Dissertation

On the use of balancing scores in cohort and matched case-control studies

submitted by

Dipl.-Ing.
Regina Riedl

for the Academic Degree of

Doctor of Medical Science

(Dr. scient. med.)

at the

Medical University of Graz

Institute for Medical Informatics, Statistics and Documentation

under the Supervision of

Univ.-Prof. Dipl.-Ing. Dr. Andrea Berghold and

Dipl.-Ing. Ruth Pfeiffer, PhD.

July 2014

Dissertation committee

Univ.-Prof. Dipl.-Ing. Dr. Andrea Berghold
Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz

Dipl.-Ing. Ruth Pfeiffer, PhD, Senior Investigator
Biostatistics Branch, Division of Cancer Epidemiology and Genetics,
National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA

Ao.Univ.-Prof. Dr. Willibald Stronegger
Institute of Social Medicine and Epidemiology
Medical University of Graz

Eidesstattliche Erklärung

Ich erkläre ehrenwörtlich, dass ich die vorliegende Arbeit selbständig angefertigt und abgefasst, und jene Personen und Institutionen, die am Zustandekommen der Forschungsdaten beteiligt waren, namentlich genannt habe. Andere als die angegebenen Quellen habe ich nicht verwendet und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen habe ich als solche kenntlich gemacht. Die Arbeit an der Dissertation und daraus entstandener Publikationen wurde gemäß den Regeln der 'Good Scientific Practice' durchgeführt.

Declaration

I hereby declare that this thesis is my own original work and that I have fully acknowledged by name all of those individuals and organisations that have contributed to the research for this thesis. Due acknowledgment has been made in the text to all other material used. Throughout this thesis and in all related publications I followed the guidelines of 'Good Scientific Practice'.

Graz, am Unterschrift

Danksagung

Ich möchte mich ganz herzlich bei meinen beiden Betreuerinnen Univ.-Prof. DI Dr. Andrea Berghold und DI Dr. Ruth Pfeiffer bedanken für die freundliche und engagierte Betreuung beim Erstellen dieser Arbeit.

Ebenso möchte ich mich bei meinen KollegInnen am Institut für medizinische Informatik, Statistik und Dokumentation und am National Cancer Institute bedanken, die mir immer mit Ratschlägen zur Seite gestanden sind.

Ein besonderer Dank gebührt auch meinen Freunden und meiner Familie für die moralische Unterstützung während meines Dissertationsstudiums.

Abstract

In observational studies the presence of many confounders, defined as variables associated with both, exposures and outcome may induce a bias in the estimates of association if not properly accounted for. However, this is challenging when the number of confounders is large and the outcome is rare. In such a case balancing scores like the propensity score (PS) or the disease risk score (DRS) have been applied to address confounding in design and analysis of observational studies. However, the performance of estimates of conditional exposure effects based on score adjusted regression models from cohort and case-control data has only been studied in limited simulations, and there is lack of theoretical results.

This thesis has two parts. First, we give an overview of the use of balancing scores in the design and analysis of observational studies. We then derive the asymptotic bias in estimates of conditional exposure effects when scores, including the PS or the DRS are used as adjustment variables in multivariable regression models for the analysis of cohort data, and as matching variables in case-control studies. The magnitude of the bias in estimates of conditional exposure effects is then computed numerically and in simulations for several settings. In cohort analysis, adjusting for the PS or the DRS instead of the confounders itself yields unbiased estimates for linear exposure effects. For non-linear exposure effects in both study designs, unbiased estimates are given for the PS only under the null hypothesis of no association. The estimates obtained from matching cases and controls on the DRS and analyzing the data using conditional logistic regression are unbiased. All other settings result in bias. The magnitude of this bias as a function of various study parameters is assessed in detail for Poisson and logistic regression for cohorts, and for matched case-control studies when conditional or unconditional logistic regression adjusted for the matching scores are used to analyze the data. We additionally investigate the bias in simulations with more complex confounder correlations based on a real dataset and observed similar results as for the analytical calculations.

In the second part of the thesis we study the impact of having received a blood transfusion on cancer risk in the U.S. elderly, based on a sample from the SEER (Surveillance, Epidemiology and End Results)-Medicare database. The risk is investigated for cancer overall and for specific sites and subtypes. Various summary score methods, including matching and adjustments are applied to account for possible confounders. We observed an increased cancer risk shortly after the receipt of a blood transfusion for cancer overall and for cancers of the stomach, colon, liver, kidney, renal pelvis and ureter, myeloma, leukemia and for Hodgkin lymphoma and non-Hodgkin lymphoma. Risk was not elevated for transfusions received long before cancer diagnosis with the exception of liver cancer. However, after careful adjustment for confounding, no long term effect of transfusion on risk was observed. Our findings suggest that transfusions in most patients that were later diagnosed with cancer may be prompted by an undiagnosed cancer or a precursor to cancer.

In summary, this thesis contributes to the understanding of results from score adjusted regression models for cohort and from case-control data where scores were used to match cases and controls. Our work also contributes to the understanding of the relationship between blood transfusions and cancer risk.

Zusammenfassung

In Beobachtungsstudien gestaltet sich die Schätzung von Zusammenhängen zwischen Expositionen und einem bestimmten Endpunkt oft als schwierig, wenn die Anzahl der möglichen Störfaktoren groß und der Endpunkt selten ist. Balancing Scores (z.B. der Propensity Score, PS, und der Disease Risk Score, DRS) werden bei der Planung von Studien und bei der statistischen Analyse verwendet um für viele Störfaktoren gleichzeitig zu kontrollieren. Allerdings wurden die statistischen Eigenschaften von Schätzern von bedingten Behandlungseffekten, die von Score adjustierten Modellen stammen, bis jetzt hauptsächlich in Simulationsstudien für Kohortenstudien und in nur einer Simulationsstudie für Fall-Kontroll-Studien untersucht, theoretische Ergebnisse fehlen.

Diese Arbeit besteht aus zwei Teilen. Der erste Teil beschäftigt sich mit dem asymptotischen Bias der in Assoziationsparametern entsteht, wenn anstatt der Störfaktoren selber der PS oder der DRS als Adjustierungsvariable in Kohortenstudien oder als Matching Variable in Fall-Kontroll-Studien herangezogen wird. Lineare Behandlungseffekte in Kohortenanalysen können mittels Adjustierung für den PS oder DRS unverzerrt geschätzt werden. Bei nicht-linearen Effekten (z.B. Odds Ratios) ist eine unverzerrte Schätzung nur für den PS als Adjustierungsvariable unter der Nullhypothese "kein Zusammenhang zwischen Exposition und Endpunkt" gegeben. Wird der PS zum Matching in Fall-Kontroll-Studien herangezogen, sind die Schätzer ebenfalls nur unter der Nullhypothese asymptotisch unverzerrt. Ein Matching nach dem DRS ergibt erwartungstreue Schätzer, wenn die Daten mittels bedingter logistischer Regression ausgewertet werden. Wir berechnen den Bias analytisch. Dieser asymptotische Bias wird mittels numerischen Beispielen und Simulationsstudien für Poisson Regression und logistische Regression in der Kohorte und in gematchten Fall-Kontroll Studien exploriert. Zusätzlich haben wir den Bias in Simulationsstudien mit einer komplexen Verteilung der Störfaktoren basierend auf realen Datensätzen untersucht und ähnliche Ergebnisse wie in den analytischen Berechnungen beobachtet.

Im zweiten Teil der Arbeit wird der Zusammenhang von Bluttransfusionen und der Entstehung von Tumorerkrankungen evaluiert. Diese Fragestellung wurde mittels einer populationsbezogenen Fall-Kontroll-Studie in der älteren US-Bevölkerung, basierend auf einer Stichprobe aus der SEER (Surveillance, Epidemiology and End Results)-Medicare Datenbank, analysiert. Das Risiko aller Tumore gemeinsam, sowie einzelner Subtypen wurde untersucht. Unter Berücksichtigung einer großen Anzahl an Störfaktoren mittels Summary Scores (u.a. PS und DRS) ergab die Analyse ein signifikant erhöhtes Krebsrisiko für alle Tumore gemeinsam und für Krebserkrankungen des Magens, Dickdarms, der Leber, Niere / Nierenbecken / Ureter, für Lymphoma, Myeloma und Leukämie, kurz nach dem Erhalt einer Bluttransfusion. Nur bei Leberkarzinomen wurde für längere Latenzzeiten ein signifikant erhöhtes Krebsrisiko beobachtet. In Sensitivitätsanalysen mit zusätzlicher Adjustierung nach akuten und chronischen Erkrankungen, die sehr stark mit der Verabreichung einer Bluttransfusion und der Entstehung einer Tumorerkrankung in Verbindung stehen, sowie in den Balancing Score basierenden Analysen wurde kein signifikant erhöhtes Risiko für Leberkarzinome über einen längeren Zeitraum beobachtet. Diese zeitlich restriktiven Zusammenhänge deuten darauf hin, dass möglicherweise ein nicht diagnostizierter Krebs oder Vorstufen von Tumorerkrankungen eine Bluttransfusion bedingen.

Die Ergebnisse dieser Arbeit liefern einen Beitrag zum besseren Verständnis der Eigenschaften von Schätzern basierend auf Beobachtungsstudien die Balancing Scores im Design und in der Auswertung verwenden. Sie tragen auch zum Verständnis des Zusammenhangs von Bluttransfusionen und Krebsrisiko bei.

Contents

1 Introduction **10**
- 1.1 Motivation and aim . . . 10
- 1.2 Structure of the thesis . . . 11

2 Observational study designs and confounding **13**
- 2.1 Overview of observational study designs . . . 13
  - 2.1.1 Cohort study . . . 13
  - 2.1.2 Subsamples of cohorts . . . 15
  - 2.1.3 Case-control study . . . 17
- 2.2 Bias and Confounding . . . 18
  - 2.2.1 Bias . . . 18
  - 2.2.2 Confounding . . . 19
- 2.3 Summary . . . 21

3 Approaches to reduce the impact of confounding in observational studies **22**
- 3.1 Design-based approaches . . . 22
  - 3.1.1 Randomization . . . 22
  - 3.1.2 Restriction . . . 23
  - 3.1.3 Matching . . . 23
- 3.2 Analysis-based approaches . . . 29
  - 3.2.1 Stratified analysis . . . 29
  - 3.2.2 Multivariable analysis . . . 29
- 3.3 Summary . . . 31

4 Balancing scores **32**
- 4.1 Potential outcome framework . . . 32
- 4.2 Balancing scores . . . 35
- 4.3 Propensity score (PS) . . . 36
  - 4.3.1 Properties of PS . . . 36
  - 4.3.2 Estimating the PS . . . 37
- 4.4 Disease risk score (DRS) . . . 38
  - 4.4.1 Properties of DRS . . . 39
  - 4.4.2 Estimating the DRS . . . 40

5 Use of balancing scores in observational studies 41
5.1 Design . . . 41
5.1.1 Restriction . . . 41
5.1.2 Matching . . . 42
5.2 Analysis . . . 43
5.2.1 Stratification . . . 43
5.2.2 Inverse probability weighting (IPW) . . . 44
5.2.3 Adjustment in regression models . . . 45

6 Analytic bias for conditional treatment effect 47
6.1 Motivation . . . 47
6.2 Models . . . 49
6.3 Bias calculations for cohort analysis . . . 51
6.4 Numerical examples for cohort analysis . . . 53
6.4.1 Scenarios and settings . . . 53
6.4.2 Comparison with the published results . . . 56
6.4.3 Results - Poisson regression . . . 58
6.4.4 Results - logistic regression . . . 62
6.5 Bias calculation for the matched case-control design . . . 66
6.6 Numerical examples for the matched case-control design . . . 68
6.6.1 Scenarios and settings . . . 68
6.6.2 Results . . . 70
6.7 Summary . . . 73

7 Simulations based on blood transfusion data 76
7.1 Data generation . . . 76
7.2 Score calculation . . . 78
7.3 Case-control design . . . 79
7.4 Outcome analysis . . . 79
7.5 Scenarios and settings . . . 80
7.6 Results . . . 82

8 Data example: a study of blood transfusions and the subsequent risk of cancers in the U.S. elderly 86
8.1 Motivation . . . 86
8.2 Materials . . . 87
8.2.1 Definition of cases and control population . . . 88
8.2.2 Blood transfusions . . . 88
8.2.3 Possible confounders . . . 89
8.3 Methods . . . 91
8.3.1 Study designs . . . 91
8.3.2 Statistical analysis . . . 93
8.4 Results . . . 95
8.4.1 Analysis of all cases and all controls . . . 95
8.4.2 Analysis of matched data sets . . . 99
8.4.3 Sensitivity analyses . . . 105
8.4.4 Transfusions and risk of liver cancer . . . 107
8.5 Discussion . . . 112

9 Summary and future work **115**

Appendices **119**

Appendix A Supplemental Tables and Figures **119**
A.1 Supplemental Figures for Chapter 6 120
A.1.1 Cohort study - Poisson regression 120
A.1.2 Cohort study - logistic regression 123
A.1.3 Case-control study 126
A.2 Supplemental Tables for Chapter 7 128
A.3 Supplemental Tables for Chapter 8 130

Appendix B R-code **136**
B.1 Analytic bias - cohort study 137
B.1.1 Poisson regression 137
B.1.2 Logistic regression 140
B.2 Analytic bias - matched case-control study 143
B.2.1 Matched analysis, conditional logistic regression 143
B.2.2 Unmatched analysis, unconditional logistic regression 146

Chapter 1

Introduction

1.1 Motivation and aim

In observational research the presence of confounders, defined as variables associated with both, main exposure of interest and disease outcome, may induce a bias in the estimates of association between exposure and outcome. Conventional statistical approaches to reduce the impact of confounding are not feasible when there are many confounders and the disease outcome is rare. One way to deal with these challenges is to calculate a confounder summary score (Miettinen, 1976b), which combines the multiple potential confounding variables into a single scalar. The propensity score (PS) and the disease risk score (DRS) are such popular scores. The PS is defined as the conditional probability of exposure given a vector of measured baseline covariates. Rosenbaum and Rubin (1984) have shown that the PS is an exposure balancing score that balances the measured confounders between the exposure groups and Rosenbaum and Rubin (1984) extensively discussed its properties. The DRS predicts the risk of disease outcome in the absence of exposure as a function of confounding variables (Miettinen, 1976b; Arbogast and Ray, 2011). Given the DRS, the distribution of the confounders can be balanced across the levels of the outcome (Hansen, 2008).

In the last decades, the application of the PS and the DRS in observational research has increased considerably (Stürmer et al., 2006; Tadrous et al., 2013). In cohort studies, Rosenbaum and Rubin (1984) proposed stratification or matching exposed and unexposed individuals on the PS to mimic a randomized study design and to remove effects by measured confounders. In the recent medical literature however, the PS has also been used as an adjustment variable in regression models for cohort analysis or as a matching variable for disease outcome in the design of case-control studies (Alarcón et al., 2007; El-Serag et al., 2009; Etminan et al., 2010; Thillemann et al., 2010; Modén et al., 2010; Howell et al., 2010; Michalia et al., 2012). In a simulation study, Austin et al. (2007b) showed that adjusting for the PS in logistic or Cox proportional hazards models for cohort data can result in biased estimates of conditional exposure effects when the estimated odds ratios or hazards ratios were different than one.

The use of the DRS has been proposed as alternative to the PS when exposure is rare or has multiple categories (e.g. Glynn et al., 2006; Arbogast and Seeger, 2012). A systematic review about the application of the DRS in medical research revealed that, since 2000 the majority of applications used the DRS as a covariate in regression models (Tadrous et al., 2013). In a simulation study, Arbogast and Ray (2011) investigated the performance of the DRS as

adjustment variable in cohort analysis in comparison to PS adjustment in Poisson regression and in comparison to conventional multivariable Poisson models (i.e. including all confounders). For an adequate number of events per confounder, adjusting for the DRS and the conventional regression models performed best in terms of bias. Slightly more, but still negligible bias was observed for the model that adjusted for the PS. However, Arbogast and Ray (2011) considered only a limited range of scenarios and suggested further research in other settings, especially in the case-control setting.

This thesis has two parts. The first part focuses on the estimation of conditional exposure effects for the following two settings: summary scores (the PS or the DRS) are used as adjustment variables in multivariable regression models for cohort data, and as matching variables a case-control studies. In the cohort setting, analytic bias calculations are presented for linear, logistic and Poisson regression models. In the case-control setting, the amount of bias is investigated analytically when matching on the scores is performed and data are analyzed using conditional or unconditional logistic regression adjusted for the score. The asymptotic bias calculations are verified by several simulations. The second part of the thesis is an empirical investigation of blood transfusions and subsequent cancer risk in the U.S. elderly. Blood transfusions are common in older adults and also modulate the immune system (Vamvakas and Blajchman, 2007; Rogers et al., 2011). Hjalgrim et al. (2007) studied cancer incidence in Swedish and Danish individuals of all ages who received transfusions between 1968 and 2002. They observed an increased cancer risk shortly after the receipt of a transfusion and long term elevated risk for specific cancer subtypes, including liver cancer. In our study, we focused on an elderly population taking into account a large number of possible confounders. We apply several summary score approaches, including the PS and the DRS to estimate the association of transfusion and cancer risk.

1.2 Structure of the thesis

Chapter 2 presents an overview of observational study designs, i.e. cohort and case-control studies, and briefly summarizes their strengths and limitations. Further, the role of bias and confounding in clinical and epidemiological research is discussed and illustrated for binary exposure and disease outcome. In Chapter 3, common approaches for addressing confounding in the design (randomization, restriction and matching) and in the analysis (stratified and multivariable analysis) of an observational study are presented. Advantages and disadvantages of those approaches to reduce the impact of confounding are highlighted. Particular focus is given on various matching types, matching algorithm and special issues for matching in case-control studies.

The next two chapters concentrate on balancing scores. Chapter 4 introduces the PS and the DRS in the context of potential outcome framework, summarizes their properties and discusses various issues about the estimation of the scores. In Chapter 5, the use of balancing scores in observational studies that have been presented in the literature to lessen the impact of confounding are summarized. Restriction, matching, stratification, weighting and using the scores as adjustment variable in a regression model are presented in the context of cohort studies.

The next chapters summarize the results of our research where we investigated the performance of summary scores for the estimation of conditional exposure effects analytically and in simulation studies. The scores are designed to estimate marginal or population-average exposure effects in the cohort setting (Rosenbaum and Rubin, 1984; Hansen, 2008). However, in the

literature, the scores are also used in non-linear regression analysis (i.e logistic regression, Poisson regression and Cox proportional hazard models) and as matching or adjusting variables in a case-control study. As shown in simulation studies, adjusting for the PS in non-linear regression models can result in biased estimates for conditional odds ratios or hazard ratios (Austin et al., 2007b; Martens et al., 2008). There is a lack of information about the performance of the scores in the case-control setting. We therefore investigate the amount of bias for conditional exposure effects introduced by adjusting on the scores instead of the covariates itself analytically. First, we analytically calculated the potential bias in estimates of exposure-outcome association estimates, when logistic regression models or Poisson regression models adjusted for summary scores instead of the true confounding variables are used to analyze cohort data. Second, we computed the bias that occurs when in a case-control study cases and controls are matched on the scores and analyzed by conditional or unconditional logistic regression. The analytic calculations and numerical examples for different scenarios and settings are presented in Chapter 6. To investigate the performance of the summary scores in cohort and case-control studies for complex confounding settings, a simulation study based on a real data example is performed and presented in Chapter 7.

In the second part of the thesis we apply PS and DRS matching to a substantial problem, namely an investigation of blood transfusions and the subsequent risk of cancers in the U.S. elderly, based on a sample from the SEER (Surveillance, Epidemiology and End Results)-Medicare database (Riedl et al., 2013). The results of this study are presented in Chapter 8. In the thesis, different methods to control for confounding by many medical diagnoses and chronic conditions incorporating several summary scores are investigated. Finally, a summarization of the findings and an outlook on further work based on this thesis is given in Chapter 9.

Chapter 2

Observational study designs and confounding

In medical research randomized controlled trials (RCTs) are considered the gold standard to estimate the effects of treatments and exposures on health outcomes. To ensure that the estimate of treatment effect is not distorted by other factors (confounders), a random allocation of the subjects to different treatment groups is one of the most important design technique for avoiding bias. But a randomized experiment is not feasible in many settings. For example in cancer research, exploring risk factors for cancer incidence, is typically based on an observational rather than experimental design. In contrast to a randomized experiment, an observational study is an empirical investigation where the investigator cannot control the assignment of treatments or exposures to individuals. So, there is a greater risk of bias than in experimental studies, due to confounding. Confounding can lead to under- or overestimation of an investigated effect or suggest an association where none exists. Therefore, methods are needed to reduce the impact of confounding on the estimates of treatment or exposure effects in observational studies.

This chapter gives an overview of different observational study designs and summarizes the role of bias and confounding in clinical and epidemiological research.

2.1 Overview of observational study designs

A detailed description of different observational study designs is given in Breslow and Day (1980, 1987); Woodward (1999); Kass and Gold (1999) and Rothman et al. (2008). This section summarizes briefly two main types of study designs in epidemiological research, cohort studies and case-control studies.

2.1.1 Cohort study

In a cohort study (longitudinal study), the association between exposure and disease is investigated by following a set of individuals (i.e. the 'cohort') forward in time and measuring the rate of occurrence of new cases (incidence) in the different exposure groups. The key feature in cohort studies is that individuals are sampled from a well defined disease-free population, that is followed forward through time and information on their exposure (e.g. the receipt of a blood transfusion or not), before their disease occurrence (e.g. developing cancer) is determined

(Breslow and Day, 1987). Cohort studies can be conducted prospectively or retrospectively. In a prospective cohort study, the exposure information is obtained at the start of follow-up and the individuals are then followed into the future. In retrospective cohort studies (historical cohort study), the exposure information is obtained at a certain defined time in the past (e.g. from historical records) and the individuals are then retrospectively followed forward through time.

The main advantage of a cohort design is that direct information on the time sequence between exposure and events of interest (e.g. occurrence of disease) is available. Since exposure status is determined before the outcome occurs, a cohort design provides direct estimates of incidence rates. Especially in prospective designs, information about exposure status and parameters that might influence disease occurrence can be assessed more accurately compared with retrospective designs. Additionally, during follow-up, many disease outcomes can be studied simultaneously. It is also a suitable design for rare exposures. However, a major disadvantage of cohort studies is that they are not suitable for rare diseases or, especially for prospective cohort studies, for diseases with long latency periods. Large sample sizes and very long follow-up would be required, which is expensive and time-consuming. A summary of strengths and limitations of cohort studies is given in Breslow and Day (1987, p.15-22) and Woodward (1999, p.192-193).

Assessing risk factors in cohort studies

Consider a simple cohort design with binary exposure (T=0 for unexposed, T=1 for exposed) and binary disease outcome (Y=0 for controls, Y=1 for cases), where all cohort members are followed-up for the same time-period. The results for this cohort study can be summarized in a 2x2 table (Table 2.1), where n_{ij} with $i, j = 0, 1$ denotes the observed frequency for the exposure (level i) and disease status (level j) combinations and N denotes the total cohort size. A measure of association of exposure and disease of interest is the relative risk (RR), i.e. the ratio of the risk of disease between exposed and unexposed individuals. Based on the notation from Table 2.1, the risks of disease for the exposed (R_1) and unexposed (R_0) individuals are given by

$$R_1 = n_{11}/n_{1\cdot} \quad \textit{and} \quad R_0 = \quad n_{01}/n_{0\cdot}.$$

The RR of disease for the exposed group, compared to the unexposed can be calculated by

$$\text{RR} = \frac{R_1}{R_0} = \frac{n_{11}/n_{1\cdot}}{n_{01}/n_{0\cdot}}.$$

Table 2.1: Summary of data in cohort study.

	Disease status Y		
Risk factor T	Disease Y=1	No disease Y=0	Total
Exposed T=1	n_{11}	n_{10}	$n_{11} + n_{10} = n_{1\cdot}$
Unexposed T=0	n_{01}	n_{00}	$n_{01} + n_{00} = n_{0\cdot}$
Total	$n_{11} + n_{01} = n_{\cdot 1}$	$n_{10} + n_{00} = n_{\cdot 0}$	N

In most cohort studies individuals are followed for different lengths of time, e.g. patients drop out before the end of the study period or die from causes other than the one of interest, i.e. are censored. To consider the different lengths of follow-up in analysis, one can calculate rates (i.e.

incidence rates) and rate ratios, instead of risks and risk ratios. In contrast to the risk of disease, rates use person-years at risk as the denominator. A common method of analyzing cohort data, which also takes into account the different durations of follow-up, is to carry out a survival analysis (for details see Woodward (1999) and Rothman et al. (2008)).

2.1.2 Subsamples of cohorts

To overcome the problem of long latency periods or rare diseases in cohort studies, to reduce costs and time efforts and to improve efficiency, different strategies to sample within a cohort have been developed, such as nested case-control and case-cohort designs. The sampling strategy of both designs is usually to include all cases that occur within a well defined source population (the cohort) but sample only a fraction of those who do not experience the events (controls). The sampling can be performed so that the advantages of a cohort design remain (i.e. information of the time sequence) and therefore similar methods as in cohort studies can be used to estimate the disease risk association (Barlow et al., 1999; Langholz, 2005). Such designs are useful for the collection of new information that was not part of the original cohort. E.g. laboratory testing might be too expensive if all of the cohort members are tested. Including only a fraction of the controls can reduce such costs.

Nested case-control study

Originally, case-control sampling strategies within a well-defined cohort were proposed to reduce computational problems in large cohort studies. Mantel (1973) suggested to convert large prospective studies with rare outcomes into a 'synthetic retrospective study' by selecting cases and controls randomly from the cohort with typically high sampling fractions for cases and small sampling fractions for the controls. Further formalizations and developments of the nested case-control design are given in Kass and Gold (1999).

Generally, cases and controls (i.e. the study population) are sampled from a large cohort in which exposure data and population characteristics are available. The sampling depends on disease status, whereat any time a new case develops a fixed number of controls '*at risk*' is randomly sampled. This sampling strategy is also called *incidence density sampling* (Miettinen, 1976a) or *risk set sampling* (Robins et al., 1986). The risk set for each case consists of all disease-free individuals at the time point or within a time interval the case arises. This sampling can also be regarded as matching cases and controls on a time variable, such as age, calender time or follow-up time, which can be combined with matching on other possible confounders (e.g. sex, race, etc.) as well. Figure 2.1 illustrates the sampling scheme of the nested case-control design with two controls (white circle) sampled for each case (black circle). The risk sets for each case are presented by the vertical dashed lines.

Since the random sampling of the controls from each risk set is independent of sampling from the other risk sets, individuals can be chosen as controls more than once. Moreover, an individual who is chosen as a control can later become a case. As in cohort studies, individuals contribute person-time before they become a case.

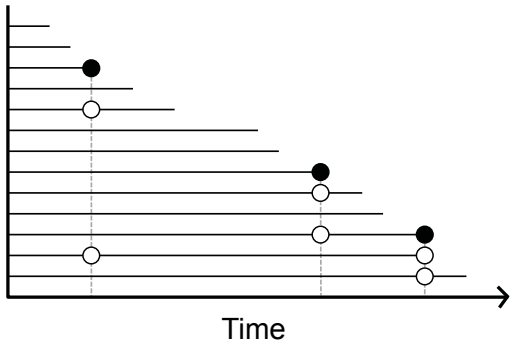


Figure 2.1: Diagram of risk set sampling. The solid lines indicate the follow-up period for each individual in the cohort. For each case (black circles), two controls (white circles) are randomly sampled from the risk sets defined by the time the case arise (vertical dashed lines).

Case-cohort study

A case-cohort (case-base) study, originally proposed by Prentice (1986), is designed to reduce the costs and efforts in large cohort studies. Instead of analyzing all data on all individuals in the cohort, a subcohort is sampled randomly from the entire cohort or within strata of the cohort at beginning of the study (baseline). Then all the cases that develop in the full cohort during follow-up are selected. The analysis is then performed on the individuals of the subcohort and all cases (Figure 2.2).

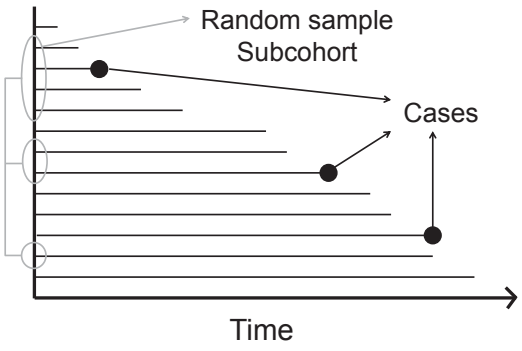


Figure 2.2: Diagram of case-base sampling. A subcohort is randomly sampled at the beginning of the follow-up period (solid lines) in the study cohort. At the end of the follow-up period, all cases (black circles) in the cohort are sampled.

In contrast to nested case-control studies, the controls are not matched to cases on time (or other variables). Thus, case-cohort studies can be seen as unmatched nested case-control studies (Kass and Gold, 1999). An advantage of case-cohort studies compared to nested case-control studies is that several disease outcomes (various case-types) can be compared to the same subcohort. Further considerations regarding the choice between the two designs are given in Wacholder (1991).

2.1.3 Case-control study

In a case-control study (case-referent study), the exposure-disease relationship is investigated by selecting individuals with a specific disease (cases) and comparable individuals without the disease (controls). One or more risk factors for the disease of interest can be obtained retrospectively and compared between the cases and controls. Usually, only one disease of interest can be investigated since individuals are selected based on disease status (examples for exceptions i.e. 'multi-disease studies' and 'para-cases' are given in Breslow and Day (1987, p.15).

The main advantage of case-control studies is that, in contrast to prospective cohort studies, they are less time-consuming and more cost-effective, since no waiting for diseases with long latency periods is needed. The case-control study starts with individuals, who have already developed the disease, i.e. are sampled retrospectively. This makes them also particularly suitable for studying rare diseases (e.g. cancer). However, in contrast to nested case-control studies, there is no matching on follow-up time. Thus only relative quantities can be estimated. A further major limitation of case-control designs is the non-comparability between cases and controls and the quality of the collected information, since individuals are selected after exposure and outcome occurs. Therefore, there is a high risk for bias. This major limitation is discussed in the next section. An overview of strengths and limitations of case-control studies is given in Breslow and Day (1987, p.20-22) and Woodward (1999, p.243-247).

Assessing risk factors in case-control studies

In case-control studies, the relative risk as an estimate of association of exposure and disease is incorrect, since the disease incidences for exposed and unexposed individuals (from Table 2.1) are unknown. In case-control studies the individuals are sampled retrospectively, i.e. on the basis of disease status, therefore a higher proportion of cases compared to what is seen in the general population is present. Assume f_1 is the sampling fraction for cases and f_2 the sampling fraction for controls with $f_1 >> f_2$, then the expected values for the case-control sample are given in Table 2.2.

Table 2.2: Summary of data in ccase-control study.

	Disease status Y		
Risk factor T	Disease Y=1	No disease Y=0	Total
Exposed T=1	$f_1 n_{11}$	$f_2 n_{10}$	$f_1 n_{11} + f_2 n_{10}$
Unexposed T=0	$f_1 n_{01}$	$f_2 n_{00}$	$f_1 n_{01} + f_2 n_{00}$
Total	$f_1 n_{.1}$	$f_2 n_{.0}$	n

The relative risk in the population from Table 2.1 and the sample relative risk from Table 2.2 are not the same:

$$\frac{f_1 n_{11}/(f_1 n_{11} + f_2 n_{10})}{f_1 n_{01}/(f_1 n_{01} + f_2 n_{00})} \neq \frac{n_{11}/(n_{11} + n_{10})}{n_{01}/(n_{01} + n_{00})} = \text{RR}.$$

Instead of relative risks, odds ratios (ORs) as measure of association can be calculated in case-control studies, since the sampling fractions f_1 and f_2 cancel out:

$$\text{OR} = \frac{f_1 n_{11}/f_2 n_{10}}{f_1 n_{01}/f_2 n_{00}} = \frac{n_{11}/n_{10}}{n_{01}/n_{00}}.$$

However, if the disease is rare, odds ratios approximate risk ratios. This is also known as 'rare disease assumption' (Cornfield, 1951). Thus, under rare disease, the quantities n_{11} and n_{01} in Table 2.2 are much smaller compared to n_{10} and n_{00} (i.e. $n_{11} << n_{10}$ and $n_{01} << n_{00}$) and moreover, the disease risks for the exposed (R_1) and unexposed individuals (R_0) is small. The relation between the relative risk and the odds ratio can be expressed as (Agresti, 1990, p.17)

$$\text{OR} = \frac{n_{11}/(n_{1\cdot} - n_{11})}{n_{01}/(n_{0\cdot} - n_{01})} = \frac{\frac{n_{11}}{n_{1\cdot}}/(1 - \frac{n_{11}}{n_{1\cdot}})}{\frac{n_{01}}{n_{0\cdot}}/(1 - \frac{n_{01}}{n_{0\cdot}})} = \frac{R_1/(1-R_1)}{R_0/(1-R_0)} = \text{RR} \times \left(\frac{1-R_0}{1-R_1}\right).$$

As $R_0 << 1$ and $R_1 << 1$, it follows that $(1 - R_0)/(1 - R_1)$ can be approximated by 1 and therefore, $OR \approx RR$. If the disease is not rare, the OR overestimates the RR (Rothman et al., 2008). If $R_0 > R_1$ then $1 - R_1 > 1 - R_0$ and OR > RR > 1. If $R_0 < R_1$ then $1 - R_1 < 1 - R_0$ and OR < RR < 1.

2.2 Bias and Confounding

2.2.1 Bias

Bias is defined as a systematic error, attributable to methodological aspects of the study design or analysis and results in a distortion of estimates of investigated treatment or exposure effects. There are different types of bias which can be classified according to where their source lies in the stages of a study (Sackett, 1979). Some examples are given in the following paragraphs (for details see Kleinbaum et al. (1982, p.195-265) and Rothman et al. (2008, p.134-147)).

Selection bias

Selection bias results from the non-representative selection of individuals in the study sample and refers to undesired systematic differences between the comparison groups. In case-control studies, which are highly susceptible for this bias, the choice of cases and maybe controls can affect the occurrence of selection bias. For example, the presence of exposure is related to more medical attention leading to those with disease having more exposure. In nested case-control or case cohort studies the selection bias can be minimized, since controls are selected from the same cohort the cases occur. In cohort studies selection bias can arise due to differential loss to follow-up (e.g. the presence of exposure leads to loss to follow-up which further affects the risk).

Information bias

Distortion of estimates can arise due to measurement error or misclassification of individuals (e.g. invalid measurements, incorrect diagnostic criteria, omissions, imprecisions). If the probability of misclassification is the same in all study groups (nondifferential misclassification) the exposure effect generally decreases (Rothman et al., 2008). If the probability of misclassification varies between the groups (differential misclassification), exposure effects can be over- or underestimated. Information bias is an issue especially in case-control studies because risk factors are obtained retrospectively. Cases may tend to remember past exposures more often or more accurately than controls (recall bias) which can lead to an overestimation of the exposure effects (Breslow and Day, 1987). This might not be an issue in cohort studies or subsamples of cohorts, since the exposure information is usually collected before disease status and therefore is not influenced by the knowledge of disease status.

Bias due to confounding or bias due to model misspecification

Bias can also arise at the analysis stage of a study if variables associated with both, the exposure and the outcome, are omitted in the outcome model or if the true outcome model requires interaction or quadratic terms and such terms are ignored in the final model (i.e. misspecification of the functional form of an outcome model). The exclusion of important covariates, i.e. confounders, can occur if variable selection strategies to reduce the number of covariates in an outcome model, like stepwise selection in regression models, are applied. If there is a large number of covariates, variables that are weakly associated with the outcome might not be part of the final outcome model which further can result in a biased estimation of an exposure effect. The concept of confounding is discussed in more detail in the next section.

2.2.2 Confounding

Confounding is a distortion of the exposure-disease relation by some other factors which are related to both, exposure and disease. Confounding can be seen as mixing of effects of the exposure of interest with effects of other extraneous factors (exposures, interventions, treatments, etc) on the disease being studied. If a confounder is not adequately controlled for, bias arises in the estimated effect of exposure on disease, whether or not there is an exposure-disease association.

Definition of confounding

Confounding is illustrated by a transfusion and liver cancer example given in Figure 2.3 whereat T denotes the binary transfusion status and Y the binary disease outcome. There are three criteria a variable X must satisfy to qualify as confounding factor (Rothman et al., 2008, p.132-134):

1. X must be associated with disease outcome (Y) among unexposed individuals (T=0).
2. X must be associated with exposure (T) in the source population.
3. X must not be affected by exposure or the disease (i.e. is not an intermediate).

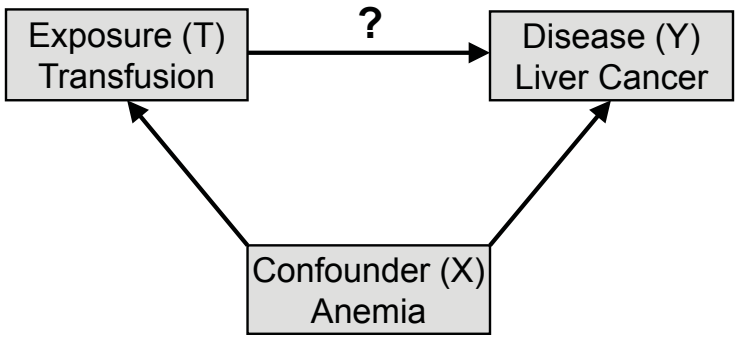


Figure 2.3: Diagram of the relationship of a confounder (X) to exposure or treatment (T) and disease outcome (Y).

Ad 1.
A confounding factor must be an extraneous risk factor for the disease of interest and therefore is associated with disease occurrence apart from its association with exposure. The confounding

factor is either a cause of the disease or a surrogate for a cause other than exposure. For example, in the investigation of the association of transfusion and liver cancer (as shown in Figure 2.3), anemia represents a possible confounder. It is known, that anemia increases the risk of liver cancer, independently of the transfusion.

Ad 2.
The association of a confounder with the exposure must be present in the source population, that is the study cohort or the source population that gives rise to the sample for a case-control study. Concerning the example in Figure 2.3, anemia increases the probability of receiving a transfusion. If the proportion of individuals with anemia is balanced between the individuals who received a transfusion and those who do not, anemia does not confound the association between transfusion and liver cancer.

Ad 3.
Any variable that is an intermediate in the causal pathway between exposure and disease should not be treated as an extraneous confounder. If such a factor is treated as a confounder, serious bias could be introduced. For example, transfusions may lead to infection which could further cause liver cancer. Infection represents an intermediate between exposure and disease (Figure 2.4) and not a confounder.

Figure 2.4: Diagram of infection that is an intermediate in the causal pathway between the exposure (transfusion) and the outcome (liver cancer).

Direction of confounding

Omitting a confounder in the analysis leads to under- or overestimation of an effect, depending on the direction of the association between the confounder with exposure and disease. There are two types of confounding situations illustrated here for a confounder with two levels (Breslow and Day, 1980, p.93-94):

Type 1: In one level of the confounder, there is a high risk for disease and a high prevalence of exposure and in the other level low risk for disease and a low prevalence of exposure. Thus the association of an exposure (T) and a confounder (X) is positive,

$$\text{corr}(X, T) > 0$$

and the association of a disease (Y) and a confounder (X) is positive as well,

$$\text{corr}(X, Y) > 0.$$

Then the association between exposure and disease will be stronger if confounding is ignored. For example, smoking increases the probability of receiving a transfusion and is also a risk factor for liver cancer. Because more smokers are in the group that received a transfusion than in the other

group, more cancer cases as well are expected in the group with transfusion. The same will happen, if

$$\text{corr}(X, T) < 0 \quad \text{and} \quad \text{corr}(X, Y) < 0.$$

Type 2: High risk for disease and low prevalence of exposure is given in one level of the confounder and vice versa for the other. So, the association between the exposure (T) and the confounder (X) is negative and the association of the disease (Y) and the confounder (X) is positive:

$$\text{corr}(X, T) < 0 \quad \text{and} \quad \text{corr}(X, Y) > 0,$$

then association of exposure and disease will be weaker if confounding is ignored. The same will happen, if

$$\text{corr}(X, T) > 0 \quad \text{and} \quad \text{corr}(X, Y) < 0.$$

2.3 Summary

Table 2.3: Summary of strength and limitations of cohort and case-control studies.

Study design	Strength	Limitation
Cohort	• direct information of time sequence of events of interest (e.g. occurrence of disease) is available which allows the estimation of incidence rates • many diseases can be studied simultaneously during the follow-up • suitable for rare exposure • prospective cohort study: exposure status and other important parameters can accurately be assessed • one can estimate absolute quantities, e.g. absolute risk for outcome	• can be expensive and time-consuming (e.g. monitoring a large number of individuals over a long time-period) • not suitable for diseases with long latency periods (e.g. developing cancer after 10-years) • not suitable for rare disease. For example, if the incidence rate of a specific cancer is 5 cases per 100 000 per year an enormous cohort sample size would be needed to ascertain adequate case numbers.
Case-control	• compared to prospective cohort studies, case-control studies are less time-consuming and more cost-effective, since no waiting for diseases with long latency periods is needed • good applicability to rare disease • it is easier to ascertain many risk factors simultaneously	• disease incidence is not known therefore only approximate estimates of relative risk available • generally, only one disease of interest can be investigated since individuals are selected based on disease status • high possibility for bias, e.g. due to different quality of information on characteristics or risk factor status between cases and controls (e.g. cases more interested in their medical history) (recall bias) or non-comparability between cases and controls (selection bias)

Chapter 3

Approaches to reduce the impact of confounding in observational studies

There are several ways to deal with confounding. Generally, the influence of confounding can be reduced by design or controlled in the analysis of a study or both. The main strategies to control for confounding are randomization, restriction, matching or stratification and adjustment when fitting regression models to analyze the data. Among these strategies randomization is the only approach that potentially balances measured and unmeasured confounders. The other strategies only address measured confounders.

3.1 Design-based approaches

Randomization, matching and restriction are applied at the study design phase to reduce the impact of confounding.

3.1.1 Randomization

Randomization is considered the gold standard to reduce bias due to potential confounding factors. A random allocation implies that each study participant has an equal chance of being assigned to any treatment/exposure group. Considering a simple randomization, e.g. flipping a coin for studies with two groups (e.g. treatment or placebo), all study participants have assignment probability of 0.5 for the treatment group regardless of their characteristics.

A random allocation to treatment groups reduces the impact of confounding by ensuring that potential confounding factors, known and unknown, are equally distributed among the study groups at the beginning (baseline) of the study. That is treatment and confounders are independent. If randomization is performed accurately, the risk of bias in the estimation of the treatment-outcome effect due to serious imbalance in potential confounding factors is reduced. However, randomization cannot rule out all risk of bias (Jadad, 1998). Covariate imbalance after randomization can arise by chance, especially in small trials. Different randomization methods, like block randomization or stratified randomization, have been developed to promote balance in important baseline characteristics (Pocock, 1983).

3.1.2 Restriction

In contrast to randomization, which is applicable only in experimental studies, restriction can be performed in all study designs the same way. The idea is, to restrict the inclusion criteria for individuals to avoid a confounder to vary and thus reducing confounding by known factors. For example, if only male individuals in a specific age range are included in the study one eliminates confounding by sex and age. Restriction is a simple and effective method to prevent or reduce confounding if there is a large pool of available individuals. However, restriction can also complicate the recruitment of individuals, if the criteria are too narrow (e.g. restriction is based on many possible confounders) and the desired sample size can therefore not be achieved without additionally effort (e.g. extension of the recruitment period). On the other hand, if the restriction criteria are too wide, there is a higher possibility of residual confounding (i.e. the effect of the confounder is not completely removed). A further limitation is that results may not be generalizable.

3.1.3 Matching

Confounding can be reduced by matching study participants on potential confounders and thus assuring similar confounder distributions among the study groups. In cohort studies matching is performed by matching exposed and unexposed individuals with similar baseline characteristics without accounting for the disease status. Matching in prospective cohort studies is uncommon, because large cohorts are often required and this will be too expensive or time-consuming (Rothman et al., 2008). But in registry-based cohort studies matching is frequently applied to reduce the impact of confounding. In case-control studies matching is performed by selecting cases and controls with similar confounding values. More detail on matching in case-control studies is given at the end of this section.

Matching in observational studies is done to mimic a randomized design. That is, achieving covariate balance for potential baseline confounding factors between the exposure groups and therefore 'breaking' the association between the confounders and the exposure. Or, in case-control matching, 'breaking' the association between confounders and disease status.

Various aspects of matching for case-control studies are now discussed. However if 'cases' are replaced by 'exposed' and 'controls' by 'unexposed', the description equally applies to cohort studies.

Types of matching

Generally, two types of matching can be performed, namely *individual matching* or *frequency matching*. In both types, the confounders on which matching is performed are equally distributed among the cases and controls.

Individual matching: here, matching is performed based on the selection of one or more controls with the same characteristics, i.e. the matching factors, for each case. If sex and age are possible confounders, one would match a 50 years old male case to a 50 years old male control. This is done for all included cases in the study, individually.

