'output; clear;';

	%let start =%sysfunc(inputn(&start,anydtdte9.));
	%put &start;
	%let end =%sysfunc(inputn(&end,anydtdte9.));
	%let dif =%sysfunc(intck(month,&start,&end));
	%put &end;

		%do f = 0 %to &dif;

			%let datea = %sysfunc(intnx(month,&start,&f,b),date9.);
			%put &datea;
			%let date = %sysfunc(inputn(&datea,anydtdte9.));

			%let date1a = %sysfunc(intnx(month,&start,&f+1,b),date9.);
			%put &date1a;
			%let date1 = %sysfunc(inputn(&date1a,anydtdte9.));

			%let date12a = %sysfunc(intnx(month,&start,&f-12,b),date9.);
			%let date12 = %sysfunc(inputn(&date12a,anydtdte9.));

			%put &date12;



data charlson_longitudinal_ppn_1 ;
	set charlson_lon ;
	where episode_end_date >= &date12 and episode_end_date < &date;
	by ppn;
	format day date9.;
	day = &date;

	iteration_month = &f;
	retain ccgrp1-ccgrp17 elixgrp1-elixgrp31;
	array tot(17) ccgrp1-ccgrp17;
	array hosp(17) cc_grp_1-cc_grp_17;

		if first.ppn then do;
			cc_total = 0;
			do i = 1 to 17;
				tot(i)=0;
			end;
		end;
```

```
                do i = 1 to 17;
                        if hosp(i)=1 then tot(i)=1;
                end;
                if last.ppn then do;
                        cc_total=sum(of ccgrp1-ccgrp17);
                        wgt_cci_total = sum(of ccgrp1-ccgrp10)+ccgrp11*2+ccgrp12*2
                                +ccgrp13*2+ccgrp14*2+ccgrp15*2+ccgrp16*6+ccgrp17*6;
                end;

        array total(31) elixgrp1-elixgrp31;
        array hospit(31) elix_grp_1-elix_grp_31;
                if first.ppn then do;
                        elix_total = 0;
                        do i = 1 to 31;
                                total(i)=0;
                        end;
                end;
                do i = 1 to 31;
                        if hospit(i)=1 then total(i)=1;
                end;
                if last.ppn then do;
                        elix_total=sum(of elixgrp1-elixgrp31);
                end;


/* proc print data =charlson_longitudinal(obs =100);run; */

        LABEL   ccgrp1 = 'Acute myocardial infarction'
                        ccgrp2 = 'Congestive heart failure'
                        ccgrp3 = 'Peripheral vascular disease'
                        ccgrp4 = 'Cerebrovascular disease'
                        ccgrp5 = 'Dementia'
                        ccgrp6 = 'Chronic Pulmonary Disease'
                        ccgrp7 = 'Connective Tissue Disease'
                        ccgrp8 = 'Peptic Ulcer Disease'
                        ccgrp9 = 'Mild Liver  Disease'
                        ccgrp10 = 'Diabetes without complications'
                        ccgrp11 = 'Diabetes with complications'
                        CCGRP12 = 'Paraplegia and Hemiplegia'
                        ccgrp13 = 'Renal Disease'
                        ccgrp14 = 'Cancer'
                        ccgrp15 = 'Moderate or Severe Liver  Disease'
                        ccgrp16 = 'Metastatic carcinoma'
                        ccgrp17 = 'HIV/AIDS';

                        * now only output last record;

                keep ppn day iteration_month elix_total cc_total wgt_cci_total;

                if last.ppn;
/*              output charlson_longitudinal_ppn_1; */

run;
                %if &f =0 %then %do;
                data charlson_monthly_index;
                        set charlson_longitudinal_ppn_1;
                run;
```

```
                %end;
        %else %do;
                        proc append data = charlson_longitudinal_ppn_1 base = charlson_monthly_index;
                        run;
                %end;
%end;
%mend;

options mprint;
%loopy_dates(01jan2006,31mar2016);

proc sort data = charlson_monthly_index;
by ppn day;
run;


proc copy  in = work out = mylib;
select charlson_monthly_index;
run;
```

Rxrixk Index

```
/*************************************************************/
/*                SETUP ANALYSIS WORKSPACE AND DATA DIRECTORIES                */
/*************************************************************/


/*Point to the Raw Data file library on the VM */
libname fmts "G:\edvlink_2019_original_extracted_data\45ANDUP_SEEF" access=readonly;
libname pbs "G:\edvlink_2019_original_extracted_data\PBS" access=readonly;
/*Point to the Formats used on the master file*/
options fmtsearch=(fmts) nofmterr;
proc format cntlin=fmts.mbs_pbs_formats library=work; run;


proc format;
        value atc 0="Unknown"
                        1 = "Alcohol dependency"
                        2 = "Allergies"
                        3 = "Anticoagulants"
                        4 = "Antiplatelets"
                        5 = "Anxiety"
                        6=  "Arryhthmia"
                        7 = "Benign prostatic hyperplasia"
                        8 = "Bipolar disorder"
                        9 = "Chronic airways disease"
                        10 = "Congestive heart failure"
                        11 = "Dementia"
                        12 = "Depression"
                        13 = "Diabetes"
                        14 = "Epilepsy"
                        15 = "Glaucoma"
                        16 = "Gastrooesophageal reflux disease"
                        17 = "Gout"
                        18 = "Hepatitis B"
```

```
                                19 = "Hepatitis C"
                                20 = "HIV"
                                21 = "Hyperkalemia"
                                22 = "Hyperlipidaemia"
                                23 = "Hypertension"
                                24 = "Hyperthyroidism"
                                25 = "Hypothyroidism"
                                26 = "Irritable bowel syndrome"
                               27 = "Ischaemic heart disease - angina"
                               28 = "Ischaemic heart disease - hypertension"
                               29 = "Incontinence"
                               30 = "Inflammation/pain"
                               31 = "Liver failure"
                               32 = "Malignancies"
                               33 = "Malnutrition"
                               34 = "Migraine"
                               35 = "Osteoporosis/Pagets"
                               36 = "Pain"
                               37 = "pancreatic insufficiency"
                               38 = "Parkinsons disease"
                               39 = "Psoriasis"
                               40 = "Psychotic illness"
                               41 = "Pulmonary hypertension"
                               42 = "Renal disease"
                               43 = "Smoking cessation"
                               44 = "Steroid responsive disease"
                               45 = "Transplant"
                               46 ="Tuberculosis"
;

run;

data _NULL_;
        if 0 then set pbs.pbs_2004to2017 nobs=n;
        call symputx('nrows',n);
run;