Frequency matching: here, the controls are chosen to ensure that the distribution of the matching factors among controls overall is the same as for cases. In contrast to individual matching, the matching is performed on a group level. For example, if there are 50 male cases in a

study, than 50 or 100 male controls are selected depending on the matching ratio. For more than one matching factor, e.g. sex and age groups, the controls are selected based on the frequency of cases within the strata formed by those matching factors, so that their joint distribution is the same among cases and controls.

Further matching variants are partial and marginal matching (Rothman et al., 2008). In partial matching, controls can be selected only for some cases, or different matching factors for different individuals can be used. Marginal matching can be considered as a form of frequency matching whereat controls are selected so that the marginal distribution of the matching factors is the same among cases and controls.

Exact and inexact matching

When individual matching is performed, one can choose between exact and inexact matching methods. Using exact matching, a case is matched to one ore more controls with exactly the same covariate values as the case which results in exact covariate balance between the matched cases and the controls. If this matching is performed on only a few confounders (e.g. sex and age groups) and an adequate number of study members is available, exact matching on those factors can easily be performed. But if the number of matching factors increases, exact matching becomes difficult. For example, if there are two binary categorized confounders there are $2^2 = 4$ combinations and it is typically easy to find two individuals with the same covariates. But, if there are 10 binary confounders, $2^{10} = 1024$ combinations are possible. For a high dimensional confounder or covariate matrix $\boldsymbol{X}$, a suitable match for each case might not be found (incomplete matching). Such incomplete matching can potentially lead to substantial bias due to large differences in the covariate distribution between matched and unmatched cases (Rosenbaum and Rubin, 1985a). Exact matching can also be difficult when there are continuous covariates.

To overcome this dimensionality problem in matching on many covariates, one can match cases and controls who are 'as similar as possible' (inexact matching) instead of finding exact matches on $\boldsymbol{X}$. Using inexact matching methods, cases and controls with different covariate values are matched. In the case of several continuous covariates, this can be done by creating categories, e.g. deciles, and use those in a matching algorithm. Further methods are matching on balancing scores (see next chapter) or matching on distance matrices that are presented next.

A distance matrix $\boldsymbol{D}$ is a matrix of dimension $n \times m$, whereat n represents the number of cases and m the number of possible controls. The matrix entry d_{ij} is the 'distance' between the i-th case ($i = 1, \ldots, n$) and the j-th control ($j = 1, \ldots, m$), which quantifies the dissimilarity in the covariate vectors $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ik})$ for the i-th case and $\boldsymbol{X}_j = (X_{j1}, \ldots, X_{jk})$ the j-th control as defined by some chosen distance metric. More precisely, the distance between two vectors is a function d:

$$d : \mathbb{R}^k \times \mathbb{R}^k \to [0, \infty],$$

whereat $d_{ij} = d(\boldsymbol{X}_i, \boldsymbol{X}_j) = d(\boldsymbol{X}_j, \boldsymbol{X}_i)$ is a nonnegative number. If the distance between two individuals is zero, then both individuals have the same covariate values. The larger the distance, the less similar they are.

A popular distance measure used for multivariate matching methods is the Mahalanobis distance (Rubin, 1980), which is defined as

$$d_{ij} = (\boldsymbol{X}_i - \boldsymbol{X}_j)'\Sigma^{-1}(\boldsymbol{X}_i - \boldsymbol{X}_j), \tag{3.1}$$

with variance-covariance-matrix Σ of $\boldsymbol{X}$. The Mahalanobis distance works well for normally distributed covariates, for which the distance was originally developed. Problems can arise for rare binary covariates or if outliers are present (Rosenbaum, 2010; Stuart, 2010). Modifications of the Mahalanobis distance are given in Rosenbaum (2010) (i.e. rank-based Mahalanobis distance for categorical variables) or in Sekhon (2011) (i.e. a generalized form of the Mahalanobis metric used in the genetic matching algorithm that employs weights).

Exact and inexact matching methods can also be combined. For example, if gender is a strong confounder, exact matching on gender can be done while for other covariates a distance matrix for male and female separately is used.

Matching algorithm

After a distance measure has been chosen, the next step is choosing a matching algorithm. This section summarizes the most common matching algorithms used in praxis. An overview of further matching algorithm are given in Caliendo and Kopeinig (2008); Rosenbaum (2010); Stuart (2010).

Greedy matching (nearest neighbor matching): this is one of the most popular matching algorithm, easy to implement and straightforward to understand (Stuart, 2010). For a case i, a control j with the smallest absolute distance measure is chosen as the match. That is, for case i choose control j such that j is

$$min_j|d_{ij}|. \tag{3.2}$$

Using a nearest neighbor algorithm, a match can always be found regardless of the dissimilarity. However, 'poor' matches can lead to bias due to remaining residual confounding.

Caliper matching: Given a distance d_{ij}, a match j for case i is selected, only if in addition to (3.2)

$$|d_{ij}| < \epsilon, \tag{3.3}$$

where ϵ is a prespecified caliper. This approach can avoid poor matches but can lead to incomplete matching which can lead to bias Rosenbaum and Rubin (1985a). The choice of an optimal caliper width is discussed in Chapter 5. A variant of caliper matching is called radius matching, proposed by Dehejia and Wahba (2002). Instead of selecting a fixed number of controls within a caliper, all available controls within the caliper or radius are selected.

Optimal matching: In the greedy or caliper matching algorithm, if there is more than one control which satisfies the conditions (3.2) or (3.3) for a case, the control is chosen randomly. This means which control is matched depends on the (random) order of the controls as well. Similar, if there is one single nearest neighbor control available for each case, the (random) order of the cases can influence the resulting matches. It might happen, that one control chosen for the first case is also the best match (i.e. the control with the smallest distance) for the second case. A different order of the dataset might result globally in better matches. Thus, an optimal matching algorithm

takes the overall set of matches into account, by minimizing a global distance measure, e.g. minimizing the sum of the distances for the matched pairs (Rosenbaum, 2010). It is possible that for a case the second best match is chosen in order to minimize the global distance for all matched sets.

Matching with and without replacement: Several variants of the greedy or caliper matching algorithm can be performed by matching with or without replacement. In contrast to matching without replacement, in matching with replacement, controls can be chosen more than once which can increase the quality of matches and therefore decreases bias. Additionally, like in an optimal matching algorithm, matching with replacement does not depend on the order in which cases and controls are matched. Thus, an optimal match is always achieved. This method is useful, when there are only a few controls available, however the disadvantage is that inference becomes more complex, since one must take into account the dependencies between the repeated samples at the analysis stage.

Matching ratio

The matching ratio is the number of controls selected for each case. One can choose between a one to one matching (1:1), one to many matching (1:M) or full matching.

In a 1:1 matching design, also called pair matching, for each case one control is selected. If there is a large number of controls available, one can also match more than one control for each case. This will increase the overall sample size and therefore increase the efficiency (i.e. the precision of the estimators). However, the gain in efficiency decreases when the number of controls chosen increases. For a relative risk close to 1 the efficiency of using a 1:M case-control ratio relative to a large number of controls (i.e. $M = \infty$) can be approximated by $M/(M+1)$ (Breslow and Day, 1987). A matching ratio of 1:1 yields to an efficiency of 50%, a ratio of 1:4 to 80%. There is only little gain in efficiency when more than four controls per case are chosen (i.e. a ratio of 1:5 yields to an efficiency of 83%). Additionally, when increasing the number of controls the chance of finding poor matches increases which might increase bias. This bias-variance trade-off has been investigated in several simulation studies (Austin, 2010; Rassen et al., 2012) which can help to choose the optimal matching ratio depending on the matching algorithm used.

The one to many matching can be modified by allowing matching on a variable ratio instead of a fixed one (Ming and Rosenbaum, 2000). This is related to many to many matching (N:M) or full matching, in which a matched set can contain one case and at least one control or in the reverse situation, one control and at least one case. In full matching the cases and controls are divided into mutually exclusive subsets so that within the subsets an average distance measure is minimized which results in an optimal matching (Gu and Rosenbaum, 1993). Details on full matching methods are given in Rosenbaum (2010, p.179-183).

Assessing quality of matches

After matching it is important to check if in the resulting matched sample an adequate covariate balance has been achieved. This can be done by using descriptive diagnostic criteria (e.g. balance metrics) or graphical diagnostics (e.g. empirical quantile-quantile plot, density plots). The use of statistical hypothesis testing is not recommended (Imai et al., 2008; Austin, 2009), because the power of the test used to detect imbalance is reduced in the matched sample due to the smaller sample size compared with the unmatched sample.

The most common descriptive balance diagnostic is comparing the means or proportions of each covariate between the groups before and after matching. The standardized difference between cases and controls for variables is defined as:

$$d_{before} = \frac{(\bar{x}_{case} - \bar{x}_{control})}{\sqrt{(s^2_{case} + s^2_{control})/2}} \tag{3.4}$$

before matching and as

$$d_{after} = \frac{(\bar{x}_{case} - \bar{x}_{control,m})}{\sqrt{(s^2_{case} + s^2_{control})/2}} \tag{3.5}$$

after matching, where $\bar{x}_{case}$ and s^2_{case} are the sample mean and variance for the cases and $\bar{x}_{control}$ and $s^2_{control}$ are the sample mean and variance for the controls. $\bar{x}_{control,m}$ denotes the mean for the controls after matching. To calculate d_{after}, the standard deviations before matching are used to reflect the standardized change of mean differences not influenced by the change of the standard deviation through the matching (Rosenbaum, 2010, p.188). A standardized difference close to zero indicates good balance of the covariate between the cases and the controls. A standardized difference of > 0.1 indicates considerable imbalance, however it is desirable that better balance is achieved for strong confounders than for weak confounders (Austin, 2009).

Some authors further suggest to compare higher order moments: e.g. variance ratios closer to unity indicating good balance (Ho et al., 2007; Austin, 2009). In addition to comparing the balance for each single covariate separately, one can check balance for all measured covariates simultaneously. Franklin et al. (2014) evaluated the association between bias and covariate imbalance measured by several different balance metrics and proposed the use of post-matching C-statistic and average standardized differences.

Matching in case-control studies

Matching in case-control studies is the most often used strategy and a direct and intuitive approach to control for confounding by design for rare diseases (Breslow and Day, 1980). In cohort and case-control studies, the main goal of matching is to increase validity (i.e. to reduce the impact of confounding) and to increase efficiency (i.e. to increase the precision of an estimate of association) (Costanza, 1995).

However, in contrast to cohort studies, matching can induce bias regardless of whether the matching factor is a real confounder or not (Breslow, 2005; Rothman et al., 2008). This happens when the controls are selected according to a matching factor associated with exposure. The distribution of exposure within controls in the source population will be made more like the distribution of exposure within the cases, when controls are matched to cases on a factor that correlates with exposure. For example, when there is a perfect correlation between the matching factor and exposure, the exposure distribution in the matched control sample would be identical to that of the cases (e.g. crude $OR = 1$). The true exposure effect would be underestimated (bias towards the null) if matching is not taken into account in the analysis. If the matching factor is not associated with exposure (note: no real confounder), no bias is introduced by matching. As long as there is an association between a variable and exposure, regardless of being associated with disease, matching on that variable will induce bias. However, this bias can be controlled in the analysis stage, i.e. stratification (Mantel Haenszel method) or adjusting (conditional logistic

regression) on the matching factor removes this bias (see next section).

Matching on a confounder can increase or decrease efficiency, depending on the strengths of the confounder-exposure and confounder-disease association. Matching, if well done, will often make stratification or equivalently regression approaches more efficient, since through matching the distribution within all strata is made more similar. Selecting cases and controls without matching, there might be no case or no control within some strata, especially when a large number of confounders is present (Rothman et al., 2008, p.176-177). With strong confounder-disease association, more gain in efficiency is observed, whereas loss in efficiency is observed for strong confounder-exposure and exposure-disease association (Breslow, 2005). Especially, matching on variables that are no real confounders will harm efficiency and results in overmatching.

Overmatching

There are three types of overmatching described in Rothman et al. (2008, p.179-181) (further descriptions are given in Breslow and Day (1980, p.104-106) and Breslow (2005)):

Type 1: 'Matching that harms statistical efficiency'

As mentioned earlier, matching on a variable associated with exposure and not with the outcome (not an independent risk factor), will make the matched variable behave like a confounder and it will be necessary to control for such a confounder in the analysis. This in turn increases the variance of an estimator without reducing bias and therefore reduces statistical efficiency. The loss of efficiency depends on the strength of the covariate-exposure association.

Type 2: 'Matching that harms validity' (bias)

If matching is performed on an intermediate between exposure and disease (i.e. the matching factor is on the pathway between exposure and outcome, see Figure 2.4), then both crude and adjusted effect estimates of treatment effect will be biased. This bias can be seen as an irreparable form of selection bias (Rothman et al., 2008).

Type 3: 'Matching that harms cost efficiency'

Matching can be expensive, make data analysis more complex and lengthen the data-collection phase of a study. Matching on variables which are not strictly necessary, can also be seen as 'matching that harms cost efficiency' (Rothman et al., 2008). For example, matching on a large number of variables can increase costs because a larger sample of controls is needed to find suitable controls for the cases.

3.2 Analysis-based approaches

Stratification and multivariable (adjusted) analysis are the two basic approaches to reduce the impact of confounding (model-based approaches).

3.2.1 Stratified analysis

Stratification is the simplest form to control for confounding in the analysis and can be performed by grouping individuals according to levels of potential confounders. Comparisons within these subgroups (i.e. strata) of the confounders cannot be confounded by a factor that is constant within strata. The analysis can then be performed within each strata separately and an effect estimate that summarizes the association across strata of the confounders may be calculated. For example, if sex is considered as potential confounder, one can calculate the association of an exposure and a disease for males and females separately. If the resulting ORs are constant across the strata defined by sex, a weighted average of the stratum-specific ORs can be calculated.

The Mantel-Haenszel (MH) estimate is most commonly used for estimation of a summary OR. Across the different strata $i = 1, \ldots, I$, the MH-estimator $\widehat{OR}_{MH}$ is given by

$$\widehat{OR}_{MH} = \frac{\sum_{i=1}^{I} n_{11,i} n_{00,i}/N_i}{\sum_{i=1}^{I} n_{10,i} n_{01,i}/N_i},$$

with stratum-specific sample sizes N_i. The MH-approach is based on the assumption that the stratum-specific estimates of association do not vary across the strata (homogeneity assumption). If the estimates are different (heterogeneous) across the levels of a variable, also called effect modification or interaction, a common measure of association should not be calculated and separate analysis by levels of the variable should be performed instead (Rothman et al., 2008). Further summary effect measures are given in Kleinbaum et al. (1982, p.359-361).

Stratification is simple to perform when the number of confounders is small and the variables are dichotomous or categorical. For continuous variables, such as age, a stratified analysis requires to categorize that variable (e.g. building age-groups). If the stratum boundaries are chosen too broad, stratification may not eliminate all of the confounding effect and residual confounding remains. On the other hand increasing the number of strata by the use of a finer categorization of continuous variables reduces stratum-specific sample sizes. Further, for a larger number of confounders the interpretation of the results may be challenging, especially when the results of the individual strata differ without any obvious pattern.

3.2.2 Multivariable analysis

Another frequently used model-based approach to control or adjust for confounding is regression modeling. With multivariable analysis the association between dependent (disease outcome) and independent variables (treatment or exposure, further risk factors of interest) can be estimated while controlling for the influence of other independent variables (i.e. the potential confounding factors) simultaneously. Which type of regression model (linear, logistic, Cox proportional hazard and Poisson regression) is appropriate depends on the study design and the outcome variable.

If the outcome of interest is continuous, linear regression to control for confounding can be applied. Logistic regression and polytomous logistic regression are standard methods for binary or

polytomous outcome variables, Cox regression for time-to-event data and Poisson regression for count data.

In epidemiological research, the focus lies on the investigation of risk factors for a disease of interest which is frequently measured as a binary variable. In cohort studies and clinical trials usually Cox proportional hazard models that account for the different durations of follow-up are applied to estimate Hazard ratios. For case-control studies typically logistic regression models are used to estimate ORs. For an individual matched case-control study conditional logistic regression should be used to account for the bias induced through the matching process.

If in studies with rare outcome events the number of confounders to control is large, multivariable analysis, like logistic regression may result in biased estimates and increased standard errors (Peduzzi et al., 1996; Cepeda et al., 2003). For ten or more events per confounder, no major problems occurred in logistic regression analysis (Peduzzi et al., 1996). Different model selection strategies (i.e. stepwise selection) to reduce computational complexity or improve the estimate's precision may lead to exclusion of a number of weak but important confounders. Making models parsimonious can result in model misspecification and hence in a biased estimate of the exposure-outcome association (Arbogast and Ray, 2011).

3.3 Summary

Table 3.1: Summary of strength and limitations of various approaches to reduce impact of confounding in design phase

Approach	Strength	Limitation
Randomization	• balances measured and unmeasured covariates	• differences in confounding factors may arise by chance • not always feasible or ethical
Restriction	• simple to perform • applicable to all study designs • very efficient when there is a large pool of individuals	• reduces confounding only by measured factors • possibility of residual confounding if restriction criteria are too imprecise • reduces the pool of available subjects especially when restriction criteria are too narrow • might limits the generalizability of findings
Matching	• easy to communicate • useful for strong confounding factors • increases power in small studies	• reduces confounding only by measured factors • does not allow one to examine the association between outcome and matching factors • accounting for matching in analysis is necessary • may be expensive and time consuming • may limit sample size • potential for overmatching

Table 3.2: Summary of strength and limitations of various approaches to reduce impact of confounding in analysis phase

Approach	Strength	Limitation
Stratification	• simple to perform if a small number of confounders is present • investigation of effect modification possible	• reduces confounding only by measured factors • not a feasible method to control for many confounding factors simultaneously • limited to categorical variables • residual confounding within strata may remain if stratum boundaries are chosen too broad • for a large number of strata the interpretation of the results may be challenging
Multivariable analysis	• adjusting simultaneously for all confounding factors • provides estimates of effects that are mutually unconfounded (i.e. adjusted for each other)	• reduces confounding only by measured factors • rare outcomes and a large number of variables can lead to bias and increased standard errors • complex statistical analysis

Chapter 4

Balancing scores

In the presence of a large number of confounders, summary score methods to reduce high dimensional confounding variables are a popular approach to control for confounding in observational research. Instead of controlling for each measured confounder separately, scoring methods combine the covariates into a single scalar variable. As balancing scores were originally developed in the causal framework (Rosenbaum and Rubin, 1983), we first give a short introduction to the counterfactual setup or the *Rubin Causal Model*. The concept of causal effects and potential outcomes was originally proposed in Neyman et al. (1990) in the context of agricultural experiments and has been further formalized for randomized and non randomized studies by Holland (1986) and Rubin (1974). This concept is nowadays applied in many fields including statistics (Rosenbaum, 2002; Rubin, 2006) and economics (Imbens, 2004; Caliendo and Kopeinig, 2008; Imbens and Wooldridge, 2009).

4.1 Potential outcome framework

The aim is to evaluate the true causal effect of a treatment of interest, T, on an outcome Y. Consider a binary treatment T, with $T = 0$ if an individual is not exposed to the treatment and $T = 1$ if the individual is exposed to the treatment. Suppose that an individual i has two potential outcomes: the first outcome $Y_i(0)$ under the control treatment and second outcome $Y_i(1)$ under the treatment of interest. In reality, typically only one of both outcomes can be observed, namely the outcome under the treatment that the individual actually received. This is also called the *fundamental problem of causal inference* (Holland, 1986) and can be regarded as a missing data problem. If an individual is exposed to the treatment and the outcome $Y_i(1)$ is observed, then $Y_i(0)$ is referred as the counterfactual outcome. So, the outcome observed for individual i under the actual treatment can be written as

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0) = \begin{cases} Y_i(1), & \text{if } T_i = 1 \\ Y_i(0), & \text{if } T_i = 0. \end{cases}$$

If the potential outcomes of an individual i differ, $Y_i(0) \neq Y_i(1)$, the treatment has a causal effect. The individual causal treatment effect can be expressed as difference $Y_i(1) - Y_i(0)$ or some other function such as ratios $Y_i(1)/Y_i(0)$. Since this effect is typically impossible to determine, average treatment effects are considered instead. The two most common quantities of interest are the population average treatment effect (ATE) and the population average treatment effect on the treated (ATT) (Imbens, 2004).

The ATE is defined as population expectation ($E[.]$) of the unit level causal effect τ_i, i.e.

$$\tau_{ATE} = E[\tau_i] = E[Y_i(1) - Y_i(0)] \tag{4.1}$$

If one wants to determine the treatment effect on those who receive or would receive the treatment instead of the whole population which includes individuals that might never be subject to treatment (e.g. one wants to investigate the influence of a drug on individuals with a specific illness) the population average treatment effect on the treated (ATT) is the parameter of interest. The ATT for the subpopulation of treated individuals is given by

$$\tau_{ATT} = E[\tau_i | T_i = 1] = E[Y_i(1) - Y_i(0) | T_i = 1]. \tag{4.2}$$

Note, that this expectation is based on distributions in the subpopulation, i.e. the treated population.

However, the problem of the definitions (4.1) and (4.2) is that the counterfactuals cannot be observed directly. Thus only a reasonably approximation of the counterfactuals can be computed in practice, and this requires some assumptions.

The first assumption is the *stable unit treatment value assumption* (*SUTVA*) (Rubin, 1978), which implies homogeneity of treatment and no interference among individuals, i.e. the potential outcome for an individual is not affected by treatment status of any other individual. This assumption is violated for example, if there is a competition for resources, e.g. only a limited number of treatments is available. For approaches to deal with such violations see Imbens and Wooldridge (2009). In many biomedical applications, this assumption is plausible. So in this section we assume that SUTVA holds.

A second essential assumption for estimating treatment effects is the *strongly ignorable treatment assignment* (*SITA*) assumption introduced by Rosenbaum and Rubin (1983), which is a combination of following two assumptions:

1. *Unconfoundedness* (Imbens and Wooldridge, 2009), also called *conditional independence assumption* (*CIA*) (Caliendo and Kopeinig, 2008):

   $$(Y(0), Y(1)) \perp T | \boldsymbol{X}.$$

   Under the unconfoundedness assumption treatment assignment is independent (denoted with '$\perp$') of potential outcomes conditional on the observed baseline covariates $\boldsymbol{X}$. This implies that all variables that affect treatment assignment have been measured. In Rosenbaum (2002), bias due to measured confounders is denoted by overt bias and bias introduced by unmeasured confounding is called hidden bias.

2. *Common support* or *overlap* assumption:

   $$0 < P(T = 1 | \boldsymbol{X}) < 1,$$

   whereat $P(T = 1|\boldsymbol{X})$ denotes the probability of receiving treatment conditional on the covariates $\boldsymbol{X}$. Under this condition, every individual has a nonzero chance of being assigned to both treatment groups. This ensures that there are both, treated and untreated individuals

for each value of $\boldsymbol{X}$. If gender is a confounder and all male individuals receive the treatment ($P(T = 1|\text{male}) = 1$), no adequate comparison between the treatment groups can be performed. So, this is an essential assumption for comparing average treatment effects between groups.

Average treatment effects in randomized studies

In randomized studies, the random treatment allocation ensures in expectation that the distributions of both observed and unobserved variables are equal in both groups or in other words: the treatment assignment is independent from the potential outcomes $(Y(1), Y(0)) \perp T$, so that

$$E\left[Y(t)|T=1\right] \quad = \quad E\left[Y(t)|T=0\right] = E\left[Y(t)\right] \qquad \forall t \in \{0,1\}\,.$$

Hence, conditioning on T is irrelevant and it follows that the ATT=ATE. Now, the ATE can be identified by:

$$\tau_{ATE} = E\left[Y(1) - Y(0)\right] = E\left[Y(1)\right] - E\left[Y(0)\right] = E\left[Y|T=1\right] - E\left[Y|T=0\right],$$

since

$$\begin{aligned} E\left[Y|\,T=1\right] &= E\left[TY(1) + (1-T)Y(0)|T_i = 1\right] \\ &= E\left[Y(1)|T=1\right] \\ &= E\left[Y(1)\right]. \end{aligned}$$

Similarly, it can be shown that $E\left[Y|T=0\right] = E\left[Y(0)\right]$. This means that on average, observations in the control group can be used to fill the missing potential outcomes for the treated observations.

Average treatment effects in observational studies

Due to the lack of random assignment in observational studies, the treated individuals typically differ from the untreated individuals and $E\left[Y(t)|T=1\right] \neq E\left[Y(t)\right]$ for $t \in \{0,1\}$, hence, $\tau_{ATE} \neq \tau_{ATT}$. But, if *CIA* holds, then

$$E\left[Y(t)|T=1,\boldsymbol{X}\right] = E\left[Y(t)|T=0,\boldsymbol{X}\right] = E\left[Y(t)|\boldsymbol{X}\right] \qquad \forall t \in \{0,1\}\,,$$

and the ATE conditional on $\boldsymbol{X}$ and the ATT conditional on $\boldsymbol{X}$ are equal ($\tau_{ATE|\boldsymbol{X}} = \tau_{ATT|\boldsymbol{X}}$). The conditional ATE can be identified by

$$\tau_{ATE|\boldsymbol{X}} = E\left[Y(1) - Y(0)|\boldsymbol{X}\right] = E\left[Y(1)|\boldsymbol{X}\right] - E\left[Y(0)|\boldsymbol{X}\right] = E\left[Y|T=1,\boldsymbol{X}\right] - E\left[Y|T=0,\boldsymbol{X}\right],$$

since

$$\begin{aligned} E\left[Y|\,T=1,\boldsymbol{X}\right] &= E\left[TY(1) + (1-T)Y(0)|T=1,\boldsymbol{X}\right] \\ &= E\left[Y(1)|T=1,\boldsymbol{X}\right] \\ &= E\left[Y(1)|\boldsymbol{X}\right] \end{aligned}$$

and $E\left[Y|T=0,\boldsymbol{X}\right] = E\left[Y(0)|\boldsymbol{X}\right]$. So, $\tau_{ATE|\boldsymbol{X}}$ can be identified for a subpopulation with a realization $\boldsymbol{X} = \boldsymbol{x}$. The average treatment effect τ_{ATE} can be assessed by averaging $\tau_{ATE|\boldsymbol{x}}$ over the distribution of $\boldsymbol{X}$, if the *overlap* assumption holds, i.e. for every realization $\boldsymbol{X} = \boldsymbol{x}$ there exist

treated and untreated individuals, then

$$\begin{aligned} E_{\boldsymbol{X}}\left[\tau_{ATE|\boldsymbol{x}}\right] &= E_{\boldsymbol{X}}\left\{E\left[Y|T=1,\boldsymbol{X}=\boldsymbol{x}\right]-E\left[Y|T=0,\boldsymbol{X}=\boldsymbol{x}\right]\right\} \\ &= E_{\boldsymbol{X}}\left\{E\left[Y(1)|\boldsymbol{X}\right]-E\left[Y(0)|\boldsymbol{X}\right]\right\} \\ &= E_{\boldsymbol{X}}\left\{E\left[Y(1)|\boldsymbol{X}\right]\right\}-E_{\boldsymbol{X}}\left\{E\left[Y(0)|\boldsymbol{X}\right]\right\} \\ &= E\left[Y(1)\right]-E\left[Y(0)\right] \\ &= \tau_{ATE}. \end{aligned}$$

Causation versus association

In the context of potential outcome framework, what is the difference between causation and association? Causation refers to a treatment effect in the entire population under the two possible treatments, i.e. an causal effect is present on average when $E\left[Y(1)\right] \neq E\left[Y(0)\right]$. In contrast, association refers to a different treatment effect in two mutually exclusive subsets of the population, defined by the actual treatment, i.e. $E\left[Y(1)|\,T=1\right] \neq E\left[Y(0)|\,T=0\right]$ (Hernán and Robins, 2015).

In general, causation does not equal association due to confounding. Under randomization and if 'unconfoundedness' really holds, the association measure equals the causal effect measure. In observational studies this is usually not the case. One way to make observational studies resemble randomized trials is to utilize balancing scores.

4.2 Balancing scores

Let Y denote the outcome, T the binary exposure and $\boldsymbol{X} = (X_1, \ldots, X_k)$ a vector of possible confounders. A summary or confounder score is a mapping $s(\boldsymbol{X}) : \mathbb{R}^k \to \mathbb{R}$ that maps the vector $\boldsymbol{X}$ to a scalar. The score $s(\boldsymbol{X})$ is called:

- *Exposure balancing* if the exposure T and the confounders $\boldsymbol{X}$ are conditionally independent given $s(\boldsymbol{X})$:

  $$\boldsymbol{X} \perp T | s(\boldsymbol{X}),$$

  or equivalently for particular values t, $\boldsymbol{x}$ and c:

  $$P(T = t|\boldsymbol{X} = \boldsymbol{x}, s(\boldsymbol{X}) = c) = P(T = t|s(\boldsymbol{X}) = c).$$

  So, the conditional distribution of $\boldsymbol{X}$ given $s(\boldsymbol{X})$ is the same for the treated ($T = 1$) and for the control units ($T = 0$) (Rosenbaum and Rubin, 1983).

- *Outcome balancing* if the outcome Y and the confounders $\boldsymbol{X}$ are conditionally independent given $s(\boldsymbol{X})$ and T, that is:

  $$\boldsymbol{X} \perp Y | s(\boldsymbol{X}), T,$$

  or equivalently for particular values y, t, $\boldsymbol{x}$ and c:

  $$P(Y = y|\boldsymbol{X} = \boldsymbol{x}, T = t, s(\boldsymbol{X}) = c) = P(Y = y|T = t, s(\boldsymbol{X}) = c).$$

  Thus, the distribution of Y is constant or balanced across the values of $\boldsymbol{X}$ given $s(\boldsymbol{X})$ and T (Rothman et al., 2008).

By using balancing scores, the aim is to reduce the impact of confounding by either breaking the association between the measured confounders and the exposure or by breaking the association between the measured confounders and the outcome. In the next section, two popular scores, i.e. the propensity score and the disease risk score are discussed in detail.

4.3 Propensity score (PS)

The propensity score (PS) is defined as conditional probability of receiving treatment or exposure T given a vector of measured baseline covariates $\boldsymbol{X}$ (Rosenbaum and Rubin, 1983),

$$e(\boldsymbol{X}) = P(T = 1|\boldsymbol{X}).$$

The PS summarizes the multiple potential confounding variables $\boldsymbol{X}$ into a single variable. The PS can also be defined in a more general way, i.e. over the conditional expectation of exposure whereas exposures or treatments can be continuous, ordinal or discrete (Robins et al., 1992; Imbens, 2000). Since the thesis focuses on a binary treatment, the definition given in Rosenbaum and Rubin (1983) is used.

The PS has several useful properties which are stated in the paper of Rosenbaum and Rubin (1983). The next section describes these properties.

4.3.1 Properties of PS

1.) The PS is an exposure balancing score, i.e. $e(\boldsymbol{X})$ *deconfounds* T from $\boldsymbol{X}$:

$$\boldsymbol{X} \perp T|e(\boldsymbol{X}),$$

or equivalently

$$P(\boldsymbol{X}|T = 1, e(\boldsymbol{X})) = P(\boldsymbol{X}|T = 0, e(\boldsymbol{X})).$$

So, conditioning on $e(\boldsymbol{X})$, the distribution of the measured covariates $\boldsymbol{X}$, is the same for treated and untreated individuals. This means that in every subset of individuals with the same PS values, the treated and untreated individuals will have the same distribution in $\boldsymbol{X}$.

2.) The PS is the coarsest balancing score. Any score that is 'finer' as the PS, i.e. $e(\boldsymbol{X}) = f(b(\boldsymbol{X}))$ for some function f, is a balancing score.

This implies that, if a subset of individuals has the same PS and the same values of a component of $\boldsymbol{X}$, then the treated and untreated individuals in this subset will still have the same distribution in $\boldsymbol{X}$. This is a useful property, if one wants to combine PS stratification or matching with some other covariates (e.g. sex).

3.) If the treatment assignment is strongly ignorable given $\boldsymbol{X}$ (*SITA*), then treatment assignment is strongly ignorable given $e(\boldsymbol{X})$, i.e.

$$(Y(0), Y(1)) \perp T|e(\boldsymbol{X}) \qquad \text{and} \qquad 0 < P(T = 1|e(\boldsymbol{X})) < 1.$$

In Rosenbaum and Rubin (1983) it is shown that this holds for any exposure balancing score $s(\boldsymbol{X})$. Moreover, from the combination of SITA and an exposure balancing score follows, that the

expected differences over the observed outcomes in the treated and the control group at $e(\boldsymbol{X})$ is equal to the average treatment effect at $e(\boldsymbol{X})$, that is

$$E\left[Y|T=1, e(\boldsymbol{X})\right] - E\left[Y|T=0, e(\boldsymbol{X})\right] = E\left[Y(1)-Y(0)|e(\boldsymbol{X})\right] = \tau_{ATE|e(\boldsymbol{X})}.$$

The overall average treatment effect can be identified by averaging $\tau_{ATE|e(\boldsymbol{X})}$ over the distribution of $e(\boldsymbol{X})$,

$$E_{e(\boldsymbol{X})}\left[\tau_{ATE|e(\boldsymbol{X})}\right] = E_{e(\boldsymbol{X})}\left\{E\left[Y(1)-Y(0)|e(\boldsymbol{X})\right]\right\} = E\left[Y(1)-Y(0)\right] = \tau_{ATE}.$$

This implies, that treated and untreated individuals with the same PS, but not necessarily with the same values in $\boldsymbol{X}$, can act as control for each other. Therefore, conditioning on the PS, an unbiased estimate of the ATE can be obtained as expected difference in the outcomes between treated and untreated individuals (see also Theorem 3 and 4 in Rosenbaum and Rubin (1983)). Further, matching treated and untreated individuals on their PS, stratification on the PS or covariate adjustment on the PS in the cohort, can produce unbiased estimates of the ATE. More details on how the PS can be applied are given in the next chapter. While the PS has been developed for the linear model in cohort studies it has also been applied for other outcome types and in case-control studies, which is summarized in Chapter 6.

4.3.2 Estimating the PS

In randomized experiments, the PS is known, since the treatment allocation mechanism is known (e.g. flipping a coin for two groups has the assignment probability of 0.5). In observational studies, the PS is unknown and has to be estimated from the data. The estimated PS is even preferred over the true PS, since the estimated PS removes bias by reducing systematic and random imbalances in the observed sample whereas the true PS only reduces systematic imbalance of measured covariates between treatment groups (Rosenbaum and Rubin, 1984; Robins et al., 1992).

The most commonly used method to estimate $e(\boldsymbol{X})$ for a binary treatment from the available data is via logistic regression, with treatment as dependent variable and the measured baseline covariates $\boldsymbol{X}$ included in the model as independent variables, i.e.

$$\hat{P}(T=1|\boldsymbol{X}) = \hat{e}(\boldsymbol{X}) = \frac{\exp\left\{\hat{\alpha}_0 + \hat{\boldsymbol{\alpha}}_X'\boldsymbol{X}\right\}}{1+\exp\left\{\hat{\alpha}_0 + \hat{\boldsymbol{\alpha}}_X'\boldsymbol{X}\right\}},$$

with $\hat{\alpha}_0$ as the intercept and $\hat{\boldsymbol{\alpha}}_X$ as the parameter vector for the covariates $\boldsymbol{X}$. The estimated PS values, ranging from 0 to 1, reflect the probability of receiving treatment, with high treatment probability for a PS value closer to one. Other methods to estimate the PS have been investigated, including neural networks or classification and regression trees (Setoguchi et al., 2008; Lee et al., 2010).

Since the aim of PS methods is to remove bias due to imbalance in baseline covariates and not to predict treatment, standard model development and diagnostic criteria for logistic regression such as c-statistic or stepwise selection procedure might not be appropriate to assess the adequacy of the PS-model (Brookhart et al., 2006; Westreich et al., 2011). Including variables based on c-statistic may lead to inclusion of variables associated only with treatment and can further lead to poor overlap between the treatment groups (Weitzen et al., 2004). Excluding variables based on stepwise selection procedures may lead to exclusion on potential important confounders.

Several guidelines and recommendations about which variables should be included in the PS-model and how adequacy of the PS-model can be assessed have been proposed (Stuart, 2010; Austin, 2011a; Williamson et al., 2012; Patorno et al., 2013). In general, all variables associated with both, treatment and outcome (i.e. true confounders) should be included in the PS-model, otherwise this would lead to bias. However, including all measured baseline covariates, regardless of whether they are true confounders or not, may lead to overfitting problems in the PS-model or loss in efficiency. In simulation studies, it has been shown that including variables in the PS-model that are only related to the treatment but not to the outcome can increase the variance of effect estimates without decreasing bias (Brookhart et al., 2006; Austin et al., 2007a). Moreover, in the presence of unmeasured confounders, including such variables may also increase bias (Brookhart et al., 2010). In contrast, including variables only associated with outcome yielded to an optimal PS-model in terms of bias and efficiency (Brookhart et al., 2006) and in the context of PS-Matching, a larger number of matched pairs can be found (Austin et al., 2007a). In practice, it might be difficult to decide whether a variable is a real confounder or not. Stuart (2010) recommended a liberal variable selection strategy since omitting important confounders can increase bias and therefore harm the validity of treatment effects. In studies with small sample sizes, one should focus on variables associated to the outcome (Brookhart et al., 2006). Further variable selection strategies for the estimation of treatment effects on multiple outcomes and using a multistep algorithm for high-dimensional PS-adjustment is given in Wyss et al. (2013) and Schneeweiss et al. (2009).

If the PS-stratification or PS-Matching is used to control for confounding, balance diagnostics introduced in Section 3.1.3 can be used to check if the PS-model is adequately specified. In Austin (2008) Goodness-of-fit diagnostics for PS-adjustment such as weighted standardized differences which is an extension of the standardized differences (3.4, 3.5) and quantile regression are proposed.

4.4 Disease risk score (DRS)

A second class of confounder scores are disease risk scores (see e.g. Stürmer et al., 2005; Tadrous et al., 2013), that relate outcome to confounders instead of modeling the association of the confounders to the exposure, as the PS does. This approach of combining the confounders was originally proposed by Miettinen (1976b), as a hybrid between multivariable analysis and stratification. Miettinen's *multivariate confounder score* is defined as risk of disease in the absence of exposure, given the observed covariates $\boldsymbol{X}$, i.e. the disease risk under no exposure or treatment, given by $E[Y|\boldsymbol{X}, T = 0]$. A more general definition of outcome scores based on potential outcomes is given by Hansen (2008), referred as *prognostic scores*. The score is defined as a sufficient statistic (i.e. a scalar or vector-based function of covariates) for the potential outcome under the control treatment $Y(0)$ in the sense that conditioning on the score induces independence between $Y(0)$ and $\boldsymbol{X}$. Assuming a generalized linear model, the unidimensional score is given by $E[Y(0)|\boldsymbol{X}]$. In the case of a binary outcome, the score can also be written as $P(Y(0)|\boldsymbol{X})$. Assuming additionally no unmeasured confounders and no effect modification, i.e. $P(Y(1)|Y(0), \boldsymbol{X}) = P(Y(1)|Y(0))$, Miettinen's multivariate confounder score and Hansson's prognostic score coincide. We maintain these assumptions throughout the thesis.

As a retrospective equivalent to the PS, Allen and Satten (2011) proposed a balancing score for case-control studies. The distribution of the confounders, given disease status and that score only depends on the score (Allen and Satten, 2011). Thus, within strata defined by the same score value, or by properly weighting, the confounders are independent of outcome and need not be considered further in the analysis of the main exposure. The association parameters obtained from such an analysis, however, are not standard odds ratios, which may hamper their interpretability.

4.4.1 Properties of DRS

The theoretical properties of the DRS were evaluated by Hansen (2008) who showed that:

1.) The DRS is outcome or prognostic balancing, in the sense that the potential outcome $Y(0)$ under no treatment and the confounders $\boldsymbol{X}$ are conditionally independent given $DRS(\boldsymbol{X})$, that is,

$$\boldsymbol{X} \perp Y(0) | DRS(\boldsymbol{X}).$$

If there is no effect modification then the potential outcome under treatment or exposure and $\boldsymbol{X}$ are conditionally independent given $DRS(\boldsymbol{X})$,

$$\boldsymbol{X} \perp Y(1) | DRS(\boldsymbol{X}).$$

If there is effect modification with an effect modifier $m(\boldsymbol{X})$, e.g. there is an interaction between exposure and a covariate in the outcome model, then

$$\boldsymbol{X} \perp Y(1) | DRS(\boldsymbol{X}), m(\boldsymbol{X}).$$

2.) If $\boldsymbol{X}$ deconfounds the potential outcome under no treatment $Y(0)$ from T ($Y(0) \perp T | \boldsymbol{X}$), also called weak unconfoundedness (Imbens and Wooldridge, 2009), then the DRS deconfounds $Y(0)$ from T as well,

$$Y(0) \perp T | DRS(\boldsymbol{X}).$$

If there is no effect modification and $Y(1) \perp T | \boldsymbol{X}$, then

$$Y(1) \perp T | DRS(\boldsymbol{X}).$$

3.) If there is weak unconfoundedness $Y(t) \perp T | \boldsymbol{X}$ for $t \in \{0, 1\}$ and overlap on the DRS, $0 < P(T = 1 | DRS(\boldsymbol{X})) < 1$, then the average treatment effect under no effect modification can be identified by

$$\begin{aligned} E[Y(1) - Y(0)] &= E_{DRS(\boldsymbol{X})} \{ E[Y(1) - Y(0) | DRS(\boldsymbol{X})] \} \\ &= E_{DRS(\boldsymbol{X})} \{ E[Y(1) | DRS(\boldsymbol{X})] - E[Y(0) | DRS(\boldsymbol{X})] \} \\ &= E_{DRS(\boldsymbol{X})} \{ E[Y | T = 1, DRS(\boldsymbol{X})] - E[Y | T = 0, DRS(\boldsymbol{X})] \}, \end{aligned}$$

since $E[Y(t) | DRS(\boldsymbol{X})] = E[Y(t) | T = t, DRS(\boldsymbol{X})] = E[Y | T = t, DRS(\boldsymbol{X})]$. This implies that the ATE can be estimated by stratification on levels of the DRS. The identification of the ATE under effect modification and the ATT is given in Hansen (2008).

4.4.2 Estimating the DRS

The DRS can be estimated by an appropriate model that links the covariates and the outcome in different ways:

1.) In the full cohort

The score can be estimated in the full cohort using a regression model, including the covariates $\boldsymbol{X}$ and exposure T. The DRS can then be computed as the fitted value from that regression model setting exposure status to unexposed for everyone. For a binary outcome Y that is,

$$\widehat{DRS}(\boldsymbol{X}) = \frac{\exp\left\{\hat{\mu} + \hat{\beta}(T=0) + \hat{\boldsymbol{\gamma}}'\boldsymbol{X}\right\}}{1 + \exp\left\{\hat{\mu} + \hat{\beta}(T=0) + \hat{\boldsymbol{\gamma}}'\boldsymbol{X}\right\}},$$

with $\hat{\mu}$ as the estimated intercept and the estimated parameters $\hat{\beta}$ and $\hat{\boldsymbol{\gamma}}$ for T and the covariates $\boldsymbol{X}$, respectively. This estimation method was proposed for the multivariate confounder score by Miettinen (1976b) and has been further used by Stürmer et al. (2005); Cadarette et al. (2010) and Arbogast and Ray (2011). For a binary outcome the score estimated in the full entire cohort is termed as 'full-cohort DRS' (Arbogast and Ray, 2011).

2.) In the unexposed group only

The DRS can be calculated in the unexposed population only, by fitting a regression model restricted to the unexposed individuals ($T = 0$), i.e. $\widehat{DRS}(\boldsymbol{X}) = E[Y|\boldsymbol{X}]$. Then the fitted values from this regression model are used to calculate the DRS for all individuals in the entire population. For a binary outcome this score is termed as 'unexposed-only DRS' (Arbogast and Ray, 2011) and is equivalent to the prognostic score given in Hansen (2008).

The full-cohort DRS benefits from a larger sample size. Restricting the estimation to the unexposed population might not be reliable if there are only few outcome events. Furthermore, the disease risk for the exposed group has to be extrapolated from the unexposed group which can lead to bias due to model misspecification, especially when the baseline risk in the exposed group differs from the baseline risk in the unexposed group (confounding by indication) (Cadarette et al., 2010). Arbogast and Ray (2011) compared the performance of the full-cohort DRS and the unexposed only DRS in simulations under several confounder-exposure associations. They observed that for moderate and strong confounder-exposure association the DRS estimated in the unexposed population only had slightly more bias.

One drawback of the DRS is that, if the score is estimated in the same sample, adjusting or stratification on the DRS resulted in an overestimation of the significance level (Pike et al., 1979; Cook and Goldman, 1989; Hansen, 2008). This is more pronounced when the confounders and treatment are highly correlated. However, Cook and Goldman (1989) noted that this extreme correlation of exposure and confounders is unlikely to occur in practice and that the same sample estimation problem is also an issue for the PS. But, they also observed in their simulations, that stratification on the PS is less affected by this high confounder-exposure correlation compared to the DRS. An alternative is to estimate the DRS in an external cohort to avoid overfitting problems due to same sample estimation (Hansen, 2008; Glynn et al., 2012; Wyss et al., 2014). However, this approaches might not be useful in practice since they can only be applied when an external cohort or historical data are available, the covariate assessment do not differ in diverse populations or the risk factors for disease do not change over time.

Chapter 5

Use of balancing scores in observational studies

We now describe how balancing scores (PS and DRS) can be used in the design and analysis of observational studies to lessen the impact of confounding. Typically approaches including restriction, matching, stratification, adjusting and weighting are discussed in more detail. A good overview of PS and DRS approaches is given in D'Agostino (1998); Austin (2011a); Heinze and Jüni (2011) and Arbogast and Seeger (2012). Especially PS is frequently used in cohort studies to control for confounding, but little is known about the performance of DRS and also about the performance of both scores in case-control studies (Tadrous et al., 2013). The Sections 5.1 and 5.2 are therefore more focused on the use of the scores (especially the PS) in cohort studies. Although the DRS can be used as matching variable, the application of DRS methods is mainly limited to stratification and adjustment (Arbogast and Ray, 2009).

5.1 Design

5.1.1 Restriction

The idea to restrict the inclusion criteria for individuals to reduce the impact of a confounder, as described in Section 3.1.2, can also be applied to the PS. One can compare the distribution of the PS in the treated and untreated group and exclude those individuals who are outside the range of the common support or exhibit extremely high or low PS-values. The exclusion of individuals based on the PS can be done in several ways.

Some authors suggested to exclude individuals outside the common support symmetrically (e.g. exclusion of 1% or 2% of the individuals who are always or never treated), to restrict the population to individuals that have a realistic chance of treatment (Glynn et al., 2006; Patorno et al., 2013). Another method suggested by Stürmer et al. (2010) is to exclude individuals from mortality analysis based on the PS to control for unmeasured confounding due to frailty. Unmeasured frailty might lead to a 'last resort' treatment or to a 'treatment withheld'. Therefore, excluding individuals contrary to prediction (asymmetric trimming), i.e. treated individuals with low PS and untreated individuals with high PS, might improve the validity of an estimator. Note, that in general matching on the PS automatically excludes the individuals outside the common support.

5.1.2 Matching

In cohort studies treated and untreated individuals with similar propensity score values are matched. The most commonly implemented method is individual pair-matching for a binary treatment ($T \in \{0,1\}$) in which a treated individual i is randomly matched to an untreated individual j with the same PS-values $e(\boldsymbol{X}_i) = e(\boldsymbol{X}_j)$ in the population where the PS was calculated. The covariates of PS matched individuals do not necessarily coincide, $\boldsymbol{X}_i \neq \boldsymbol{X}_j$, but, because of the balancing property of the PS, on average, the distribution of the covariates $\boldsymbol{X}$ is the same for treated ($T = 1$) and untreated ($T = 0$) individuals in the whole matched study population. If the strongly ignorable treatment assignment (SITA) assumption holds and pair matching for the treated and untreated individuals on their exact PS-values is performed then the mean of the matched pair differences of the outcome is an unbiased estimate of the ATE (Rosenbaum and Rubin, 1983, Corollary 4.1).

In practice, exact matching on the PS is rarely possible. Therefore, one has to define the 'closeness' on which the matching should be performed. Generally and as stated in Gu and Rosenbaum (1993), if matching to control for confounding is performed, one has to choose an appropriate *distance* measure, a matching *algorithm* and the *structure* of the matched samples - which means the matching ratio. In Chapter 3 (Section 3.1.3) several options concerning distance, algorithm and matching ratios are discussed in a broader context. This section includes some additions referring in particular to the PS.

The difference of the PS can be used as distance measure, that is

$$d_{ij} = |e(\boldsymbol{X}_i) - e(\boldsymbol{X}_j)|,$$

to match treated individuals i ($i = 1, \ldots, n$) to untreated individuals j ($j = 1, \ldots, m$). This can also be combined with the Mahalanobis distance (3.1), by matching on the Mahalanobis distance within propensity score calipers as suggested in Rosenbaum and Rubin (1985b). In Gu and Rosenbaum (1993) the performance of different distance measures has been investigated by simulations in combination with several matching algorithm (nearest neighbor and optimal matching) and matching ratios (1:M and full matching). For a large number of covariates, they observed better covariate balance if the PS is used as distance measure compared with the Mahalanobis distance matching or matching on the Mahalanobis distance within propensity score calipers. Similar, better performance concerning balance was observed for full matching compared with 1:M matching, and no substantial differences between the balance under optimal and nearest neighbor matching were observed. Austin (2014) compared several 1:1 PS-matching algorithms including caliper matching and matching with replacement and concluded that in most situations caliper matching without replacement is most suitable for matching in cohort studies.

Several simulation studies have been performed to investigate the optimal caliper width. Rosenbaum and Rubin (1985b) recommended a caliper width of 0.25 times the standard deviation of the logit of the PS. They also suggested to match on the logit of the PS, that is

$$\text{logit}(\hat{e}(\boldsymbol{X})) = \log[(1 - \hat{e}(\boldsymbol{X}))/\hat{e}(\boldsymbol{X})],$$

as the distribution of $\text{logit}(\hat{e}(\boldsymbol{X}))$ better approximates the normal distribution. Austin (2011b) recommended a tightened caliper of 0.2 times the standard deviation of the logit of the PS. He also observed a smaller impact of the caliper width on the performance of the estimation of risk

differences or mean differences, when there are only binary matching factors. More recently, Lunt (2014) investigated the influence of the caliper width and the order in which matches are made on quality of matching using a 1:1 nearest-neighbor matching without replacement. He observed, that matching without a caliper can result in substantial bias, especially when a small pool of unexposed individuals is available for a match and there is strong exposure-confounder association (i.e. poor overlap).

To assess the quality of the matching, i.e. if the covariates are balanced between the matched groups, the diagnostic criteria described in Section 3.1.3. can be applied. To check if the overlap condition holds (i.e. every individual has a chance of being assigned to treatment), one can simply compare the distribution of the PS (i.e. the conditional probability of receiving treatment given $\boldsymbol{X}$) between treated and untreated individuals. This can be done graphically by using for example boxplots, histograms or density plots.

In practice, the same matching methods developed for the PS can be applied to the DRS. However, further methodical research concerning matching on DRS is needed to understand the resulting properties. It is also possible to combine both scores in the matching. While the PS focuses on the relationship between the confounders and the treatment, the DRS focuses on the relationship between the confounders and the outcome. Therefore, matching on the PS tends to achieve covariate balance rather for confounders associated with treatment and matching on the DRS tends to achieve covariate balance rather for confounders associated with outcome. In a simulation study, Leacy and Stuart (2013) investigated several matching approaches based on balancing scores, including matching on the PS, matching on the DRS unexposed-only and matching on both scores based on the Mahalanobis distance. They observed better performance for the matching methods that jointly use the PS and the DRS especially when one of the score models was misspecified.

5.2 Analysis

5.2.1 Stratification

As mentions earlier in Chapter 3 (Section 3.2.1), stratification or subclassification becomes challenging when the number of confounders one wants to control for is large. To overcome this dimensionality problem, stratification on a balancing score can be performed.

Using stratification or subclassification on the PS, groups of individuals are sampled so that all individuals within a group have the same PS-values. If individuals are stratified into mutually exclusive subsets based on their estimated PS and the PS is correctly specified, the distribution of the measured baseline covariates will be similar between the treated and untreated individuals within the same stratum. The treatment effects within the strata can then be estimated and a pooled treatment effect across the strata can be calculated. If the strongly ignorable treatment assignment (SITA) assumption holds, an unbiased estimate of the ATE can be obtained for linear treatment effects by calculating weighted averages of the stratum-specific treatment effects whereat the weights are given by the fraction of the population within PS-strata (Rosenbaum and Rubin, 1983, Corollary 4.2). For $j = 1, \ldots, J$ strata, an estimate for the stratum-specific ATE is

given by

$$\hat{\tau}_{S_j} = \frac{1}{n_{1j}} \sum_{i=1}^{n_{1j}} Y_{ij} - \frac{1}{n_{0j}} \sum_{i=1}^{n_{0j}} Y_{ij},$$

with n_{1j} individuals receiving the treatment and n_{0j} individuals not receiving the treatment in the j-th stratum. The overall ATE can be estimated by

$$\hat{\tau}_S = \sum_{j=1}^{J} \frac{n_{1j} + n_{0j}}{n} \hat{\tau}_{S_j}.$$

As required for continuous variables, one has to divide the PS into categories. Rosenbaum and Rubin (1984) showed, that dividing the PS into five equally sized strata, i.e. PS-quintiles, removes approximately 90% of the bias due to confounding by the measured covariates. In practice, the strata may not be exactly homogenous in the PS which can lead to residual bias due to remaining imbalances in the covariates (Lunceford and Davidian, 2004). Increasing the number of strata can lead to an improved bias reduction, but residual confounding can also remain if the PS-model is not adequately specified. Thus, checking the balance within the strata and if necessary a modification of the PS-model (e.g. including additional covariates or adding interactions) is recommended (Rosenbaum and Rubin, 1984; Lunceford and Davidian, 2004). Additionally, stratification can also be combined with model-based adjustments, i.e. within-stratum regression adjustment, to further reduce the impact of residual confounding (Rosenbaum and Rubin, 1984; Lunceford and Davidian, 2004).

Stratification on the DRS can be performed the same way as described for the PS, i.e. the individuals are divided into mutually exclusive subsets based on their estimated DRS. Within the strata, the individuals have the same baseline risk for disease in the absence of the treatment or exposure of interest. This approach is straightforward and allows one the possibility to investigate possible effect modification across the levels of baseline disease risk. As suggested by Strauss (1998), stratification on the DRS provides a useful descriptive or graphical comparison between exposure groups.

5.2.2 Inverse probability weighting (IPW)

Here, weights based on the PS are used to create a sample, i.e. a 'pseudo-population', in which the distribution of measured baseline covariates is independent of treatment assignment. This was originally proposed by Rosenbaum (1987) as model-based direct standardization. Using IPW, the weights equate to the inverse of the treatment probability for treated individuals and to the inverse of one minus the treatment probability for untreated individuals, that is

$$W = \frac{1}{e(\mathbf{X})} T + \frac{1}{1 - e(\mathbf{X})} (1 - T) = \begin{cases} \frac{1}{e(\boldsymbol{X})}, & \text{if } T_i = 1 \\ \frac{1}{1-e(\boldsymbol{X})}, & \text{if } T_i = 0. \end{cases}$$

Imbens (2004) showed, that the ATE can be written as

$$\tau_{ATE} = E\left[\frac{TY}{e(\boldsymbol{X})} - \frac{(1-T)Y}{1-e(\boldsymbol{X})}\right],$$

since

$$E\left[\frac{TY}{e(\boldsymbol{X})}\right] = E\left[Y(1)\right] \qquad \text{and} \qquad E\left[\frac{(1-T)Y}{1-e(\boldsymbol{X})}\right] = E\left[Y(0)\right]. \tag{5.1}$$

Using that $TY = TY(1)$, for the first equation in (5.1) it follows

$$\begin{aligned} E\left[\frac{TY}{e(\boldsymbol{X})}\right] &= E\left[\frac{TY(1)}{e(\boldsymbol{X})}\right] = E_X\left\{E\left[\frac{TY(1)}{e(\boldsymbol{X})}\right]|\boldsymbol{X}\right\} = E_X\left\{\frac{E\left[T|\boldsymbol{X}\right]E\left[Y(1)|\boldsymbol{X}\right]}{e(\boldsymbol{X})}\right\} \\ &= E_X\left\{\frac{e(\boldsymbol{X})E\left[Y(1)|\boldsymbol{X}\right]}{e(\boldsymbol{X})}\right\} = E_X\left\{E\left[Y(1)|\boldsymbol{X}\right]\right\} = E\left[Y(1)\right], \end{aligned}$$

by using the law of iterated expectation in the second line and the unconfoundedness assumption in the third line. The same argument holds for $E\left[Y(0)\right]$. An estimator for the ATE is given by,

$$\hat{\tau}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{e(\boldsymbol{X})} - \frac{(1-T_i)Y_i}{1-e(\boldsymbol{X})}.$$

The estimation of the ATE via the difference of weighted averages of the outcomes for treated and untreated individuals can be problematic since the weights do not necessarily add up to one (Imbens, 2004). They only add up to one in expectation. Other estimators, for example normalizing the weights to unity are summarized in Imbens (2004); Lunceford and Davidian (2004). A further disadvantage is, that the IPW-estimator is more affected by outliers compared to other PS-methods (e.g. stratification). Individuals with PS-values close to zero or one can have large weights. The IPW-estimator gives more weight to those individuals and thus gets imprecise (Imbens and Wooldridge, 2009).

There exist several further applications of IPW-methods, for example a combination of weighting and regression adjustment is described in Hirano and Imbens (2001) and IPW, in the context of marginal structural models for estimation of time-varying exposure or treatments effects is given in Robins et al. (2000).

5.2.3 Adjustment in regression models

Another analysis based approach to control for confounding is to use the PS as adjustment variable in a regression model in which the outcome Y is the dependent variable and exposure T and the PS are included as independent variables. In contrast to the other approaches, i.e. matching and stratification which are focused on mimicing the features of a randomized design, a functional form of the association of $e(\boldsymbol{X})$ and Y is assumed. The functional form depends on the outcome type, e.g. for continuous outcomes a linear model and for binary outcomes a logistic regression model is used (see also Section 3.2.2). If the strongly ignorable treatment assignment (SITA) assumption holds and for a linear relationship of $e(\boldsymbol{X})$ and Y, i.e.

$$E\left[Y|e(\boldsymbol{X}), T=t\right] = \mu_t + \gamma_t e(\boldsymbol{X}) \qquad \forall t \in \{0,1\},$$

then an unbiased estimator for the ATE at $e(\boldsymbol{X})$ is given by

$$\hat{\tau}_{A|e(\boldsymbol{X})} = (\hat{\mu}_1 - \hat{\mu}_0) + (\hat{\gamma}_1 - \hat{\gamma}_0)e(\boldsymbol{X}).$$

The overall ATE can be estimated by

$$\hat{\tau}_A = (\hat{\mu}_1 - \hat{\mu}_0) + (\hat{\gamma}_1 - \hat{\gamma}_0)\bar{e}(\boldsymbol{X}),$$

with $\bar{e}(\boldsymbol{X}) = 1/n\sum_{i=1}^{n} e(\boldsymbol{X}_i)$ (Rosenbaum and Rubin, 1983, Corollary 4.3).

Adjusting for the PS in a regression model instead of the confounders $\boldsymbol{X}$ itself is beneficial if the dimension of $\boldsymbol{X}$ is large and the disease outcome is rare ((Braitman and Rosenbaum, 2002; Cepeda et al., 2003)). In a simulation study for binary outcome and with seven or fewer events per confounder, adjusting for the PS resulted in less biased, more robust and precise estimates compared with a logistic regression model including the confounders. However, better performance was given for the latter model if there were eight or more events per confounder (Cepeda et al., 2003).

The DRS can be used as adjustment variable in a regression model similar to the PS adjustment, i.e. Y is the dependent variable and exposure T and the DRS are included as independent variables. The score can be used on a continuous scale or in categories. Adjusting on DRS categories (e.g. quintile or deciles) avoids assuming a functional form of the DRS and the outcome model and enables to investigate effect modification as well (Arbogast et al., 2008).

For the models that estimate the PS or DRS, the precision of the estimates are of minor importance. Therefore a larger number of possible confounders can be included in the score estimation models compared with conventional multivariable outcome models. Further, DRS (as well as PS) reduces the dimensionality of analysis and provides a more objective way for variable selection than traditional selection procedures (Arbogast et al., 2008). Stürmer et al. (2005) investigated this issue only in an empirical example and observed no major differences when adjusting for the PS, for the full-cohort DRS or a conventional multivariable outcome model was used to estimate the exposure effect. In their study the exposure and disease probabilities were approximately 20%. Stürmer et al. (2005) suggested that for rare disease, the DRS might be advantageous over conventional multivariable outcome models. But further work is needed to investigate this issue.

Chapter 6

Analytic bias for conditional treatment effect

We illustrate analytic bias calculations and numerically examples for the conditional treatment effect when summary scores are used to control for confounding in observational studies. This chapter is based on the paper Pfeiffer and Riedl (2014), submitted for publication.

6.1 Motivation

To estimate the conditional treatment or exposure effect in observational studies, while controlling for a large number of confounders, balancing scores are frequently applied. Especially the use of the PS has greatly increased since 1999 (Stürmer et al., 2006). Compared to the PS, the application of the DRS is less frequent, however, the use of this score has also increased in the last decade (Tadrous et al., 2013).

Use of balancing scores in cohort analysis

General recommendations for the application of balancing scores in cohort analysis, i.e. settings that favor PS and settings where DRS methods might be better applicable, are given in Arbogast and Seeger (2012). In summary, DRS methods can be useful for the investigation of exposure effects with more than two exposure categories. Although the PS can be applied for several exposure categories (Robins et al., 1992; Imbens, 2000), PS methods might not be practicable in such settings especially when some categories are infrequent. In contrast, the PS might be preferable over the DRS for the investigation of multiple outcomes and when the exposure is common and the disease is rare. Systematic reviews about the application of balancing scores in medical research have shown that the scores are frequently used in cohort analysis as adjustment variable in a regression model (Shah et al., 2005; Tadrous et al., 2013). A comparison of the estimators obtained from PS methods and from conventional multivariable regression models revealed that, in general, the results of both methods were similar. Nevertheless, adjusting on the PS resulted in estimators slightly closer to the null (Shah et al., 2005). This issue was further investigated via simulation study by Austin et al. (2007b). They showed that adjusting for the PS in logistic or Cox proportional hazards models for cohort data can result in estimates of conditional treatment effects (odds ratios (ORs) or hazards ratios (HRs)) that are biased towards the null. Only when there was no treatment effect was the null association estimated without bias. No bias was seen for rate ratio estimates

(RRs) when Poisson regression was used to analyze the data (Austin et al., 2007b). Arbogast and Ray (2011) investigated the performance of the DRS as adjustment variable in cohort analysis in simulations. For an adequate number of events per confounder, including the DRS (calculated in the full cohort) as an adjustment variable in a Poisson regression model resulted in no substantial bias in all scenarios.

Use of balancing scores in case-control analysis

Propensity score methods to control for confounding were originally designed for cohort analysis. However, propensity scores have been used in a few case-control settings as well. For example, to assess the relationship between receipt of acid suppression therapy and nosocomial C difficile infection (used as a binary outcome) Howell et al. (2010) included the PS as an adjustment variable for a cohort analysis using unconditional logistic regression, and also used it as a matching variable to create a matched case-control data set that was analyzed with conditional logistic regression. To capture the effect of confounders not included in the matching process, Michalia et al. (2012) included the PS in a conditional logistic regression model as adjustment variable in an analysis that investigated the impact of packed red blood cell transfusion on the occurrence of bloodstream infections. Other examples in which the PS was included as an adjustment variable in the analysis of case-control data include Alarcón et al. (2007) and El-Serag et al. (2009), and examples where the PS was used as a matching variable for case-control studies include Etminan et al. (2010); Thillemann et al. (2010) and Modén et al. (2010).

Scoring in case-control studies can be challenging and additional considerations, especially concerning PS estimation are required. Some properties for the use of PS in case-cohort and case-control studies are given in Joffe and Rosenbaum (1999) and Månsson et al. (2007). In case-control studies, typically all cases and only a fraction of controls are sampled from the population. This sampling can produce a distortion of the relation between exposure and the covariates and therefore the estimated PS can be biased which can further lead to residual confounding (Månsson et al., 2007). In simulations, Månsson et al. (2007) investigated several PS estimation approaches in case-cohort and case-control studies, including the estimation of the PS in the full cohort, in a subcohort (i.e. a random sample from the full cohort), estimated the PS in a sample resulted from weighted (weights are based on the sampling fraction of controls) and unweighted case-control sampling and from controls only. The association of exposure to the outcome was estimated by stratification on the quintiles of the PS, adjusting for the PS as continuous variable in a logistic regression model and using weighting methods. Matching on the PS was not considered in their simulation. In nearly all variants of PS estimation residual confounding was observed if an exposure-outcome effect was present. However, the observed magnitude of residual confounding was modest and nearly unbiased results were observed if the scores were estimated in the subcohort and in the unweighted case-control sample and stratification on the PS or logistic regression adjusting for the PS was used.

In contrast to the PS, it is more intuitive to use the DRS as matching variable in a case-control study to balance the confounders between cases and controls. However, there has been little theoretical work about the performance of this score in such setting (Arbogast and Ray, 2009; Tadrous et al., 2013). Examples for applications of the DRS in case-control studies are given in Tadrous et al. (2013).

We focus now on what happens when the PS or DRS is used as an adjustment variable in the analysis of cohort or matched case-control data to obtain estimates of conditional exposure effects. Since, multivariable regression methods (e.g. logistic regression) including the PS as covariate are frequently used methods to reduce the impact of confounding in observational research, we investigated the amount of bias due to this model-misspecification analytically. In the cohort setting, we calculated the asymptotic bias for logistic regression and Poisson regression if adjustment on the PS or the DRS is performed. In the case-control setting, exact matching on the true summary scores with matching ratio 1:1 is investigated. We study the behavior of the estimates when conditional logistic regression and standard logistic regression, adjusting for the matching variables is used to analyze the matched data.

Additionally, we verified our analytic calculations by simulation studies. In the analytic calculations the true scores are used. In simulations, additionally the estimated scores are investigated.

6.2 Models

In our calculations, Y denotes the disease outcome and T a binary exposure of interest with $T = 1$ for exposed and $T = 0$ for unexposed individuals. The potential confounders are investigated as binary variables and denoted with $\boldsymbol{X} = (X_1, \ldots, X_k)$. The probability $P(T = 1|\boldsymbol{X})$ is a function of some or all of the components of $\boldsymbol{X}$.

True outcome model F

The conditional expectation, i.e. the true outcome model of Y given T and $\boldsymbol{X}$ with data generating probability distribution F is given by

$$E_F\left[Y|T, \boldsymbol{X}, \boldsymbol{\theta}\right] = h(\mu + \beta T + \boldsymbol{\gamma'X}), \tag{6.1}$$

with h as known function and $\boldsymbol{\theta} = (\mu, \beta, \boldsymbol{\gamma})$ as parameter vector with dimension k+2, whereat μ denotes the intercept, β the parameter for exposure T and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)$, the parameter vector of the covariates $\boldsymbol{X}$. We assume that the distribution of Y follows a distribution in the exponential family, i.e.

$$f(y|\eta, \phi) = \exp\left\{\frac{y\eta - b(\eta)}{a(\phi)} + c(y, \phi)\right\}, \tag{6.2}$$

with a(.), b(.) and c(.) as some specific functions and ϕ is a dispersion parameter. η is called the canonical parameter and is related to the moment of Y through (McCullagh and Nelder, 1989)

$$E_F(Y) = b'(\eta) \qquad \text{and} \qquad \text{Var}_F(Y) = a(\phi)b''(\eta).$$

Letting $\eta = \mu + \beta T + \boldsymbol{\gamma'X}$, h in (6.1) is given by $h(x) = b'(x)$.

False outcome model G

If a summary score $s(\boldsymbol{X})$ is used to adjust for confounding instead of $\boldsymbol{X}$, the conditional expectation, i.e. the misspecified outcome model of Y given T and $\boldsymbol{X}$ under data generating probability distribution G is given by

$$E_G\left[Y|T, \boldsymbol{X}, \boldsymbol{\theta}_s\right] = h(\mu_s + \beta_s T + \gamma_s s(\boldsymbol{X})), \tag{6.3}$$

where h denotes the same function as in (6.1) and $\boldsymbol{\theta}_s = (\mu_s, \beta_s, \gamma_s)$ is a parameter vector with dimension three. Again, we assume that the misspecified probability model of Y has a correctly specified distributional form (6.2). As summary scores $s(\boldsymbol{X})$ (introduced in Chapter 4) we consider the PS, i.e. the conditional probability of exposure given $\boldsymbol{X}$,

$$e(\boldsymbol{X}) = P(T = 1|\boldsymbol{X})$$

and the DRS, i.e. the disease risk under no exposure given $\boldsymbol{X}$,

$$DRS(\boldsymbol{X}) = E\left[Y|\boldsymbol{X}, T = 0\right].$$

We assume that there is no unmeasured confounding and that the score models are correctly specified.

General approach to bias calculation

We now analytically compute the asymptotic bias in maximum likelihood estimates (MLEs) β_s when the mean model (6.3) is used instead of the true mean model (6.1) for various study designs. More precisely, we are interested in the bias $\beta_s - \beta$, whereat the estimator of β_s is obtained by maximizing

$$l(y_1, \ldots, y_n; \boldsymbol{\theta}_s, \phi) = \prod_{i=1}^{n} f_G(y_i|T_i, s(\boldsymbol{X}_i); \boldsymbol{\theta}_s, \phi). \tag{6.4}$$

White (1982) has shown that the MLE under model misspecification converges to a parameter value $\boldsymbol{\theta}_s^*$ which minimizes the Kullback-Leibler divergence between the misspecified model G with respect to the true model F,

$$\boldsymbol{\theta}_s^* = \underset{\boldsymbol{\theta}_s}{\operatorname{argmin}}\ \mathrm{KL}(F||G), \tag{6.5}$$

with

$$\mathrm{KL}(F||G) := E_F\left[\log\frac{P_F(Y|\boldsymbol{\theta})}{P_G(Y|\boldsymbol{\theta}_s)}\right] = E_F\left[\log(P_F(Y|\boldsymbol{\theta})) - \log(P_G(Y|\boldsymbol{\theta}_s))\right],$$

where the expectation is taken with respect to the true model F. Under certain regularity conditions for F and G which are fulfilled for distributions from the exponential family, one can obtain $\boldsymbol{\theta}_s^*$ by solving

$$\frac{\partial}{\partial\boldsymbol{\theta}_s} E_F\left[\log(P_F(Y|\boldsymbol{\theta})) - \log(P_G(Y|\boldsymbol{\theta}_s))\right] = -E_F\left[\frac{\partial}{\partial\boldsymbol{\theta}_s}\log(P_G(Y|\boldsymbol{\theta}_s))\right] = 0. \tag{6.6}$$

6.3 Bias calculations for cohort analysis

We first assume that a cohort is analyzed to estimate the treatment effect and its standard errors. Instead of the true mean model (6.1) the misspecified mean model (6.3) is used when maximizing the likelihood based on the exponential distribution (6.2). We assume below that the score $s(\boldsymbol{X})$ is not equal to $\boldsymbol{\gamma}'\boldsymbol{X}$, for $\boldsymbol{\gamma}$ given in (6.1) which would be equivalent to adjusting for the true confounders in the correctly specified form.

THEOREM 1. *If $\beta = 0$ in (6.1), then the MLE $\hat{\beta}_s$ based on mean model (6.3) when the distribution of Y is in the exponential family (6.2) consistently estimates zero if the score in (6.3) is the PS, i.e. $s(\boldsymbol{X}) = e(\boldsymbol{X})$. If the function h in (6.1) is the identity function, i.e. under a linear model, using the PS in mean model (6.3) leads to consistent estimates of β, i.e. $\beta_s = \beta$, even when $\beta \neq 0$.*

Proof. Under the exponential family, differentiating the likelihood (6.4) under the misspecified mean model (6.3) yields the set of equations

$$\frac{\partial f}{\partial \mu_s} : E_X E_{T|X} E_{Y|X,T}\Big\{Y - h(\mu_s + \beta_s T + \gamma_s s(\boldsymbol{X}))\Big\} = 0 \tag{6.7}$$

$$\frac{\partial f}{\partial \beta_s} : E_X E_{T|X} E_{Y|X,T}\Big\{T\big[Y - h(\mu_s + \beta_s T + \gamma_s s(\boldsymbol{X}))\big]\Big\} = 0 \tag{6.8}$$

$$\frac{\partial f}{\partial \gamma_s} : E_X E_{T|X} E_{Y|X,T}\Big\{s(\boldsymbol{X})\big[Y - h(\mu_s + \beta_s T + \gamma_s s(\boldsymbol{X}))\big]\Big\} = 0, \tag{6.9}$$

with expectation under the true probability distribution F as $E_F[\cdot] = E_X E_{T|X} E_{Y|T,X}[\cdot]$. When $\beta = 0$, the distribution of Y under F does not depend on T, i.e. $E_{Y|X,T}[\cdot] = E_{Y|X}[\cdot]$. Assume now that $\beta_s = 0$ in the misspecified mean model. Then equations (6.8) and (6.9) reduce to

$$\frac{\partial f}{\partial \beta_s} : E_X E_{Y|X}\Big\{E_{T|X}(T)\big[Y - h(\mu_s + \gamma_s s(\boldsymbol{X}))\big]\Big\} = 0 \tag{6.10}$$

$$\frac{\partial f}{\partial \gamma_s} : E_X E_{Y|X}\Big\{s(\boldsymbol{X})\big[Y - h(\mu_s + \gamma_s s(\boldsymbol{X}))\big]\Big\} = 0. \tag{6.11}$$

Subtracting (6.11) from (6.10) yields

$$E_X E_{Y|X}\Big\{\big[E_{T|X}(T) - s(\boldsymbol{X})\big]\big[Y - h(\mu_s + \gamma_s s(\boldsymbol{X}))\big]\Big\} = 0.$$

Thus if the confounder score is the PS, i.e. $s(\boldsymbol{X}) = e(\boldsymbol{X})$, then as $E_{T|X}(T) = e(\boldsymbol{X})$, the above equation is satisfied.

Special case: linear regression

If the function h in equation (6.1) is the identity function then we obtain a standard linear regression model. In that case, after taking the conditional expectation of Y, $E_{Y|X,T}[\cdot]$, equations (6.7) - (6.9) correspond to

$$\frac{\partial f}{\partial \mu_s} : E_X E_{T|X}\Big\{\mu + \beta T + \boldsymbol{\gamma}'\boldsymbol{X} - \mu_s - \beta_s T - \gamma_s s(\boldsymbol{X})\Big\} = 0 \tag{6.12}$$

$$\frac{\partial f}{\partial \beta_s} : E_X E_{T|X}\Big\{T(\mu + \beta T + \boldsymbol{\gamma}'\boldsymbol{X} - \mu_s - \beta_s T - \gamma_s s(\boldsymbol{X}))\Big\} = 0 \tag{6.13}$$

$$\frac{\partial f}{\partial \gamma_s} : E_X E_{T|X}\Big\{s(\boldsymbol{X})(\mu + \beta T + \boldsymbol{\gamma}'\boldsymbol{X} - \mu_s - \beta_s T - \gamma_s s(\boldsymbol{X}))\Big\} = 0. \tag{6.14}$$

Using that $E_{T|X}\left[T^2\right] = E_{T|X}\left[T\right] = e(\boldsymbol{X})$, we solve equations (6.12) - (6.14) explicitly for $\boldsymbol{\theta}_s = (\mu_s, \beta_s, \gamma_s)$. After some algebra, it can be seen that β_s satisfies the equation

$$\beta_s \left[Cov(s(\boldsymbol{X}), T)^2 - Var(T)Var(s(\boldsymbol{X}))\right] = \beta \left[Cov(s(\boldsymbol{X}), T)^2 - Var(T)Var(s(\boldsymbol{X}))\right] + \sum_{i=1}^{k} \gamma_i \left[Cov(s(\boldsymbol{X}), X_i)Cov(s(\boldsymbol{X}), T) - Var(s(\boldsymbol{X}))Cov(T, X_i)\right]$$

Note that $Cov(s(\boldsymbol{X}), T) = Cov(s(\boldsymbol{X}), e(\boldsymbol{X}))$ and $Cov(X_i, T) = Cov(X_i, e(\boldsymbol{X}))$. Thus as $Cov(s(\boldsymbol{X}), T)^2 < Var(T)Var(s(\boldsymbol{X}))$, we obtain

$$\beta_s = \beta + \frac{\sum_{i=1}^{k} \gamma_i \left[Cov(s(\boldsymbol{X}), X_i)Cov(s(\boldsymbol{X}), e(\boldsymbol{X})) - Var(s(\boldsymbol{X}))Cov(e(\boldsymbol{X}), X_i)\right]}{Cov(s(\boldsymbol{X}), e(\boldsymbol{X}))^2 - Var(T)Var(s(\boldsymbol{X}))}$$

When $s(\boldsymbol{X}) = e(\boldsymbol{X})$ then

$$Cov(s(\boldsymbol{X}), X_i)Cov(s(\boldsymbol{X}), e(\boldsymbol{X})) - Var(s(\boldsymbol{X}))Cov(e(\boldsymbol{X}), X_i) = 0, \qquad i = 1, \ldots, k$$

and thus $\beta_s = \beta$. When the score used in the linear model is the DRS, then estimates β_s are trivially unbiased, as mentioned above, as $s(\boldsymbol{X})$ equal to $\boldsymbol{\gamma}'\boldsymbol{X}$, is equivalent to adjusting for the true confounders in the correctly specified form. For general scores s the bias in β_s depends on all the possible correlations between $\boldsymbol{X}$, s and $e(\boldsymbol{X})$ and the strength of confounding, $\gamma_i, i = 1, \ldots, k$. □

However, when h in equation (6.1) is not the identity function and $\beta \neq 0$, then adjusting for the PS in the mean model generally results in biased estimates of the exposure effect, $\beta_s \neq \beta$. Additionally, when $s(\boldsymbol{X})$ in model (6.3) is not the PS then the estimate β_s is typically biased even when $\beta = 0$. This implies that adjusting for a monotone transformation of the PS also results in biased estimates β_s.

We now solve the equations (6.7) - (6.9) numerically for Poisson and logistic regression models to assess the magnitude of the bias when the PS or the DRS is used as confounder score.

6.4 Numerical examples for cohort analysis

In our numerical calculations we assume that the relationship of the confounders $\boldsymbol{X}$ to the exposure T is given by the logistic regression model,

$$\text{logit}(P(T = 1|\boldsymbol{X})) = \alpha_0 + \boldsymbol{\alpha}'\boldsymbol{X},$$

which is also used as the true PS.

For Poisson count outcome, $Y \sim \text{Poi}(\lambda)$, the true outcome model (6.1) is given by

$$E[Y|T, \boldsymbol{X}] = \lambda = \exp\{\mu + \beta T + \boldsymbol{\gamma}'\boldsymbol{X}\},$$

and the DRS is computed as $DRS(\boldsymbol{X}) = \exp\{\mu + \boldsymbol{\gamma}'\boldsymbol{X}\}$.

If Y is a binary outcome, $Y \sim \text{Bernoulli}(p)$, the true outcome model (6.1) is given by a logistic regression model,

$$\text{logit}(P(Y = 1|\boldsymbol{X}, T)) = \text{logit}(p) = \mu + \beta T + \boldsymbol{\gamma}'\boldsymbol{X},$$

which is further used to calculate the DRS by setting $T = 0$, i.e. $DRS(\boldsymbol{X}) = P(Y = 1|\boldsymbol{X}, T = 0)$.

We calculated the bias when the DRS or the PS is used as adjusting variable on both scales, i.e. $e(\boldsymbol{X})$ and $\text{logit}(e(\boldsymbol{X}))$, $DRS(\boldsymbol{X})$ and $\log(DRS(\boldsymbol{X}))$ for Poisson regression and $\text{logit}(DRS(\boldsymbol{X}))$ for logistic regression.

6.4.1 Scenarios and settings

Various scenarios (i.e. various confounder associations $\boldsymbol{X}$ to T and Y) under different settings (i.e. varying disease, exposure and confounder prevalences and different exposure effects) are considered for the bias calculations. The first scenario is based on the simulation study from Austin et al. (2007b), where the performance of the PS on the estimation of conditional exposure effects is investigated. The second scenario is motivated by Arbogast and Ray (2011) who performed a simulation study to investigate the performance of DRS and PS in comparison with conventional multivariable outcome regression. Here, we briefly summarize the data generating process and parameter setting from these simulation studies, which we also used for our bias calculations. More details are given in Austin et al. (2007b) and Arbogast and Ray (2011).

Austin et al. (2007b) performed a simulation study based on nine independent binary covariates with prevalence of 50% for each covariate. A binary exposure was generated using a logistic regression, binary outcomes and count outcomes were generated using a logistic and a Poisson model, respectively. The intercepts of the models were chosen to achieve an exposure prevalence of 50% ($\alpha_0 = -3.5$) and a disease prevalence of 25% under the null hypothesis ($\mu = -5$ in logistic regression and $\mu = -3$ in Poisson regression). The associations of the covariates to exposure and outcome are given in Table (6.1, scenario Austin). To estimate the conditional exposure effect, the PS was used as an adjustment variable on probability scale instead of the covariates. A logistic regression was used for binary outcomes and a Poisson regression for count outcomes. In the simulations, the true exposure effect β varied between -2.3 and 2.3. In our bias calculations, the same parameter values described above are used. We additionally calculated the bias when the DRS is used as adjusting variable.

Arbogast and Ray (2011) performed a simulation study based on 10 independent binary covariates with 10% prevalence. A binary disease outcome with fixed period of follow-up was simulated using a Poisson model with baseline disease rate (all covariate values and exposure zero) of 0.01 per person and a binary exposure was simulated using a logistic regression with baseline exposure prevalence (all covariate values zero) of 10%. The associations of the covariates to exposure and outcome are given in Table (6.1, scenario Arbogast). To estimate the exposure effect, Poisson regression was used including the respective summary score as a covariate. However, no clear information about how the scores are used as adjustment variable (i.e. on which scale) is given. In the simulations, the true exposure effect β varied between 0.22 and 1.1. In our calculations, the intercept of the exposure model was $\alpha_0 = -2.2$ and for the outcome model $\mu = -4.6$ to achieve exposure and disease prevalences given above.

Additionally, we investigated scenarios where the confounder associations $\boldsymbol{X}$ to T and Y are 'misaligned' (Table 6.1, scenario A-D). In the misaligned scenario, 5 covariates are weakly associated with T ($\alpha_i = 0.22$ for $i = 1, \ldots, 5$) and strongly associated with Y ($\gamma_i = 1.1$ for $i = 1, \ldots, 5$) and the other 5 are strongly associated with T and weakly associated with Y ($\alpha_i = 1.1$ and $\gamma_i = 0.22$ for $i = 6, \ldots, 10$). Additionally, to investigate the direction of bias under this scenario, we varied the sign of the covariate-associations to T and Y (denoted with (T,Y)) as follows: A: (+,+), B: (-,+), C: (-,-), D: (+,-), where the first sign corresponds to all $\boldsymbol{\alpha}$ and the second to all $\boldsymbol{\gamma}$.

Table 6.1: Relationship of covariates, exposure and outcome. Parameters for the exposure (T) model: α_i, parameters for the outcome (Y) model: γ_i, i=1,...,10

	Scenario Austin		Scenario Arbogast	Scenarios							
				A		B		C		D	
Parameter	α_i	γ_i	$\alpha_i = \gamma_i$	α_i	γ_i	α_i	γ_i	α_i	γ_i	α_i	γ_i
X_1	1.6	1.6	0.69	0.22	1.1	-0.22	1.1	-0.22	-1.1	0.22	-1.1
X_2	0.69	1.6	0.69	0.22	1.1	-0.22	1.1	-0.22	-1.1	0.22	-1.1
X_3	0	1.6	0.69	0.22	1.1	-0.22	1.1	-0.22	-1.1	0.22	-1.1
X_4	1.6	0.69	0.69	0.22	1.1	-0.22	1.1	-0.22	-1.1	0.22	-1.1
X_5	0.69	0.69	0.69	0.22	1.1	-0.22	1.1	-0.22	-1.1	0.22	-1.1
X_6	0	0.69	0.69	1.1	0.22	-1.1	0.22	-1.1	-0.22	1.1	-0.22
X_7	1.6	0	0.69	1.1	0.22	-1.1	0.22	-1.1	-0.22	1.1	-0.22
X_8	0.69	0	0.69	1.1	0.22	-1.1	0.22	-1.1	-0.22	1.1	-0.22
X_9	0	0	0.69	1.1	0.22	-1.1	0.22	-1.1	-0.22	1.1	-0.22
X_{10}	-	-	0.69	1.1	0.22	-1.1	0.22	-1.1	-0.22	1.1	-0.22

For the scenarios A-D we considered the following (Table 6.2):

1.) The influence of the exposure effect β
The exposure effect varies between $\beta = -1.6$ and $\beta = 1.6$, with fixed confounder prevalences of $p_X = P(X_i = 1) = 0.10$, $i = 1, \ldots, 10$, fixed exposure prevalence $p_T = P(T = 1) = 0.10$ and fixed disease prevalence $p_Y = P(Y > 0) = 0.10$ for count outcomes or $p_Y = P(Y = 1) = 0.10$ for binary outcomes.

2.) The influence of disease prevalence p_Y
The disease prevalence varies between $p_Y = 0.25$ and $p_Y < 0.01$. The exposure and confounder prevalences for all 10 covariates are fixed by $p_T = 0.10$ and $p_X = 0.10$. The true exposure effect $\beta = 1$.

3.) The influence of exposure prevalence p_T
The exposure prevalence varies between $p_T < 0.01$ and $p_T = 0.50$. The exposure and confounder prevalences for all 10 covariates are fixed by $p_Y = 0.10$ and $p_X = 0.10$. The true exposure effect $\beta = 1$.

4.) The influence of confounder prevalence p_X
For all 10 covariates, the confounder prevalences are the same but vary between $p_X = 0.05$ and $p_X = 0.50$. The exposure and disease prevalences are fixed by $p_T = p_Y = 0.10$. The true exposure effect $\beta = 1$.

5.) The influence of the number of confounders $\dim(\boldsymbol{X}) = k$
For the parameter values from scenario A, we vary the number of confounding variables $\boldsymbol{X}$ from k=2 to k=10 in steps of two, i.e. two confounders $\boldsymbol{X} = (X_1, X_6)$ from Table 6.1, four confounders $\boldsymbol{X} = (X_1, X_2, X_6, X_7)$ from Table 6.1 and so on. The exposure, disease and confounder prevalences are fixed by $p_T = p_Y = p_X = 0.10$ and the true exposure effect $\beta = 1$.

Table 6.2: Summary of different settings with varying exposure effect β, disease prevalences p_Y, exposure prevalences p_T, confounder prevalences p_X and number of confounders k.

Influence of	p_X	p_T	p_Y	β	k
1.) exposure effect β	10%	10%	10%	-1.6 - 1.6	10
2.) disease prevalence p_Y	10%	10%	<1% - 25%	0 and 1	10
3.) exposure prevalence p_T	10%	<1% - 50%	10%	0 and 1	10
4.) confounder prevalence p_X	5% - 50%	10%	10%	0 and 1	10
5.) number of confounders k	10%	10%	10%	1	2-10

Comparison of asymptotic bias to estimates obtained from simulations

We additionally compared the results of our asymptotic bias calculations with estimates obtained from simulation studies with practically relevant sample sizes. In the simulations, the data was generated under the true outcome model with the same parameter values from scenario A. In our misspecified outcome models we investigated the estimated exposure effect if the estimated summary scores are used to adjust for confounding. This was done for a part of the settings described above. The simulations were based on 1000 repetitions for cohort sizes of $N = 10\,000$ and $N = 2500$ individuals.

Presentation of the results

The bias is reported as difference $\beta_s - \beta$ between the exposure effect under the misspecified outcome model β_s and the true exposure effect β . If no adjustment is performed the index s is replaced by C for the misspecified crude outcome model. The index P is used to indicate the exposure effect under the misspecified outcome model adjusting for the PS, (lP if the score is used on the logit scale) and D is used for the model that adjusts for the DRS (lD if the score is used on the log or the logit scale).

All asymptotic bias calculations and simulations were performed using the statistical software R-2.15.3. The R-code for our bias calculations is given in the Appendix B.1.

6.4.2 Comparison with the published results

The results of our asymptotic bias calculations are compared with the simulation results from Austin et al. (2007b) in Table 6.3.

Table 6.3: Comparison of the exposure effect β obtained from the asymptotic bias calculations with the simulation results from Austin et al. (2007b).