%put &nrows;
data _NULL_;
        y = &nrows;
        call symputx('n',y);


run;

data Rx-Risk;
        set pbs.pbs_2004to2017 (obs= &n);
        format atc_cat atc.;
        if mod(_N_,10000) = 0 then do;
                x = round(_N_/&n*100,0.1);
                put x '%';
        end;
        drop x;

        /*

        */
```

```
/*          if rand('uniform') < 0.05 then atc_code = 'C03CA01';*/
/*          else atc_code = 'C09AA01';*/


            atc_cat = 0;
            if 'N07BB01' <=: atc_code <=: 'N07BB99' then do;
                        atc_cat = 1;
                        weight = 6;
            end;
            if ('R01AC01' <=: atc_code <=: 'R01AD60') or ('R01AD02' <=: atc_code <=: 'R06AX27') or atc_code=:
'R06AB04' then do;
                        atc_cat = 2;
                        weight = -1;
            end;
            if ('B01AA03' <=: atc_code <=: 'B01AB06') or atc_code in: ('B01AE07','B01AF01','B01AF02','B01AX05')
then do;
                        atc_cat = 3;
                        weight =1;
            end;
            if ('B01AC04' <=: atc_code <=: 'B01AC30') then do;
                        atc_cat = 4;
                        weight =1;
            end;
            if ('N05BA01' <=: atc_code <=: 'N05BA12') or atc_code =: 'N05BE01' then do;
                        atc_cat = 5;
                        weight =1;
            end;
            if atc_code =: 'C01AA05' or ('C01BA01' <=: atc_code <=: 'C01BD01') or atc_code =: 'C07AA07' then do;
                        atc_cat = 6;
                        weight =1;
            end;
            if atc_code =: 'G04CA01' <=: atc_code =: 'G04CA99' or atc_code in: ('G04CB01','G04CB02') then do;
                        atc_cat = 7;
                        weight = 0;
                        * need to add in a check that the person is male;
            end;
            if atc_code =: ('N05AN01')  then do;
                        atc_cat = 8;
                        weight = -1;
            end;
            if ('R03AC02') <=: atc_code <=: ('R03AC03') or atc_code =: 'R03DX05' then do;
                        atc_cat = 9;
                        weight = 2 ;
            end;


            if 'C03DA02' <=: atc_code <=: 'C03DA99'
                                    or (atc_code=: 'C07AB02' and pbs_item_code  in : ('8732N', '8733P', '8734Q',
'8735R'))
                                    or atc_code in: ('C07AB07', 'C07AG02', 'C07AB12', 'C03DA04')

                                    then do;
                                                atc_cat = 10;
                                                weight  = 2;
                                    end;

            if ('N06DA02' <=: atc_code <=: 'N06DA04') or atc_code =: 'N06DX01' then do;
```

```
                atc_cat= 11;
                weight  =2;
        end;
        if ('N06AA01' <=: atc_code <=:'N06AG02') or ('N06AX03' <=: atc_code <=:'N06AX011')
                or ('N06AX13' <=: atc_code <=: 'N06AX18') or  ('N06AX21' <=: atc_code <=:'N06AX26') then
do;
                atc_cat = 12;
                weight  =2;
        end;
        if ('A10AA01' <=: atc_code <=:'A10BX99') then do;
                atc_cat = 13;
                weight  =2;
        end;
        if ('N03AA01' <=: atc_code <=:'N03AX99') then do;
                atc_cat = 14;
                weight  =0;
        end;
        if ('S01EA01' <=: atc_code <=:'S01EB03') or ('S01EC03' <=: atc_code <=:'S01EX99') then do;
                atc_cat = 15;
                weight  =0;
        end;
        if ('A02BA01' <=: atc_code <=:'A02BX05') then do;
                atc_cat = 16;
                weight  =0;
        end;
        if ('M04AA01' <=: atc_code <=:'M04AC01') then do;
                atc_cat = 17;
                weight  =1;
        end;
        if atc_code in: ('J05AF08', 'J05AF10', 'J05AF11') then do;
                atc_cat = 18;
                weight = 0;
        end;
        if atc_code in: ('J05AB54', 'L03AB10', 'L03AB11', 'L03AB60', 'L03AB61', 'J05AE14')
                                or 'J05AE11' <=: atc_code <=: 'J05AE12'
                                or atc_code in: ('J05AX14', 'J05AX15', 'J05AX65','J05AB04')
        then do;
                atc_cat = 19;
                weight = 0;
        end;

        if 'J05AE01'  <=: atc_code <=: 'J05AE10' or 'J05AF12' <=: atc_code <=: 'J05AG05'
                                                or 'J05AR01' <=: atc_code <=: 'J05AR99'  or 'J05AX07' <=: atc_code
<=: 'J05AX09'
                        or atc_code =: 'J05AX12' or  'J05AF01' <=: atc_code <=: 'J05AF07' or atc_code =:
'J05AF09'
        then do;
                atc_cat = 20;
                weight  =0;
        end;


        if (atc_code =:'V03AE01') then do;
                atc_cat = 21;
                weight  =4;
        end;
        if (atc_code =:'A10BH03') or ('C10AA01' <=: atc_code <=: 'C10BX99' )then do;
```

```
                atc_cat = 22;
                weight  =-1;
        end;
        if ('C03AA01' <=: atc_code <=: 'C03BA11') or atc_code in: ('C03DB01','C03DB99','C03EA01') or
                ('C09BA02' <=: atc_code <=: 'C09BA09') or
                ('C09DA02' <=: atc_code <=: 'C09DA08') or
                ('C02AB01' <=: atc_code <=: 'C02AC05') or
                ('C02DB02' <=: atc_code <=: 'C02DB99') /*need to check*/
        then do;
                atc_cat = 23;
                weight  =-1;
        end;
        if (atc_code =:'H03BA02') or atc_code =: 'H03BB01' then do;
                atc_cat = 24;
                weight  =2;
        end;
        if  ('H03AA01' <=: atc_code <=: 'H03AA02' )then do;
                atc_cat = 25;
                weight  =0;
        end;
        if  ('A07EC01' <=: atc_code <=: 'A07EC04' ) or ('A07EA01' <=: atc_code <=: 'A07EA02' ) or
                atc_code in: ('A07EA06','L03AA33')
        then do;
                atc_cat = 26;
                weight  =0;
        end;
        if  ('C01DA02' <=: atc_code <=: 'C01DA14' ) or atc_code in:('C01DX16', 'C08EX02')
        then do;
                atc_cat = 27;
                weight  =2;
        end;



        if 'C07AA01' <=: atc_code <=: 'C07AA06'
                        or (atc_code=: 'C07AB02' and pbs_item_code not in : ('8732N', '8733P', '8734Q',
'8735R'))
                        or atc_code =: 'C07AG01'
                        or 'C08CA01' <=: atc_code <=: 'C08DB01'
                        or 'C09DB01' <=: atc_code <=: 'C09DB04'
                        or atc_code =: 'C09DX01'
                        or 'C09BB02' <=: atc_code <=: 'C09BB10'
                        or atc_code in: ('C07AB03', 'C09DX03', 'C10BX03')
                        then do;
                                atc_cat = 28;
                                weight  = -1;
                        end;


        if  ('G04BD01' <=: atc_code <=: 'G04BD99' )
        then do;
                atc_cat = 29;
                weight  =0;
        end;
        if  ('M01AB01' <=: atc_code <=: 'M01AH06' )
        then do;
                atc_cat = 30;
```

```
                weight  =-1;
        end;
        if  ('A06AD11' <=: atc_code <=: 'A07AA11' )
        then do;
                atc_cat = 31;
                weight  =3;
        end;
        if  ('L01AA01' <=: atc_code <=: 'L01XX41' )
        then do;
                atc_cat = 32;
                weight  =2;
        end;
        if  ('B05BA01' <=: atc_code <=: 'B05BA10' )
        then do;
                atc_cat = 33;
                weight  =2;
        end;
        if  ('B05BA01' <=: atc_code <=: 'B05BA10' )
        then do;
                atc_cat = 33;
                weight  =0;
        end;
        if  ('N02CA01' <=: atc_code <=: 'N02CX01' )
        then do;
                atc_cat = 34;
                weight  =-1;
        end;
        if  ('M05BA01' <=: atc_code <=: 'M05BB05' ) or atc_code in:
('M05BX03','M05BX04','G03XC01','H05AA02')
        then do;
                atc_cat = 35;
                weight  =-1;
        end;
        if ('N02AA01' <=: atc_code <=: 'N02AX02') or atc_code in: ('N02AX06','N02AX52','N02BE51')
        then do;
                atc_cat = 36;
                weight  =3;
        end;
        if atc_code in: ('A09AA02') then do;
                atc_cat = 37;
                weight = 0;
        end;
        if ('N04AA01' <=: atc_code <=: 'N04BX02')
        then do;
                atc_cat = 38;
                weight  =3;
        end;
        if ('D05AA01' <=: atc_code <=: 'D05AA99') or atc_code in: ('D05BB01','D05BB02','D05AX02','D05AX52')
                or  ('D05AC01' <=: atc_code <=: 'D05AC99')
        then do;
                atc_cat = 39;
                weight  =0;
        end;
        if ('N05AA01' <=: atc_code <=: 'N05AB02') or ('N05AB06' <=: atc_code <=: 'N05AL07') or ('N05AX07' <=:
atc_code <=: 'N05AX13')
        then do;
                atc_cat = 40;
```

```
                weight  = 6;
        end;
        *check if this meant to be either or;
        if ('C02KX01' <=: atc_code <=: 'C02KX05') and pbs_item_code in: ('9547L','9605M')  then do;
                atc_cat = 41;
                weight  = 6;
        end;
        if ('B03XA01' <=: atc_code <=: 'B03XA03') or ('A11CC01' <=: atc_code <=: 'A11CC04')
                or atc_code in  ('V03AE02','V03AE03','V03AE05')
        then do;
                atc_cat = 42;
                weight  = 6;
        end;
        if ('N07BA01' <=: atc_code <=: 'N07BA03') or atc_code in: ('N06AX12') then do;
        atc_cat = 43;
        weight = 6;
end;

if ('H02AB01' <=: atc_code <=: 'H02AB10')  then do;
        atc_code = 44;
        weight = 2;
end;
if atc_code in: ('L04AA06', 'L04AA10', 'L04AA18', 'L04AD01', 'L04AD02') then do;
        atc_cat = 45;
        weight = 2;
end;
if  'J04AC01' <=: atc_code  <=: 'J04AC51'
        or 'J04AM01' <=: atc_code  <=:'J04AM99' then do;
        atc_cat = 46;
        weight = 0;
end;


        if 'C03CA01' <=: atc_code <=: 'C03CC01' then diur = 1; else diur = 0;
        if 'C09AA01' <=: atc_code <=: 'C09AX99' then ace = 1; else ace = 0;
        if 'C09CA01' <=: atc_code <=: 'C09CX99' then arb = 1; else arb= 0;
        yr_supp = year(date_of_supply);
        yr_pres = year(date_of_prescribing);
run;

proc sort data = mylib.Rx-Risk out = risk_sorted;
        by ppn date_of_supply /* diur ace arb */;
run;


%macro loopy_dates(start,end);

dm 'clear log';
dm 'output; clear;';

        %let start =%sysfunc(inputn(&start,anydtdte9.));
        %put &start;
        %let end =%sysfunc(inputn(&end,anydtdte9.));
        %let dif =%sysfunc(intck(month,&start,&end));
        %put &end;

                %do f = 0 %to &dif;
```

```
%let datea = %sysfunc(intnx(month,&start,&f,b),date9.);
%put &datea;
%let date = %sysfunc(inputn(&datea,anydtdte9.));

%let date1a = %sysfunc(intnx(month,&start,&f+1,b),date9.);
%put &date1a;
%let date1 = %sysfunc(inputn(&date1a,anydtdte9.));

%let date12a = %sysfunc(intnx(month,&start,&f-12,b),date9.);
%let date12 = %sysfunc(inputn(&date12a,anydtdte9.));

%put &date12a;
```

```
data Rx-Risk_last_MR  ;
        set risk_sorted;
        where date_of_supply >= &date12  and date_of_supply < &date;

        by PPN;

        format day date9.;
        day = &date;

        iteration_month = &f;
        * also update the records for the chronic heart failure and hypertension records;
        retain atcgrp1-atcgrp46 wgt1-wgt46 chf hypertension diur_l ace_l arb_l ;
        array tot(46) atcgrp1-atcgrp46;
        array wgt(46) wgt1-wgt46;


        if first.ppn then do;
                        diur_l =0;
                        ace_l =0;
                        arb_l =0;
                        chf =0;
                        hypertension = 0;
        end;

        do i = 1 to 46;
                tot(i)=0;
                wgt(i) = 0;
        end;
        if diur =1 then diur_l = 1;
        if ace = 1 then ace_l = 1;
        if arb  = 1 then arb_l = 1;