	True exposure effect β						
model	$\beta = -2.3$	$\beta = -1.6$	$\beta = -0.69$	$\beta = 0$	$\beta = 0.69$	$\beta = 1.6$	$\beta = 2.3$
	Results simulation Austin for Poisson regression						
Crude	-1.43	-0.76	0.17	0.86	1.55	2.47	3.16
Adjusted for $e(\boldsymbol{X})$	-2.30	-1.61	-0.67	0.00	0.68	1.59	2.28
	Results asymptotic bias calculations for Poisson regression						
Crude	-1.44	-0.75	0.17	0.86	1.55	2.47	3.16
Adjusted for $e(\boldsymbol{X})$	-2.28	-1.59	-0.68	0.00	0.69	1.59	2.28
Adjusted for logit$(e(\boldsymbol{X}))$	-2.34	-1.64	-0.70	0.00	0.71	1.64	2.33
Adjusted for $DRS(\boldsymbol{X})$	-2.19	-1.44	-0.44	0.30	1.02	1.96	2.66
Adjusted for log$(DRS(\boldsymbol{X}))$	-2.30	-1.61	-0.69	0.00	0.69	1.61	2.30
	Results simulation Austin for logistic regression						
Crude	-1.14	-0.53	0.20	0.72	1.22	1.87	2.37
Adjusted for $e(\boldsymbol{X})$	-1.97	-1.31	-0.56	0.00	0.55	1.25	1.79
	Results asymptotic bias calculations for logistic regression						
Crude	-1.44	-0.75	0.17	0.86	1.55	2.47	3.16
Adjusted for $e(\boldsymbol{X})$	-1.93	-1.33	-0.56	0.00	0.54	1.25	1.78
Adjusted for logit$(e(\boldsymbol{X}))$	-1.96	-1.35	-0.57	-0.01	0.53	1.23	1.77
Adjusted for $DRS(\boldsymbol{X})$	-2.39	-1.64	-0.67	0.05	0.75	1.65	2.32
Adjusted for logit$(DRS(\boldsymbol{X}))$	-2.30	-1.61	-0.69	0.00	0.69	1.61	2.30

For count outcomes, Austin et al. (2007b) observed a positive and nearly constant bias ($\beta_C - \beta \approx 0.85$) for the unadjusted exposure effect obtained from the model including only the exposure T (crude model), for each of the true exposure effects β. If the estimated PS is used to adjust for confounding, a negligible negative bias was observed only for strong exposure effects.

For logistic regression, the crude model yielded to positive bias ($\beta_C - \beta > 0$), which is decreasing with increasing β (i.e. from 1.2 to 0.1 for $\beta = -2.3$ to $\beta = 2.3$). If the estimated PS is used to adjust for confounding, a bias towards the null ($\beta_P - \beta > 0$ for $\beta < 0$ and $\beta_P - \beta < 0$ for $\beta > 0$) is introduced into the estimation of conditional exposure effects, whereat the amount of

bias is increasing with increasing true exposure effect size (0.3 for $\beta = -2.3$ and -0.5 for $\beta = 2.3$). No bias was observed only under no exposure effect ($\beta = 0$).

The results of our asymptotic bias calculations for the crude model and the PS adjusted regression models agree with the simulation results from Austin et al. (2007b). Adjusting for the PS on the logit scale yielded similar results for count data and binary outcomes. If the DRS is used as summary score, our calculations resulted in bias ranging from approximately -0.1 to 0.1 for binary outcomes and in larger bias ranging from approximately 0.1 to 0.4 for count outcomes. If the DRS on the log scale for count data and on the logit scale for binary data is used, no bias was observed.

In Arbogast and Ray (2011), the results are presented as percent bias for rate ratios only in figures, which makes the comparison to our asymptotic bias calculations difficult. Additionally, as mentioned earlier, no clear information about how the scores are used as adjustment variable is given. However, what can be seen from the figures is, that the estimates from the conventional multivariable outcome model (i.e. the true outcome model) and from the model that adjusts on the DRS agree and were unbiased. For the PS model little bias was observed which increased to a percent bias of $<3\%$ with increasing rate ratio.

The results of our calculations under the Poisson model for the Arbogast and Ray (2011) settings are presented in Table 6.4. If no adjustment is performed, this confounding scenario yields a positive bias ($\beta_C - \beta = 0.6$). No bias was observed for the model that adjusts for the DRS on the log scale and for the PS on the logit scale as well, since the confounder-exposure and confounder-outcome associations are the same ($\boldsymbol{\alpha} = \boldsymbol{\gamma}$) and therefore the misspecified model and the true outcome model coincide. The parameters of the misspecified model are given as follows: $\mu_{lP} = \mu - \alpha_0$, $\beta_{lP} = \beta$ and $\boldsymbol{\gamma}_{lP} = 1$. The same happens if the parameter values are chosen as $\boldsymbol{\alpha} = f\boldsymbol{\gamma}$, with f as multiplicative factor (i.e. $\mu_{lP} = \mu - f\alpha_0$, $\beta_{lP} = \beta$ and $\boldsymbol{\gamma}_{lP} = f$).

If the scores are used on their original scale, negligible bias was observed for the PS model and a bias of similar magnitude compared with the crude model was observed for the DRS model ($\beta_D - \beta = 0.5$). The results of our asymptotic bias calculations are comparable with the simulation results from Arbogast and Ray (2011) only for the PS where we obtained a percent bias of $<3\%$ (calculated on exponential scale) and for the DRS on log scale, where we observed no bias.

Similar results are given for our bias calculations under the logistic model except for the DRS model which resulted in a bias of $\beta_D - \beta = 0.1$ (data not shown).

Table 6.4: Results of the asymptotic bias calculations based on the scenario from Arbogast and Ray (2011).

	True exposure effect β			
model	$\beta = 0.22$	$\beta = 0.41$	$\beta = 0.69$	$\beta = 1.1$
	Poisson outcome model			
Crude	0.78	0.97	1.25	1.66
Adjusted for e($\boldsymbol{X}$)	0.23	0.41	0.71	1.12
Adjusted for logit(e($\boldsymbol{X}$))	0.22	0.41	0.69	1.10
Adjusted for $DRS(\boldsymbol{X})$	0.75	0.93	1.22	1.62
Adjusted for log($DRS(\boldsymbol{X})$)	0.22	0.41	0.69	1.10

6.4.3 Results - Poisson regression

The results of the bias calculations ($\beta_s - \beta$) under the Poisson model for our scenarios A-D and for varying exposure effect are summarized in Figure 6.1. For scenario A, the calculated β_s under varying disease, exposure and confounder prevalences and under different number of confounders (settings 2-5) are presented in Figure 6.2. The plots for the scenarios B-C are given in the Appendix A.1.1.

In general, we observed no bias for the log-transformed DRS, since this corresponds to adjusting for all the covariates itself (i.e. $\boldsymbol{\gamma}'\boldsymbol{X}$) and for the PS under no exposure effect ($\beta = 0$) under all scenarios and settings. In the other scenarios and settings, adjusting for the DRS or for the PS on the logit scale can result in a biased estimation of the exposure effect β with bias in different directions and of different magnitude. For all scenarios and settings studied, substantial bias was observed if no adjustment for the confounders was performed (crude model). If the summary scores were used to adjust for confounding, the bias was strongest for the DRS, i.e. $>10\%$ in nearly all scenarios and settings. Little bias was observed for the PS on both scales, i.e. $<5\%$ in nearly all scenarios and settings.

Influence of direction of confounder association $\boldsymbol{X}$ to T and Y

The direction of the bias strongly depends on the different confounder associations to T and to Y (scenarios A - D). Without any adjustment (crude model), a large positive bias was observed, if the confounders were either positively associated or negatively associated with both T and Y (scenarios A and C). If the confounder-exposure and confounder-outcome associations were in different directions, the bias was negative (scenarios B and D). This represents Type 1 and Type 2 bias in Section 2.2.2. For the crude model, this was observed in all our settings and scenarios.

There was no clear direction of the bias when adjustment for the PS or the logit of the PS was performed. However, under varying exposure effects and disease prevalences, we observed a higher exposure effect compared with the PS on the logit scale ($\beta_P > \beta_{lP}$), if the confounders were positively associated with Y (i.e. scenarios A and B). For a negative confounder-outcome association (i.e. scenarios C and D) the opposite was observed ($\beta_P < \beta_{lP}$). This relation of the PS on probability and on logit scale turned around when the exposure prevalence was high ($>40\%$).

Adjusting for the DRS, the direction of the bias depends on the confounder associations to T. For a positive association (scenario A and D) the bias is positive ($\beta_D - \beta > 0$) and for a negative association (scenario B and C) the bias is usually negative ($\beta_D - \beta < 0$). Deviations from this were observed only for higher confounder prevalences.

Influence of the exposure effect β

Figure 6.1 illustrates the calculated bias ($\beta_s - \beta$) for exposure effects between -1.6 and 1.6. There was no influence of the exposure effect on the crude model. For the PS model, there was a bias away from the null in Scenario A (i.e. $0 < \beta < \beta_P$ and $0 > \beta > \beta_P$) and a bias towards the null in scenario D (i.e. $0 < \beta_P < \beta$ and $0 > \beta_P > \beta$) with increasing magnitude for increasing exposure effect. However, this bias was less than 0.1. For the DRS model and in scenario A, the magnitude of the bias increased with increasing β (i.e. from 0.1 to 0.25), for scenarios B-D the bias was nearly constant (i.e. scenario B: $\beta_D - \beta = -0.2$, scenario C and D: $\beta_D - \beta < 0.1$).

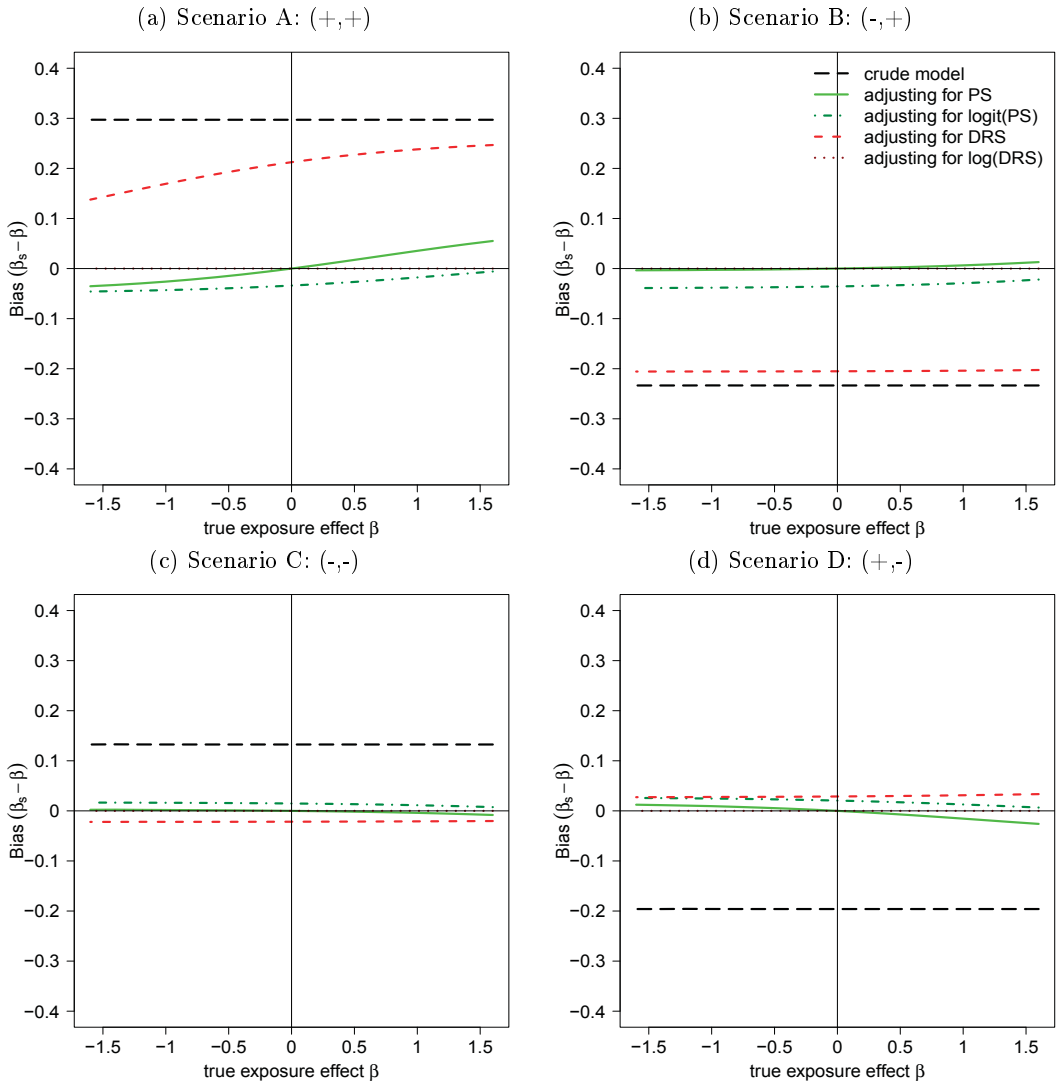


Figure 6.1: Analytic bias for varying exposure effects β, with $P(X = 1) = P(T = 1) = P(Y > 0) = 0.1$.

Influence of prevalences of Y, T and $\boldsymbol{X}$

The disease prevalence had no influence on the magnitude of the bias in all models and scenarios. For a true exposure effect of $\beta = 1$, the calculated exposure effects in scenario A for the misspecified outcome models are as follows: $\beta_C = 1.30$, $\beta_P = 1.03$, $\beta_{lP} = 0.98$ and $\beta_D = 1.24$ (Figure 6.2 (a)). Similar, little influence was observed if the exposure prevalence varied between 50% and 10% with similar β_s as stated before (Figure 6.2 (b)). However, if the exposure prevalence decreased to less than 10%, β_P slightly increased to 1.06 and β_{lP} slightly decreased to 0.95. For the DRS we observed a sharp drop on β_D towards 1. In other words, the bias for the PS increased over 5% and for the DRS under 10% if the exposure prevalence was rare. Varying the confounder prevalence (Figure 6.2 (c)), the major influence was observed on the crude model (i.e. the bias increased with increasing confounder prevalence) and on the DRS model (i.e. the bias increased for confounder prevalences up to 15% and then decreased with increasing confounder prevalence).

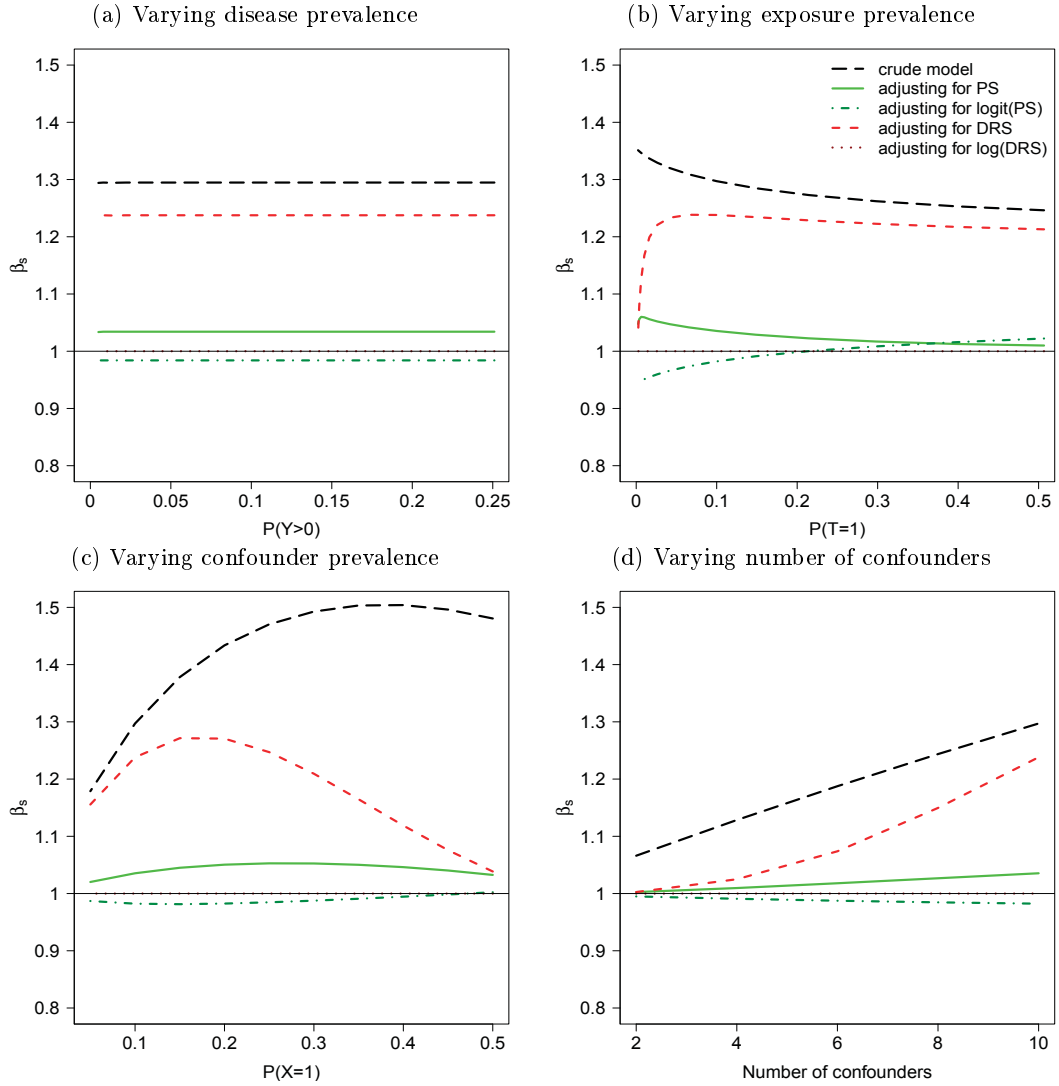


Figure 6.2: Analytic bias under scenario A as prevalence of Y, T, $\boldsymbol{X}$ and k vary with true exposure effect $\beta = 1$.

Influence of the number of confounders k

Figure 6.2 (d) shows the calculated β_s when the number of confounders varied between 2 and 10 for scenario A and with fixed prevalences of 10% for $\beta = 1$. The magnitude of bias increased for all misspecified models (except for the log-transformed DRS). The calculated β_s ranged from 1.1 for $\boldsymbol{X} = (X_1, X_6)$ to 1.3 for $\boldsymbol{X} = (X_1, \ldots, X_{10})$ for the crude model, from 1 to 1.04 for the PS adjusted model, from 1 to 0.98 for the logit-PS model and from 1 to 1.24 for the DRS adjusted model.

Estimates obtained from simulations

Table 6.5 shows the calculated β_s and the simulation based estimates $\hat{\beta}_s^1$ and $\hat{\beta}_s^2$ for cohort sizes of N=10 000 and N=2500, respectively, and for different disease and exposure prevalences when the estimated scores were used to adjust for confounding under scenario A. For N=10 000, the results β_s of our calculations and the simulation based estimates $\hat{\beta}_s^1$ agree very well. Only when the exposure prevalence got very low (around 1%), the simulation results slightly differ from the analytic ones. For N=2500 the differences in $\hat{\beta}_s^2$ and β_s were larger, and sometimes the bias changed sign. For example, for a Poisson model with $P(T = 1) = 0.01$ and $P(Y > 0) = 0.1$, the mean over the PS adjusted estimates was $\hat{\beta}_s^2 = 0.975$ while $\beta_s = 1.059$, and the mean over the logit-PS adjusted estimates was $\hat{\beta}_s^2 = 1.054$, while $\beta_s = 0.951$.

Table 6.5: Comparison of mean estimates $\hat{\beta}_s^1$ based on 1000 repetitions from a simulated cohort of N=10 000 individuals and $\hat{\beta}_s^2$ based on 1000 repetitions from a simulated cohort of N=2500 individuals to asymptotic estimates β_s computed analytically with a true value $\beta = 1$.

Prevalences		Estimates of β	Outcome models adjusted for			
P(T=1)	P(Y>0)	from	PS	logit-PS	DRS	log-DRS
0.100	0.100	calculation β_s	1.035	0.982	1.238	1.000
		simulation $\hat{\beta}_s^1$	1.033	0.981	1.205	0.999
		simulation $\hat{\beta}_s^2$	1.029	0.980	1.169	0.993
0.042	0.100	calculation β_s	1.047	0.966	1.233	1.000
		simulation $\hat{\beta}_s^1$	1.042	0.969	1.206	0.998
		simulation $\hat{\beta}_s^2$	1.031	0.979	1.170	0.992
0.010	0.100	calculation β_s	1.059	0.951	1.167	1.000
		simulation $\hat{\beta}_s^1$	1.034	0.969	1.210	0.994
		simulation $\hat{\beta}_s^2$	0.975	1.054	1.120	0.949
0.108	0.008	calculation β_s	1.034	0.984	1.238	1.000
		simulation $\hat{\beta}_s^1$	1.022	0.972	1.178	0.986
		simulation $\hat{\beta}_s^2$	0.893	0.849	1.000	0.868

6.4.4 Results - logistic regression

The results of the bias calculations under the logistic regression model for scenarios A-D and for varying exposure effects are summarized in Figure 6.3. For scenario A, the calculated β_s under varying disease, exposure and confounder prevalences and for different number of confounders are presented in Figure 6.4. Additional plots are given in the Appendix A.1.2.

As previously observed for the Poisson regression, adjusting for the DRS on the logit scale resulted in no bias in all scenarios and settings and for the PS under no exposure effect $\beta = 0$ as well. In general, the bias was stronger for the logistic regression model than for Poisson regression, when the PS was used to adjust for confounding and weaker for the DRS as summary score. The largest bias was observed for the crude model ranging from approximately 5% to >60%. For the summary scores, the amount of bias varied between 2% and 18%.

Influence of direction of confounder association $\boldsymbol{X}$ to T and Y

The direction of the bias for the crude model is the same as for the Poisson model, a positive bias for scenarios A and C, a negative bias for scenarios B and D. For the logistic models that adjusted on the summary scores, a similar pattern to the Poisson model was observed as well. The bias was smaller when the confounders were negatively associated with the outcome Y (scenarios C and D) compared with the bias observed for a positive confounder-outcome association (scenarios A and B).

Influence of the exposure effect β

The bias for the crude model ($\beta_C - \beta$) was decreasing for scenario A (from 0.4 to 0.1) and increasing for scenario B (from -0.1 to -0.3), when the exposure effects varied between -1.6 to 1.6 (Figure 6.3). For the PS model, there was an underestimation of the true magnitude of the association for all scenarios. The bias ($\beta_P - \beta$) varied between -0.1 and 0.1. The bias for the DRS model was smaller and ranged from -0.08 to 0.05.

Influence of prevalences of Y, T and $\boldsymbol{X}$

In contrast to the bias calculations under the Poisson model, the magnitude of bias under the logistic model depends on the disease prevalence (Figure 6.4 (a)). For a true exposure effect of $\beta = 1$, adjusting for the PS or for the the logit-PS, a bias was present and with increasing disease prevalence, the amount of bias increased. The direction of bias was negative in all scenarios but for very rare disease prevalences ($< 1\%$) a small positive bias was present. This is in good agreement with the Poisson model, that the logistic model approaches for rare disease. Concerning influence of disease prevalence, the opposite was observed for the DRS model: with increasing disease prevalence, the bias decreased. For example in scenario A, $\beta_P = 1.02$ and $\beta_D = 1.2$ for disease prevalences of $< 1\%$, $\beta_P = 0.9$ and $\beta_D = 1.02$ for disease prevalences of 25%.

The changes in the calculated β_s were small, if the exposure prevalence varied between $< 1\%$ to 50%, for $\beta = 1$ and a disease prevalence of 10%. With decreasing exposure prevalence, β_{lP} decreases to 0.85 in scenario A and the other β_s remained nearly constant, i.e. $\beta_P = 0.9$ and $\beta_D = 1.05$ (Figure 6.4 (b)). With increasing confounder prevalence, the amount of bias slightly increased (Figure 6.4 (c)).

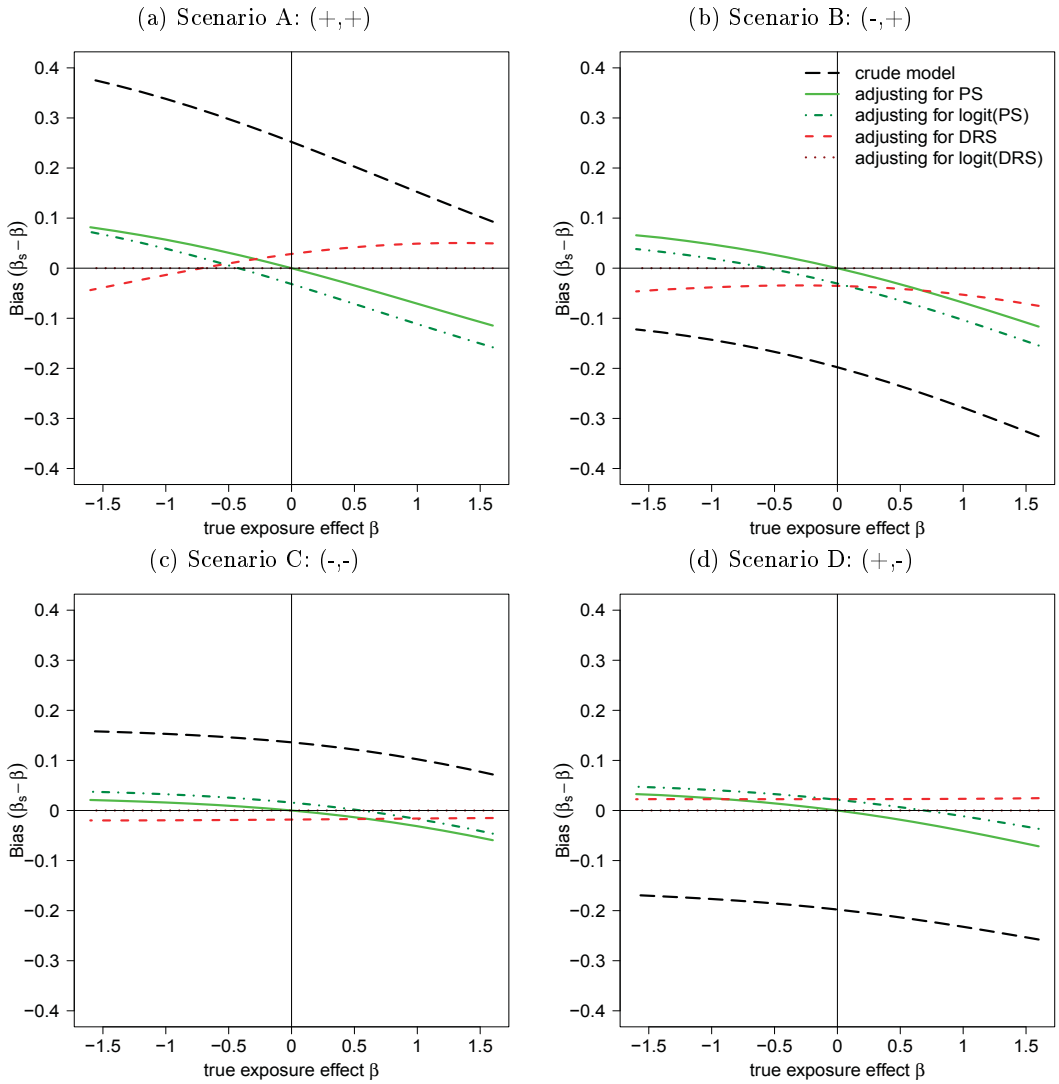


Figure 6.3: Analytic bias for different exposure effects β, with $P(X = 1) = P(T = 1) = P(Y = 1) = 0.1$.

Influence of the number of confounders k

Figure 6.4 (d) shows the calculated β_s when the number of confounders varied between 2 and 10 for scenario A and with disease, exposure and confounder prevalences of 10% for $\beta = 1$. The bias increased for all misspecified models (except for the logit-DRS). The calculated β_s ranged from 1.04 to 1.14 for the crude model, from 0.98 to 0.92 for the PS adjusted model, from 0.97 to 0.88 for the logit-PS model and from 1.01 to 1.04 for the DRS adjusted model.

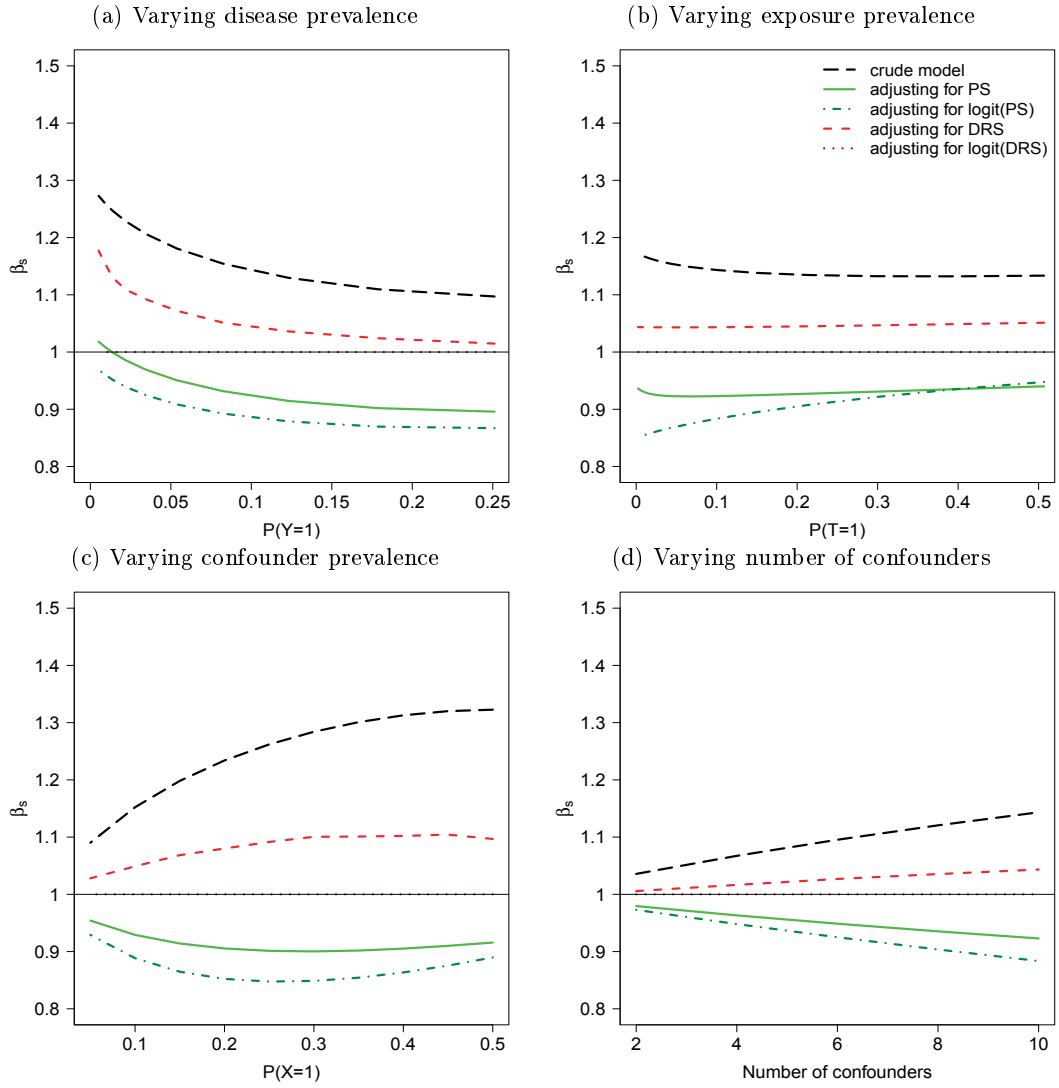


Figure 6.4: Analytic bias under scenario A as prevalence of Y, T, $\boldsymbol{X}$ and k vary with true exposure effect $\beta = 1$.

Estimates obtained from simulations

Table 6.6 shows the calculated β_s and the simulation based estimates $\hat{\beta}_s^1$ and $\hat{\beta}_s^2$ for cohort sizes of N=10 000 and N=2500, respectively. For N=10 000, the values of β_s and the simulation based estimates $\hat{\beta}_s^1$ agree very well. As observed for the Poisson regression, the differences between the simulation based estimates $\hat{\beta}_s^2$ and the calculated β_s increased with a cohort size of N=2500. For low exposure or disease prevalences (around 1%) the largest differences were observed.

Table 6.6: Comparison of mean estimates $\hat{\beta}_s^1$ based on 1000 repetitions from a simulated cohort of N=10 000 individuals and $\hat{\beta}_s^2$ based on 1000 repetitions from a simulated cohort of N=2500 individuals to asymptotic estimates β_s computed analytically with a true value $\beta = 1$.

Prevalences		Estimates of β	Outcome models adjusted for			
P(T=1)	P(Y>0)	from	PS	logit-PS	DRS	log-DRS
0.100	0.100	calculation β_s	0.923	0.883	1.043	1.000
		simulation $\hat{\beta}_s^1$	0.924	0.886	1.045	1.002
		simulation $\hat{\beta}_s^2$	0.913	0.881	1.038	0.999
0.042	0.100	calculation β_s	0.923	0.867	1.043	1.000
		simulation $\hat{\beta}_s^1$	0.920	0.869	1.042	1.000
		simulation $\hat{\beta}_s^2$	0.911	0.878	1.036	0.998
0.010	0.100	calculation β_s	0.930	0.855	1.043	1.000
		simulation $\hat{\beta}_s^1$	0.928	0.880	1.042	1.001
		simulation $\hat{\beta}_s^2$	0.814	0.906	0.931	0.900
0.108	0.008	calculation β_s	1.011	0.962	1.159	1.000
		simulation $\hat{\beta}_s^1$	1.003	0.956	1.141	0.999
		simulation $\hat{\beta}_s^2$	0.908	0.865	1.045	0.923

6.5 Bias calculation for the matched case-control design

For rare outcomes case-control studies are often used to estimate exposure effects. To eliminate the impact of confounding controls can be matched to cases on covariates $\boldsymbol{X}$. However, when the dimension of $\boldsymbol{X}$ is large, matching may not be feasible. We thus consider the setting where cases and controls are matched based on a summary score $s(\boldsymbol{X})$. To be explicit, we assume that in the cohort that gives rise to the case-control data Y has a binomial distribution, where the binomial probability is given by a logistic model with mean model (6.1) (i.e. h is the logistic function) We then sample a fixed number of cases, and compute the score $s(\boldsymbol{X})$ for each case. We match each case to a control that has the same score as the case.

We compute the asymptotic bias when β_s is estimated using conditional logistic regression, accounting for the matched design, or based on unconditional logistic regression adjusted for the matching variable, i.e. the score $s(\boldsymbol{X})$.

Matched analysis

THEOREM 2. *If cases and controls are matched on the PS, i.e. $s(\boldsymbol{X}) = e(\boldsymbol{X})$ then the MLE $\hat{\beta}_s$ based on conditional logistic regression consistently estimates $\beta = 0$ in mean model (6.1). When cases and controls are matched on the disease risk score, i.e. $s(\boldsymbol{X}) = DRS(\boldsymbol{X})$, then the MLE $\hat{\beta}_s$ based on conditional logistic regression is an unbiased estimate of β, i.e. $\beta_s = \beta$, even when $\beta \neq 0$.*

Proof. We assume the matched pairs (Y_1, Y_2) are analyzed using a conditional logistic model conditioning on the matched set,

$$P_G(Y_1, Y_2 | Y_1 + Y_2 = 1, T_1, T_2, s(\boldsymbol{X}_1) = s(\boldsymbol{X}_2)) = \frac{\exp(\beta_s T_1 Y_1 + \beta_s T_2 Y_2)}{\exp(\beta_s T_1) + \exp(\beta_s T_2)}. \tag{6.15}$$

To find the asymptotic value of the estimate β_s obtained by maximum likelihood estimation based on (6.15), we again solve equation (6.5). As (6.15) does not depend on μ_s and γ_s we only have to solve one equation for β_s:

$$E_F\Big\{T_1 Y_1 + T_2 Y_2 - \frac{T_1 \exp(\beta_s T_1) + T_2 \exp(\beta_s T_2)}{\exp(\beta_s T_1) + \exp(\beta_s T_2)}\Big\} = 0. \tag{6.16}$$

The expectation E_F under the true model F is computed accounting for the matched retrospective sampling design under the distribution

$$\begin{aligned} &P(T_1, T_2, \boldsymbol{X}_1, \boldsymbol{X}_2 | Y_1 = 1, Y_2 = 0, s(\boldsymbol{X}_1) = s(\boldsymbol{X}_2)) = \\ &\qquad P(\boldsymbol{X}_1, T_1 | Y_1 = 1) P(\boldsymbol{X}_2, T_2 | Y_2 = 0) I_s(\boldsymbol{X}_1, \boldsymbol{X}_2) = \\ &\frac{P(Y_1 = 1 | \boldsymbol{X}_1, T_1) P(T_1 | \boldsymbol{X}_1) P(\boldsymbol{X}_1)}{P(Y_1 = 1)} \frac{P(Y_2 = 0 | \boldsymbol{X}_2, T_2) P(T_2 | \boldsymbol{X}_2) P(\boldsymbol{X}_2) I_s(\boldsymbol{X}_1, \boldsymbol{X}_2))}{P(Y_2 = 0, s(\boldsymbol{X}_2) = s(\boldsymbol{X}_1))}, \end{aligned} \tag{6.17}$$

where

$$P(Y_2 = 0, s(\boldsymbol{X}_2) = s(\boldsymbol{X}_1)) = \sum_{\boldsymbol{X}_2} \sum_{T_2} P(Y_2 = 0 | T_2, \boldsymbol{X}_2) P(T_2 | \boldsymbol{X}_2) P(\boldsymbol{X}_2) I_s(\boldsymbol{X}_1, \boldsymbol{X}_2)$$

and $I_s(\boldsymbol{X}_1, \boldsymbol{X}_2)$ as indicator function which takes the value one if $s(\boldsymbol{X}_1) = s(\boldsymbol{X}_2)$ and zero otherwise.

If $\beta = 0$ and $\beta_s = 0$, Y does not depend on T, and after taking the expectation with respect to T equation (6.16) reduces to

$$E_{\boldsymbol{X}_1,\boldsymbol{X}_2|s(\boldsymbol{X}_1)=s(\boldsymbol{X}_2)}\Big\{\frac{P(Y_1=1|\boldsymbol{X}_1)}{P(Y_1=1)}\frac{P(Y_2=0|\boldsymbol{X}_2)}{P(Y_2=0,s(\boldsymbol{X}_2)=s(\boldsymbol{X}_1))}\Big(e(\boldsymbol{X}_1)-\frac{e(\boldsymbol{X}_1)+e(\boldsymbol{X}_2)}{2}\Big)\Big\}=0. \quad (6.18)$$

If the score in model (6.3) is the PS then for each pair $e(\boldsymbol{X}_1) = e(\boldsymbol{X}_2)$, and thus $(e(\boldsymbol{X}_1) + e(\boldsymbol{X}_2))/2 = e(\boldsymbol{X}_1)$. Therefore equation (6.18) is satisfied, and the MLE $\hat{\beta}_s$ based on the conditional model (6.15) consistently estimates the true effect $\beta = 0$. In other words, under the null hypothesis of no association between Y and T, matching on the PS yields unbiased estimates, $\beta_s = \beta = 0$.

Next we assume that the case and the control are matched on their disease risk score, i.e. $s(\boldsymbol{X}) = DRS(\boldsymbol{X})$. If $DRS(\boldsymbol{X}_1) = DRS(\boldsymbol{X}_2)$ then $\boldsymbol{\gamma}'X_1 = \boldsymbol{\gamma}'X_2$ as the logistic function is a monotone transformation of $\boldsymbol{\gamma}'\boldsymbol{X}$. Thus $\boldsymbol{\gamma}'\boldsymbol{X}_1$ and $\boldsymbol{\gamma}'\boldsymbol{X}_2$ cancel out of the conditional logistic probability, and (6.16) reduces to

$$E_{\boldsymbol{X}}E_{T_1,T_2|\boldsymbol{X}_1,\boldsymbol{X}_2,s(X_1)=s(\boldsymbol{X}_2)}\Big\{\frac{T_1\exp(\beta T_1)+T_2\exp(\beta T_2)}{\exp(\beta T_1)+\exp(\beta T_2)}-\frac{T_1\exp(\beta_s T_1)+T_2\exp(\beta_s T_2)}{\exp(\beta_s T_1)+\exp(\beta_s T_2)}\Big\}=0$$

which is satisfied when $\beta_s = \beta$. Thus matching on the disease risk score leads to unbiased estimates of the conditional treatment effect, β. □

Unmatched analysis adjusted for matching variables

Here we study the behavior of estimates when cases and controls are matched based on a summary score $s(\boldsymbol{X})$ but then are analyzed using unconditional logistic regression adjusted for the matching variables, i.e. based on the model

$$P_G(Y_1,Y_2|T_1,T_2,s(\boldsymbol{X}_1),s(\boldsymbol{X}_2)) =$$
$$\frac{\exp[Y_1(\mu_s+\beta_s T_1+\gamma_s s(\boldsymbol{X}_1))+Y_2(\mu_s+\beta_s T_2+\gamma_s s(\boldsymbol{X}_2))]}{[1+\exp(\mu_s+\beta_s T_1+\gamma_s s(\boldsymbol{X}_1))][1+\exp(\mu_s+\beta_s T_2+\gamma_s s(\boldsymbol{X}_2))]}=P_G(Y_1)P_G(Y_2),$$

where

$$P_G(Y_c)=\exp\{Y_c(\mu_s+\beta_s T_c+\gamma_s s(\boldsymbol{X}_c)\}/[1+\exp\{\mu_s+\beta_s T_c+\gamma_s s(\boldsymbol{X}_c)\}] \qquad c\in\{1,2\}\,.$$

THEOREM 3. *If cases and controls are matched on a score $s(\boldsymbol{X})$ and then analyzed by unconditional logistic regression adjusting for $s(\boldsymbol{X})$, then the MLE $\hat{\beta}_s$ consistently estimates $\beta = 0$ in (6.1) when s is the PS. When $\beta \neq 0$ or any other function $s(\boldsymbol{X})$ is used (e.g. the DRS) then the MLE β_s is biased.*

Proof. The system of equations needed to find (6.5) is

$$\frac{\partial f}{\partial \mu_s}: E_F\Big\{Y_1+Y_2-p_G(Y_1=1)-p_G(Y_2=1)\Big\}=0 \quad (6.19)$$
$$\frac{\partial f}{\partial \beta_s}: E_F\Big\{Y_1T_1+Y_2T_2-p_G(Y_1=1)T_1-p_G(Y_2=1)T_2\Big\}=0 \quad (6.20)$$
$$\frac{\partial f}{\partial \gamma_s}: E_F\Big\{Y_1s(\boldsymbol{X}_1)+Y_2s(\boldsymbol{X}_2)-p_G(Y_1=1)s(\boldsymbol{X}_1)-p_G(Y_2=1)s(\boldsymbol{X}_2)\Big\}=0. \quad (6.21)$$

The expectation under model F is computed under the retrospective sampling accounting for the matching using the distribution given in equation (6.17).

When $\beta = 0$ and $\beta_s = 0$, Y does not depend on T and after taking the expectation with respect to T equations (6.20) and (6.21) reduce to

$$\frac{\partial f}{\partial \beta_s} : E_{\boldsymbol{X}_1,\boldsymbol{X}_2,|s(\boldsymbol{X}_1)=s(\boldsymbol{X}_2)}\Big\{\frac{P(Y_1=1|\boldsymbol{X}_1)}{P(Y_1=1)}\frac{P(Y_2=0|\boldsymbol{X}_2)}{P(Y_2=0,s(\boldsymbol{X}_2)=s(\boldsymbol{X}_1))}\Big(e(\boldsymbol{X}_1) - p_G(Y_1=1)e(\boldsymbol{X}_1) - p_G(Y_2=1)e(\boldsymbol{X}_2)\Big)\Big\} = 0$$

$$\frac{\partial f}{\partial \gamma_s} : E_{\boldsymbol{X}_1,\boldsymbol{X}_2,|s(\boldsymbol{X}_1)=s(\boldsymbol{X}_2)}\Big\{\frac{P(Y_1=1|\boldsymbol{X}_1)}{P(Y_1=1)}\frac{P(Y_2=0|\boldsymbol{X}_2)}{P(Y_2=0,s(\boldsymbol{X}_2)=s(\boldsymbol{X}_1))}\Big(s(\boldsymbol{X}_1) - p_G(Y_1=1)s(\boldsymbol{X}_1) - p_G(Y_2=1)s(\boldsymbol{X}_2)\Big)\Big\} = 0.$$

The same argument as for the bias computations for the full cohort analysis applies. If the MLE $\hat{\gamma}_s$ satisfies the last equation, then if in addition $s(\boldsymbol{X}_i) = e(\boldsymbol{X}_i)$, the equation for $\partial f/\partial \beta_s$ also holds. Thus if $\beta = 0$, then $\beta_s = 0$ when the PS is used to match cases and controls and then as an adjustment variable in an unconditional logistic model.

However, when cases and controls are matched on the DRS analyzing the data using unconditional logistic regression will result in biased estimates of β.

□

6.6 Numerical examples for the matched case-control design

6.6.1 Scenarios and settings

We calculated the analytic bias numerically for various confounder associations $\boldsymbol{X}$ to T and Y. Due to the computational burden we studied scenarios with two confounders $\boldsymbol{X} = (X_1, X_2)$. The R-code, given in the Appendix B.2, is written for an arbitrary number of binary confounder variables, but ran over 6 hours for one setting and one summary score when $k = 10$.

Based on our analytic calculations, no bias for the DRS in the matched analysis is expected. Therefore, we focused in the choice of our scenarios on the case-control matching on the PS. In previous simulations and calculations (results not shown), we observed a bias for PS-matching if two conditions are met:

1. Different confounder combinations lead to the same PS (i.e. there exist 'mismatches')
2. The 'mismatches' are differently associated with the outcome Y

In scenario I the true association of the confounders to exposure T is given by $\boldsymbol{\alpha} = (1.6, 1.6)$ and to outcome Y by $\boldsymbol{\gamma} = (2, -0.6)$. In this scenario, we get three different propensity scores for four different confounder combinations, i.e. the combinations $X_1 = 1$, $X_2 = 0$ and $X_1 = 0$, $X_2 = 1$ have the same PS. This leads to imbalance in the covariates when cases and controls are matched on the PS. For a larger number of confounders and different confounder-exposure associations, combinations of several confounders can also lead to the same PS (e.g. for $\boldsymbol{X} = (X_1, X_2, X_3, X_4)$ with $\boldsymbol{\alpha} = (1.1, 1.2, 1.3, 1.4)$ the same PS is given for $X_1 = X_4 = 1$, $X_2 = X_3 = 0$ and $X_2 = X_3 = 1$, $X_1 = X_4 = 0$). For the confounder-outcome associations, one variable is positively associated with Y and the other negatively.

In scenario II the association of the confounders to exposure T is unequal i.e. $\boldsymbol{\alpha} = (1.6, 1)$, which leads to four different PS values for the four different confounder combinations. The confounder-outcome association is given by $\boldsymbol{\gamma} = (2, -0.6)$.

For both scenarios the same settings (i.e. varying disease, exposure and confounder prevalences and different exposure effects) as described in Section 6.1.2 are considered (see also Table 6.7).

Table 6.7: Summary of different settings with varying exposure effect β, disease prevalences p_Y, exposure prevalences p_T, confounder prevalences p_X and number of confounders k.

Influence of	p_X	p_T	p_Y	β	k
1.) exposure effect β	10%	10%	10%	-1.6 - 1.6	2
2.) disease prevalence p_Y	10%	10%	<1% - 25%	0 and 1	2
3.) exposure prevalence p_T	10%	<1% - 50%	10%	0 and 1	2
4.) confounder prevalence p_X	5% - 50%	10%	10%	0 and 1	2
5.) number of confounders k	10%	10%	10%	1	2-10

Comparison of asymptotic bias to estimates obtained from simulations

As for the cohort study, we compared the results of our asymptotic bias calculations with estimates obtained from simulation studies. To create a case-control study, we first simulated a cohort of size $N = 55\,000$ by using logistic regressions to generate the binary exposure T and the binary outcome Y with the same parameter values from our first scenario (unequal confounder-exposure association). The summary scores were estimated from this cohort. Then, the cases and controls were sampled from the simulated cohort by using individual matching (1:1 ratio) on a fixed number (i.e. $n = 1000$) of randomly drawn cases ($Y = 1$). Exact matching was performed on the summary scores on both scales (logit and probability based). The exposure effect $\hat{\beta}_s$ was estimated by using a conditional logistic regression and by using an unmatched analysis adjusted for the matching variables. We investigated the estimated exposure effect if the true summary scores are used to control for confounding and also if the estimated summary scores are used. This was done for a part of the settings described above. The results of our simulations are based on 1000 simulation runs.

All asymptotic bias calculations and simulations were performed using the statistical software R-2.15.3. The R-code for our bias calculations is given in the Appendix B.2.

6.6.2 Results

The results of the bias $(\beta_s - \beta)$ in the case-control setting using matched and unmatched analysis for scenario I and II under varying exposure effect are summarized in the Figure 6.5. For scenario I, the calculated β_s for varying disease, exposure and confounder prevalences and for a varying number of confounders are plotted in Figure 6.6. The corresponding results for scenario II are given in Figures A.7-A.9 in the Appendix A.1.3.

There was no bias in the exposure effect β_s when the DRS was used to match cases and controls and a conditional logistic regression (CLR) was used to analyze the data. When the true exposure effect $\beta = 0$, no bias was observed for all models and in all scenarios and settings. If $\beta \neq 0$ in scenario I, matching on the summary scores and using conditional or unconditional logistic regression with the PS and unconditional logistic regression adjusted for the DRS resulted in an underestimation of the true exposure effect (bias towards the null) in all scenarios and settings. Since in scenario I the equal confounder-exposure associations lead to imbalance of the matched pairs, a larger bias was observed compared with scenario II. Moreover, in the second scenario there was no bias for matching on the PS and using CLR. For the PS matched cases and controls the estimates β_s are very similar for the matched and unmatched analysis. Adjusting for the logit of the PS resulted in a slightly larger bias than for the PS on probability scale. Adjusting for the DRS or the logit of the DRS yielded a smaller bias in β_s than for the PS adjusted analysis.

Influence of the exposure effect β

With increasing exposure effect, the amount of bias $(\beta_s - \beta)$ increased (Figure 6.5). Matching on the PS and using matched or unmatched analysis resulted in bias ranging from 0.12 to -0.23 in Scenario I. Matching on the DRS and adjusting for this score led to a bias ranging from 0.05 to -0.13 in Scenario I and from 0.02 to -0.06 in Scenario II.

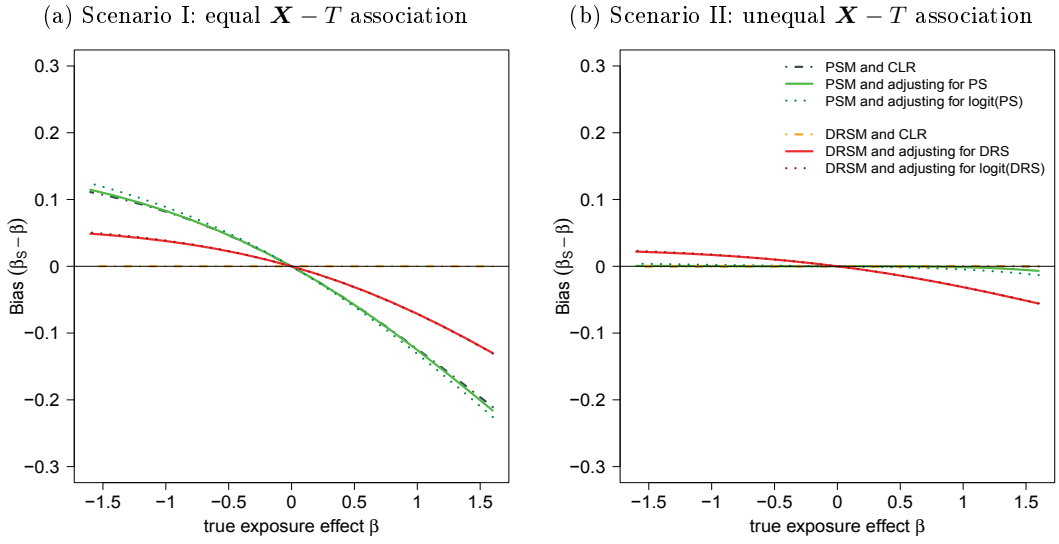


Figure 6.5: Analytic bias under unequal and equal confounder-exposure association for different exposure effects β, with $P(X = 1) = P(T = 1) = P(Y = 0) = 0.1$. PSM denotes PS matching, DRSM denotes DRS matching and CLR conditional logistic regression.

Influence of prevalences of Y, T and $\boldsymbol{X}$

For a true exposure effect of $\beta = 1$, with exposure and confounder prevalences of 10% and an increasing disease prevalence, the bias in β_s calculated from the PS matched cases and controls increased up to 17%. For the DRS matched cases and controls in unmatched analysis the bias was constant (approximately 7%) (Figure 6.6 (a)). Only for rare disease (<0.5%) PS matching shows a smaller bias compared with DRS matching and unmatched analysis. When the exposure prevalence decreased and fixing the other prevalences at 10%, the bias for the PS matching increased to 13% and for the DRS matching the bias remained under 10% (Figure 6.6 (b)). Slightly over 10% bias in β_s was observed for the PS matching when the confounder prevalences were <25% and for the DRS matching when the confounder prevalences were >20% (Figure 6.6 (c)).

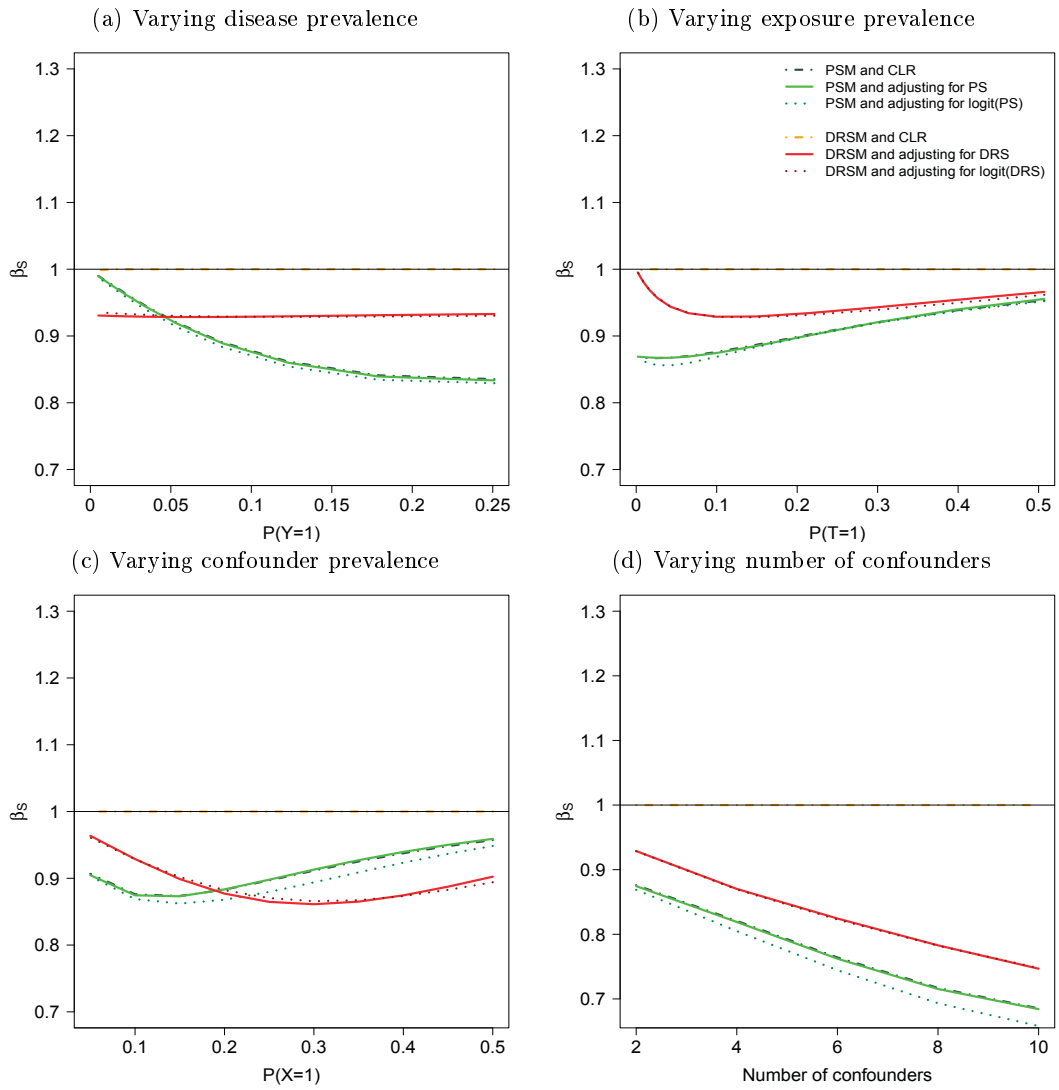


Figure 6.6: Analytic bias under equal confounder-exposure association association (Scenario I) with true exposure effect $\beta = 1$. PSM denotes PS matching, DRSM denotes DRS matching and CLR conditional logistic regression.

Influence of the number of confounders k

Figure 6.6 (d) shows the calculated β_s in scenario I, when the number of confounders varied between 2 and 10 (all prevalences are 10% and $\beta = 1$). The bias linearly increased with increasing number of variables. Adjusting for the logit of the PS resulted in a slightly larger bias than for the PS. This difference was more pronounced for a larger number of confounders.

Estimates obtained from simulations

Table 6.8 shows the calculated β_s and the estimates obtained from the simulations for different exposure prevalences under scenario I with $P(Y = 1) = P(X = 1) = 0.01$ and $\beta = 1$. $\hat{\beta}_s$ denotes the simulation results for the true summary scores and $\hat{\beta}_{\hat{s}}$ for the estimated ones. The results of our calculations compared with the simulation based estimates agree very well. Only the simulation results based on the estimated PS differed from our calculations, i.e. there was no bias in the estimates $\hat{\beta}_{\hat{s}}$. However, this difference is plausible, because due to the estimation of the PS, the confounder-exposure association is no longer equal which results in a nearly unbiased estimation. This was also observed in the calculations with unequal confounder-exposure associations (see Figures A.7-A.9 for scenario II).

Table 6.8: Comparison of asymptotic estimates β_s obtained from calculations with estimates $\hat{\beta}_s$ obtained from simulations based on the true summary scores $s(\boldsymbol{X})$ and the results $\hat{\beta}_{\hat{s}}$ based on the estimated ones.

		Matched analysis			**Unmatched analysis adjusted for $s(\boldsymbol{X})$**					
					probability based			logit based		
Matching on	P(T=1)	β_s	$\hat{\beta}_s$	$\hat{\beta}_{\hat{s}}$	β_s	$\hat{\beta}_s$	$\hat{\beta}_{\hat{s}}$	β_s	$\hat{\beta}_s$	$\hat{\beta}_{\hat{s}}$
PS	0.50	0.954	0.962	1.001	0.955	0.962	1.001	0.952	0.958	0.999
	0.10	0.876	0.879	0.995	0.875	0.876	0.992	0.869	0.870	0.981
	0.04	0.867	0.882	1.018	0.867	0.882	1.014	0.856	0.869	0.996
DRS	0.50	1.000	1.003	1.003	0.966	0.969	0.969	0.962	0.965	0.965
	0.10	1.000	0.997	0.997	0.929	0.927	0.927	0.929	0.926	0.926
	0.04	1.000	1.019	1.019	0.944	0.959	0.959	0.945	0.960	0.960

6.7 Summary

Here, the main results of our bias calculations for the cohort study and for the case-control study are briefly summarized in Tables 6.9, 6.10 and 6.11.

Table 6.9: Summary of the asymptotic bias calculation results observed in all scenarios and settings for Poisson regression in cohort studies.

Cohort study - Poisson regression			
Model	**Bias** $\beta = 0$	$\beta \neq 0$	**Influence**
Crude	Yes	Yes	• Largest bias observed in all scenarios and settings.
PS	**No**	Yes	• minor bias (<5%) in nearly all scenarios and settings • the bias slightly increases with increasing exposure effect • for an exposure prevalence of <10% the bias increases to >5% • the disease and confounder prevalences have no or only little influence on the magnitude of the bias
logit-PS	Yes	Yes	• minor bias (<5%) in nearly all scenarios and settings • bias slightly increases with increasing exposure effect • for an exposure prevalence of <10% the bias increases to >5% • the disease and confounder prevalences have no or only little influence on the magnitude of the bias
DRS	Yes	Yes	• bias >10% in nearly all scenarios and settings • for rare exposures (<1%) and for common confounders (>40%) the bias decreases to <10% • the disease prevalence has no influence on the magnitude of the bias
log-DRS	**No**	**No**	-

Table 6.10: Summary of the asymptotic bias calculation results observed in all scenarios and settings for logistic regression in cohort studies.

Cohort study - Logistic regression			
Model	**Bias** $\beta = 0$	$\beta \neq 0$	**Influence**
Crude	Yes	Yes	• Largest bias observed in all scenarios and settings.
PS	**No**	Yes	• bias towards null (ranging from <5% to <15%) in nearly all scenarios • positive bias for very rare disease prevalences (<1%) • bias increases with increasing exposure effect and disease prevalence • exposure prevalence and confounder prevalence have little influence on the magnitude of the bias
logit-PS	Yes	Yes	• bias ranging from <5% to 15% in nearly all scenarios • bias increases with increasing exposure effect and disease prevalence • for an exposure prevalence of <10% the bias increases to >10% • confounder prevalence has little influence on the magnitude of the bias
DRS	Yes	Yes	• bias ranging from <5% to <20% in nearly all scenarios • bias increases with increasing exposure effect • bias increases with decreasing disease prevalence • exposure prevalence and confounder prevalence have little influence on the magnitude of the bias
logit-DRS	**No**	**No**	-

Table 6.11: Summary of the asymptotic bias calculation results observed in all scenarios and settings for case-control studies when matching on a summary score is performed.

Case-control matching on the PS			
Analysis	**Bias** $\beta = 0$	$\beta \neq 0$	**Influence**
matched	**No**	Yes	• bias towards null (<20%) • No bias for $\beta \neq 0$ if confounder-exposure associations are unequal • bias increases with increasing exposure effect and disease prevalence • bias increases with decreasing exposure prevalence
unmatched and adjusted on PS	**No**	Yes	• bias towards null (<20%) • very small bias for $\beta \neq 0$ if confounder-exposure associations are unequal • bias increases with increasing exposure effect and disease prevalence • bias increases with decreasing exposure prevalence • for confounder prevalences <20% the bias increases to > 10%
unmatched and adjusted on logit-PS	**No**	Yes	• bias towards null (<20%) • very small bias for $\beta \neq 0$ if confounder-exposure associations are unequal • slightly larger bias than for adjusting on PS probability based • bias increases with increasing exposure effect and disease prevalence • bias increases with decreasing exposure prevalence • for confounder prevalences <20% the bias increases to > 10%
Case-control matching on the DRS			
Analysis	**Bias** $\beta = 0$	$\beta \neq 0$	**Influence**
matched	**No**	**No**	-
unmatched and adjusted on DRS	**No**	Yes	• bias towards null (<10% in nearly all settings) • bias increases with increasing exposure effect • the disease and exposure prevalences have minor influence on the magnitude of the bias • for confounder prevalences >20% the bias increases to > 10%
unmatched and adjusted on logit-DRS	**No**	Yes	• bias towards null (<10% in nearly all settings) • bias increases with increasing exposure effect • the disease and exposure prevalences have minor influence on the magnitude of the bias • for confounder prevalences >20% the bias increases to > 10%

Chapter 7

Simulations based on blood transfusion data

In the previous chapter, we investigated the performance of summary scores concerning bias in the estimation of conditional exposure effects analytically under simplified scenarios for independent confounding variables. The aim of the following simulation study is to investigate the performance of propensity score and disease risk score methods to control for confounding in large cohort and population-based case-control studies under 'real-life' conditions. Therefore, we studied the setting when a large number of confounders is present with confounder correlation structure and confounder prevalences based on controls from a study that investigated the association of blood transfusions with cancer risk in the US-elderly (Riedl et al., 2013). We generate cohort data that we analyze using logistic regression with various scores used as adjusting variables. We also generate matched case-control data using various matching strategies.

7.1 Data generation

Random sampling of the confounders

To simulate a cohort with similar covariate associations as seen in the blood transfusion data example, we use a subsample of 100 000 controls from a 5% random sample of Medicare beneficiaries (N=1 237 153), from which the the cancer-free controls for the analysis in (Riedl et al., 2013) were selected. A detailed description of the data set used for the simulation here is given in the next chapter. For the simulation we considered the demographic parameters sex, age in four categories (70-74, 75-79, 80-84 and 85-99 years), race in three categories (white, black and others) and the selection year in four categories (1997-1999, 2000-2001, 2002-2003 and 2004-2005). The selection year referred to the end of a twelve month period for which the receipt of a blood transfusion was ascertained. For the demographic parameters dummy coding is used for the various categories. These covariates are thus summarized in a vector $\boldsymbol{D}$ that has nine entries.

We further included the 20 most common medical conditions in the 5% random sample of Medicare controls as binary baseline covariates (i.e. condition present or not within 12 to 24 months prior to transfusion or the mid year point, as an index date, if no transfusion was received). These conditions (e.g. heart failure, anemias and iron deficiency) are combined in the vector $\boldsymbol{C} = (C_1, \ldots, C_{20})$.

The covariate distributions in the subsample of 100 000 Medicare controls for the demographic and medical conditions are summarized in Table 7.1. From this subsample, in every simulation step a cohort of size N=55 000 was created by sampling the covariates $\boldsymbol{X} = (\boldsymbol{D}, \boldsymbol{C})$ with replacement. In the next step the exposure and the outcome were generated.

Table 7.1: In the simulation included demographic and medical conditions with their distributions in the subsample of 100 000 controls.

Parameter	Description	Distribution
D_1	sex (female)	64.2
D_2	age in categories	
	70-74	30.1
	75-79	29.2
	80-84	21.7
	84-99	19.0
D_3	Selection year in categories	
	1997-1999	20.8
	2000-2001	25.5
	2002-2003	26.2
	2004-2005	27.5
D_4	race in categories	
	white	85.5
	black	6.7
	others	7.9
C_1	Disorders of lipoid metabolism	52.4
C_2	Cataract	47.1
C_3	Special screening for malignant neoplasms	35.2
C_4	Other and unspecified anemias	22.6
C_5	Other dermatoses	22.2
C_6	Heart failure	17.4
C_7	Hyperplasia of prostate	16.4
C_8	Disorders of fluid, electrolyte, and acid-base balance	16.1
C_9	Menopausal and postmenopausal disorders	12.4
C_{10}	Contact dermatitis and other eczema	12.4
C_{11}	Atherosclerosis	12.3
C_{12}	Disorders of refraction and accommodation	9.9
C_{13}	Iron deficiency anemias	9.8
C_{14}	General medical examination	9.6
C_{15}	Pneumonia, organism unspecified	9.4
C_{16}	Acute sinusitis	8.3
C_{17}	Benign neoplasm of skin	8.2
C_{18}	Other disorders of kidney and ureter	8.0
C_{19}	Other disorders of breast	7.8
C_{20}	Gastrointestinal hemorrhage	7.5

Generating exposure T

Given the covariate vector $\boldsymbol{X}$, the exposure variable T was generated from a Bernoulli distribution, Bernoulli(p_T), with exposure probability p_T based on the logistic model

$$\text{logit}(p_T) = \text{logit}(P(T = 1|\boldsymbol{X})) = \alpha_0 + \boldsymbol{\alpha}'_X \boldsymbol{X}, \tag{7.1}$$

which includes all measured covariates. The parameter α_0 denotes the intercept and $\boldsymbol{\alpha}_X = (\boldsymbol{\alpha}_D, \boldsymbol{\alpha}_C)$ is the parameter vector for $\boldsymbol{X}$ with parameter $\boldsymbol{\alpha}_D$ and $\boldsymbol{\alpha}_C$ for the components $\boldsymbol{D}$ and $\boldsymbol{C}$, respectively.

Generating outcome Y

Given $\boldsymbol{X}$ and T, the binary outcome Y was generated from a Bernoulli distribution, Bernoulli(p_Y), with p_Y defined as

$$\text{logit}(p_Y) = \text{logit}(P(Y = 1|\boldsymbol{X}, T)) = \mu + \beta T + \boldsymbol{\gamma}'_X \boldsymbol{X}, \tag{7.2}$$

with intercept μ, parameter of interest β and parameters $\boldsymbol{\gamma}_X = (\boldsymbol{\gamma}_D, \boldsymbol{\gamma}_C)$, whereat $\boldsymbol{\gamma}_D$ and $\boldsymbol{\gamma}_C$ denote the parameters for the covariates $\boldsymbol{D}$ and $\boldsymbol{C}$.

7.2 Score calculation

All summary measures in the simulation were calculated in the entire simulated cohort of size N=55 000. In summary, three different summary measures were considered defined as follows:

Any condition

As also done in (Riedl et al., 2013), we investigated the performance of the summary score $S(\boldsymbol{X})$ called 'any condition', defined as one, when any of the 20 conditions is present and as zero, otherwise, that is

$$S(\boldsymbol{X}) \quad = \quad \begin{cases} 0 & \text{if} \quad C_i = 0 \quad \forall \quad i = 1, \ldots, 20 \\ 1 & \text{if any } C_i = 1,\ i = 1, \ldots, 20. \end{cases}$$

Propensity score (PS)

The PS was estimated from the entire cohort as the predicted probability of having the exposure T given $\boldsymbol{X}$ using a logistic regression model, i.e. model (7.1) as

$$\text{logit}(\hat{e}(\boldsymbol{X})) = \text{logit}(P(T = 1|\boldsymbol{X})) = \hat{\alpha}_0 + \hat{\boldsymbol{\alpha}}'_X \boldsymbol{X}.$$

Disease risk score (DRS)

The DRS, was calculated as the predicted probability of disease outcome, using the logistic regression model including all covariates $\boldsymbol{X}$ and exposure T, in the whole cohort (i.e. model (7.2)). The score was then computed for all individuals as the fitted value from the logistic regression model, setting $T = 0$. The logit of the estimated '*full cohort*' DRS is given by

$$\text{logit}(\widehat{DRS}(\boldsymbol{X})) = \text{logit}(P(Y = 1|\boldsymbol{X}, T = 0)) = \hat{\mu} + \hat{\boldsymbol{\beta}}(T = 0) + \hat{\boldsymbol{\gamma}}'_X \boldsymbol{X}.$$

7.3 Case-control design

To create a case-control study, we individually matched a fixed number of randomly drawn cases $Y = 1$ (n=1000) to controls $Y = 0$ (n=1000). For each case, one control is chosen (1:1 ratio) based on the following matching variables: the demographic parameters $\boldsymbol{D}$, the estimated PS, $\hat{e}(\boldsymbol{X})$, the estimated PS on logit scale, the estimated DRS, $\widehat{DRS}(\boldsymbol{X})$ and the estimated DRS on the logit scale. We used a nearest neighbor algorithm implemented in the R-package 'Matching' (Sekhon, 2011) instead of an exact matching algorithm, which can lead to slight differences in scores of the matched sets. For covariate matching, a Mahalanobis distance matrix was used and the score matching matched each case to the nearest control based on the unidimensional metric of the score vector.

Assessing the quality of matches

To determine if the matching procedure has achieved covariate balance the standardized differences in the covariates $\boldsymbol{X}$ between cases ($Y = 1$) and controls ($Y = 0$) before and after matching were calculated based on the formulas 3.4 and 3.5 introduced in Chapter 3.

To investigate the balance of all measured covariates in the simulation, the mean of their standardized differences for each simulation was calculated. A standardized difference close to zero indicates good balance of the covariate between the cases and the controls. As proposed in Austin (2009), standardized differences of > 0.1 indicate considerable imbalance.

7.4 Outcome analysis

To control for confounding in the cohort design and to assess bias and efficiency of estimates of association, i.e. the log odds ratio β, multivariable logistic regression models were investigated. The following adjustments were considered in the simulation study: adjusting for all covariates $\boldsymbol{X}$ that were used to generate the data (Model 1), adjusting for the demographic parameters sex, age, selection year and race (i.e. $\boldsymbol{D}$) and the summary variable 'any condition' (Model 2), adjusting for the estimated PS on probability and on logit scale (Models 3a and 3b) and adjusting for the estimated DRS on both scales (Models 4a and 4b). Additionally, the crude outcome model, which includes only T without further adjustment, was fit to the data (Model 0). Model 1 that adjusts for all covariates represents the true outcome model. All other Models (2-4b), represent different misspecified outcome models.

To analyze the matched case-control data, conditional and unconditional analyses were investigated. Conditional logistic regression for matched pairs is termed Model 5. For unconditional analysis, the logistic regression models described above (Model 1 - Model 4b) were implemented.

7.5 Scenarios and settings

The associations of the demographic parameters $\boldsymbol{D}$ to exposure and outcome are based on the observed associations in SEER-Medicare data set and do not vary across the scenarios. These associations are summarized in Table 7.2. For the exposure and outcome model, the parameters $\boldsymbol{\alpha}_D$ and $\boldsymbol{\gamma}_D$ ranged from 0.2 to 0.9 and from -0.6 to 0.1, respectively.

Table 7.2: Parameters $\boldsymbol{\alpha}_D$ and $\boldsymbol{\gamma}_D$ for the demographic covariates for the exposure and outcome model.

Variable	$\boldsymbol{\alpha}_D$	$\boldsymbol{\gamma}_D$
D_1	0.20	-0.60
D_2		
D_{21}	Reference	Reference
D_{22}	0.30	0.10
D_{23}	0.60	0.10
D_{24}	0.90	-0.01
D_3		
D_{31}	Reference	Reference
D_{32}	0.30	0.10
D_{33}	0.40	0.05
D_{34}	0.40	-0.03
D_4		
D_{41}	Reference	Reference
D_{42}	0.30	0.10
D_{43}	0.40	-0.40

The relationship of the 20 conditions to exposure and outcome are summarized in Table 7.3. Two scenarios are used, i.e. in scenario I, strong confounder associations to exposure and outcome were chosen in both directions ($\boldsymbol{\alpha}_C$ and $\boldsymbol{\gamma}_C$ were set to ± 1), scenario II is based on similar associations of the conditions to exposure and outcome as observed in the real SEER-Medicare data set for liver cancer cases and ranged from -0.3 to 1.5 for $\boldsymbol{\alpha}_C$ and from -0.5 to 0.7 for $\boldsymbol{\gamma}_C$.

Both scenarios were investigated under three different settings:

- **Setting 1:** no exposure-outcome association ($\beta = 0$) and a disease prevalence of approximately 10% ($\mu = -3.5$ for scenario I and $\mu = -2.8$ for scenario II).
- **Setting 2:** exposure-outcome association present ($\beta = 1$) and a disease prevalence of 10% ($\mu = -3.5$ for scenario I and $\mu = -2.8$ for scenario II).
- **Setting 3:** exposure-outcome association present ($\beta = 1$) and a disease prevalence of 5% ($\mu = -5$ for scenario I and $\mu = -3.8$ for scenario II).

The prevalence for exposure was fixed to be approximately 10% for all scenarios ($\alpha_0 = -5$ for scenario I and $\alpha_0 = -5.1$ for scenario II).

Table 7.3: Parameters $\boldsymbol{\alpha}_C$ and $\boldsymbol{\gamma}_C$ for the conditions for the exposure and outcome model for scenario I and II.

		Scenario I		Scenario II	
Parameter	Description	$\boldsymbol{\alpha}_C$	$\boldsymbol{\gamma}_C$	$\boldsymbol{\alpha}_C$	$\boldsymbol{\gamma}_C$
C_1	Disorders of lipoid metabolism	1	1	0.1	-0.1
C_2	Cataract	1	1	-0.1	0.0
C_3	Special screening for malignant neoplasms	1	1	-0.3	-0.5
C_4	Other and unspecified anemias	1	1	1.5	0.6
C_5	Other dermatoses	1	1	-0.2	-0.1
C_6	Heart failure	-1	-1	1.3	0.5
C_7	Hyperplasia of prostate	-1	-1	-0.1	0.5
C_8	Disorders of fluid, electrolyte, and acid-base balance	-1	-1	1.2	0.3
C_9	Menopausal and postmenopausal disorders	-1	-1	-0.3	-0.7
C_{10}	Contact dermatitis and other eczema	-1	-1	0.1	0.1
C_{11}	Atherosclerosis	1	-1	0.9	0.3
C_{12}	Disorders of refraction and accommodation	1	-1	-0.2	0.0
C_{13}	Iron deficiency anemias	1	-1	1.4	0.7
C_{14}	General medical examination	1	-1	-0.3	-0.2
C_{15}	Pneumonia, organism unspecified	1	-1	1.0	0.3
C_{16}	Acute sinusitis	-1	1	0.0	-0.3
C_{17}	Benign neoplasm of skin	-1	1	-0.2	-0.1
C_{18}	Other disorders of kidney and ureter	-1	1	1.2	0.3
C_{19}	Other disorders of breast	-1	1	-0.2	-0.3
C_{20}	Gastrointestinal hemorrhage	-1	1	0.7	0.7

Presentation of the results

To summarize the results across the n=1000 simulation runs, the mean of the log odds ratios of the estimated exposure effects $\hat{\beta}_{s,j}$, $j = 1, \ldots, n$ for the different outcome models was calculated. That is

$$\text{mean}(\hat{\beta}_s) = \frac{1}{n} \sum_j \hat{\beta}_{s,j}.$$

The model-based variance (VarM), i.e. the average of the model based variances for $\hat{\beta}_{s,j}$ on the log scale and the empirical variance (VarE) across the simulation runs were calculated as follows:

$$\begin{aligned} \text{VarM} &= \frac{1}{n} \sum_j \widehat{Var}(\hat{\beta}_{s,j}), \\ \text{VarE} &= \frac{1}{n-1} \sum_j (\hat{\beta}_{s,j} - \text{mean}(\hat{\beta}_s))^2 \end{aligned}$$

We further calculated the relative bias for $\text{mean}(\hat{\beta}_s)$,

$$\text{relBias} = (\text{mean}(\hat{\beta}_s) - \beta)/\beta$$

and the coverage probability as percentage of estimated 95% confidence interval that include the true β:

$$P(\beta \in 95\%\ \text{CI}_j), \quad j = 1, \ldots, n,$$

with 95% confidence interval for the j-th simulation run

$$95\%\ \text{CI}_j = \left[\hat{\beta}_{s,j} \pm 1.96 * \sqrt{\widehat{Var}(\hat{\beta}_{s,j})}\right].$$

Proper coverage is indicated by a coverage probability close to 95%.

The simulation study was performed using the statistical software R-2.15.3. For matching the R-package 'Matching' (Sekhon, 2011) was used.

7.6 Results

For scenario I the results of the simulations under setting 1 and 2 for the cohort design and the matched case-control designs are summarized in Table A.1, for scenario II in Table 7.4. The results for both scenarios under setting 3 are given in Table A.2. The results for scenario I are similar to our analytic bias calculations for the cohort and the matched case-control study. We therefore focused in this section on the results for scenario II with more realistic confounder-exposure and confounder-disease associations.

Cohort study

There was no substantial bias in the exposure effect $\hat{\beta}_s$ estimated with the model that adjusts for all confounders (Model 1), for the PS (Model 3a) and for the logit DRS (Model 4b) in all three settings. If the summary variable 'any condition' (Model 2) was used as an adjustment variable instead of the 20 conditions, the estimated exposure effect was 1.288 for setting 1 ($\beta = 0$) and more than twice as large as the true effect for setting 2 and 3 ($\beta = 1$). The bias was even larger compared with the model that includes only the exposure as covariate (Model 0). In addition to the models in Table 7.4, including only the characteristics $\boldsymbol{D}$ resulted in largest bias compared with all investigated models with estimated exposure effects of 1.324, 2.201 and 2.3 for the three settings, respectively.

Adjusting for the PS on the logit scale (Model 3b) resulted in an underestimation of 6% for $\beta = 1$ and a common disease prevalence of 10% (setting 2). However, decreasing the disease prevalence to 5% (setting 3) the bias for this model decreased to $<$3%. Adjusting on the DRS (Model 4a) resulted in an overestimation of 12% in setting 2 and with decreasing the disease prevalence in setting 3, the bias for this model increased to $>$30%.

Matched case-control study

When cases and controls were matched on the demographic parameters $\boldsymbol{D}$, the adjustment for all covariates $\boldsymbol{X}$ by unconditional logistic regression (Model 1) yielded to unbiased results of the true β. Conditional logistic regression or adjusting on $\boldsymbol{D}$ and the summary variable 'any condition' resulted in substantial bias of similar amount compared with the bias in the cohort analysis. If unconditional logistic regression adjusted for the PS or DRS was used to analyze the covariate matched data, only a small bias was given for the PS or logit PS models (Model 3a and 3b). A bias up to 30% was observed for the DRS models (Model 4a and 4b) with coverage probabilities ranging from 16% to 62%.

Table 7.4: Mean estimates $\hat{\beta}_s$ from simulations based on blood transfusion data, with disease prevalence of 10% and exposure prevalence of 10% for $\beta = 0$ and $\beta = 1$ under scenario II.

	$\beta = 0$				$\beta = 1$				
Model	mean $\hat{\beta}_s$	VarM	VarE	coverage	mean $\hat{\beta}_s$	VarM	VarE	relBias	coverage
	Full cohort analysis								
0	1.286	0.001	0.002	0	2.122	0.001	0.001	1.122	0
1	0.001	0.003	0.003	93.9	1.000	0.002	0.002	0	95.6
2	1.288	0.002	0.002	0	2.165	0.001	0.001	1.165	0
3a	0.001	0.004	0.004	93.6	1.000	0.002	0.002	0	96.1
3b	0.029	0.003	0.003	92.8	0.942	0.002	0.002	-0.058	79.2
4a	0.046	0.003	0.004	79.1	1.116	0.002	0.002	0.116	22.3
4b	0.001	0.002	0.003	90.4	1.000	0.002	0.002	0	91.8
	Case-control analysis, matching on sex, age, race and selection year								
1	0.010	0.031	0.030	95.5	1.015	0.030	0.030	0.015	95.5
2	1.281	0.017	0.018	0	2.149	0.019	0.019	1.149	0
3a	0.010	0.030	0.032	95.2	0.985	0.028	0.027	-0.015	95.2
3b	0.075	0.028	0.027	93.2	1.021	0.027	0.026	0.021	95.1
4a	0.346	0.023	0.021	35.3	1.249	0.021	0.020	0.249	61.7
4b	0.425	0.021	0.019	16.2	1.279	0.021	0.019	0.279	52.0
5	1.319	0.020	0.020	0	2.191	0.027	0.026	1.191	0
	Case-control analysis, matching on $\hat{e}(\boldsymbol{X})$								
1	-0.023	0.023	0.021	96.1	0.952	0.020	0.020	-0.048	93.1
3a	0.002	0.022	0.021	95.1	0.930	0.019	0.019	-0.070	92.0
5	0.002	0.022	0.021	95.1	0.937	0.021	0.021	-0.063	92.0
	Case-control analysis, matching on logit($\hat{e}(\boldsymbol{X})$)								
1	-0.014	0.023	0.021	96.0	0.943	0.020	0.021	-0.057	93.0
3b	0.011	0.020	0.018	96.0	0.869	0.018	0.017	-0.131	82.9
5	0.012	0.022	0.022	95.2	0.928	0.021	0.022	-0.072	90.7
	Case-control analysis, matching on $\widehat{DRS}(\boldsymbol{X})$								
1	0.004	0.020	0.020	94.4	1.009	0.019	0.020	0.009	94.7
4a	0.004	0.015	0.015	94.4	0.956	0.013	0.014	-0.044	92.8
6	0.006	0.015	0.017	94.0	1.007	0.015	0.016	0.007	93.8
	Case-control analysis, matching on logit($\widehat{DRS}(\boldsymbol{X})$)								
1	0.002	0.020	0.021	94.1	1.014	0.019	0.020	0.014	94.1
4b	0.001	0.014	0.015	93.1	0.967	0.013	0.013	-0.033	94.9
5	0.003	0.015	0.017	92.4	1.011	0.015	0.016	0.011	94.5

Model 0: logistic regression including only exposure T (crude model)
Model 1: logistic regression adjusted for all covariates $\boldsymbol{X}$
Model 2: logistic regression adjusted for sex, age race, selection year, and 'any condition'
Model 3a: logistic regression adjusted for estimated PS, $\hat{e}(\boldsymbol{X})$
Model 3b: logistic regression adjusted for estimated PS on logit scale, logit($\hat{e}(\boldsymbol{X})$)
Model 4a: logistic regression adjusted for estimated DRS, $\widehat{DRS}(\boldsymbol{X})$
Model 4b: logistic regression adjusted for estimated DRS on logit scale, logit($\widehat{DRS}(\boldsymbol{X})$)
Model 5: conditional logistic regression

When the case-control matching was performed on the PS or on the logit PS, no bias was observed for $\beta = 0$ for all models. In the settings with $\beta = 1$, there was a slight underestimation of the true exposure effect with relative bias $< 10\%$. This underestimation was more pronounced when matching on the logit-PS followed by unconditional logistic regression adjusting for the logit-PS (Model 3b) was performed, with a relative bias of 13% and coverage probability of 83%. For the more rare disease scenario 3, the bias was less pronounced and ranged from 3% to 9%. Matching on the DRS or on the logit DRS resulted in no substantial bias for all models. Only in scenario 3, an underestimation was observed when matching on the DRS followed by unconditional logistic regression adjusting for the DRS (Model 4a) was performed (relative bias of 10% and coverage probability of 84%).

Quality of matches

To check if the different matching designs resulted in good balance of the covariates between the cases and the controls we compared the standardized differences for the covariates $\boldsymbol{D}$ and $\boldsymbol{C}$. For scenario I under setting 2 the results are plotted in Figure 7.1. For the other scenarios the results are similar (data not shown).

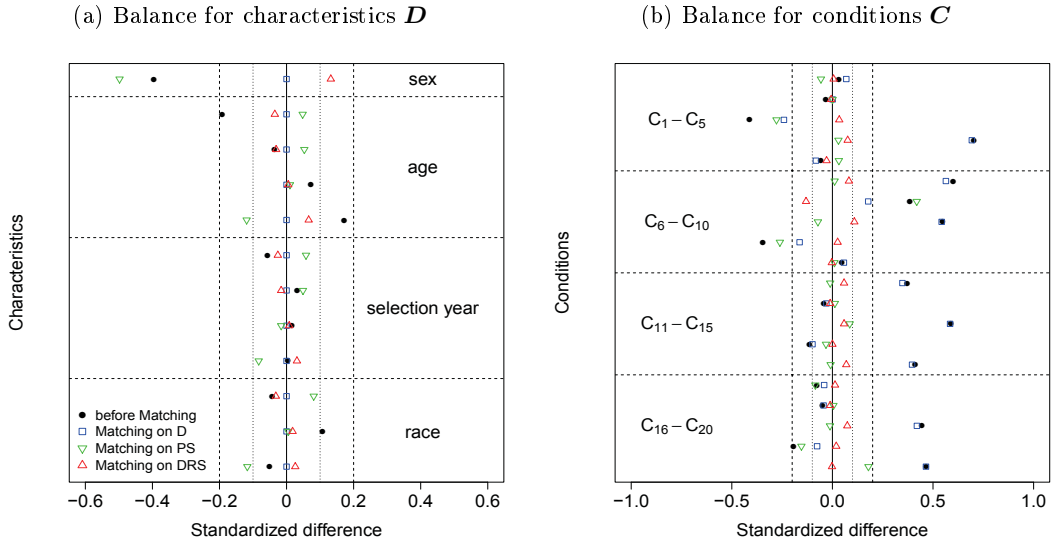


Figure 7.1: Standardized differences before and after case-control matching on $\boldsymbol{X}$, PS and DRS for scenario II under setting 2.

Before matching, large imbalances were seen for more than half of the conditions (n=9 with a standardized difference >0.1 and n=4 with a standardized difference <-0.1) and for the characteristics $\boldsymbol{D}$, the largest imbalance was present for sex with a standardized difference of -0.4.

Matching on the covariates $\boldsymbol{D}$ resulted in perfect balance only for the matching variables $\boldsymbol{D}$ (Figure 7.1 (a)). No substantial reduction in the standardized differences for the conditions $\boldsymbol{C}$ after matching could be achieved compared with the standardized differences before matching (Figure 7.1 (b)). For n=11 conditions large imbalances were still present (n=9 with a standardized difference >0.1 and n=2 with a standardized difference <-0.1). Nearly all of these conditions were either positively associated or negatively associated with both exposure and outcome which explains the remaining large positive bias after case-control matching on only the characteristics $\boldsymbol{D}$.

Comparing the balance for matching on the summary scores PS and DRS, somewhat better balance was achieved for the DRS. Only 3 baseline characteristics and conditions (sex, hyperplasia of prostate (C7) and disorders of fluid (C8)) showed a standardized difference greater than ± 0.1 if case-control matching was performed on the DRS. After PS-matching, the standardized difference was greater than ± 0.1 for 8 baseline characteristics and conditions. For 2 variables (sex and hyperplasia of prostate C_7) the balance was worse after matching compared to before. However, in other simulation studies (data not shown) we observed, that if matching on the PS in combination with selected covariates was performed, good balance could be achieved which further reduced bias. Concerning balance between cases and controls of the PS itself, we observed a standardized difference of zero. Thus, the balance issue is not driven by the specific matching algorithm and even if a caliper matching is used instead of a nearest neighbor, the quality of the matches would not be improved.

Comparing the results of scenario I and II, the models that adjust on $\boldsymbol{D}$ and 'any condition' (Model 2) and on the DRS (Models 4a and 4b) for cohort analysis in scenario I resulted in less bias compared with scenario II. For example, the relative bias was approximately 20% in setting 3 for the Models 2 and 4a in scenario I and $>30\%$ in scenario II (Table A.2). In contrast, larger negative bias was observed for the PS Models 3a and 3b ranging from 8% up to 30% for $\beta = 1$ in scenario I, whereas in scenario II the bias remained under 6%. The same was observed for the different case-control matched designs: less bias for matching on the demographic covariates or the DRS and larger bias for matching on the PS.

Chapter 8

Data example: a study of blood transfusions and the subsequent risk of cancers in the U.S. elderly

We illustrate the summary score methods with an investigation of the associations of blood transfusion with risk of cancer in the US-elderly (this chapter is based on the publication Riedl et al. (2013)). To answer this question, we designed a matched case-control study based on the SEER-Medicare database (Engels et al., 2011). First we matched cases and controls on various covariates and then we repeated the matching using propensity scores and disease risk scores. This chapter describes the corresponding results.