/* proc print data = Rx-Risk_last; run; */
        if diur_l = 1 and (ace_l = 1 or arb_l=1) then do ;
                atc_cat = 10;
                weight = 2;
        end;

        if ((diur_l =1 or ace_l =1) or (diur_l =1 or arb_l =1)) and not ((diur_l =1 and ace_l =1) or (diur_l=1 and
arb_l =1)) then do;
```

```
                        atc_cat = 23;
                        weight = -1;

        end;


        do i = 1 to 46;
                if atc_cat = i then do;
                        tot(i)=1;
                        wgt(i) = weight;
                end;
        end;


        retain w_gt1-w_gt46;
        array wt(46) w_gt1-w_gt46;


        if first.ppn then do;
        wgtsum = 0;
        do i = 1 to 46;
                wt(i) = 0;
        end;
end;

        do i = 1 to 46;
        if tot(i) =1 then wt(i) = wgt(i);
        end;

        keep ppn day iteration_month wgtsum;
                if last.ppn then do;
                                wgtsum = sum(of w_gt1-w_gt46);
        output;
                end;

                        %if &f =0 %then %do;
                data Rx-Risk_monthly_index;
                        set Rx-Risk_last_mr;
                run;
                %end;
         %else %do;
                        proc append data = Rx-Risk_last_mr base = Rx-Risk_monthly_index;
                        run;
                %end;
%end;
%mend;

options mprint;
%loopy_dates(01jan2006,31dec2016);

        proc copy  in = work out = mylib;
        select Rx-Risk_monthly_index;
        run;
```

/* SAS Program for target trial for all cause mortality */

```
/* Load Datafiles and Code Up Basic DemoTable */
libname _45data 'G:\edvlink_2019_original_extracted_data\45ANDUP_SEEF' access=readonly;
libname ITT 'G:\morshad_sas\conference';
libname saved 'G:\morshad_sas\all_cause_mortality\New_data';
libname _neg 'G:\morshad_sas\negative control';

options nodate nocenter fmtsearch=(infect);

data itt_prep;
set saved.itt_analysis_5years;
run;

data itt_prep1;
set itt_prep;
base_age =  age+ intck('month',datentoday,baseline_trial)/12;

by ppn trial;


if first.trial then trial_age = base_age;
else trial_age+(1/12);

trial_age = round(trial_age,0.01);
run;


data itt_prep2;
merge itt_prep1(in = a) _45data._45andupdata_broad_geog (in = b keep = ppn RA_CODE_2011);
by ppn;
if a;
run;

data itt_analysis_5;
set itt_prep2;

location = .;
if RA_CODE_2011 eq 10 then location = 1; /* 1 = major city, 2 = regional, remote */
if  RA_CODE_2011 eq 11 or RA_CODE_2011 eq 12 or  RA_CODE_2011 eq 13 or RA_CODE_2011 eq 14
then location = 2;
if location eq . then location = 0; /* Missing */

/*age_trial_square = age_trial*age_trial;*/

/*if smok_stat eq 0 then delete;*/
run;

data itt_analysis_5;
set itt_analysis_5;
by ppn;
 retain over_age_base;
```

```
 if first.ppn then over_age_base = trial_age;
 run;

 data itt_analysis_6;
 set itt_analysis_5;
 by ppn trial;
  retain trial_age_base;
  if first.trial then trial_age_base = trial_age;
  run;


 proc sql;
  create table itt_analysis_5years as
  select ppn, trial,period,followup_month, intervention, cont_donor, event, died_fup, total_don_fup_year, sex,
cat_age_trial2, bmi_c, smok_stat, ratehealth_cat, alcohol_cat_day, education, income_cat,
            vigour_act, cat_fruit_veg,  cci_base, Rx-Risk_base, gp_visit_base, referral_base,trial_cci_base,
trial_Rx-Risk_base, trial_gp_visit_base, trial_referral_base, cci, Rx-Risk, gp_visit,
            referral,  end_month_itt_negcontrol, followup_month_itt_negcontrol, baseline_trial, age_trial,
            inj_fup, inj_month, diabetes, treated_for_highbp, treated_for_highch,
treated_with_aspirin, prior_high_bp, weight_excluded, location, over_age_base, trial_age_base,
trial_square, period_square

from itt_analysis_6;
quit;

proc copy in = work out = itt;
select itt_analysis_5years;
run;

/*data itt_analysis_5years;*/
/*set itt.itt_analysis_5years;*/
/*run;*/

/* ITT analysis baseline adjusted */

proc genmod data = itt_analysis_5years descending;

            class ppn intervention sex  bmi_c smok_stat ratehealth_cat alcohol_cat_day education income_cat
            vigour_act cat_fruit_veg cci_base Rx-Risk_base gp_visit_base referral_base trial_cci_base
trial_Rx-Risk_base trial_gp_visit_base trial_referral_base location ;


            model event = intervention sex over_age_base trial_age_base bmi_c smok_stat ratehealth_cat
alcohol_cat_day education income_cat
            vigour_act cat_fruit_veg  period period_square trial trial_square cci_base Rx-Risk_base gp_visit_base
referral_base trial_cci_base
trial_Rx-Risk_base trial_gp_visit_base trial_referral_base location
            /dist = binomial link=logit;

            repeated subject = ppn/type = ind;
```

```
        title " ITT adjusted with baseline covariates";

        estimate "Donor" intervention -1 1 /exp; /* ref is 1(non donor)*/

        store out = itt_model;
        run;

/* ITT analysis age-sex adjusted */

proc genmod data = itt_analysis_5years descending;

        class ppn intervention sex  ;


        model event = intervention sex over_age_base trial_age_base period period_square trial trial_square
        /dist = binomial link=logit;
        repeated subject = ppn/type = ind;

        title " ITT adjusted with age-sex covariates";

        estimate "Donor" intervention -1 1 /exp; /* ref is 1(non donor)*/

/*      store out = itt_model;*/
        run;


/* ITT analysis unadjusted */

proc genmod data = itt_analysis_5years descending;

        class ppn intervention  ;


        model event = intervention period period_square trial trial_square
        /dist = binomial link=logit;
        repeated subject = ppn/type = ind;

        title " ITT adjusted with age-sex covariates";

        estimate "Donor" intervention -1 1 /exp; /* ref is 1(non donor)*/

/*      store out = itt_model;*/
        run;