8.1 Motivation

The contributions of inflammation, immune disturbances, and infection to the etiology of cancer are increasingly recognized. The process of cancer-related inflammation is complex, involving numerous cell types and signaling molecules, but it is posited that these pathways lead to chronic tissue damage that, in turn, promotes cancer through the accumulation of genetic changes, cell proliferation, and angiogenesis (Coussens and Werb, 2002). Immunosuppression, due to infection with human immunodeficiency virus (HIV) or from medications used in organ transplantation, increases the risk for numerous malignancies (Grulich et al., 2007). Worldwide, approximately 16% of all cancers (23% in less developed countries and 7% in developed countries) are attributable to viruses and other infectious agents (De Martel et al., 2012).

In this context, it is important to consider that blood transfusions may be a cause of some malignancies. Receipt of a blood transfusion may increase risk of cancer by causing immune suppressive or proinflammatory changes in the recipient's immune system, a constellation of effects termed 'transfusion-related immunomodulation' (TRIM) (Vamvakas and Blajchman, 2007). Blood transfusions might also increase risk by transmitting infectious agents associated with cancer risk, including Epstein-Barr virus, hepatitis C virus (HCV), and HIV, although transfused blood is typically screened now for most of these agents (Buddeberg et al., 2008; Dwyre et al., 2011).

The incidence of many types of cancer increases with age, and blood transfusions are also disproportionately more common among elderly adults (Vamvakas and Taswell, 1994; Anderson et al., 2007; Rogers et al., 2011). While an assessment of the contribution of blood transfusions to risk of cancer in the elderly is important, it is also possible that the presence of an undiagnosed cancer or a precursor to cancer in an individual causes anemia. These undiagnosed conditions could result in associations between blood transfusion and subsequent risk of cancer due to reverse causation. Along these lines, reverse causation was the explanation Hjalgrim et al. (2007) gave for a marked increase in cancer risk they found in blood transfusion recipients from Denmark and Sweden, shortly after the transfusion. The overall standardized incidence ratios (SIRs) of cancer 6 months after transfusion were 4.80 (95% CI: 4.69-4.91) for recipients aged 70-79 years and 4.11 (95% CI: 4.01-4.22) for recipients 80 years or older. The SIRs of most cancers decreased with longer time periods since transfusion to near unity. However, the risk of tobacco- and alcohol- related cancers, including cancers of the tongue, mouth, pharynx, esophagus, liver and respiratory and urinary tracts, was still increased 10-20 years after blood transfusion suggesting that lifestyle-related factors correlated with conditions prompting transfusions may cause the observed cancer occurrence.

In the U.S., a detailed assessment of the association of transfusions with cancer risk has so far only been undertaken for hematologic malignancies. Chang et al. (2010) investigated the risk of hematologic malignancies after blood transfusion among the U.S. elderly using the Surveillance, Epidemiology and End Results (SEER)-Medicare database. They found that 7.9% of hematologic malignancy cases compared to 5.9% controls had a history of transfusion. Blood transfusions were associated with increased risk of multiple hematologic malignancy subtypes and risk following blood transfusion remained elevated for two non-Hodgkin lymphoma (NHL) subtypes, lymphoplasmacytic lymphoma and marginal zone lymphoma, at latency periods of several years, consistent with a possible etiologic relationship. However, Chang et al. (2010) excluded transfusions occurring within one year of cancer diagnosis, and only studied the impact of latencies of one or more years.

To further characterize the associations of blood transfusion with risk of cancer overall and for specific sites, we conducted a large case-control study in the U.S. elderly population using SEER-Medicare data. In our study (Riedl et al., 2013), we addressed the possibility of reverse causation by examining cancer risk in several intervals of time following blood transfusion, and also controlled for confounding of the associations between transfusion and cancer risk by other diagnosed medical conditions.

8.2 Materials

We used the SEER-Medicare database that links data from the National Cancer Institute's SEER program cancer registries and the U.S. government health insurance program Medicare. Currently, the SEER cancer registries cover approximately 26% of the U.S. population. Within their catchment areas, the SEER registries are highly complete, recovering 97-98% of cancer cases (Zippin et al., 1995). Medicare provides health insurance Part A coverage (hospital coverage includes inpatient care) for approximately 97% of people aged 65 or older in the U.S. Medicare Part B covers physician and outpatient services for approximately 96% of beneficiaries (Engels et al., 2011). Full details about the linkage are described in Potosky et al. (1993). In brief, the SEER-Medicare data are linked based on a deterministic matching algorithm that has been used by the National Center

for Health Statistics to match to the National Death Index. The matching variables include each person's social security number, first name, last name, and date of birth. For the creation of the algorithm, hundreds of records were reviewed to determine the presence of false positives or false negatives. The data are linked every two years. Each linkage includes millions of Medicare beneficiaries, with an overall match rate of 94% for persons age 65 and older.

8.2.1 Definition of cases and control population

Cases were defined as persons diagnosed with a first tumor during 1997-2005 in the SEER database. Solid tumors were defined using the SEER site recode (http://seer.cancer.gov/siterecode/icdo3_d01272003/) based on the International Classification of Diseases for Oncology, 3rd edition (Fritz et al., 2000) site and morphology codes. To ensure that subjects had enough time to provide exposure and confounder information, we included only cases with a minimum of 60 continuous months of Medicare coverage before diagnosis. Due to this restriction, only individuals ages 70 years or older were selected. We excluded cases who were 100 years or older at diagnosis. We excluded cases diagnosed only on death certificate or at autopsy, those with missing month of diagnosis and cases whose diagnosis date was after their date of death. To ensure complete claims were available, cases had to have continuous Part A and B coverage and no health maintenance organization (HMO) enrollment during the 60 month period. After these exclusions, our study was based on a total of 552 951 cases diagnosed with a first cancer during 1997-2005 in SEER.

Controls were selected from a 5% random sample of Medicare beneficiaries living in the SEER catchment areas. To avoid biases due to coding changes or gaps in Medicare coverage, individuals who were alive and cancer free as of July 1 in the calendar year of selection of the cases were eligible as controls. Similarly to the cases, we included only individuals with a minimum of 60 continuous months of Medicare coverage before selection and continuous Part A and B coverage and no HMO enrollment during the 60 month period. Controls could have been sampled multiple times in different calendar years or could later have become a case. After the same exclusions as applied to cases, there were 267 165 unique controls in the 5% random sample from Medicare beneficiaries during 1997-2005. These controls were eligible for selection in multiple years. In summary, the control data set includes 1 237 153 records for 267 165 unique subjects. 15.9% of controls were selected once, 13.3% twice, 10.6% three times and 60.2% four or more times.

8.2.2 Blood transfusions

We reviewed Medicare claims from hospitals and physicians and other noninstitutional medical care providers to identify any blood transfusions administered. For Medicare hospital claims, blood transfusions were defined as receipt of packed cells (International Classification of Diseases [version 9, ICD-9] procedure code 9904) or an indication that the number of transfused units (BLDPNTS variable) during the hospitalization was greater than zero. For provider claims, we used the Healthcare Common Procedure Codes of P9016, P9021, P9022, P9038, P9039, P9040, P9058, C1020, C1021, C9504, C9505, P9057. The analysis only includes blood that was administered. Blood products that were prepared, but not used, are not included as they cannot be billed to Medicare.

To address the latency of any blood transfusion effect on cancer risk, we investigated the receipt of transfusions in three time periods: 0-12 months, 13-30 months and 31-48 months prior

to cancer diagnosis or control selection. If a person had more than one transfusion in a given interval, we selected the transfusion closest to the midpoint of the interval, otherwise we used the actual transfusion date in that interval. For an un-transfused person the index date for each transfusion interval was the midpoint of the interval. Figure 8.1 gives an example of the three time intervals and the corresponding diagnosis/selection and transfusion/index dates.

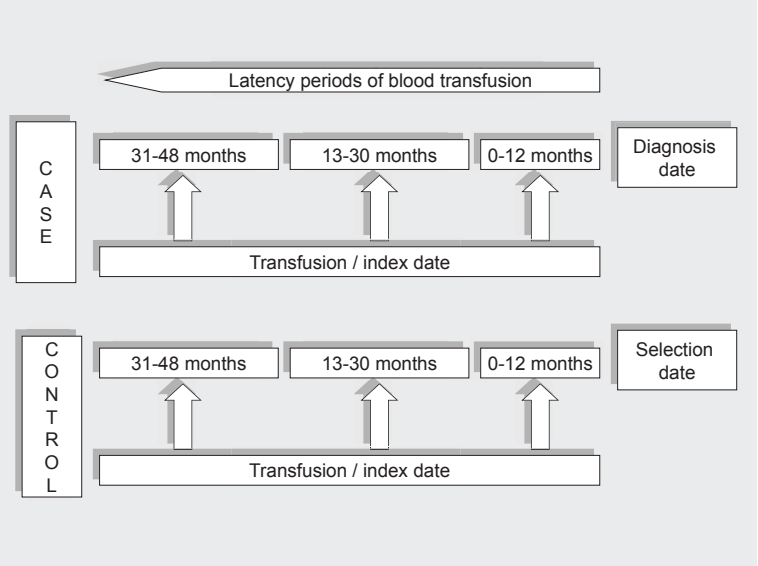


Figure 8.1: Description of ascertainment of blood transfusion in relation to date at diagnosis (cases) or selection (controls) for a particular case and the corresponding matched control.

8.2.3 Possible confounders

Using Medicare claims, we identified a total of 1 822 conditions recorded in the Medicare database for cases diagnosed from 1997-2005. As many of those conditions were extremely rare, we focused on those that potentially could have noticeable impact on the findings, given their frequencies and strengths of association with blood transfusion. Ninety-six acute medical diagnoses and chronic conditions including gastrointestinal hemorrhage, gastric ulcer, duodenal ulcer, bronchopneumonia, and vascular insufficiency of intestine, were associated with the probability of receiving a blood transfusion and fulfilled the following criteria: the condition was present in at least 75 transfused cases and the condition was associated with transfusion, with an odds ratio larger or equal than 1.4 or smaller or equal than 0.7. The 96 medical diagnoses and chronic conditions are summarized in Table 8.1. We considered medical conditions reported within 12 to 24 months prior to the transfusion/index date to be relevant for the transfusion in that interval.

Table 8.1: Medical diagnoses and chronic conditions considered as possible confounders and their frequency and association (OR) with transfusion in cancer cases.

	No transfusion	Transfusion		
Condition based on ICD9 Classification	Percent	Count	Percent	OR
Other congenital anomalies of nervous system	0.01	127	0.04	3.67
Aplastic anemia and other bone marrow failure syndromes	0.04	343	0.12	3.22
Acquired hemolytic anemias	0.01	77	0.03	3.03

Acute renal failure	0.06	511	0.17	2.75
Other and unspecified protein-calorie malnutrition	0.03	242	0.08	2.56
Renal failure, unspecified	0.08	587	0.20	2.43
Hypertensive chronic kidney disease	0.06	403	0.14	2.43
Chronic kidney disease (CKD)	0.15	1062	0.36	2.41
Pneumonitis due to solids and liquids	0.02	152	0.05	2.40
Septicemia	0.08	524	0.18	2.35
Iron deficiency anemias	0.33	2144	0.73	2.21
Other disorders of stomach and duodenum	0.03	179	0.06	2.19
Nephritis and nephropathy, not specified as acute or chronic	0.02	130	0.04	2.19
Other and unspecified anemias	0.74	4555	1.55	2.10
Osteomyelitis, periostitis, and other infections involving bone	0.03	187	0.06	1.90
Gastrointestinal hemorrhage	0.34	1876	0.64	1.87
Other congenital anomalies of circulatory system	0.04	201	0.07	1.85
Fracture of other and unspecified parts of femur	0.03	136	0.05	1.85
Pulmonary congestion and hypostasis	0.12	657	0.22	1.83
Care involving use of rehabilitation procedures	0.03	154	0.05	1.80
Gastric ulcer	0.06	340	0.12	1.78
Transient mental disorders due to conditions classified elsewhere	0.04	214	0.07	1.77
Pleurisy	0.20	1056	0.36	1.76
Other diseases of blood and blood-forming organs	0.06	322	0.11	1.75
Other deficiency anemias	0.21	1052	0.36	1.74
Pneumococcal pneumonia [Streptococcus pneumoniae pneumonia]	0.03	161	0.05	1.73
Other bacterial pneumonia	0.08	381	0.13	1.69
Chronic ulcer of skin	0.13	637	0.22	1.68
Complications peculiar to certain specified procedures	0.14	668	0.23	1.67
Other disorders of kidney and ureter	0.32	1577	0.54	1.67
Purpura and other hemorrhagic conditions	0.10	471	0.16	1.64
Heart failure	0.98	4678	1.59	1.64
Fracture of pelvis	0.02	117	0.04	1.63
Diseases of white blood cells	0.10	466	0.16	1.63
Intestinal obstruction without mention of hernia	0.11	550	0.19	1.63
Duodenal ulcer	0.04	188	0.06	1.62
Other disorders of pancreatic internal secretion	0.05	253	0.09	1.61
Fracture of neck of femur	0.10	452	0.15	1.61
Disorders of fluid, electrolyte, and acid-base balance	0.48	2256	0.77	1.59
Hypertensive heart and chronic kidney disease	0.04	169	0.06	1.59
Arterial embolism and thrombosis	0.08	368	0.13	1.59
Other complications of procedures, NEC	0.07	329	0.11	1.57
Other ill-defined and unknown causes of morbidity and mortality	0.18	835	0.28	1.56
Diseases of aortic valve	0.03	137	0.05	1.55
Bronchopneumonia, organism unspecified	0.03	135	0.05	1.54
Viral pneumonia	0.02	103	0.04	1.53
Infections of kidney	0.03	150	0.05	1.51
Hypotension	0.12	543	0.18	1.50
Acute myocardial infarction	0.16	688	0.23	1.50
Disorders of plasma protein metabolism	0.03	149	0.05	1.50
Other disorders of arteries and arterioles	0.09	378	0.13	1.49
Diseases of other endocardial structures	0.05	221	0.08	1.49
Other nonorganic psychoses	0.08	329	0.11	1.48
Chronic pulmonary heart disease	0.08	359	0.12	1.46
Nonspecific abnormal findings in other body substances	0.08	346	0.12	1.44
Hydronephrosis	0.04	163	0.06	1.44
Secondary hypertension	0.02	80	0.03	1.43
Vascular insufficiency of intestine	0.02	75	0.03	1.43
Other congenital anomalies of heart	0.11	452	0.15	1.42

Chronic liver disease and cirrhosis	0.07	280	0.10	1.41
Atherosclerosis	0.36	1491	0.51	1.41
Cardiomyopathy	0.24	992	0.34	1.40
Pneumonia, organism unspecified	0.45	1840	0.63	1.40
Other venous embolism and thrombosis	0.09	355	0.12	1.40
Disorders of lipoid metabolism	2.68	5613	1.91	0.71
Inflammation of eyelids	0.28	580	0.20	0.70
Acute sinusitis	0.37	758	0.26	0.69
Need for prophylactic vaccination and inoculation against combinations of diseases	0.07	141	0.05	0.69
Superficial injury of other, multiple, and unspecified sites	0.04	85	0.03	0.69
Cataract	2.37	4816	1.64	0.69
Open wound of finger(s)	0.05	106	0.04	0.68
Special screening for other conditions	0.06	117	0.04	0.68
Contact dermatitis and other eczema	0.59	1165	0.40	0.67
Thyrotoxicosis with or without goiter	0.09	183	0.06	0.67
Migraine	0.06	109	0.04	0.67
Acute pharyngitis	0.21	412	0.14	0.66
Need for isolation and other prophylactic measures	0.05	104	0.04	0.66
Lipoma	0.04	76	0.03	0.66
Retinal detachments and defects	0.08	142	0.05	0.62
Inflammatory diseases of prostate	0.19	347	0.12	0.62
Erythematous conditions	0.15	264	0.09	0.61
Benign neoplasm of skin	0.42	747	0.25	0.61
Disorders of refraction and accommodation	0.35	634	0.22	0.61
Other hypertrophic and atrophic conditions of skin	0.14	240	0.08	0.60
OTHER DIS VIRUS	0.11	189	0.06	0.59
Hyperplasia of prostate	1.30	2270	0.77	0.59
Need for prophylactic vaccination and inoculation against single diseases	0.06	104	0.04	0.59
Other dermatoses	1.47	2550	0.87	0.59
Special screening for malignant neoplasms	1.41	2428	0.83	0.58
Other disorders of prostate	0.06	92	0.03	0.56
Other disorders of breast	0.34	550	0.19	0.56
Menopausal and postmenopausal disorders	0.47	764	0.26	0.55
Disorders of penis	0.16	257	0.09	0.55
Family history of malignant neoplasm	0.08	127	0.04	0.54
Benign mammary dysplasias	0.21	277	0.09	0.46
General medical examination	0.30	396	0.13	0.45

8.3 Methods

8.3.1 Study designs

Following two study designs were performed:

A. Cancer cases and all controls

We analyzed the data based on all cases (n=552 951) and all eligible controls (n=1 237 153) from the 5% random sample of Medicare beneficiaries.

B. Matching

To ensure that controls are at risk a similar length of time as cases, we selected controls matched to cases. The controls were selected from the control population defined in Section 8.2.1. as follows:

1. **Covariate matching**

   Here we sampled a total number of 100 000 controls based on a frequency matched to cases by sex, age in four categories (70-74, 75-79, 80-84, and 85-99 years), and calendar year of selection (in single years). The key feature of this matching is that controls (in the same age group and gender) had to be alive and cancer free on July 1 of the calendar year of selection of the cases. Therefore, similar follow-up time is ensured.

2. **Propensity score matching**

   Here controls were frequency matched to cases by calendar year of selection, sex and propensity scores. Three propensity scores, defined as the probability of receiving a blood transfusion in the three lag-time intervals (0-12 months, 13-30 months and 31-48 months) given covariates, were computed for each person from the 5% random Medicare sample and for each case, for men and women and separately for every calendar year, also ensuring similar follow-up time. The propensity score for each interval was calculated using a logistic regression model that included age in four categories, race and any of the 96 transfusion associated conditions identified using stepwise selection (based on a significance level for entry of 5%). Cases and controls were then frequency matched by sex and the deciles of the three propensity scores together for each period, with a matching ratio of 1:0.2.

3. **Disease risk score matching**

   Similar to the propensity score matching, we calculated three disease risk scores defined as conditional probability of cancer in the absence of transfusion given the measured baseline covariates for the three investigated lag-time intervals for receiving a blood transfusion (0-12 months, 13-30 months and 31-48 months). The disease risk score for each interval was calculated separately for men and women and separately for every calendar year by using a logistic regression model that included transfusion, age in four categories, race and any of the 96 transfusion associated conditions identified using stepwise selection (based on a significance level for entry of 5%). The scores were then computed for all individuals in the entire cohort as the fitted value from the logistic regression, setting transfusion to zero. Cases and controls were then frequency matched by sex and the deciles of the disease risk scores for each period, with a matching ratio of 1:0.2.

For all three matching designs, we used frequency matching to select controls to achieve overall balance of the matching factors in our data set, especially gender and time at risk. The characteristics of the cases and the controls in the population were somewhat different (Table 8.2). For example, among cases 50.6% were male, while in the Medicare random sample 35.7% were male. Our original plan was to select one control for each case, but due to the appreciable gender imbalance we were not able to find one control for each case for the overall cancer analysis.

Table 8.2: Characteristics of all eligible SEER-Medicare cases and controls

Characteristic	Cases (n = 552 951)		Controls (n = 267 165)	
	N	Percent %	N	Percent %
Sex				
Male	280 034	50.64	100 844	37.75
Female	272 917	49.36	166 321	62.25
Age at diagnosis or selection (years)				
70-74	159 254	28.80	129 772	48.57
75-79	167 536	30.30	60 221	22.54
80-84	126 165	22.82	40 846	15.29
85-99	99 996	18.08	36 326	13.60
Median (Range)	78 (70-99)		75 (70-99)	

8.3.2 Statistical analysis

We used unconditional logistic regression models to estimate odds ratios (ORs) and 95% confidence intervals (CIs) for the associations of cancer overall with transfusions received in the three time periods. To assess the dose-response relationship of blood transfusions and cancer risk, we created a variable defined as having had 0, 1, 2 or 3 periods with blood transfusions. We also combined transfusions over all transfused periods prior to cancer diagnosis/selection into a variable with eight categories (no transfusions, transfusions only during 0-12 months, only during 13-30 months, only during 31-48 months, transfusions in 0-12 and 13-30 months, 0-12 and 31-48 months, 13-30 and 31-48 months, and transfusions received in all three time periods). For the analysis of cancers at specific sites, we used polytomous logistic regression models. We accounted for the repeated sampling of controls and the fact that some controls later became cases using a robust variance estimator (Engels et al., 2011).

A. Cancer cases and all controls

In the analysis based on the cancer cases and all controls, all models were adjusted for sex, age in four categories (70-74, 75-79, 80-84 and 85-99 years), year of diagnosis or selection in four categories (1997-1999, 2000-2001, 2002-2003 and 2004-2005), race (white, black, Asian, others/ unknown) and the summary variable 'any conditions', that was defined to be 1 when any of the 96 conditions associated with blood transfusion was diagnosed and zero otherwise.

The logistic model for overall cancer risk was adjusted for age, race, selection year and any condition as categorical variables. The logistic model for the site specific cancer groups included sex, age, race and any condition as categorical variables and selection year as a continuous variable. The models for subtypes of the site specific cancers are adjusted for sex and race as categorical variables and age, selection year and any condition as continuous variables. We performed these different adjustments to reduce the number of parameters to estimate.

B. Matching

For the two matching approaches, described above, the following adjustments were used:

1. **For covariate matched data**

   All models were adjusted for the matching factors sex, age and year of diagnosis or selection and for race and 'any conditions' in the same way as for the analysis under design A. Further, all models were additionally stratified by sex.

2. **For propensity score matched data**

   Odds ratios for the association of cancer risk with blood transfusion for the three time periods were estimated using unconditional logistic regression models adjusted for sex, age, year of diagnosis or selection and race and the estimated propensity scores. The OR for overall cancer risk was adjusted for sex, age, race and selection year as categorical variables and for the estimated propensity scores as continuous variables. The ORs for the site specific cancers were adjusted for sex, age and race as categorical variables and selection year and the estimated propensity scores as continuous variables. The ORs for subtypes of the site specific cancer groups are adjusted for sex and race as categorical variables and age, selection year and the estimated propensity scores as continuous variables.

   Because the matching on the propensity score was performed in deciles and the adjustment by using a continuous variable, we investigated the impact of different adjustments (continuous propensity score, continuous propensity score on the logit scale, propensity score in deciles with a trend and in ten separate categories) on the estimates of the overall cancer risk.

3. **For disease risk score matched data**

   The same adjustments were used as for the propensity score matched data with the exception that, instead of the estimated propensity scores, the estimated disease risk scores on the logit scale were used. For overall cancer risk, we also investigated different adjustments, i.e. we adjusted for the continuous disease risk scores, the continuous disease risk scores on the logit scale, in deciles with a trend and in ten separate categories.

Sensitivity analyses

We performed a sensitivity analysis based on the covariate matched dataset as follows:

- For all cancers that were significantly associated with transfusion, we fitted logistic regression models additionally adjusted for each of the individual 96 conditions separately and combinations of the conditions. For the combined adjustment, the special conditions where selected based on expert opinions and on the results of the single logistic regressions for the 96 conditions, i.e. the conditions that mostly influenced the ORs were considered for inclusion.
- We estimated the overall cancer risk for a subgroup of cases and controls from whom no conditions were present and for a subgroup of individuals for whom at most one condition was present.

For discrepant results for site-specific cancers and subtypes we repeated our different matching approaches by restricting the cancer cases to the investigated subgroup. The controls were matched to the investigated cancer cases out of the 5% random sample of the US elderly (n=1 237 153 controls) by using the three different matching designs: covariate matching (frequency matching on sex, age in four categories (70-74, 75-79, 80-84, and 85-99 years) and calendar year of selection (in single years) with a matching ratio of 1:1), propensity score matching and disease risk score matching. Matching on the summary scores was performed using frequency matching by sex and the deciles of the scores for each period, with a matching ratio of 1:1. The scores were computed as described before, but with the restriction to the investigated cancer subgroup.

We present ORs and 95% CIs and we focus on results that are still significant after a Bonferroni adjustment for 129 comparisons (43 sites for 3 time periods) with a type one error rate of 5%. After this correction findings with $p<0.0004$ are considered significant. All analyses were performed using SAS, version 9.1 (SAS Institute, Inc., Cary, NC).

8.4 Results

8.4.1 Analysis of all cases and all controls

Characteristics of all cases and all controls

The characteristics of all cases (n = 552 951) and all eligible controls (n = 1 237 153) in the 5% random sample of Medicare beneficiaries are summarized in Table 8.3. For all characteristics significant differences between the cases and controls were observed, except for any condition in the time period 13-30 months. The largest imbalance was observed for sex. Among cancer cases 49.4% were female, whereas among the controls 64.3% were female. Compared to the cases slightly more controls were observed in the age groups 70-74 and 85-99 years. The median age at diagnosis or selection was 78 years. The majority of the study population was white, and cases were slightly more likely to be white or black than controls.

The presence of a medical condition associated with blood transfusion was somewhat higher for the time periods closer to diagnosis/selection. Slightly more cases with any condition than controls were observed in the time period 0-12 months (92.3% and 92.0 % respectively) and slightly fewer cases in the time period 31-48 months (89.0% and 89.4 % respectively).

Table 8.3: Characteristics of cases and controls in the SEER-Medicare database from 1997 to 2005.

Characteristic	Cases (n = 552 951) N (%)	Controls (n = 1 237 153) N (%)	p-value
Sex			
Male	280 034 (50.6)	441 484 (35.7)	<.0001
Female	272 917 (49.4)	795 669 (64.3)	
Age at diagnosis or selection (years)			
70-74	159 254 (28.8)	371 234 (30.0)	<.0001
75-79	167 536 (30.3)	361 533 (29.2)	
80-84	126 165 (22.8)	268 046 (21.7)	
85-99	99 996 (18.1)	236 340 (19.1)	
Selection year			
1997-1999	110 534 (20.0)	255 985 (20.7)	<.0001
2000-2001	153 125 (27.7)	317 376 (25.7)	
2002-2003	147 348 (26.7)	323 550 (26.2)	
2004-2005	141 944 (25.7)	340 242 (27.5)	
Race			
White	482 412 (87.2)	1 057 306 (85.5)	<.0001
Black	39 249 (7.1)	82 241 (6.7)	
Asian	13 105 (2.4)	40 753 (3.3)	
Other or unknown	18 185 (3.3)	56 853 (4.6)	
Any Condition 0-12 months			
No	42 331 (7.7)	99 684 (8.1)	<.0001
Yes	510 620 (92.3)	1 137 469 (92.0)	
Any Condition 13-30 months			
No	49 773 (9.0)	111 779 (9.0)	0.465
Yes	503 178 (91.0)	1 125 374 (91.0)	
Any Condition 31-48 months			
No	60 592 (11.0)	130 915 (10.6)	<.0001
Yes	492 359 (89.0)	1 106 238 (89.4)	
Total number of conditions 0-12 months			
0	42 331 (7.7)	99 684 (8.1)	<.0001
1	40 163 (7.3)	92 193 (7.5)	
2	50 909 (9.2)	119 573 (9.7)	
3+	419 548 (75.9)	925 703 (74.8)	
Total number of conditions 13-30 months			
0	49 773 (9.0)	111 779 (9.0)	<.0001
1	46 054 (8.3)	103 878 (8.4)	
2	56 767 (10.3)	130 131 (10.5)	
3+	400 357 (72.4)	891 365 (72.1)	
Total number of conditions 31-48 months			
0	60 592 (11.0)	130 915 (10.6)	<.0001
1	54 677 (9.9)	120 371 (9.7)	
2	64 258 (11.6)	144 755 (11.7)	
3+	373 424 (67.5)	841 112 (68.0)	

Association of transfusion with overall cancer risk, site-specific cancers and subgroups

The results of the analysis based on all cases and all eligible controls from the 5% random sample of Medicare beneficiaries are summarized in Table 8.4. The overall cancer risk after blood transfusion was significantly increased (after Bonferroni adjustment) during the first 12 months after transfusion (OR=1.95, 95% CI: 1.91-1.99) and then risk decreased to OR=1.03 (95% CI: 1.01-1.05) and OR=1.05 (95% CI: 1.02-1.07) for blood transfusions received 13-30 months and 31-48 months before diagnosis or selection, respectively.

Also for most of the site-specific cancers and subgroups, a stronger risk during the first 12 months after blood transfusion and a decreasing risk with increasing time since transfusion was observed. Pronounced statistically significantly higher risk during the time period 0-12 months after transfusion was found for cancers of the digestive system overall (OR=2.97, 95% CI: 2.88-3.06) and specifically for cancers of the stomach (OR=3.74, 95% CI: 3.46-4.04), colon (OR=3.53, 95% CI: 3.40-3.66), liver (OR=3.12, 95% CI: 2.76-3.53) and pancreas (OR=2.04, 95% CI: 1.88-2.21); for kidney, renal pelvis, and ureter (OR=2.30, 95% CI: 2.10-2.51) for myeloma (OR=4.48, 95% CI: 4.11-4.88), leukemia (OR=6.09, 95% CI: 5.67-6.55), Kaposi Sarcoma (OR=3.00, 95% CI: 1.88-4.78) and for Hodgkin lymphoma (OR=2.97, 95%CI: 2.20-4.01) and NHL (OR=2.25, 95%CI: 2.11-2.39). From these cancer groups, the risk remained significantly elevated in the time periods 13-30 months and 31-48 months only for liver cancer, with OR=1.66 (95% CI: 1.43-1.92) and OR=1.63 (95% CI: 1.38-1.92), respectively and for leukemia with OR=1.58 (95% CI: 1.42-1.76) and OR=1.34 (95% CI: 1.18-1.52), respectively.

Table 8.4: Associations of blood transfusions with cancer risk, overall and for specific sites and subtypes.

		OR[1] (95% CI)		
	Total[2]	0-12 months	13-30 months	31-48 months
All cancers	552 951	1.95 (1.91-1.99)[3]	1.03 (1.01-1.05)	1.05 (1.02-1.07)
Oral Cavity and Pharynx	8 897	1.30 (1.14-1.50)[3]	1.43 (1.26-1.62)[3]	1.17 (1.01-1.37)
Lip	1 119	0.99 (0.63-1.55)	1.33 (0.92-1.92)	1.13 (0.73-1.75)
Tongue	2 042	1.21 (0.90-1.62)	1.58 (1.23-2.02)	1.31 (0.97-1.77)
Salivary Gland	1 262	1.44 (1.02-2.03)	1.35 (0.97-1.88)	0.81 (0.52-1.28)
Mouth	2 418	1.46 (1.14-1.86)	1.36 (1.07-1.73)	1.25 (0.95-1.65)
Pharynx	2 056	1.28 (0.95-1.73)	1.48 (1.14-1.94)	1.20 (0.87-1.67)
Digestive System	125 386	2.97 (2.88-3.06)[3]	1.05 (1.01-1.09)	1.01 (0.97-1.06)
Esophagus	5 581	1.68 (1.44-1.97)[3]	1.29 (1.09-1.51)	1.41 (1.18-1.68)[3]
Stomach	10 961	3.74 (3.46-4.04)[3]	1.10 (0.98-1.24)	1.25 (1.11-1.42)
Colon	58 480	3.53 (3.40-3.66)[3]	1.06 (1.01-1.12)	0.92 (0.86-0.98)
Rectum	17 608	1.91 (1.76-2.08)[3]	0.71 (0.62-0.80)[3]	0.84 (0.74-0.96)
Anus	1 281	1.70 (1.23-2.34)	0.74 (0.48-1.13)	1.31 (0.89-1.91)
Liver	5 039	3.12 (2.76-3.53)[3]	1.66 (1.43-1.92)[3]	1.63 (1.38-1.92)[3]
Pancreas	17 114	2.04 (1.88-2.21)[3]	0.93 (0.83-1.03)	1.01 (0.90-1.13)
Respiratory System	92 639	1.62 (1.55-1.69)[3]	1.18 (1.12-1.23)[3]	1.24 (1.18-1.30)[3]
Lung	88 275	1.64 (1.57-1.71)[3]	1.16 (1.11-1.22)[3]	1.24 (1.18-1.30)[3]
Bones and Joints	396	1.18 (0.60-2.33)	1.26 (0.68-2.33)	0.99 (0.47-2.10)
Soft Tissue	2 504	1.88 (1.52-2.33)[3]	0.92 (0.70-1.20)	0.97 (0.72-1.32)
Skin excluding Basal and Squamous	16 520	0.98 (0.87-1.10)	0.96 (0.86-1.07)	1.04 (0.93-1.17)
Melanoma	14 259	0.97 (0.85-1.10)	0.91 (0.81-1.03)	1.04 (0.91-1.18)
Breast[4]	64 701	0.87 (0.81-0.92)[3]	0.85 (0.80-0.91)[3]	0.92 (0.86-0.98)
Female Genital System	24 812	1.42 (1.31-1.54)[3]	0.91 (0.83-0.99)	0.95 (0.86-1.05)
Cervix	1 727	1.62 (1.24-2.13)	1.30 (0.98-1.74)	1.25 (0.90-1.75)
Uterus	12 210	1.07 (0.94-1.21)	0.83 (0.72-0.95)	0.86 (0.74-1.01)
Ovary	7 940	1.82 (1.61-2.06)[3]	0.83 (0.71-0.98)	0.93 (0.78-1.11)
Vagina / Vulva	2 221	1.40 (1.10-1.79)	1.18 (0.92-1.51)	1.23 (0.94-1.62)
Male Genital System	91 147	0.88 (0.82-0.94)[3]	0.79 (0.74-0.84)[3]	0.84 (0.78-0.90)[3]
Prostate	90 546	0.87 (0.82-0.93)[3]	0.79 (0.74-0.84)[3]	0.84 (0.78-0.90)[3]
Urinary System	45 343	1.68 (1.59-1.78)[3]	1.04 (0.98-1.11)	1.10 (1.02-1.18)
Urinary Bladder	31 799	1.44 (1.34-1.55)[3]	1.04 (0.96-1.12)	1.11 (1.02-1.21)
Kidney, Renal Pelvis, and Ureter	13 166	2.30 (2.10-2.51)[3]	1.02 (0.91-1.15)	1.07 (0.94-1.21)
Eye and Orbit	728	0.69 (0.36-1.35)	0.80 (0.45-1.42)	1.03 (0.58-1.82)
Brain and Other Nervous System	4 824	1.02 (0.82-1.27)	0.96 (0.78-1.18)	0.83 (0.65-1.06)
Endocrine System	3 236	1.20 (0.94-1.53)	0.70 (0.52-0.94)	0.85 (0.63-1.15)
Thyroid	2 871	1.07 (0.81-1.40)	0.68 (0.50-0.93)	0.81 (0.58-1.12)
Lymphoma	29 991	2.27 (2.14-2.41)[3]	0.94 (0.87-1.02)	1.07 (0.98-1.17)
Hodgkin Lymphoma	909	2.97 (2.20-4.01)[3]	1.03 (0.67-1.57)	0.95 (0.58-1.56)
NHL, including CLL[5]	29 082	2.25 (2.11-2.39)[3]	0.93 (0.86-1.01)	1.07 (0.98-1.17)
Myeloma	7 792	4.48 (4.11-4.88)[3]	1.10 (0.96-1.26)	1.06 (0.90-1.24)
Leukemia	8 863	6.09 (5.67-6.55)[3]	1.58 (1.42-1.76)[3]	1.34 (1.18-1.52)[3]
Mesothelioma	1 703	1.74 (1.31-2.30)[3]	0.86 (0.61-1.22)	1.10 (0.77-1.57)
Kaposi Sarcoma	293	3.00 (1.88-4.78)[3]	1.54 (0.88-2.72)	0.61 (0.22-1.69)

[1] The OR for overall cancer risk is adjusted for sex, age, race, selection year, and any condition as categorical variables. The ORs for the site-specific cancer groups like digestive system are adjusted for sex, age, race, and any condition as categorical variables and selection year as continuous variable. The ORs for subtypes of the site-specific cancer groups are adjusted for sex and race as categorical variables and age, selection year, and any condition as continuous variables.

[2] The number of cases may not add up to the totals for cancer specific sites and for cancer overall because miscellaneous cancers were excluded.

[3] Association is significant at $p<0.0004$ (Bonferroni correction for all 129 comparisons, 43 sites for 3 time periods).

[4] Women only (537 males are excluded from analysis).

[5] CLL = chronic lymphocytic leukemia

8.4.2 Analysis of matched data sets

Characteristics after matching

The characteristics of the 552 951 cases and the frequency matched controls are presented in Table 8.5.

In the covariate matched data set there were no significant differences between the cases and controls for the matching variables sex, age at diagnosis/selection, and selection year. The distribution of race did not notably change after matching compared with before (Table 8.3). For the time periods 0-12 and 13-30 months before diagnosis/selection, there were significantly more cases with any condition than controls. The median number of conditions reported for the 0-12 month time period for both cases and controls was five, and for the earlier time periods the median number was four for both groups, although significantly more conditions were reported for cases than controls (Table 8.5).

Using the frequency matching based on the three propensity scores, separately for sex and diagnosis or selection year, we found a total of 110 351 matched controls for the 552 951 cases. We observed similar distributions for sex, selection year and race after propensity score matching, compared with the distributions for these characteristics after the covariate matching (Table 8.5). Also, for the 0-12 and 13-30 months time periods, there were significantly more cases with any condition than controls, but the balance for the presence of any condition between cases and controls is somewhat better after propensity score matching compared with the results after the covariate matching. For age at diagnosis/selection better balance was seen in the covariate matched data.

Frequency matching on the disease risk scores resulted in a total of 110 456 matched controls. The balance for the characteristics between the cases and the matched controls was similar compared with the propensity score matching results. Only for race, better balance was observed. For example 87.2% cases were white and disease risk score matching resulted in 87.4% white controls (Table 8.5) whereas in the covariate and in the propensity score matched sample 85.4% and 85.7% controls were white, respectively.

Table 8.5: Characteristics of cases and controls in the SEER-Medicare database from 1997 to 2005 after covariate matching (CVM), propensity score matching (PSM) and disease risk score matching (DRSM).

Characteristic	Cases (n = 552 951) N (%)	Controls CVM (n = 100 000) N (%)	Controls PSM (n = 110 351) N (%)	Controls DRSM (n = 110 456) N (%)
Sex				
Male	280 034 (50.6)	50 644 (50.6)	55 889 (50.7)	55 915 (50.6)
Female	272 917 (49.4)	49 356 (49.4)	54 462 (49.4)	54 541 (49.4)
Age at diagnosis or selection (years)				
70-74	159 254 (28.8)	28 800 (28.8)	32 000 (29.0)†	30 984 (28.1)†
75-79	167 536 (30.3)	30 299 (30.3)	32 904 (29.8)	33 693 (30.5)
80-84	126 165 (22.8)	22 817 (22.8)	24 763 (22.4)	25 748 (23.3)
85-99	99 996 (18.1)	18 084 (18.1)	20 684 (18.7)	20 031 (18.1)
Selection year				
1997-1999	110 534 (20.0)	19 989 (20.0)	21 969 (19.9)	21 983 (19.9)
2000-2001	153 125 (27.7)	27 692 (27.7)	30 596 (27.7)	30 641 (27.7)
2002-2003	147 348 (26.7)	26 648 (26.7)	29 439 (26.7)	29 458 (26.7)
2004-2005	141 944 (25.7)	25 671 (25.7)	28 347 (25.7)	28 374 (25.7)
Race				
White	482 412 (87.2)	85 394 (85.4)†	94 601 (85.7)†	96 523 (87.4)*
Black	39 249 (7.1)	6 462 (6.5)	7 470 (6.8)	7 937 (7.2)
Asian	13 105 (2.4)	3 378 (3.4)	3 465 (3.1)	2 509 (2.3)
Other or unknown	18 185 (3.3)	4 766 (4.8)	4 815 (4.4)	3 487 (3.2)
Any Condition 0-12 months				
No	42 331 (7.7)	8 488 (8.5)†	8 895 (8.1)†	8 989 (8.1)†
Yes	510 620 (92.3)	91 512 (91.5)	101 456 (91.9)	101 467 (91.9)
Any Condition 13-30 months				
No	49 773 (9.0)	9 496 (9.5)†	10 141 (9.2)*	10 102 (9.2)
Yes	503 178 (91.0)	90 504 (90.5)	100 210 (90.8)	100 354 (90.9)
Any Condition 31-48 months				
No	60 592 (11.0)	11 153 (11.2)	12 090 (11.0)	11 897 (10.8)*
Yes	492 359 (89.0)	88 847 (88.9)	98 261 (89.0)	98 559 (89.2)
Total number of conditions 0-12 months				
0	42 331 (7.7)	8 488 (8.5)†	8 895 (8.1)†	8 989 (8.1)†
1	40 163 (7.3)	7 537 (7.5)	8 165 (7.4)	7 760 (7.0)
2	50 909 (9.2)	9 711 (9.7)	10 407 (9.4)	10 343 (9.4)
3+	419 548 (75.9)	74 264 (74.3)	82 884 (75.1)	83 364 (75.5)
Total number of conditions 13-30 months				
0	49 773 (9.0)	9 496 (9.5)†	10 141 (9.2)*	10 102 (9.2)
1	46 054 (8.3)	8 570 (8.6)	9 367 (8.5)	8 975 (8.1)
2	56 767 (10.3)	10 657 (10.7)	11 524 (10.4)	11 279 (10.2)
3+	400 357 (72.4)	71 277 (71.3)	79 319 (71.9)	80 100 (72.5)
Total number of conditions 31-48 months				
0	60 592 (11.0)	11 153 (11.2)*	12 090 (11.0)	11 897 (10.8)*
1	54 677 (9.9)	9 923 (9.9)	10 785 (9.8)	10 661 (9.7)
2	64 258 (11.6)	11 913 (11.9)	12 957 (11.7)	12 752 (11.5)
3+	373 424 (67.5)	67 011 (67.0)	74 519 (67.5)	75 146 (68.0)

* Statistically significant at $p<0.05$, † Statistically significant at $p<0.0001$.

Association of transfusion with overall cancer risk

Among all cases, 3.5% had a blood transfusion within 0-12 months prior to their cancer diagnosis, while within 13-30 and 31-48 months prior to diagnosis, 2.3% and 1.8% of the cases respectively were recipients of a transfusion. Among the covariate matched controls, 1.7%, 2.1% and 1.6% had a transfusion within the three time periods, respectively. In the the propensity score matched data set, 2.0%, 2.4% and 1.7% controls and in the disease risk score matched data set, 1.9%, 2.3% and 1.7% controls with transfusions within the three time periods were observed (Table 8.6). Fewer than 1% of cases and controls received blood transfusions in two or all three time periods.

Table 8.6: Number of recipients of blood transfusions for cases and controls according to the three time periods after covariate matching (CVM), propensity score matching (PSM) and disease risk score matching (DRSM).

Characteristic	Cases (n = 552 951) N (%)	Controls CVM (n = 100 000) N (%)	Controls PSM (n = 110 351) N (%)	Controls DRSM (n = 110 456) N (%)
Blood transfusions 0-12 months				
No	533 536 (96.5)	98 265 (98.3)	108 169 (98.0)	108 351 (98.1)
Yes	19 415 (3.5)	1 735 (1.7)	2 182 (2.0)	2 105 (1.9)
Blood transfusions 13-30 months				
No	540 050 (97.7)	97 936 (97.9)	107 757 (97.7)	107 961 (97.7)
Yes	12 901 (2.3)	64 (2.1)	2 594 (2.4)	2 495 (2.3)
Blood transfusions 31-48 months				
No	542 970 (98.2)	98 376 (98.4)	108 479 (98.3)	108 628 (98.4)
Yes	9 981 (1.8)	1 624 (1.6)	1 872 (1.7)	1 828 (1.7)
Number of transfused periods				
0	514 930 (93.1)	95 046 (95.1)	104 329 (94.5)	104 642 (94.7)
1	34 147 (6.2)	4 516 (4.5)	5 441 (4.9)	5 247 (4.8)
2	3 472 (0.6)	407 (0.4)	536 (0.5)	520 (0.5)
3	402 (0.1)	31 (0.0)	45 (0.0)	47 (0.0)
Combinations of transfused periods				
none	514 930 (93.1)	95 046 (95.1)	104 329 (94.5)	104 642 (94.7)
only in 0-12 months	16 402 (3.0)	1 454 (1.5)	1 774 (1.6)	1 719 (1.6)
only in 13-30 months	9 858 (1.8)	1 714 (1.7)	2 127 (1.9)	2 026 (1.8)
only in 31-48 months	7 887 (1.4)	1 348 (1.4)	1 540 (1.4)	1 502 (1.4)
in 0-12 and 13-30 months	1 780 (0.3)	162 (0.2)	249 (0.2)	241 (0.2)
in 0-12 and 31-48 months	831 (0.2)	88 (0.1)	114 (0.1)	98 (0.1)
in 13-30 and 31-48 months	861 (0.2)	157 (0.2)	173 (0.2)	181 (0.2)
in all periods	402 (0.1)	31 (0.0)	45 (0.0)	47 (0.0)

In the covariate matched data set, the overall cancer risk after blood transfusion was significantly increased during the first 12 months after transfusion (OR=2.05, 95% CI: 1.95-2.16) and then risk decreased to OR=1.04 (95% CI: 0.99-1.09) and OR=1.05 (95% CI: 1.00-1.11) for blood transfusions received 13-30 months and 31-48 months before diagnosis or selection, respectively (Table 8.7). The number of transfused periods was significantly associated with overall cancer risk, with OR=1.39 (95% CI: 1.35-1.44), OR=1.58 (95% CI: 1.42-1.75) and OR=2.39 (95% CI: 1.66-3.45) for one, two or three transfused periods, respectively (p-trend < 0.0001). If the receipt of a transfusion occurred only in one period, the cancer risk was highest for 0-12 months prior to cancer diagnosis or control selection (OR=2.09, 95% CI: 1.98-2.20). The same pattern was

seen for the combinations of two transfused periods with highest risk for periods 0-12 and 13-30 months (OR=2.02, 95% CI: 1.72-2.38) (Riedl et al., 2013).