        /* Negative control analysis */
        /* Negative control analysis */
        /* Negative control analysis */
        /* Negative control analysis */


data neg_analysis_60trials;
```

```
set _neg.neg_analysis_60years;
run;

/* ITT analysis baseline adjusted */

proc genmod data = neg_analysis_60trials descending;

	class ppn intervention sex  bmi_c smok_stat ratehealth_cat alcohol_cat_day education income_cat
	vigour_act cat_fruit_veg cci_base Rx-Risk_base gp_visit_base referral_base trial_cci_base
trial_Rx-Risk_base trial_gp_visit_base trial_referral_base location ;


	model inj_event = intervention sex over_age_base trial_age_base bmi_c smok_stat ratehealth_cat
alcohol_cat_day education income_cat
	vigour_act cat_fruit_veg  period period_square trial trial_square cci_base Rx-Risk_base gp_visit_base
referral_base trial_cci_base
trial_Rx-Risk_base trial_gp_visit_base trial_referral_base location
	/dist = binomial link=logit;

	repeated subject = ppn/type = ind;

	title " ITT adjusted with baseline covariates";

	estimate "Donor" intervention -1 1 /exp; /* ref is 1(non donor)*/

/*	store out = itt_model;*/
	run;

/* ITT analysis age-sex adjusted */

proc genmod data = neg_analysis_60trials descending;

	class ppn intervention sex  ;


	model inj_event = intervention sex over_age_base trial_age_base period period_square trial
trial_square
	/dist = binomial link=logit;
	repeated subject = ppn/type = ind;

	title " ITT adjusted with age-sex covariates";

	estimate "Donor" intervention -1 1 /exp; /* ref is 1(non donor)*/

/*	store out = itt_model;*/
	run;


/* ITT analysis unadjusted */

proc genmod data = neg_analysis_60trials descending;
```

```
class ppn intervention  ;


model inj_event = intervention period period_square trial trial_square
/dist = binomial link=logit;
repeated subject = ppn/type = ind;

title " ITT adjusted with age-sex covariates";

estimate "Donor" intervention -1 1 /exp; /* ref is 1(non donor)*/

/*      store out = itt_model;*/
run;
```

R codes for Exposure window all-cause mortality Analysis

```
library(dplyr)
require(haven)
require(foreign)
library(splitstackshape)
library(survminer)

# work_data <- read.spss("cox_model_1.sav", to.data.frame =TRUE,stringsAsFactors=FALSE)
# head(work_data)

work_data <- read_sas("final_analysis_mortality.sas7bdat",catalog_file = "formats.sas7bcat")

# Estimating IP weights

# First calculate denominator model

func_ipweights <- function(data){

p_denom <-glm(interven_0 ~

        as.factor(sex)+
        as.factor(bmi_c)+
        as.factor(smok_stat)+
        as.factor(ratehealth_cat)+
        as.factor(alcohol_cat_day)+
        as.factor(education)+
        as.factor(income_cat)+
        as.factor(vigour_act)+
        as.factor(cat_fruit_veg)+
        as.factor(location)+
        as.factor(cci_0)+
        as.factor(Rx-Risk_0)+
        as.factor(cat_age_base)+
        as.factor(blood_grp)+
        as.factor(tot_don_bef_exp)+
        #age_baseline+I(age_baseline^2)+
        tot_visit_0 + I(tot_visit_0^2)+
        tot_referral_0 + I(tot_referral_0^2)+
        mean_hb +I(mean_hb^2)+
```

```
              mean_systolic +I(mean_systolic^2)+
              mean_diastolic + I(mean_diastolic^2),
            family = binomial(), data = data

                )

  # Now calculate numerator model

p_num <- glm(interven_0 ~
              1,
            family = binomial(), data = data
            )

# Compute predicted probabilities

  data$p_den_exp <- predict(p_denom,data,type="response")
  data$p_num_exp <- predict(p_num,data,type="response")


# Calculating the weights

  data$sw <- ifelse(data$interven_0 ==1,
                    data$p_num_exp/data$p_den_exp,
                    (1-data$p_num_exp)/(1-data$p_den_exp))
  data <- data %>%
    mutate(sw = ifelse(sw >= quantile(sw,0.99),quantile(sw,0.99),sw)) %>%
    select(PPN, sw)

  ipweights <<- data
}

  func_ipweights(work_data)
  #

  # # Expand rows per person per year
  #
  # write.foreign(work_data,"H:/work_cox_data.txt", "H:/work_cox_data.sas",package="SAS")
  #
  # # again read from sas
  #
  #
  # work_data_expan <- read.spss("work_data_expanded.sav", to.data.frame =TRUE,stringsAsFactors=FALSE)

  # Expand data for the main model

  data_expanded <- work_data %>%
                merge(ipweights, by = "PPN") %>%
                  mutate(survtime = f_time+1) %>%
                expandRows("survtime", drop=F)  %>%
                mutate(time = sequence(rle(PPN)$lengths)-1) %>%
                mutate(event_case = ifelse (time ==f_time & event_7 ==1,1,0)) %>%
                mutate(time_sq = time^2)

  # save.image(file = "coxregression")
  # load(file="coxregression")

  # Model for hazard ratio for marginal structural model

  ipw_model_hazard <- glm(event_case==1  ~ interven_0+ time,
                      family = binomial(link = "logit"), weight = sw, data = data_expanded)
```

```
# Model to predict ip weighted survival curve
# Model to predict ip weighted survival curve


ipw_model <- glm(event_case==0  ~ interven_0+time+time_sq
                 +I(interven_0*time)+I(interven_0*time_sq),
                 family = binomial, weight = sw, data = data_expanded)

summary(ipw_model)

# create data set with all time point under each treatment level

ipw_exp0 <- data.frame(cbind(seq(0,7),0,(seq(0,7))^2))
ipw_exp1 <- data.frame(cbind(seq(0,7),1,(seq(0,7))^2))


colnames(ipw_exp0) <- c("time", "interven_0","time_sq")

colnames(ipw_exp1) <- c("time", "interven_0","time_sq")

# estimating survival for each person year

ipw_exp0$p_noevent0 <- predict(ipw_model, ipw_exp0,type = "response")
ipw_exp1$p_noevent1 <- predict(ipw_model, ipw_exp1,type = "response")


# compute cumulative survival

ipw_exp0$surv0 <- cumprod(ipw_exp0$p_noevent0)
ipw_exp1$surv1 <- cumprod(ipw_exp1$p_noevent1)

# Giving 1 probability for first time point

ipw_exp0$surv0 <- ifelse( ipw_exp0$time==0,1,ipw_exp0$surv0)
ipw_exp1$surv1 <- ifelse( ipw_exp1$time==0,1,ipw_exp1$surv1)
# merge both data together

ip_graph <- merge (ipw_exp0, ipw_exp1, by = c("time","time_sq"))


# create the differenc in survival

ip_graph$survdiff <- ip_graph$surv1- ip_graph$surv0

ip_graph <- ip_graph %>%
  arrange(time)


surv0 <- ip_graph$surv0[8]
surv1 <- ip_graph$surv1[8]

Y_0 <- 1-surv0
Y_1 <- 1-surv1
risk_diff <- Y_1 -Y_0

risk_ratio <- (Y_1/Y_0)
```

```
# plot survival curves

plot_weighted <-ggplot(mortality_surv_curve, aes(x=time))+
  geom_line(aes(y=surv0,color = "Lower-frequency" , linetype = "Lower-frequency"),linewidth = 0.8)+
  geom_line(aes(y=surv1,color = "Higher-frequency",  linetype = "Higher-frequency"), linewidth =0.8)+

  # xlab("years")+
  scale_x_continuous(limits=c(0,7), breaks = seq(0,7,1))+
  scale_y_continuous(limits=c(0.98,1), breaks = seq(0.98,1,0.01))+
  # ylab("Survival probability")+
  labs (x = "Years",
        y = "Survival probability",
        color = "", linetype = "")+



  scale_linetype_manual(values = c( "Lower-frequency"="dotted","Higher-frequency" = "solid"),
                        guide = guide_legend(reverse = TRUE)
                       )+
  scale_color_manual(values = c(  "Lower-frequency"= "red","Higher-frequency" = "#00BFC4"),
                     guide = guide_legend(reverse = TRUE)
  )+
  theme_survminer(font.x=15, font.y = 15) +
   theme(legend.position = c(0.2,0.3), legend.text = element_text(size=13))



## calculate bootstraping confidence intervals

numboot <-500

result <-NULL

set.seed(100)

      # set data frame where sample will be taken.
        # Outcome and covariates should be here

dat_boot <- work_data

for (z in 1:numboot){

index <- sample(1:nrow(dat_boot),nrow(dat_boot),replace = T)
boot_dat <-dat_boot[index,]

# Create ip weights
func_ipweights(boot_dat)

# merge in weights anc create person year data

boot_dat1 <- boot_dat %>%
  merge(ipweights, by = "PPN") %>%
  mutate(survtime = f_time + 1) %>%
  expandRows("survtime", drop = F) %>%
  mutate(time = sequence(rle(PPN)$lengths) - 1) %>%
  mutate(event_case = ifelse (time == f_time &
                              event_7 == 1, 1, 0)) %>%
  mutate(time_sq = time ^ 2)
```

```
# fit glm

ipw_model_boot <- glm(event_case==0  ~ interven_0+time+time_sq
              +I(interven_0*time)+I(interven_0*time_sq),
              family = binomial, weight =sw , data = boot_dat1)

# create data set with all time point under each treatment level

ipw_exp0_boot <- data.frame(cbind(seq(0,7),0,(seq(0,7))^2))
ipw_exp1_boot <- data.frame(cbind(seq(0,7),1,(seq(0,7))^2))


colnames(ipw_exp0_boot) <- c("time", "interven_0","time_sq")

colnames(ipw_exp1_boot) <- c("time", "interven_0","time_sq")


# estimating survival for each person year

ipw_exp0_boot$p_noevent0 <- predict(ipw_model_boot, ipw_exp0_boot,type = "response")
ipw_exp1_boot$p_noevent1 <- predict(ipw_model_boot, ipw_exp1_boot,type = "response")

# compute cumulative survival

ipw_exp0_boot$surv0 <- cumprod(ipw_exp0_boot$p_noevent0)
ipw_exp1_boot$surv1 <- cumprod(ipw_exp1_boot$p_noevent1)


# Add both data together
ipw_boot <- merge(ipw_exp0_boot,ipw_exp1_boot, by = c("time","time_sq"))

ipw_boot <- ipw_boot %>%
  filter(time==7) %>%
  mutate(risk0 = 1-surv0) %>%
  mutate(risk1 = 1-surv1) %>%
  mutate(riskdiff = risk1-risk0) %>%
  mutate(logriskratio = log(risk1/risk0)) %>%
  select(risk0, risk1,riskdiff, logriskratio)

result <- rbind(result, cbind (ipw_boot$risk0, ipw_boot$risk1, ipw_boot$riskdiff, ipw_boot$logriskratio))

}

#### Create estimates and CIS ###

result_sd <- apply(result,2,sd)



lclY_0 <- Y_0 -1.96*result_sd[1]
uclY_0 <- Y_0 +1.96*result_sd[1]

lclY_1 <- Y_1-1.96*result_sd[2]
uclY_1 <- Y_1+1.96*result_sd[2]

lcldiff <- risk_diff-1.96*result_sd[3]

ucldiff <- risk_diff+1.96*result_sd[3]
```

```
lclratio <- exp(log(risk_ratio))-1.96*result_sd[4]

uclratio <- exp(log(risk_ratio))+1.96*result_sd[4]

outcome <- "Death"

final_result_1 <- cbind("Y_1 = 1",
                  paste0(format(round(Y_1*100,1),nsmall = 1),
                       "(",format(round(lclY_1*100,1),nsmall = 1), ",",
                       format(round(uclY_1*100,1),nsmall = 1),")"))

final_result_2 <- cbind( "Y_0 = 1",
                    paste0(format(round(Y_0*100,1),nsmall = 1),
                         "(",format(round(lclY_0*100,1),nsmall = 1), ",",
                         format(round(uclY_0*100,1),nsmall = 1),")"))


final_result_3 <- cbind( "RD",
                     paste0(format(round(risk_diff*100,1),nsmall = 1),
                          "(",format(round(lcldiff*100,1),nsmall = 1), ",",
                          format(round(ucldiff*100,1),nsmall = 1),")"))



final_result_4 <- cbind( "RR",
                     paste0(format(round(risk_ratio,2),nsmall = 2),
                          "(",format(round(lclratio,2),nsmall = 2), ",",
                          format(round(uclratio,2),nsmall = 2),")"))

final_result <- rbind(final_result_1,final_result_2,final_result_3,final_result_4)