In Table 8.7 the OR (95% CI) for the overall cancer risks are displayed for the different score adjustments (continuous score, continuous logit-score, in deciles coded with a trend, and in deciles, coded with dummy variables). All adjustments did not yield different results for the scores, respectively. Adjusting on the continuous PS resulted in a significantly increased overall cancer risk during 0-12 months after transfusion (OR=1.84, 95% CI: 1.76-1.92) and then, as observed before, the risk decreased to OR=0.94 (95% CI: 0.90-0.98) and OR=1.04 (95% CI: 0.99-1.10) during the two earlier time periods. If the disease risk score is used as continuous variable on logit scale, the overall cancer risk was slightly lower than the observed risk in the covariate matched data and slightly higher or similar than the observed risk in the propensity score matched data with OR=1.87 (95% CI: 1.79-1.96), OR=0.96 (95% CI: 0.91-1.00) and OR=1.04 (95% CI: 0.99-1.10) for the three time periods, respectively.

Table 8.7: Associations of blood transfusions with overall cancer risk models, adjusted for gender, race, selection year and additionally adjusted for any condition or the summary scores (PS and DRS) using different codings: continuous score, continuous logit-score, in deciles with trend, and in deciles with dummy coding (i.e. nine degrees of freedom) for the different matching designs.

		OR (95% CI)	
Additionally adjusted for	0-12 months	13-30 months	31-48 months
		Covariate matched data	
Any Condition	2.054 (1.954-2.159)	1.042 (0.993-1.093)	1.052 (0.996-1.110)
		PS matched data	
Continuous PS	1.838 (1.756-1.924)	0.938 (0.895-0.983)	1.043 (0.989-1.100)
Continuous logit-PS	1.845 (1.763-1.931)	0.945 (0.902-0.989)	1.043 (0.990-1.100)
PS in deciles with trend	1.840 (1.758-1.925)	0.939 (0.898-0.982)	1.036 (0.984-1.091)
PS in deciles with 9 df	1.849 (1.767-1.935)	0.951 (0.908-0.995)	1.039 (0.986-1.094)
		DRS matched data	
Continuous DRS	1.875 (1.791-1.962)	0.957 (0.915-1.000)	1.045 (0.992-1.100)
Continuous logit-DRS	1.874 (1.790-1.962)	0.956 (0.914-1.000)	1.045 (0.992-1.100)
DRS in deciles with trend	1.884 (1.799-1.972)	0.964 (0.922-1.008)	1.050 (0.998-1.106)
DRS in deciles with 9 df	1.887 (1.803-1.975)	0.969 (0.926-1.013)	1.056 (1.003-1.112)

Association of transfusion with site-specific cancers for covariate matched controls.

Similarly, the association of transfusion with risk of most specific cancer sites was stronger during the first 12 months after blood transfusion and decreased with increasing time since transfusion. A statistically significantly higher risk after Bonferroni adjustment during the first 12 months after transfusion was found for cancers of the digestive system overall (OR=3.14, 95% CI: 1.97-4.99) and specifically for cancers of the stomach (OR=3.92, 95% CI: 2.46-6.25), colon (OR=3.73, 95% CI: 2.46-5.65), and liver (OR=3.29, 95% CI: 2.01-5.40); for kidney, renal pelvis, and ureter (OR=2.42, 95% CI: 1.50-3.92) for myeloma (OR=4.71, 95% CI: 2.72-8.16), leukemia (OR=6.51, 95% CI: 3.81-11.13) and for Hodgkin lymphoma (OR=3.14, 95%CI: 1.75-5.62) and NHL (OR=2.38, 95%CI: 1.51-3.76) (Table 8.8). Risk for the first 12 months after transfusion

was also significantly elevated (but not after multiple comparisons adjustment) for cancers at the following sites: esophagus (OR=1.78, 95% CI: 1.06-2.99), rectum (OR=2.00, 95% CI: 1.24-3.22), pancreas (OR=2.16, 95% CI: 1.34-3.47), respiratory system overall (OR=1.71, 95% CI: 1.04-2.79) and specifically for cancers of the lung (OR=1.72, 95% CI: 1.13-2.64), soft tissue (OR=1.99, 95% CI: 1.07-3.70), the urinary system overall (OR=1.79, 95% CI: 1.06-3.00) lymphoma overall (OR=2.41, 95% CI: 1.43-4.08) and Kaposi sarcoma (OR=3.22, 95% CI: 1.51-6.86) (Table A.3).

After Bonferroni adjustment, cancer risk remained significantly elevated 13-30 months after transfusion only for liver cancer, with OR=1.72 (95% CI: 1.27-2.32). None of the site-specific cancers was significantly associated with transfusions after Bonferroni adjustment over a latency period of 31-48 months (Table A.3).

Cancers of the stomach, colon, liver, kidney, renal pelvis, and ureter, myeloma and leukemia were also significantly associated with the number of transfused periods (p-trend$<$0.0001), as were cancers of the esophagus (p-trend=0.001), lymphoma (p-trend=0.002), respiratory system (p=0.003), pancreas (p-trend=0.004), mouth (p-trend=0.008), tongue (p-trend=0.009) and Kaposi sarcoma (p-trend=0.009). Separate analysis for men and women showed similarly increased risks during the first 12 months after blood transfusion for the cancers of the digestive system, stomach, colon, liver, myeloma and leukemia for both males and females. The risks for NHL and Kaposi sarcoma were significantly increased only for men and the risk for Hodgkin lymphoma was significantly increased only for women (Tables A.4 and A.5).

These results are very similar to those based on all controls (Table 8.4). The odds ratios based on all controls are slightly lower than those using the matched controls. This could be explained by the fact that, using all controls does not ensure that cases and controls are at the same time at risk after a transfusion. For example, sicker controls might have had more blood transfusions and also have died sooner. We therefore would expect an underestimation of the true exposure-effect (bias towards null).

Association of transfusion with site-specific cancers for propensity score matched controls.

The propensity score analysis-based results for the cancer-specific sites and subtypes are presented in Table A.6. Generally, the propensity score analysis yielded similar results compared with the covariate matched results (Table A.3). We observed the same pattern for the association of transfusion with cancer risk of most specific cancer sites: a stronger risk during the first 12 months after blood transfusion and a decreasing risk with increasing time since transfusion. For the cancers, we previously observed a statistically significantly higher risk (after Bonferroni adjustment) during the first 12 months after transfusion, we observed a statistically significantly higher risk under the propensity score analysis as well, but the risk was somewhat smaller (Table 8.8). The propensity score analysis yielded following elevated risks: cancers of the digestive system overall (OR=2.74, 95% CI: 1.99-3.78) and specifically for cancers of the stomach (OR=3.35, 95% CI: 2.40-4.68), colon (OR=3.30, 95% CI: 2.47-4.39), and liver (OR=2.52, 95% CI: 1.76-3.62); kidney, renal pelvis, and ureter (OR=2.13, 95% CI: 1.51-3.00), myeloma (OR=3.90, 95% CI: 2.63-5.80), leukemia (OR=4.84, 95% CI: 3.29-7.12) and Hodgkin lymphoma (OR=2.91, 95%CI: 1.83-4.63) and NHL (OR=2.14, 95%CI: 1.55-2.95).

In contrast to the results in the standard matched dataset, none of the site-specific cancers was significantly associated with transfusions (after Bonferroni adjustment) over a latency period of 13-31 months. For liver cancer, the propensity score adjusted OR decreased to 1.09 (95%CI: 0.84-1.41) for this time period (Table A.6).

Association of transfusion with site-specific cancers for disease risk score matched controls.

The disease risk score analysis-based results for the cancer-specific sites and subtypes are presented in Table A.7. Similar results were observed as in the propensity score analysis. The risk was elevated for most specific cancer sites only during the first 12 months after blood transfusion. However, the ORs for the site-specific cancers that were significantly associated with transfusions (after Bonferroni adjustment) were somewhat lower than the observed risk in the covariate matched data and and somewhat higher than the observed risk in the propensity score matched data (Table 8.8). The disease risk score yielded following elevated risks: cancers of the digestive system overall (OR=2.89, 95% CI: 2.05-4.06) and specifically for cancers of the stomach (OR=3.62, 95% CI: 2.55-5.15), colon (OR=3.45, 95% CI: 2.55-4.68), and liver (OR=2.77, 95% CI: 1.86-4.05); kidney, renal pelvis, and ureter (OR=2.24, 95% CI: 1.56-3.22), myeloma (OR=4.19, 95% CI: 2.76-6.35), leukemia (OR=5.31, 95% CI: 3.54-7.96) and Hodgkin lymphoma (OR=2.81, 95%CI: 1.75-4.53) and NHL (OR=2.16, 95%CI: 1.54-3.03). For liver cancer OR decreased to 1.34 (95%CI: 1.05-1.72) for the latency period of 13-31 months.

Table 8.8: Statistically significant associations[1] of blood transfusions in the 0-12 months time period with cancer risk, overall and for specific sites and subtypes after covariate matching (CVM), propensity score matching (PSM) and disease risk score matching (DRSM).

		OR (95% CI)		
		CVM	PSM	DRSM
	Total	0-12 months	0-12 months	0-12 months
All cancers	552 951	2.05 (1.95-2.16)	1.84 (1.76-1.92)	1.87 (1.79-1.96)
Digestive System	125 386	3.14 (1.97-4.99)	2.74 (1.99-3.78)	2.89 (2.05-4.06)
Stomach	10 961	3.92 (2.46-6.25)	3.35 (2.40-4.68)	3.62 (2.55-5.15)
Colon	58 480	3.73 (2.46-5.65)	3.30 (2.47-4.39)	3.45 (2.55-4.68)
Liver	5 039	3.29 (2.01-5.40)	2.52 (1.76-3.62)	2.77 (1.89-4.05)
Kidney, Renal Pelvis, Ureter	13 166	2.42 (1.50-3.92)	2.13 (1.51-3.00)	2.24 (1.56-3.22)
Hodgkin Lymphoma	909	3.14 (1.75-5.62)	2.91 (1.83-4.63)	2.81 (1.75-4.53)
NHL, including CLL[2]	29 082	2.38 (1.51-3.76)	2.14 (1.55-2.95)	2.16 (1.54-3.03)
Myeloma	7 792	4.71 (2.72-8.16)	3.90 (2.63-5.80)	4.19 (2.76-6.35)
Leukemia	8 863	6.51 (3.81-11.13)	4.84 (3.29-7.12)	5.31 (3.54-7.96)

[1] Significant at $p<0.0004$ after Bonferroni correction for all 129 comparisons, 43 sites for 3 time periods.

[2] CLL = chronic lymphocytic leukemia

8.4.3 Sensitivity analyses

Sensitivity analysis for significant cancers in the covariate matched dataset

For specific cancers that were significantly associated with transfusions in the covariate matched dataset, we adjusted all models additionally for individual medical conditions. In these models, the ORs for association with blood transfusions did not change (Table A.8). For example the odds ratios for liver cancer varied between 2.94 and 3.36 for the time period 0-12 months and between 1.32-1.75 and 1.27-1.66 for the time periods 13-30 months and 31-48 months, respectively. For cancers of the stomach, colon rectum, liver and lung, we fitted logistic regression models additionally adjusted for combinations of the conditions (Table 8.9). The results for cancers of the stomach, colon rectum and lung did not really change. But for liver cancer, using a model that adjusted jointly for chronic liver disease and cirrhosis, various types of anemias, iron deficiency, gastrointestinal hemorrhage, disorders of fluid, electrolyte, and acid-base balance, purpura and other hemorrhagic conditions and heart failure, the OR was 2.6 (95%CI: 2.23-3.04) for the 0-12 month period, 0.95 (95%CI: 0.78-1.15) for the 13-30 month period, and 1.04 (95%CI: 0.83-1.29) for the latency period of 31-48 months. Therefore, we confirmed the lower estimate for liver cancer in the propensity score matched dataset also in the covariate matched dataset.

Table 8.9: Comparison of the results for selected cancer subtypes in the main analysis and in sensitivity analysis with additional adjustment for special conditions in the covariate matched dataset.

	OR (95% CI)		
	0-12 months	13-30 months	31-48 months
		Main analysis[1]	
Stomach	3.92 (2.46-6.25)	1.16 (0.88-1.55)	1.29 (0.94-1.76)
Colon	3.73 (2.46-5.65)	1.11 (0.88-1.38)	0.95 (0.73-1.22)
Rectum	2.00 (1.24-3.22)	0.72 (0.54-0.95)	0.86 (0.63-1.16)
Liver	3.29 (2.01-5.40)	1.72 (1.27-2.32)	1.65 (1.18-2.31)
Lung	1.72 (1.13-2.64)	1.17 (0.95-1.45)	1.24 (0.98-1.57)
		Sensitivity analysis	
Stomach[2]	3.80 (3.15-4.60)	1.00 (0.85-1.18)	1.30 (1.09-1.55)
Colon[2]	3.64 (3.17-4.19)	1.04 (0.95-1.15)	0.99 (0.88-1.11)
Rectum[2]	1.96 (1.60-2.40)	0.70 (0.59-0.82)	0.90 (0.76-1.08)
Liver[3]	2.60 (2.23-3.04)	0.95 (0.78-1.15)	1.04 (0.83-1.29)
Lung[4]	1.68 (1.45-1.96)	1.10 (1.01-1.20)	1.18 (1.08-1.30)

[1]Odds ratios are adjusted for sex, age, race, selection year, and any condition.

[2]Odds ratios for stomach, colon and rectum are additional adjusted for gastrointestinal hemorrhage, gastric ulcer, duodenal ulcer, vascular insufficiency of intestine and special screening for other conditions.

[3]Odds ratios for liver are additional adjusted for chronic liver disease and cirrhosis, various types of anemias, iron deficiency, gastrointestinal hemorrhage, disorders of fluid, electrolyte, and acid-base balance, purpura and other hemorrhagic conditions and heart failure.

[4]Odds ratios for lung are additional adjusted for bronchopneumonia.

Subgroup analysis

In the covariate matched dataset there are 3.8% cases (n=20 825) and 4.3% controls (n=4 277) with no condition presented in all three time periods. At most one condition is presented for 3.0% cases (n=44 147) and 8.6% controls (n=8 623). The results for the overall cancer risk for cases and controls, based on these two subgroups are summarized in Table 8.10. The only significant associations were seen for the first time interval with OR=6.06 (95%CI: 3.31-11.09) and OR=3.39 (95%CI: 2.50-4.60) for the subgroups with no condition and at most one condition, respectively. Blood transfusions received 13 or more months before cancer diagnosis were not associated with cancer risk.

Table 8.10: Associations of blood transfusions in subgroups of individuals with no condition, or at most one condition adjusted for age, gender, race, selection year as categorical variables in the covariate matched dataset.

	OR (95% CI)		
	0-12 months	13-30 months	31-48 months
No conditions in any of the 3 time intervals	6.06 (3.31-11.09)	0.17 (0.01-2.49)	1.32 (0.17-10.16)
At most one condition over all 3 time intervals	3.39 (2.50-4.60)	0.48 (0.20-1.17)	0.35 (0.35-2.04)

Case-control matching for cancer subgroups

To further investigate the discrepant results concerning liver cancer risk, we repeated our matching only for the 5 039 liver cancer cases and selected our controls out of the 5% random sample of the US elderly (n=1 237 153 controls) by using a frequency matching with matching ratio of 1:1. We have done this by using the three different matching designs (covariate matching, propensity score matching and disease risk score matching) described in Section 8.3.2. The results are summarized in the next section.

8.4.4 Transfusions and risk of liver cancer

Characteristics after matching

The characteristics for the liver cancer cases and the matched controls are presented in Table 8.11. Comparing the characteristics of the liver cancer cases with all cancer cases in Table 8.5, we observed a higher percentage for male cases, i.e. 59.0% of liver cancer cases were male and 50.6% of all cancer cases were male. Further, the liver cancer cases were somewhat younger, lesser white cases and more Asian, other and unknown cases were observed and the presence of any condition was slightly higher compared with all cancer cases.

Using covariate, propensity score and disease risk score matching, a total number of n=5039, n=5031 and n=5036 controls were found, respectively. No imbalance was observed for sex and selection year for all three different matching designs. For age, best balance was given in the covariate matched data and worst balance was given for the propensity score matched controls with slightly lower percentages in the first two age groups. Compared to liver cancer cases, more matched controls were white if the covariate matching and the propensity score matching was performed and less controls were white under the disease risk score matched design. The balance for the presence of any condition between cases and controls was best after propensity score matching, followed by covariate matching and disease risk score matching.

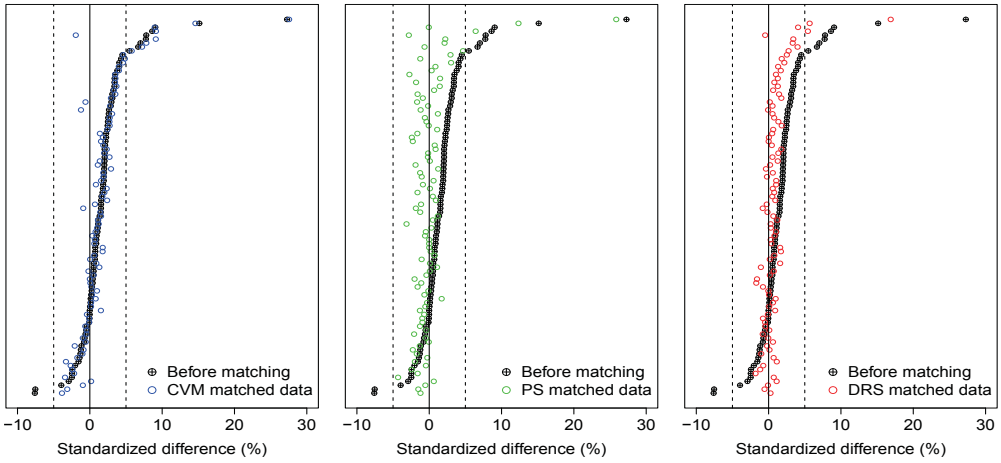
We also considered the balance for the 96 medical conditions separately by calculating standardized differences between cases and controls for the three investigated time periods. The results for the three time periods are plotted in Figure 8.2. In the time period 0-12 months, nearly all conditions showed a standardized difference of $<10\%$, only for chronic liver disease and purpura the standardized differences before matching were 27% and 15%, respectively. Covariate matching and propensity score matching did not substantially reduce this differences. However, after disease risk score matching the standardized differences were reduced to 17% and 6%, respectively. Before matching, standardized differences $>5\%$ were observed for 11 conditions including various types of anemias, iron deficiency, gastrointestinal hemorrhage and heart failure. If matching on the summary scores was performed, the standardized difference remained $>5\%$ only for 3 conditions (chronic liver disease, purpura and aplastic anemia). Similar results concerning balance were observed for the other two time periods. Overall, the disease risk score matching yielded best balance in the single conditions, followed by propensity score matching and the covariate matching.

Table 8.11: Characteristics of liver cancer cases and controls in the SEER-Medicare database from 1997 to 2005 after covariate matching (CVM), propensity score matching (PSM) and disease risk score matching (DRSM).

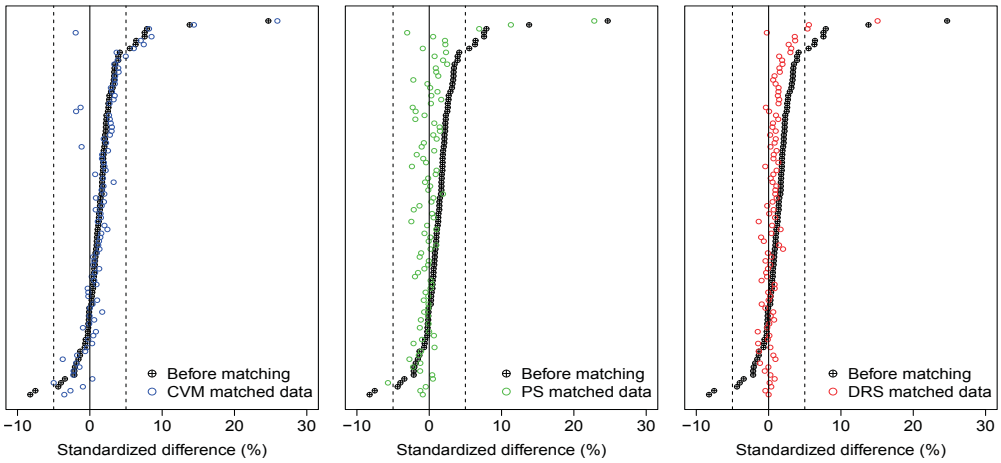
Characteristic	Cases (n = 5 039) N (%)	Controls CVM (n = 5 039) N (%)	Controls PSM (n = 5 031) N (%)	Controls DRSM (n = 5 036) N (%)
Sex				
Male	2 972 (59.0)	2 972 (59.0)	2 964 (58.9)	2 969 (59.0)
Female	2 067 (41.0)	2 067 (41.0)	2 067 (41.1)	2 067 (41.0)
Age at diagnosis or selection (years)				
70-74	1 542 (30.6)	1 542 (30.6)	1 359 (27.0)†	1 451 (28.8)
75-79	1 591 (31.6)	1 591 (31.6)	1 495 (29.7)	1 582 (31.4)
80-84	1 130 (22.4)	1 130 (22.4)	1 175 (23.4)	1 142 (22.7)
85-99	776 (15.4)	776 (15.4)	1 002 (19.9)	861 (17.1)
Selection year				
1997-1999	953 (18.9)	953 (18.9)	948 (18.8)	952 (18.9)
2000-2001	1 303 (25.9)	1 303 (25.9)	1 302 (25.9)	1 303 (25.9)
2002-2003	1 389 (27.6)	1 389 (27.6)	1 388 (27.6)	1 387 (27.5)
2004-2005	1 394 (27.7)	1 394 (27.7)	1 393 (27.7)	1 394 (27.7)
Race				
White	3 717 (73.8)	4 275 (84.8)†	4 122 (81.9)†	3 284 (65.2)†
Black	373 (7.4)	308 (6.1)	372 (7.4)	333 (6.6)
Asian	492 (9.8)	199 (4.0)	244 (4.9)	763 (15.2)
Other or unknown	457 (9.1)	257 (5.1)	293 (5.8)	656 (13.0)
Any Condition 0-12 months				
No	335 (6.7)	421 (8.4)*	292 (5.8)	461 (9.2)†
Yes	4 704 (93.4)	4 618 (91.7)	4 739 (94.2)	4 575 (90.9)
Any Condition 13-30 months				
No	393 (7.8)	449 (8.9)*	374 (7.4)	533 (10.6)†
Yes	4 646 (92.2)	4 590 (91.1)	4 657 (92.6)	4 503 (89.4)
Any Condition 31-48 months				
No	484 (9.6)	555 (11.0)*	458 (9.1)	610 (12.1)†
Yes	4 555 (90.4)	4 484 (89.0)	4 573 (90.9)	4 426 (87.9)
Total number of conditions 0-12 months				
0	335 (6.7)	421 (8.4)†	292 (5.8)	461 (9.2)†
1	286 (5.7)	355 (7.1)	280 (5.6)	369 (7.3)
2	399 (7.9)	524 (10.4)	388 (7.7)	482 (9.6)
3+	4 019 (79.8)	3 739 (74.2)	4 071 (80.9)	3 724 (74.0)
Total number of conditions 13-30 months				
0	393 (7.8)	449 (8.9)†	374 (7.4)	533 (10.6)†
1	355 (7.1)	444 (8.8)	314 (6.2)	422 (8.4)
2	496 (9.8)	547 (10.9)	458 (9.1)	512 (10.2)
3+	3 795 (75.3)	3 599 (71.4)	3 885 (77.2)	3 569 (70.9)
Total number of conditions 31-48 months				
0	484 (9.6)	555 (11.0)*	458 (9.1)*	610 (12.1)†
1	438 (8.7)	523 (10.4)	441 (8.8)	520 (10.3)
2	594 (11.8)	590 (11.7)	508 (10.1)	554 (11.0)
3+	3 523 (69.9)	3 371 (66.9)	3 624 (72.0)	3 352 (66.6)

* Statistically significant at $p<0.05$, † Statistically significant at $p<0.0001$.

(a) Standardized differences of the 96 conditions for the time period 0-12 months.



(b) Standardized differences of the 96 conditions for the time period 13-30 months.



(c) Standardized differences of the 96 conditions for the time period 31-48 months.

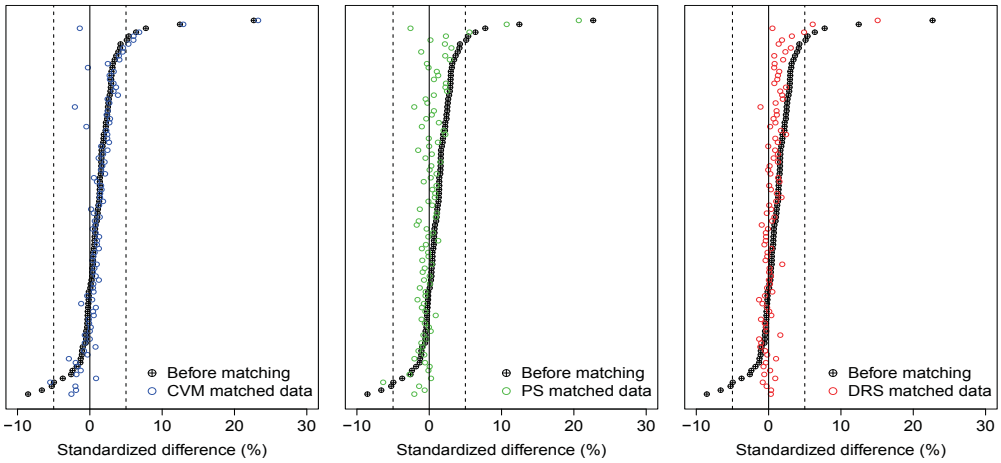


Figure 8.2: Standardized differences of the 96 conditions for the three time periods before and after covariate matching (CVM), propensity score matching (PSM) and disease risk score matching (DRSM). For better comparability the conditions are ordered by their standardized differences before matching.

Transfusions and liver cancer risk

Among the liver cancer cases, 6.0% had a blood transfusion within 0-12 months prior to their cancer diagnosis, while within 13-30 and 31-48 months prior to diagnosis, 4.2% and 3.1% of the cases respectively were recipients of a transfusion. Among the covariate matched controls, 2.1%, 1.9% and 1.3% had a transfusion within the three time periods, respectively. In the the propensity score matched data set, 3.1%, 3.8% and 2.3% controls and in the disease risk score matched data set, 2.5%, 3.2% and 2.1% controls with transfusions within the three time periods were observed.

We analyzed the matched data sets using conditional logistic regression (Model 1) and unconditional logistic regression with various adjustments. All models were adjusted for the characteristics sex and age, race, selection year (in 4 categories respectively). Under the covariate matched design, we additionally adjusted for any condition (Model 2) and for special conditions (chronic liver disease and cirrhosis, various types of anemias, iron deficiency, gastrointestinal hemorrhage, disorders of fluid, electrolyte, and acid-base balance, purpura and other hemorrhagic conditions and heart failure) where also large imbalances between cases and controls were observed (Model 3), (Figure 8.2). For the score matched designs, we additionally adjusted for the propensity scores on both scales (Model 4a and Model 4b) and for the disease risk scores (Model 5a and Model 5b) as continuous variables.

After matching on sex, age and selection year we observed a significantly increased liver cancer risk after blood transfusion in all three time periods if conditional logistic regression (Model 1) or logistic regression with adjustment on the characteristics and any condition (Model 2) was used to analyze the data. The ORs for Model 1, i.e. OR=2.75 (95%CI: 2.20-3.45), OR=1.88 (95%CI: 1.46-2.41) and OR=2.19 (95%CI: 1.63-2.95) for 0-12 months, 13-30 months and 31-48 months before cancer diagnosis/selection, respectively, did not substantially differ from the results from Model 2. Adjusting additionally for the special conditions, the liver cancer risk remained significant only in the first time period.

Matching on the summary scores and using conditional logistic regression resulted in a significantly increased risk 0-12 months and 31-48 months after transfusion. However, adjusting for the scores together with sex, age, race and selection year, no significantly increased risk during the time period 31-48 months was observed (Models 4 and 5). This might be explained due to residual confounding after matching concerning age and race. Adjusting on the scores on the logit or probability scale yielded similar results.

For the first time period, the liver cancer risk was lowest if matching on the PS was performed (OR=1.96 (95%CI: 1.60-2.40), Model 4a), somewhat higher risk was observed after covariate matching and adjusting on special conditions (OR=2.09 (95%CI: 1.62-2.69), Model 3), and after disease risk score matching using Model 5b (OR=2.13, 95%CI: 1.69-2.67) (Table 8.12).

The liver cancer risk during the first time period was lower after case-control matching restricted to liver cancer cases compared with the risk obtained from subgroup analysis after case-control matching for all cancers (Tables 8.8 and 8.9).

Table 8.12: Associations of blood transfusions with liver cancer risk after covariate matching, propensity score matching and disease risk score matching using different adjustments.

		OR (95% CI)	
Model	0-12 months	13-30 months	31-48 months
		Covariate matched controls	
1	2.747 (2.190 - 3.445)†	1.878 (1.463 - 2.410)†	2.193 (1.630 - 2.949)†
2	2.690 (2.140 - 3.380)†	1.814 (1.411 - 2.331)†	2.175 (1.613 - 2.935)†
3	2.087 (1.620 - 2.688)†	1.068 (0.788 - 1.448)	1.277 (0.903 - 1.804)
		PS matched controls	
1	1.994 (1.630 - 2.440)†	1.033 (0.833 - 1.282)	1.321 (1.023 - 1.704)*
4a	1.961 (1.602 - 2.401)†	0.936 (0.753 - 1.164)	1.206 (0.929 - 1.564)
4b	1.966 (1.605 - 2.408)†	0.953 (0.767 - 1.184)	1.222 (0.945 - 1.582)
		DRS matched controls	
1	2.471 (1.991 - 3.067)†	1.175 (0.945 - 1.460)	1.420 (1.101 - 1.832)*
5a	2.203 (1.761 - 2.757)†	0.986 (0.788 - 1.232)	1.161 (0.890 - 1.516)
5b	2.127 (1.693 - 2.673)†	0.917 (0.729 - 1.153)	1.091 (0.830 - 1.433)

* Statistically significant at $p<0.05$, † Statistically significant at $p<0.0001$.

Model 1: conditional logistic regression
Model 2: logistic regression adjusted for sex and age, race, selection year (in 4 categories respectively) and any condition as categorical variables.
Model 3: logistic regression adjusted for sex and age, race, selection year (in 4 categories respectively) and selected special conditions, i.e. chronic liver disease and cirrhosis, various types of anemias, iron deficiency, gastrointestinal hemorrhage, disorders of fluid, electrolyte, and acid-base balance, purpura and other hemorrhagic conditions and heart failure.
Model 4: logistic regression adjusted for sex and age, race, selection year (in 4 categories respectively) and the propensity scores probability scale (a) and on logit scale (b) as continuous variables.
Model 5: logistic regression adjusted for sex and age, race, selection year (in 4 categories respectively) and the disease risk scores on probability scale (a) and on logit scale (b) as continuous variables.

8.5 Discussion

In our case-control study of U.S. adults aged 70 years or older, we found that receipt of a blood transfusions was associated with significantly increased risk in the subsequent 0-12 month period, for cancer overall and (after multiple testing adjustment) risk of cancers of the stomach, colon, liver, kidney, renal pelvis and ureter, myeloma, leukemia and for Hodgkin lymphoma and non-Hodgkin lymphoma (NHL). Risk was also increased for cancer of the esophagus, rectum, pancreas, soft tissue, Kaposi sarcoma, lung and the urinary system, although the level of statistical significance was lower. There was no elevated overall cancer risk in the latency periods 13-30 months and 31-48 months after a transfusion. In site specific analyses, only the risk of liver cancer remained significantly elevated (after multiple testing adjustment) for the 13-30 month time period, however, this association disappeared after jointly adjusting for all possible confounding conditions, and when controls were matched based on balancing scores. We also found a significant association of the number of transfused periods with overall cancer risk.

Our study is unique in focusing on an elderly population, among whom both blood transfusions and cancer are common. Hjalgrim et al. (2007) previously studied cancer incidence in Swedish and Danish individuals of all ages who received transfusions between 1968 and 2002. In agreement with our findings, they also observed the strongest associations for cancer diagnoses that occurred shortly after the first transfusion, and weaker effects for longer latency periods. Four years after transfusion, the relative risk for cancer decreased to SIR=1.07 in Hjalgrim's study and to OR=1.05 in our population. Similarly to Hjalgrim we found elevated risks shortly after transfusion for cancers of the esophagus, stomach, colon, rectum, pancreas, lung, and the urinary system, as well as NHL, Hodgkin lymphoma, myeloma and leukemias combined. We did not replicate Hjalgrim's findings of increased risk for cancers of the tongue, mouth and, pharynx. We did also not replicate Hjalgrim's findings of long term elevated risk for cancers of the tongue, mouth, pharynx, esophagus, liver, and respiratory and urinary tracts. Two explanations for these discrepancies are possible. First, we compared transfusion in cancer cases and controls aged 70 years or older, while Hjalgrim studied blood transfusion recipients of all ages. As Hjalgrim found that the association between transfusion and cancer was weaker for a first transfusion received at older ages, one would expect more modest associations in an older population. Second, we were able to adjust our analyses for medical conditions that could confound the relationship of transfusion and cancer risk. Specifically for liver cancer, we found that after careful adjustment for confounding, no long term effect of transfusion on risk was observed, while the short term effect persisted.

Similar to our NHL results, a recent meta-analysis of NHL (Castillo et al., 2010) reported an increased NHL risk after a blood transfusion. However, due to differences in lag time intervals across the studies, transfusion latency periods were not investigated and the risk estimate was somewhat weaker than seen in our data, especially for the 0-12 month period after transfusion. Chang et al. (2010) investigated the risk of hematologic malignancies in the SEER-Medicare population one or more years after a blood transfusion and found elevated risks for most subtypes only for latency periods closest to cancer diagnosis. However, for two NHL subtypes (lymphoplasmacytic lymphoma and marginal zone lymphoma), associations at longer latency intervals were consistent with an etiologic relationship. Chang et al. (2010) attributed the elevated risk of these lymphoma subtypes to "transfusion-related immunomodulation" (TRIM) (Vamvakas and Blajchman, 2007) and gave a detailed discussion of the mechanistic underpinnings for a possible

relationship.

Reverse causation likely plays a role in our findings, as the associations with cancer risk overall and for specific sites were strongest for transfusions received in the 0-12 months before diagnosis/selection. Many cancers or their precursors can cause anemia, and the work-up of the anemia can lead to the diagnosis of cancer. Cancers of the digestive system, as well as precancerous colon polyps, commonly cause chronic occult blood loss and iron deficiency anemia (Rockey, 2010; Bross et al., 2010), which would explain our results for cancer of the digestive system overall and for cancers of the stomach and colon. For other cancers that we observed to be associated with transfusion, anemia could have been caused by decreased red cell production which may accompany precursor conditions, such as liver disease (a precursor to liver cancer) or bone marrow disorders (which may precede leukemias and lymphomas) (Davenport, 1996; Cazzola and Malcovati, 2005). This explanation could also be the reason for the significant association of the number of transfusions with kidney cancer, as renal disease causes anemia, which leads to transfusion and also increases the risk of kidney cancer (Bross et al., 2010).

While reverse causation is the most likely explanation for our findings, we cannot entirely rule out that blood transfusions could affect or shorten the time of transitions of pre-cancers to cancer through TRIM, for example, the transition of myelodysplasia to detectable AML.

Using Medicare claims, we identified acute medical diagnoses and chronic conditions that potentially could act as confounders, given their frequency of occurrence and strengths of association with blood transfusion among the cases. We did not select conditions in relation to their importance for individual cancers. However our criteria likely ensure that all potentially important confounding conditions were selected. Confounding by documented medical conditions thus does not explain our findings, as ORs for the first latency period remained significantly elevated when we adjusted our models for individual conditions that were selected on the basis of being associated with blood transfusions. We also repeated analyses by matching controls to cases based on propensity scores and disease risk scores. This matching and analysis approaches allowed us to more efficiently control for measured confounding. Results were changed only for liver cancer for the intermediate latency period, but not for any other cancer sites, which reassures us that we did not miss confounding by medical conditions captured by Medicare in the main analysis. In a sensitivity analysis we estimated associations of blood transfusion with overall cancer risk in individuals without any conditions and obtained results as seen for the whole study population. Although we did not see significant associations with cancer risk for the longer periods following a transfusion, we did observe significant associations for the number of intervals in which transfusions were administered for cancer overall and several specific sites. This finding may point to an especially elevated risk in the small subgroup of people with chronic anemia, who are most likely to have a longstanding precursor condition to cancer or perhaps an undiagnosed malignancy (reverse causation). The possibility of reverse causation is also supported by Edgren et al. (2010), who found that in Scandinavian blood donors, for most hematopoietic, lymphopoietic, and gastrointestinal malignancies, hemoglobin concentrations began to decrease two to three years before cancer diagnosis. However, the study was limited to hemoglobin measurements up to five years preceding the cancer diagnosis.

Unmeasured confounding by lifestyle factors, including tobacco and alcohol consumption, could also partially explain the elevated risk for cancer overall. Information on alcohol and tobacco

use were not available in our data. Rogers et al. (2011) found significant associations between blood transfusions and alcohol and tobacco use in a cohort of older Americans. As noted above, liver disease related to alcoholism could cause both anemia and liver cancer (Bross et al., 2010). Likewise, tobacco use is a risk factor for peptic ulcer disease (Rosenstock et al., 2003), which is associated with gastrointestinal blood loss and could explain the association we observed between blood transfusions and lung cancer. However, given that we adjusted for conditions associated with alcohol and smoking, additional adjustment for those factors likely would not strongly impact our findings. Transfusion-transmissible infections are an unlikely explanation for the elevated risk of liver cancer, because screening for hepatitis B virus began in 1969 and for hepatitis C virus in the early 1990s (Dwyre et al., 2011), leading to an extremely low residual infection risk (Buddeberg et al., 2008).

A further limitation of our study is that we did not have information on transfusions received before age 65. Also, we lacked information on medical conditions before age 65 that could possibly confound associations. However, as controls were matched to the cases by age and calendar year, any lack of sensitivity in exposure assessment that arises from the limited duration of claims data is non-differential and would bias estimates towards the null.

Strengths of our study are the large population based case-control design and the nearly complete ascertainment of cancer cases from the source population in the SEER catchment areas. In addition, we were able to adjust for confounding by documented medical conditions, which likely impacted the results of earlier studies. Our findings thus can be generalized to the U.S. elderly population. We also confirmed our results in sensitivity analyses using propensity scores and disease risk scores based approaches to matching and analysis, which more efficiently controlled for confounding. We exclusively compared the impact of these different matching and adjustment approaches and found in general similar results. However, lower risk was observed when propensity score and disease risk score based approaches were used, whereat for the propensity score methods the lowest estimates were observed. For liver cancer, the results of our covariate and score matching approaches differed the most. Only after careful adjustment for additional conditions we observed similar results in the design based on matching on selected covariates. Additionally, we repeated the three different matchings restricted to the liver cancer cases and observed better balance for the conditions between the cases and the controls if the matching was based on the propensity or the disease risk scores. The significantly increased liver cancer risk was comparable with the results in the main analysis, but slightly lower for all three matching designs indicating possible residual confounding in the main analysis for the site-specific cancers and subtypes.

In summary, we found that the receipt of a blood transfusion was associated with an increased short-term risk of cancer overall and for specific cancer types. Risk was not elevated over longer time periods, suggesting that transfusions in most patients that were later diagnosed with cancer are prompted by an undiagnosed cancer or a precursor to cancer. Our results do not provide support for a model in which transfusion contributes to the development of cancer over longer intervals, through pathways related to inflammation, immune modulation, or infection. Nonetheless, given the strong associations over shorter intervals, the possibility of an undiagnosed cancer should be considered for elderly patients with unexplained anemia, with further medical evaluation guided by individuals' overall health status (Rockey, 2010).

Chapter 9

Summary and future work

In the first part, the thesis focuses on bias in the estimates of conditional exposure effects when summary scores, including the propensity score (PS) and the disease risk score (DRS), are used as adjustment variables in regression models to control for confounding in observational research instead of the confounders themselves. The use of summary scores was investigated in cohort studies and in matched case-control studies analytically and in simulations. These summary score approaches were also applied in the substantive part of the thesis where we investigated the association of blood transfusions and the subsequent risk of cancers in the U.S. elderly.

For cohort studies, we showed that when the true PS is used as adjustment variable in multivariable regression models based on the exponential family, the conditional exposure estimate is unbiased only under the null hypothesis of no association. When linear regression is used to analyze the cohort data, adjusting for the true PS results in no bias even when there is an exposure effect. Otherwise, adjusting for summary scores, including adjusting for a monotone transformation of the PS, can result in a biased estimation of conditional exposure effects. We investigated the amount of bias numerically when Poisson regression or logistic regression is used to obtain rate ratios or odds ratios for several scenarios and for varying exposure effects, varying disease, exposure and confounder prevalences and different number of confounders. When there is an exposure effect, using the PS as adjustment variable in Poisson regression resulted in only minor bias. In logistic regression, we observed a bias towards null in nearly all scenarios. The bias increased with increasing exposure effect, increasing disease prevalence and decreasing exposure prevalence. Adjusting for the DRS resulted in large bias for Poisson models and somewhat less bias for logistic regression models. In logistic regression, the bias increased with increasing exposure effect and decreasing disease prevalence. When the log transformed DRS or the logit of the DRS is used as adjustment variable in Poisson models or logistic regression, the misspecified outcome model corresponds to adjusting for the confounders itself, i.e. the true outcome model, and trivially results in no bias.

We also studied the amount of bias analytically for case-control studies, when cases and controls are matched on the true summary scores and conditional or unconditional logistic regression models adjusted for the matching variables are used to analyze the matched data. Again, when matching on the PS is performed, estimates of exposure association are unbiased under the null hypothesis of no exposure outcome association. When the DRS is used, unbiased estimates are obtained even under the alternative when data are analyzed using conditional

logistic regression, but are biased for unconditional regression adjusted for the matching DRS. Using the PS as matching variable, different confounder combinations with the same PS values can lead to mismatches and, when those mismatches are differently associated with the outcome, this can further lead to bias. In our numerical bias calculations we observed a substantial bias towards null, which increased with increasing exposure effect, increasing disease prevalence and decreasing exposure prevalence.

We checked our analytic bias calculations in simulations and compared our numerically obtained results for cohort studies with simulation results from the literature (Austin et al., 2007b; Arbogast and Ray, 2011). The results of the calculations compared with our simulation based estimates for larger cohorts agree very well. Likewise, our computed results for the PS adjusted models in cohort analysis, coincide with results from Austin et al. (2007b), who assessed the performance of adjusting for the PS in logistic regression and Poisson regression models. In another simulation study, Arbogast and Ray (2011) investigated the performance of the PS and the DRS as adjustment variables in a Poisson model. For an adequate number of events per confounder (i.e. 5 events per confounder), only negligible bias was observed for all models in all scenarios. In agreement with Arbogast and Ray (2011), based on our analytic computations, we observed no or very little bias for models adjusted for the DRS on the log scale or for the PS on either scale. This is due to the very special choice of parameters, i.e. the confounder-exposure and confounder-outcome associations are the same, that causes the misspecified model and the true outcome model to be the same and leads to very narrow conclusions of limited practical relevance. When we adjusted models for the DRS however, there was a substantial bias in estimates of association. We thus assume that the results in Arbogast and Ray (2011) are based on models adjusted for the log-transformed DRS.