# Drawing kaplan meyer survival curve

install.packages("survminer", repos = "http://cran2.sure.local/")
install.packages("survival", repos = "http://cran2.sure.local/")
install.packages("splitstackshape", repos = "http://cran2.sure.local/")
library(splitstackshape)
library(survminer)

work_data<-work_data %>%
       mutate(f_time_2 = ifelse(event==1,f_time-1,f_time))



fit <- survfit(Surv(f_time,event)~interven_0,data=work_data)



plot_unweighted<- ggsurvplot(fit,
          data = work_data,
          ylim = c(0.98,1),
          xlim = c(0,7),
          risk.table =FALSE,

          break.time.by = 1,
          break.y.by = 0.01,
           palette = c("red","#00BFC4"),
          surv.geom = geom_line,
```

```
    linetype = c("dotted","solid"),
    ggtheme = theme_survminer(font.x=15, font.y = 15),
    censor=FALSE,
    legend = c(0.2,0.3),
    legend.title = "",
    legend.labs = c("Lower-frequency","Higher-frequency"),
    xlab = "Years",
  size = 1,
    font.legend = list(size = 13)
)
```

R code for gastrointestinal, colorectal and haematological cancer analysis#########

```
library(dplyr)
require(haven)
require(foreign)
library(splitstackshape)
library(survminer)
library(foreign)
library(ipw)

work_data_2 <- read_sas("cancer_analysis_fix_period.sas7bdat",catalog_file = "formats.sas7bcat")
#work_data_1 <- read_sas("test_male.sas7bdat",catalog_file = "my_format.sas7bcat")
head(work_data,5)

# work_data_male <- work_data %>%
#   filter(sex == 1 ) %>%
#   select(cancer_group!="Prostate (ICD-O-3 C61)")
#
# work_data_female <- work_data %>%
#   filter(sex == 2)



# Estimating IP weights

# First calculate denominator model

set.seed(1000)

func_ipweights <- function(data){

  p_denom <-glm(interven_4_yr ~

              as.factor(sex)+
              as.factor(bmi_c)+
              as.factor(cat_age_base)+
              # as.factor(bowel_his)+
              # as.factor(other_can_his)+
              # as.factor(bowel_screen)+
              # as.factor(other_can_screen)+
              as.factor(famcanhis)+
              as.factor(can_screen)+

              as.factor(smok_stat)+
              as.factor(ratehealth_cat)+
```

```
            as.factor(alcohol_cat_day)+
            as.factor(education)+
            as.factor(income_cat)+
            as.factor(vigour_act)+
            as.factor(cat_fruit_veg)+
            as.factor(location)+
            as.factor(blood_grp)+
            as.factor(redmeat_c)+

            as.factor(process_meat)+

            as.factor(vit_mineral)+

            # tot_don_3year_prior_exp+
            #
            # I(tot_don_3year_prior_exp^2)+

            # as.factor(cci_0)+
            # as.factor(Rx-Risk_0)+
            #
            tot_visit_0 +
            I(tot_visit_0^2)+
            tot_referral_0 +
            I(tot_referral_0^2)+
            mean_hb +I(mean_hb^2)+
            mean_systolic +I(mean_systolic^2)+
            mean_diastolic + I(mean_diastolic^2),

          family = binomial(), data = data

  )


  # Now calculate numerator model

  p_num <- glm(interven_4_yr ~
               1,
             family = binomial(), data = data
  )

  # Compute predicted probabilities

  data$p_den_exp <- predict(p_denom,data,type="response")
  data$p_num_exp <- predict(p_num,data,type="response")


  # Calculating the weights

  data$sw <- ifelse(data$interven_4_yr ==1,
                    data$p_num_exp/data$p_den_exp,
                    (1-data$p_num_exp)/(1-data$p_den_exp))
  data <- data %>%
    mutate(sw = ifelse(sw >= quantile(sw,0.99),quantile(sw,0.99),sw)) %>%
    select(PPN, sw)

  ipweights <<- data
}

func_ipweights(work_data_2)
#
```

```
# # Expand rows per person per year
#
# write.foreign(work_data,"H:/work_cox_data.txt", "H:/work_cox_data.sas",package="SAS")
#
# # again read from sas
#
#
# work_data_expan <- read.spss("work_data_expanded.sav", to.data.frame =TRUE,stringsAsFactors=FALSE)

# Expand data for the main model



data_expanded <- work_data_2 %>%
  merge(ipweights, by = "PPN") %>%
  mutate(survtime = followup_month_fixed+1) %>%
  expandRows("survtime", drop=F)  %>%
  mutate(time = sequence(rle(PPN)$lengths)-1) %>%
  mutate(event_case = ifelse (time ==followup_month_fixed & gastro_can ==1,1,0)) %>%
  mutate(time_sq = time^2)


# Model for hazard ratio for marginal structural model

# ipw_model_hazard <- glm(event_case==1  ~ as.factor(interven_2_yr)+ time+ time_sq,
#                         family = binomial(link = "logit"),weight = sw, data = data_expanded)
# summary(ipw_model_hazard)


# Model to predict ip weighted survival curve
# Model to predict ip weighted survival curve

# run outcome model for all agecategories

# data_expanded_1 <- data_expanded %>%
#                 filter(cat_age_base ==4)

ipw_model <- glm(event_case==0  ~ interven_4_yr+time+time_sq
                 +I(interven_4_yr*time)+I(interven_4_yr*time_sq),
                 family = binomial(),weight =sw,  data = data_expanded)

summary(ipw_model)



# create data set with all time point under each treatment level

ipw_exp0 <- data.frame(cbind(seq(0,59),0,(seq(0,59))^2))
ipw_exp1 <- data.frame(cbind(seq(0,59),1,(seq(0,59))^2))
#
# ipw_exp0 <- data.frame(cbind(seq(0,83),0,(seq(0,83))^2))
# ipw_exp1 <- data.frame(cbind(seq(0,83),1,(seq(0,83))^2))

colnames(ipw_exp0) <- c("time", "interven_4_yr","time_sq")

colnames(ipw_exp1) <- c("time", "interven_4_yr","time_sq")


# estimating survival for each person year
```

```
ipw_exp0$p_noevent0 <- predict(ipw_model, ipw_exp0,type = "response")
ipw_exp1$p_noevent1 <- predict(ipw_model, ipw_exp1,type = "response")


# compute cumulative survival

ipw_exp0$surv0 <- cumprod(ipw_exp0$p_noevent0)
ipw_exp1$surv1 <- cumprod(ipw_exp1$p_noevent1)

# Giving 1 probability for first time point

# ipw_exp0$surv0 <- ifelse( ipw_exp0$time==0,1,ipw_exp0$surv0)
# ipw_exp1$surv1 <- ifelse( ipw_exp1$time==0,1,ipw_exp1$surv1)
# merge both data together

ip_graph <- merge (ipw_exp0, ipw_exp1, by = c("time","time_sq"))


# create the differenc in survival

ip_graph$survdiff <- ip_graph$surv1- ip_graph$surv0

ip_graph <- ip_graph %>%
  arrange(time)


surv0 <- ip_graph$surv0[60]
surv1 <- ip_graph$surv1[60]

Y_0 <- 1-surv0
Y_1 <- 1-surv1
risk_diff <- Y_1 -Y_0

risk_ratio <- (Y_1/Y_0); risk_ratio


# plot survival curves

ggplot(ip_graph, aes(x=time))+
  geom_line(aes(y=surv0,color = "Less than 2" , linetype = "Less than 2"),size = 0.8)+
  geom_line(aes(y=surv1,color = "At least 2",  linetype = "At least 2"), size =0.8)+

  # xlab("years")+
  scale_x_continuous(limits=c(0,60), breaks = seq(0,60,10))+
  scale_y_continuous(limits=c(0.98,1), breaks = seq(0.98,1,0.01))+
  # ylab("Survival probability")+
  labs (x = "Years",
        y = "Survival probability",
        color = "", linetype = "")+



  scale_linetype_manual(values = c( "Less than 2"="dotted","At least 2" = "solid"),
                        guide = guide_legend(reverse = TRUE)
  )+
  scale_color_manual(values = c(  "Less than 2"= "red","At least 2" = "#00BFC4"),
                     guide = guide_legend(reverse = TRUE)
  )+
```

```
  theme_survminer()+
  theme(legend.position = c(0.2,0.3))



## calculate bootstraping confidence intervals

numboot <-1000

result <-NULL

set.seed(1000)

# set data frame where sample will be taken.
# Outcome and covariates should be here

dat_boot <- work_data_2

for (z in 1:numboot){

  index <- sample(1:nrow(dat_boot),nrow(dat_boot),replace = T)
  boot_dat <-dat_boot[index,]

  # Create ip weights
  func_ipweights(boot_dat)

  # merge in weights anc create person year data

  boot_dat1 <- boot_dat %>%
    merge(ipweights, by = "PPN") %>%
    mutate(survtime = followup_month_fixed + 1) %>%
    expandRows("survtime", drop = F) %>%
    mutate(time = sequence(rle(PPN)$lengths)-1) %>%
    mutate(event_case = ifelse (time ==followup_month_fixed & gastro_can ==1,1,0)) %>%
    mutate(time_sq = time ^ 2)


  # fit glm

  ipw_model_boot <- glm(event_case==0  ~ interven_4_yr+time+time_sq
                        +I(interven_4_yr*time)+I(interven_4_yr*time_sq),
                        family = binomial, weight =sw , data = boot_dat1)

  summary(ipw_model_boot)
  # create data set with all time point under each treatment level

  ipw_exp0_boot <- data.frame(cbind(seq(0,59),0,(seq(0,59))^2))
  ipw_exp1_boot <- data.frame(cbind(seq(0,59),1,(seq(0,59))^2))


  colnames(ipw_exp0_boot) <- c("time", "interven_4_yr","time_sq")

  colnames(ipw_exp1_boot) <- c("time", "interven_4_yr","time_sq")


  # estimating survival for each person year

  ipw_exp0_boot$p_noevent0 <- predict(ipw_model_boot, ipw_exp0_boot,type = "response")
  ipw_exp1_boot$p_noevent1 <- predict(ipw_model_boot, ipw_exp1_boot,type = "response")
```

```
  # compute cumulative survival

  ipw_exp0_boot$surv0 <- cumprod(ipw_exp0_boot$p_noevent0)
  ipw_exp1_boot$surv1 <- cumprod(ipw_exp1_boot$p_noevent1)


  # Add both data together
  ipw_boot <- merge(ipw_exp0_boot,ipw_exp1_boot, by = c("time","time_sq"))

  ipw_boot <- ipw_boot %>%
    filter(time==59) %>%
    mutate(risk0 = 1-surv0) %>%
    mutate(risk1 = 1-surv1) %>%
    mutate(riskdiff = risk1-risk0) %>%
    mutate(riskratio = (risk1/risk0)) %>%
    select(risk0, risk1,riskdiff, riskratio)

  result <- rbind(result, cbind (ipw_boot$risk0, ipw_boot$risk1, ipw_boot$riskdiff, ipw_boot$riskratio))

}

#### Create estimates and CIS ###

result_sd <- apply(result,2,sd)



lclY_0 <- Y_0 -1.96*result_sd[1]
uclY_0 <- Y_0 +1.96*result_sd[1]

lclY_1 <- Y_1-1.96*result_sd[2]
uclY_1 <- Y_1+1.96*result_sd[2]

lcldiff <- risk_diff-1.96*result_sd[3]

ucldiff <- risk_diff+1.96*result_sd[3]

lclratio <- risk_ratio-1.96*result_sd[4]

uclratio <- risk_ratio+1.96*result_sd[4]

outcome <- "Gastro Cancer"

final_result_1 <- cbind("Y_1 = 1",
                    paste0(format(round(Y_1*100,1),nsmall = 1),
                          "(",format(round(lclY_1*100,1),nsmall = 1), ",",
                          format(round(uclY_1*100,1),nsmall = 1),")"))

final_result_2 <- cbind( "Y_0 = 1",
                     paste0(format(round(Y_0*100,1),nsmall = 1),
                           "(",format(round(lclY_0*100,1),nsmall = 1), ",",
                           format(round(uclY_0*100,1),nsmall = 1),")"))


final_result_3 <- cbind( "RD",
                     paste0(format(round(risk_diff*100,1),nsmall = 1),
                           "(",format(round(lcldiff*100,1),nsmall = 1), ",",
                           format(round(ucldiff*100,1),nsmall = 1),")"))
```

```
final_result_4 <- cbind( "RR",
                         paste0(format(round(risk_ratio,2),nsmall = 2),
                              "(",format(round(lclratio,2),nsmall = 2), ",",
                              format(round(uclratio,2),nsmall = 2),")"))

final_result <- rbind(final_result_1,final_result_2,final_result_3,final_result_4)
```