We also performed a simulation study based on real data to obtain realistic correlations between the covariates. In that example 24 variables, i.e. 4 demographic parameters and 20 medical conditions, as potential confounders were randomly sampled from a subcohort of Medicare beneficiaries. Additionally to the PS and the DRS, a summary score called 'any condition' defined as one, when any of the 20 medical conditions was present and zero otherwise, was calculated. The amount of bias, using logistic regression adjusted for the summary scores in cohort analysis and matching on the PS, on the DRS or partial matching on demographic parameters for case-control studies was investigated. Similar results for the PS and DRS approaches in cohort or matched case-control studies were observed compared with our analytic bias calculations. Using the summary score 'any condition' as an adjustment variable in various models resulted in a very large bias (up to 120%). We also observed that partial matching on demographic parameters without further adjustment led to a large bias in conditional logistic analyses.

The second, substantive part of the thesis is an investigation of blood transfusions and the subsequent risk of cancers in the U.S. elderly (Riedl et al., 2013). For this study we performed several summary score approaches to reduce the impact of confounding by demographic parameters and several medical diagnoses and chronic conditions associated with transfusion and cancer. Due to the large number of confounders and the repeated sampling of controls which requires a more complex variance estimation, the inclusion of all confounders in a regression model was impossible. In our first approach, case-control matching on selected covariates was performed combined with adjustment for the summary measure any condition, as defined above, and adjustment for selected conditions. We also matched cases and controls on the PS and on the DRS. In general, the results were similar in all our approaches. However, lower risk estimates were observed for the PS

matching approach than for matching on the DRS and matching on selected covariates.

The thesis focuses on conditional instead of the marginal estimates of exposure association, as anyone considering a preventive or clinical action would also account for person-specific covariates that impact outcome and thus exposure estimates conditional on covariates are the most relevant for a particular person or patient. For randomized trials, i.e. when treatment allocation is independent of the covariates, Gail et al. (1984) showed that when covariates associated with outcome are omitted, estimates of the conditional treatment effect are only unbiased for linear models, but not for any other outcome that has a density in the exponential family. Gail et al. (1984) also showed that for rare disease the bias in logistic regression and Poisson models was negligible. These results do not apply here, since they are based on the assumption of independence of the treatment (or exposure) and the additional adjustment variables. In our setting the scores are correlated with the main exposure of interest. The problematic of the estimation of non-linear conditional exposure effects by using balancing scores, especially for the PS which is designed to estimate marginal linear exposure effects, has been investigated by Austin et al. (2007b) in simulation studies. We further provide analytic bias calculations which have several advantages over simulations: they are are very fast, they are not prone to random variation and allow to assess the impact of several quantities, including exposure and disease prevalence, confounder distributions and effect sizes comprehensively. The computations can be used for any score and are not limited to special settings. The analytic approach can thus also be used to assess the impact of misspecification of the summary scores.

A further novel aspect of our work is that we considered the setting when summary scores are used as matching variables in a case-control study. While the PS has been developed for the analysis of cohort data, it has nonetheless been used as a matching variable for cases and controls in medical investigations (e.g. Etminan et al., 2010; Thillemann et al., 2010; Modén et al., 2010; Howell et al., 2010; Michalia et al., 2012). We found that matching cases and controls on the PS could lead to substantial underestimation of true effect sizes, based on conditional or unconditional logistic regression models. While matching cases and controls based on the DRS lead to unbiased estimates of associations when data were analyzed using conditional logistic regression, this approach might be best suited only for special applications as one has to fit logistic models with all covariates to the data first, to obtain estimates of the DRS. In that case often using a full covariate adjusted model for analysis may be a more straight forward approach. Some authors have argued that in the estimation of summary scores, like the DRS, a larger number of confounders can be included in the score model compared with conventional multivariable outcome models and therefore more precise and less biased estimates of effects of the main exposure on outcome might can be obtained (Stürmer et al., 2005; Arbogast et al., 2008). However, as our focus was the asymptotic bias, we did not investigate the DRS in such settings. In our simulations we only investigated scenarios with an adequate number of events per confounder (i.e. >7). Thus, adjusting for all covariates in a regression model resulted in no substantial bias.

There are several other problems that were not considered in this thesis. In our analytic bias calculations we only assessed the bias for the true summary scores and assumed that maximum likelihood or methods of moments are used to estimate the exposure effects. Hade and Lu (2014) showed that there is a bias in adjusting for the PS that exists even in a linear regression model when the estimated PS is used and parameters are estimated using ordinary least squares in small samples. Further, for cohort studies, we did not present bias calculations

for the estimation of conditional hazard ratios (HRs). When the PS is used as adjustment variable, in Austin et al. (2007b) slightly larger bias towards the null was observed for the estimation of conditional HRs compared with results obtained for logistic regression models. We also did also address the question of what happens when the PS or DRS models are misspecified. However, the analytic calculations are easily modified for the investigation of such settings. We also investigated the scores under the assumption of no unmeasured confounding and no effect modification. Additionally and as suggested by Leacy and Stuart (2013) for matching in cohort studies, a further investigation of the combination of the PS and the DRS might be interesting.

In summary, in the estimation of non-linear conditional exposure effects from models adjusted for balancing scores, substantial bias can occur in cohort studies and in matched case-control studies analyzing the data using conditional logistic regression or adjusted unconditional logistic regression. However, not accounting for potential confounders or accounting for them by very simplistic summary scores without any balancing properties can result in even larger bias. Using the PS as an adjustment variable in the analysis of cohort data based on Poisson or logistic regression models or as a matching variable for case-control studies yields unbiased estimates of null associations, and thus could be useful for testing especially when the disease is rare and the exposure is somewhat more common. The impact of this adjustment on variance estimates and thus power of association tests will be part of future work.

Appendix A

Supplemental Tables and Figures

A.1 Supplemental Figures for Chapter 6

A.1.1 Cohort study - Poisson regression

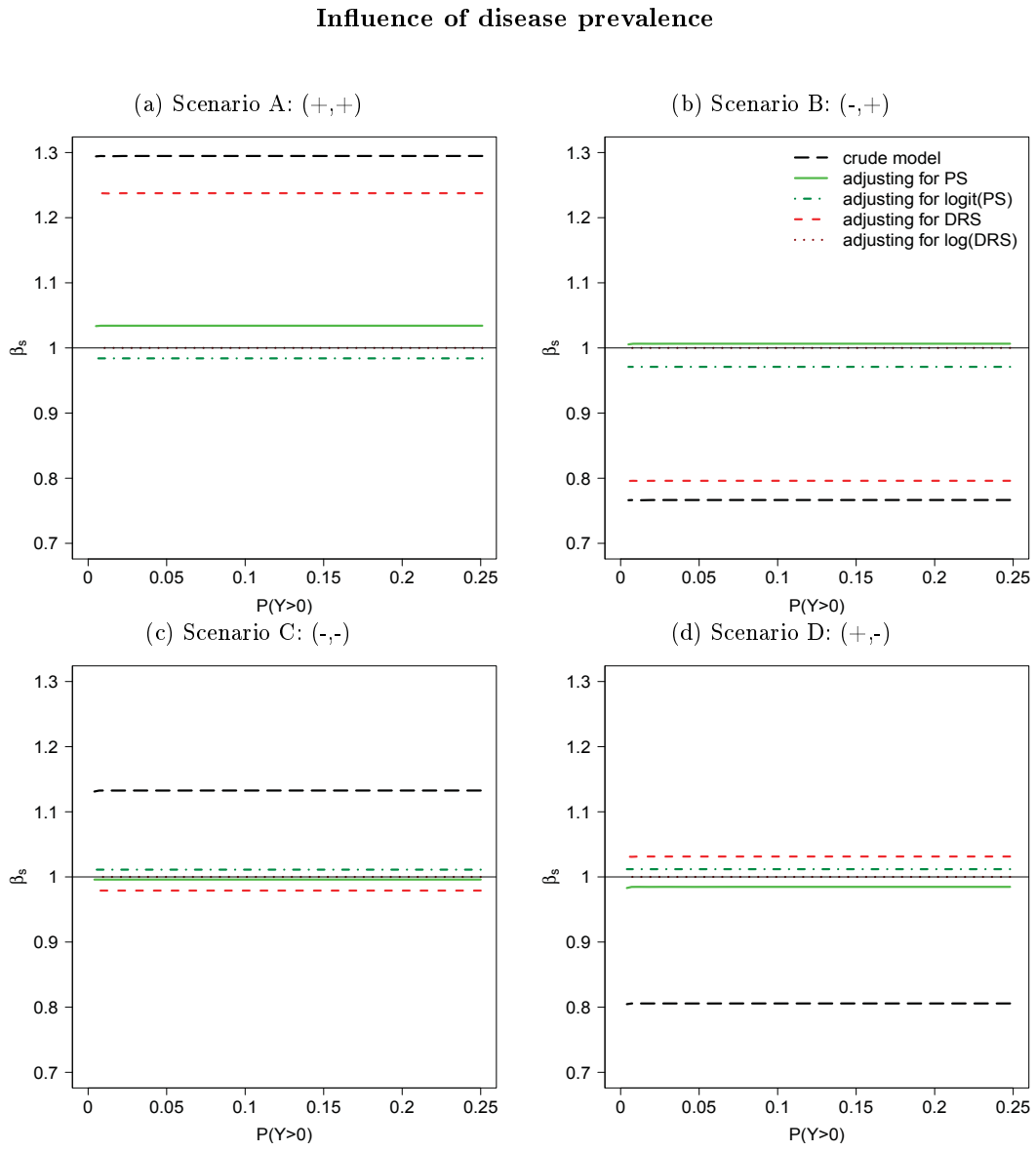


Figure A.1: Analytic bias for different disease prevalences with $P(T = 1) = P(X = 1) = 0.1$ and $\beta = 1$.

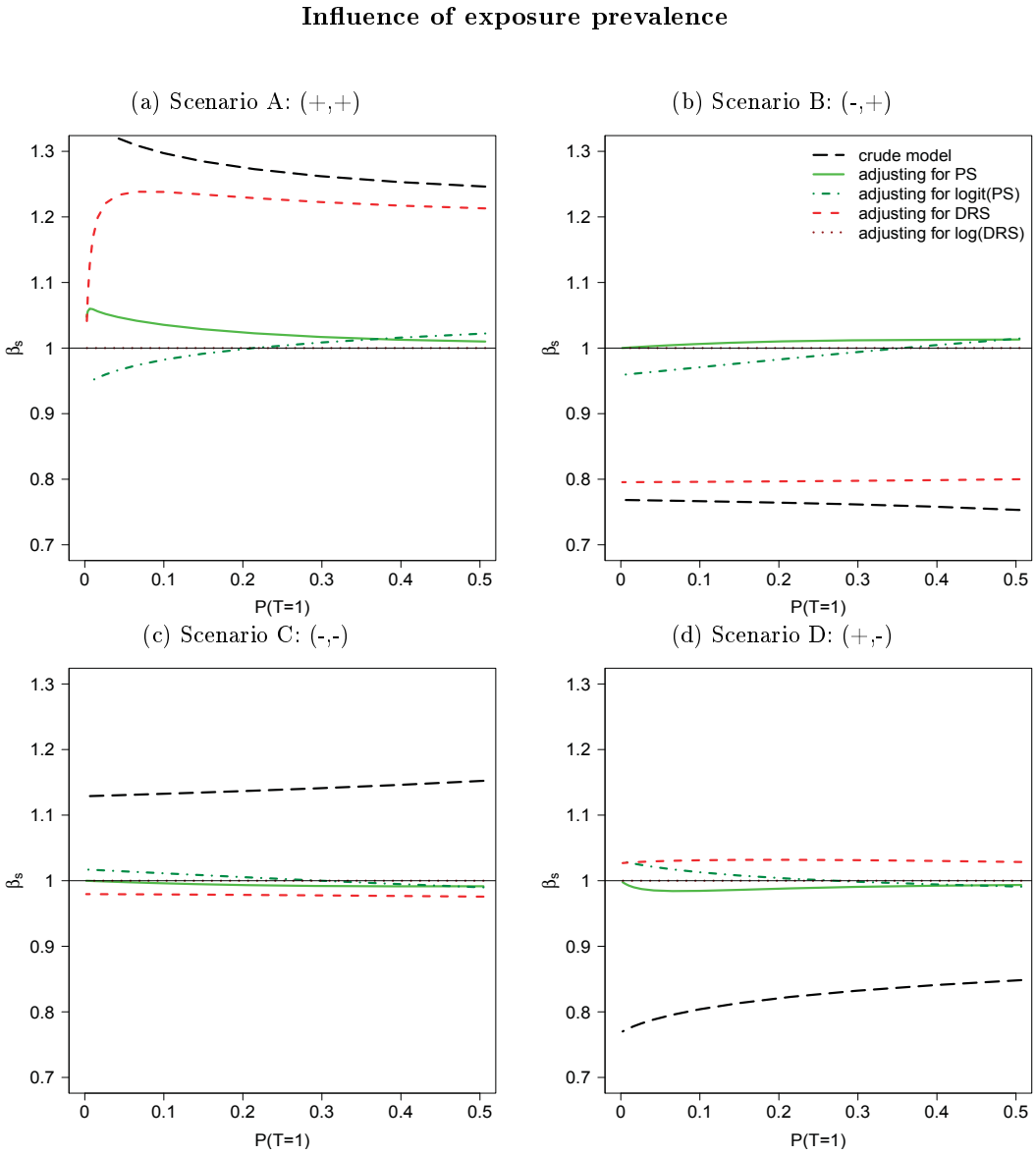


Figure A.2: Analytic bias for different exposure prevalences with $P(Y > 0) = P(X = 1) = 0.1$ and $\beta = 1$.

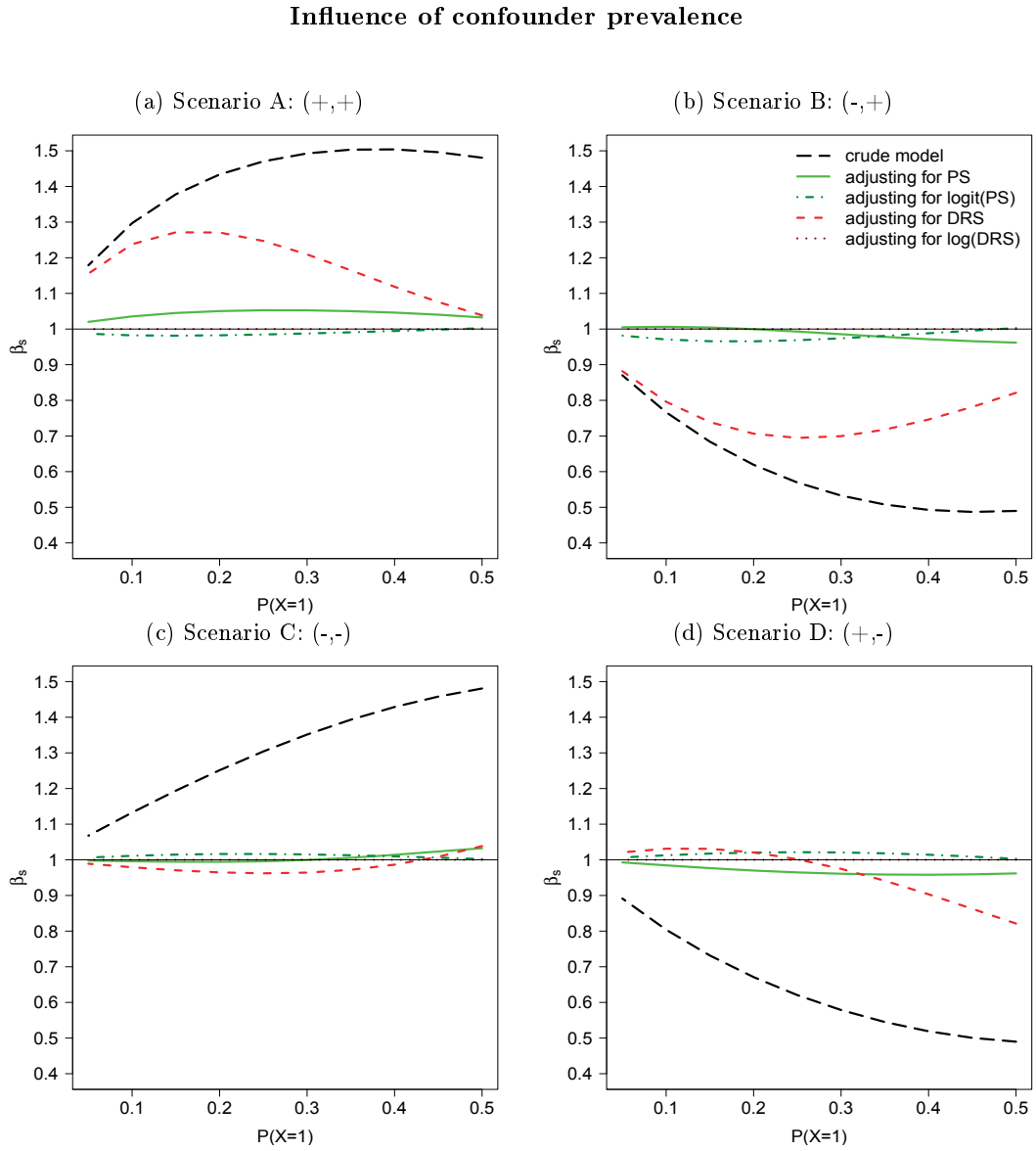


Figure A.3: Analytic bias for different confounder prevalences with $P(Y > 0) = P(T = 1) = 0.1$ and $\beta = 1$.

A.1.2 Cohort study - logistic regression

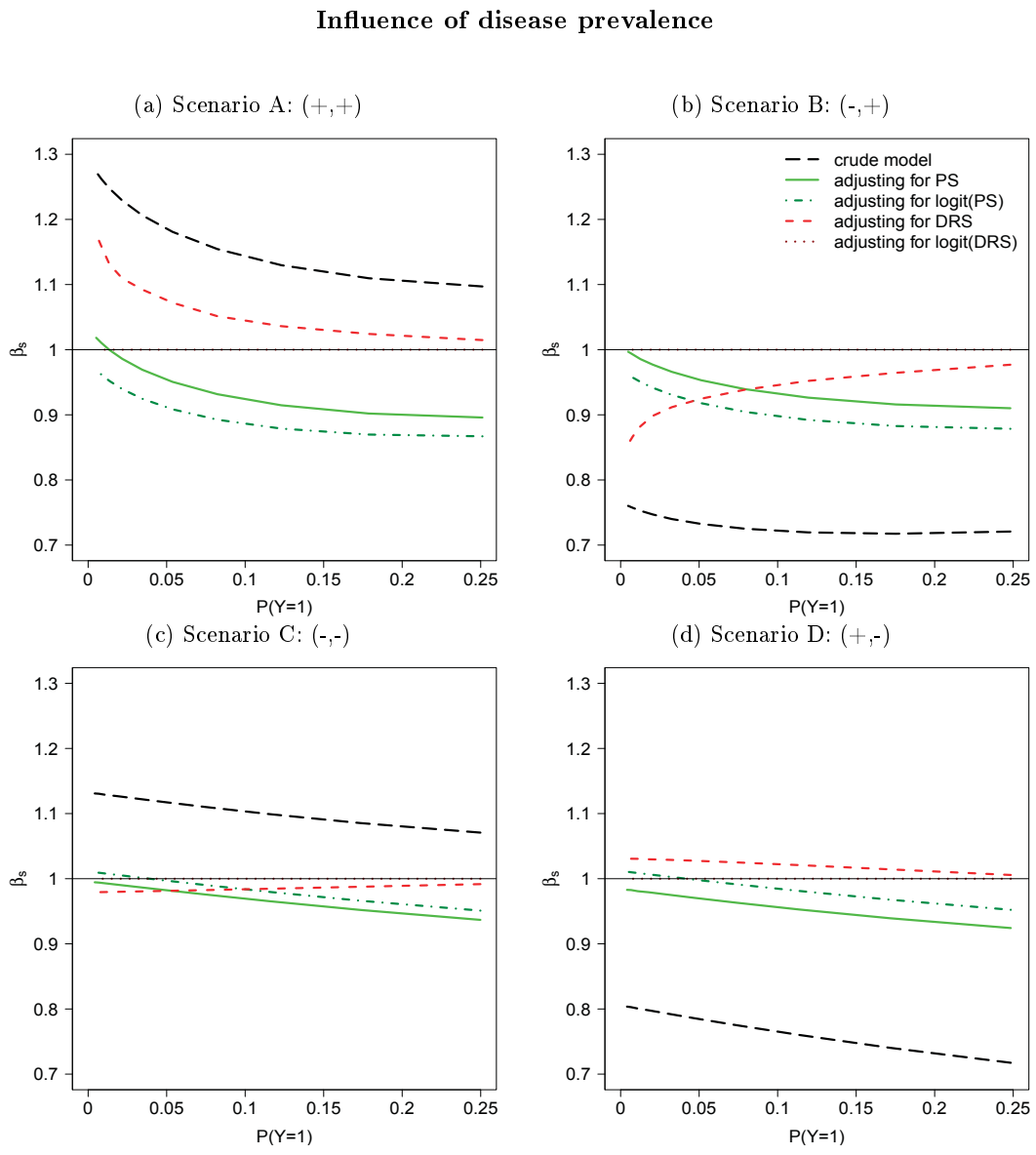


Figure A.4: Analytic bias for different disease prevalences with $P(T = 1) = P(X = 1) = 0.1$ and $\beta = 1$.

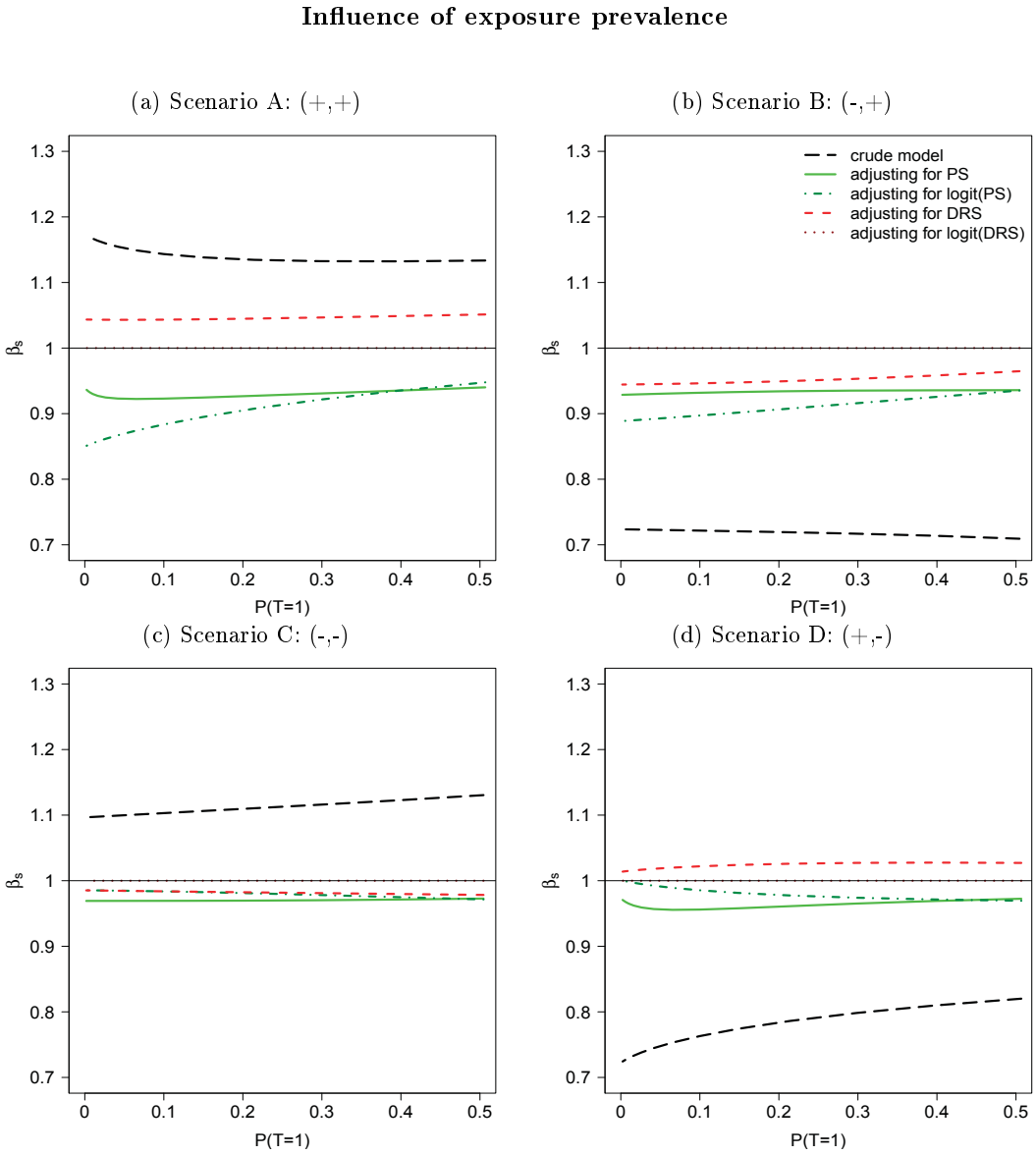


Figure A.5: Analytic bias for different exposure prevalences with $P(Y = 1) = P(X = 1) = 0.1$ and $\beta = 1$.

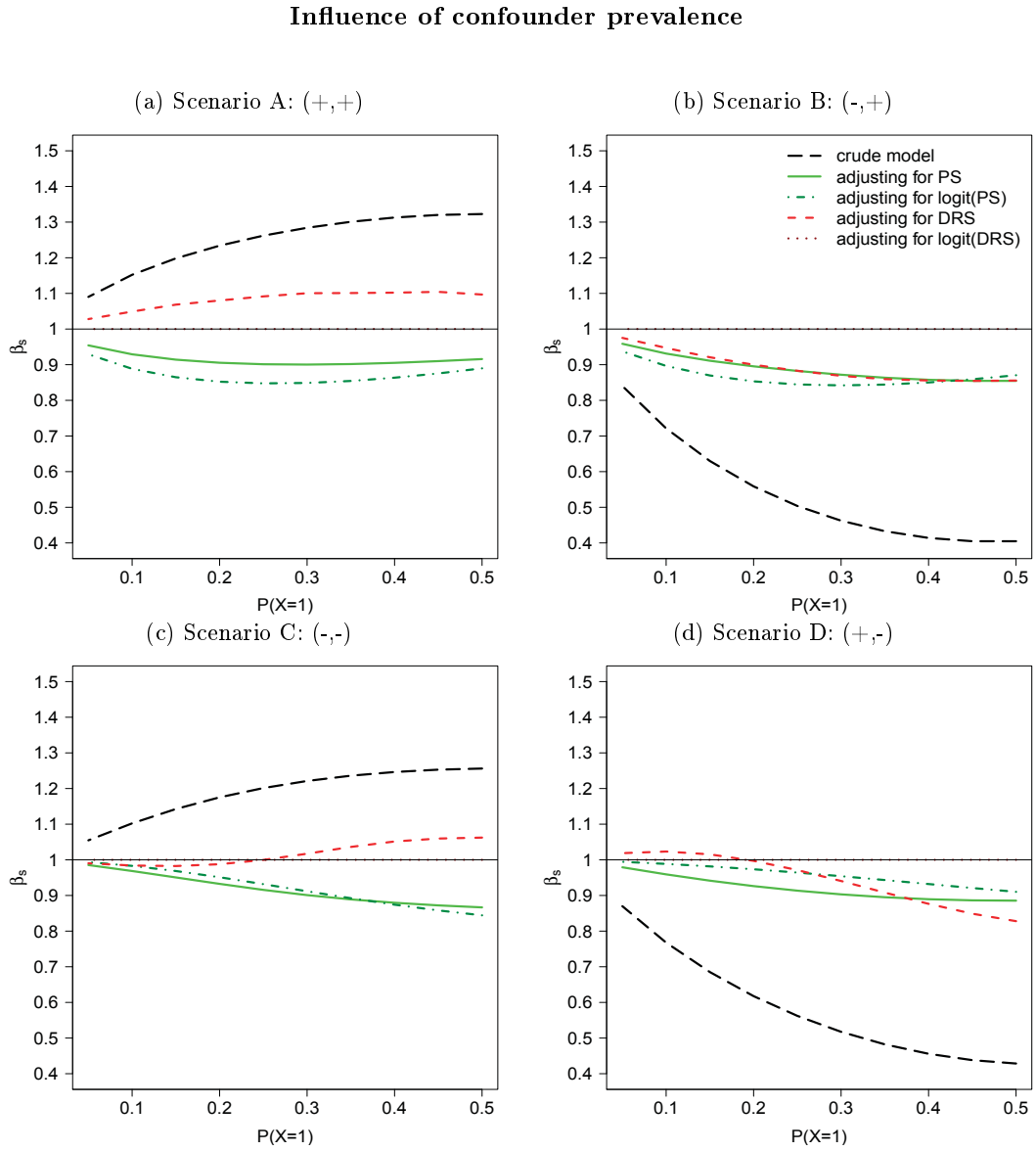


Figure A.6: Analytic bias for different confounder prevalences with $P(Y = 1) = P(T = 1) = 0.1$ and $\beta = 1$.

A.1.3 Case-control study

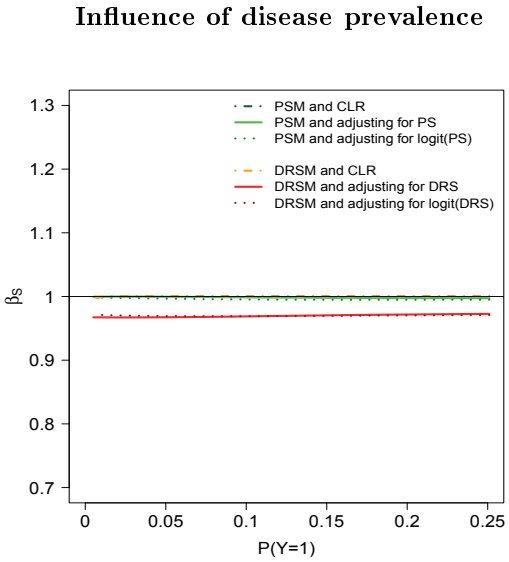


Figure A.7: Analytic bias under unequal confounder-exposure association (scenario II) with $P(T = 1) = 0.1$, $P(X = 1) = 0.1$ and true exposure effects $\beta = 1$. PSM denotes PS matching, DRSM denotes DRS matching and CLR conditional logistic regression.

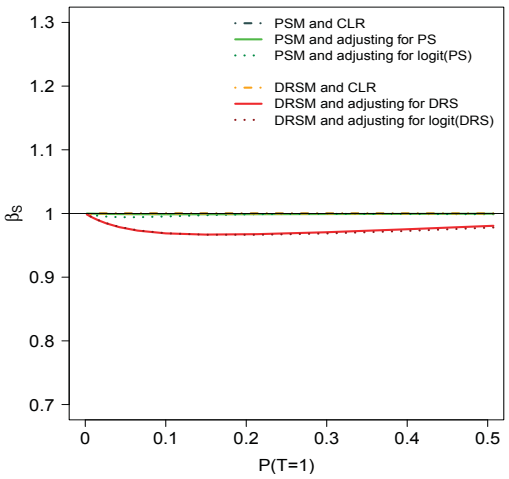


Figure A.8: Analytic bias under unequal confounder-exposure association (scenario II) with $P(Y = 1) = 0.1$, $P(X = 1) = 0.1$ and true exposure effects $\beta = 1$. PSM denotes PS matching, DRSM denotes DRS matching and CLR conditional logistic regression.

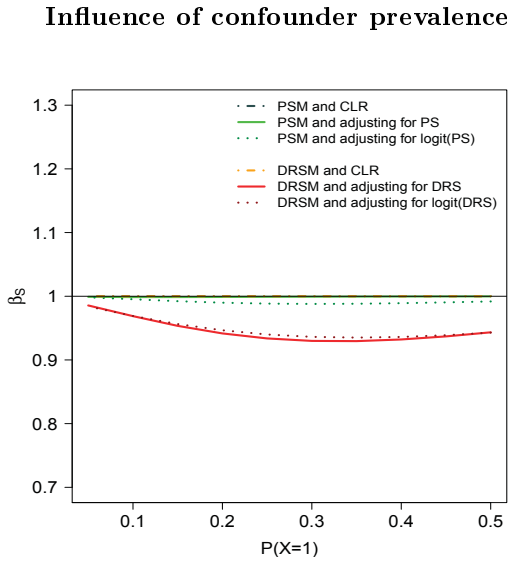


Figure A.9: Analytic bias under unequal confounder-exposure association (scenario II) with $P(Y = 1) = 0.1$, $P(T = 1) = 0.1$ and true exposure effects $\beta = 1$. PSM denotes PS matching, DRSM denotes DRS matching and CLR conditional logistic regression.

A.2 Supplemental Tables for Chapter 7

Table A.1: Mean estimates $\hat{\beta}_s$ from simulations based on blood transfusion data, with disease prevalence 10% and exposure prevalence 10% for $\beta = 0$ and $\beta = 1$ under scenario I.

	$\beta = 0$				$\beta = 1$				
Model	mean $\hat{\beta}_s$	VarM	VarE	coverage	mean $\hat{\beta}_s$	VarM	VarE	relBias	coverage
	Full cohort analysis								
0	0.472	0.002	0.002	0	1.209	0.001	0.001	0.209	0
1	0.001	0.002	0.003	94.6	1.001	0.002	0.002	0.001	94.7
2	0.375	0.002	0.002	0	1.127	0.001	0.001	0.127	3.4
3a	0.001	0.002	0.002	96.5	0.802	0.002	0.001	-0.198	0
3b	-0.067	0.002	0.002	66.7	0.700	0.001	0.001	-0.300	0
4a	0.086	0.002	0.003	52.2	1.105	0.001	0.002	0.105	22.5
4b	0.001	0.002	0.003	91.8	1.001	0.001	0.002	0.001	91.9
	Case-control analysis, matching on sex, age, race and selection year								
1	0.004	0.035	0.035	95.2	1.034	0.032	0.034	0.034	94.3
2	0.381	0.019	0.021	22.8	1.141	0.019	0.019	0.141	83.1
3a	0.005	0.025	0.025	95.0	0.793	0.023	0.023	-0.207	73.1
3b	-0.044	0.023	0.023	94.3	0.731	0.022	0.022	-0.269	56.8
4a	0.074	0.027	0.028	92.6	1.061	0.023	0.023	0.061	94.2
4b	-0.002	0.027	0.027	94.8	0.992	0.025	0.026	-0.008	94.7
5	0.459	0.019	0.020	8.7	1.214	0.020	0.021	0.214	67.3
	Case-control analysis, matching on $\hat{e}(\boldsymbol{X})$								
1	0.035	0.028	0.026	95.4	1.029	0.025	0.023	0.029	95.5
3a	-0.001	0.020	0.018	96.5	0.766	0.018	0.016	-0.234	58.4
5	-0.001	0.020	0.018	96.5	0.768	0.019	0.017	-0.232	60.9
	Case-control analysis, matching on logit$(\hat{e}(\boldsymbol{X}))$								
1	0.039	0.028	0.027	95.6	1.022	0.025	0.022	0.022	95.8
3b	0.002	0.019	0.017	96.1	0.718	0.017	0.014	-0.282	41.6
5	0.002	0.021	0.020	95.6	0.764	0.019	0.017	-0.236	59.4
	Case-control analysis, matching on $\widehat{DRS}(\boldsymbol{X})$								
1	0.003	0.020	0.018	95.7	1.017	0.020	0.020	0.017	95.4
4a	-0.002	0.016	0.014	96.1	0.979	0.016	0.016	-0.021	94.8
5	-0.001	0.016	0.015	96.2	1.008	0.018	0.018	0.008	94.9
	Case-control analysis, matching on logit$(\widehat{DRS}(\boldsymbol{X}))$								
1	0.002	0.020	0.019	94.7	1.023	0.020	0.019	0.023	94.9
4b	-0.002	0.016	0.015	96.1	0.990	0.016	0.015	-0.010	94.6
5	0	0.016	0.016	95.8	1.010	0.018	0.017	0.010	94.7

Model 0: logistic regression including only exposure T (crude model)
Model 1: logistic regression adjusted for all covariates $\boldsymbol{X}$
Model 2: logistic regression adjusted for sex, age race, selection year, and 'any condition'
Model 3a: logistic regression adjusted for estimated PS, $\hat{e}(\boldsymbol{X})$
Model 3b: logistic regression adjusted for estimated PS on logit scale, logit$(\hat{e}(\boldsymbol{X}))$
Model 4a: logistic regression adjusted for estimated DRS, $\widehat{DRS}(\boldsymbol{X})$
Model 4b: logistic regression adjusted for estimated DRS on logit scale, logit$(\widehat{DRS}(\boldsymbol{X}))$
Model 5: conditional logistic regression

Table A.2: Mean estimates $\hat{\beta}_s$ from simulations based on blood transfusion data, with disease prevalence 5% and exposure prevalence 10% for $\beta = 1$ under scenario I and II.

	Scenario I					Scenario II				
Model	mean $\hat{\beta}_s$	VarM	VarE	relBias	coverage	mean $\hat{\beta}_s$	VarM	VarE	relBias	coverage
	Full cohort analysis									
0	1.301	0.002	0.002	0.301	0	2.246	0.002	0.002	1.246	0
1	1.004	0.004	0.004	0.004	95.7	1.004	0.004	0.004	0.004	95.8
2	1.203	0.002	0.003	0.203	2.2	2.259	0.002	0.002	1.259	0
3a	0.918	0.004	0.003	-0.082	73.3	1.039	0.005	0.004	0.039	90.2
3b	0.791	0.003	0.003	-0.209	3.1	0.969	0.004	0.004	-0.028	94.1
4a	1.235	0.003	0.004	0.235	1.7	1.350	0.003	0.004	0.350	0
4b	1.004	0.003	0.004	0.004	90.5	1.004	0.003	0.004	0.006	90.0
	Case-control analysis, matching on sex, age, race and selection year									
1	1.024	0.030	0.032	0.024	94.3	1.017	0.029	0.029	0.017	94.8
2	1.206	0.017	0.017	0.206	63.9	2.248	0.017	0.018	1.248	0
3a	0.866	0.021	0.020	-0.134	83.3	1.012	0.027	0.026	0.023	95.5
3b	0.798	0.020	0.019	-0.202	69.3	1.040	0.026	0.025	0.040	95.1
4a	1.118	0.020	0.020	0.118	86.5	1.317	0.019	0.019	0.317	37.6
4b	0.978	0.023	0.024	-0.022	93.7	1.282	0.019	0.019	0.282	47.2
5	1.282	0.018	0.019	0.282	44.1	2.288	0.025	0.027	1.288	0
	Case-control analysis, matching on $\hat{e}(\boldsymbol{X})$									
1	1.026	0.023	0.022	0.026	95.5	0.949	0.020	0.018	-0.051	94.0
3a	0.841	0.017	0.016	-0.159	76.2	0.957	0.019	0.018	-0.043	93.9
5	0.843	0.018	0.017	-0.157	78.1	0.969	0.021	0.020	-0.031	94.5
	Case-control analysis, matching on logit($\hat{e}(\boldsymbol{X})$)									
1	1.016	0.024	0.023	0.016	95.7	0.941	0.020	0.020	-0.048	93.5
3b	0.791	0.015	0.013	-0.209	60.7	0.895	0.018	0.017	-0.092	88.0
5	0.840	0.018	0.016	-0.160	76.8	0.960	0.021	0.021	-0.028	93.7
	Case-control analysis, matching on $\widehat{DRS}(\boldsymbol{X})$									
1	1.013	0.018	0.018	0.013	95.5	1.013	0.019	0.018	0.013	95.7
4a	0.971	0.014	0.015	-0.029	92.2	0.898	0.012	0.012	-0.102	84.1
5	1.013	0.016	0.017	0.013	93.8	1.016	0.015	0.016	0.016	94.2
	Case-control analysis, matching on logit($\widehat{DRS}(\boldsymbol{X})$)									
1	1.010	0.018	0.018	0.010	95.5	1.015	0.019	0.019	0.025	94.7
4b	0.979	0.014	0.015	-0.021	94.6	0.971	0.013	0.014	-0.021	93.3
5	1.006	0.015	0.016	0.006	94.9	1.013	0.015	0.017	0.022	93.1

Model 0: logistic regression including only exposure T (crude model)
Model 1: logistic regression adjusted for all covariates $\boldsymbol{X}$
Model 2: logistic regression adjusted for sex, age race, selection year, and 'any condition'
Model 3a: logistic regression adjusted for estimated PS, $\hat{e}(\boldsymbol{X})$
Model 3b: logistic regression adjusted for estimated PS on logit scale, logit($\hat{e}(\boldsymbol{X})$)
Model 4a: logistic regression adjusted for estimated DRS, $\widehat{DRS}(\boldsymbol{X})$
Model 4b: logistic regression adjusted for estimated DRS on logit scale, logit($\widehat{DRS}(\boldsymbol{X})$)
Model 5: conditional logistic regression

A.3 Supplemental Tables for Chapter 8

Table A.3: Associations of blood transfusions with cancer risk, overall and for specific sites and subtypes for covariate matched controls.

		OR[1] (95% CI)		
	Total[2]	0-12 months	13-30 months	31-48 months
All cancers	552 951	2.05 (1.95-2.16)[3]	1.04 (0.99-1.09)	1.05 (1.00-1.11)
Oral Cavity and Pharynx	8 897	1.38 (0.77-2.49)	1.44 (1.04-1.98)	1.18 (0.83-1.68)
Lip	1 119	1.06 (0.54-2.10)	1.33 (0.85-2.08)	1.13 (0.67-1.90)
Tongue	2 042	1.29 (0.72-2.30)	1.58 (1.11-2.25)	1.32 (0.87-1.98)
Salivary Gland	1 262	1.54 (0.84-2.85)	1.37 (0.90-2.08)	0.82 (0.48-1.41)
Mouth	2 418	1.55 (0.89-2.72)	1.38 (0.97-1.95)	1.27 (0.85-1.90)
Pharynx	2 056	1.34 (0.75-2.39)	1.49 (1.03-2.15)	1.19 (0.77-1.84)
Digestive System	125 386	3.14 (1.97-4.99)[3]	1.08 (0.85-1.38)	1.04 (0.80-1.35)
Esophagus	5 581	1.78 (1.06-2.99)	1.30 (0.96-1.77)	1.40 (1.00-1.96)
Stomach	10 961	3.92 (2.46-6.25)[3]	1.16 (0.88-1.55)	1.29 (0.94-1.76)
Colon	58 480	3.73 (2.46-5.65)[3]	1.11 (0.88-1.38)	0.95 (0.73-1.22)
Rectum	17 608	2.00 (1.24-3.22)	0.72 (0.54-0.95)	0.86 (0.63-1.16)
Anus	1 281	1.79 (0.98-3.25)	0.75 (0.45-1.24)	1.33 (0.83-2.14)
Liver	5 039	3.29 (2.01-5.40)[3]	1.72 (1.27-2.32)[3]	1.65 (1.18-2.31)
Pancreas	17 114	2.16 (1.34-3.47)	0.95 (0.73-1.24)	1.03 (0.77-1.39)
Respiratory System	92 639	1.71 (1.04-2.79)	1.18 (0.92-1.52)	1.24 (0.95-1.62)
Lung	88 275	1.72 (1.13-2.64)	1.17 (0.95-1.45)	1.24 (0.98-1.57)
Bones and Joints	396	1.26 (0.52-3.09)	1.27 (0.64-2.52)	1.00 (0.44-2.29)
Soft Tissue	2 504	1.99 (1.07-3.70)	0.94 (0.62-1.42)	0.99 (0.63-1.55)
Skin excluding Basal and Squamous	16 520	1.05 (0.59-1.87)	0.96 (0.70-1.30)	1.04 (0.75-1.45)
Melanoma	14 259	1.04 (0.62-1.72)	0.91 (0.69-1.19)	1.03 (0.76-1.39)
Breast[4]	64 701	0.94 (0.36-2.46)	0.83 (0.52-1.33)	0.93 (0.54-1.59)
Female Genital System	24 812	1.53 (0.57-4.14)	0.89 (0.53-1.49)	0.97 (0.54-1.75)
Cervix	1 727	1.69 (0.61-4.68)	1.29 (0.73-2.28)	1.30 (0.66-2.53)
Uterus	12 210	1.15 (0.44-3.02)	0.81 (0.49-1.32)	0.88 (0.50-1.56)
Ovary	7 940	1.96 (0.75-5.14)	0.83 (0.49-1.38)	0.96 (0.53-1.74)
Vagina / Vulva	2 221	1.52 (0.55-4.24)	1.17 (0.67-2.03)	1.28 (0.67-2.43)
Male Genital System	91 147	0.89 (0.36-2.22)	0.80 (0.48-1.36)	0.81 (0.47-1.40)
Prostate	90 546	0.89 (0.40-1.96)	0.81 (0.51-1.27)	0.81 (0.50-1.32)
Urinary System	45 343	1.79 (1.06-3.00)	1.05 (0.80-1.39)	1.09 (0.81-1.47)
Urinary Bladder	31 799	1.53 (0.96-2.44)	1.05 (0.82-1.34)	1.10 (0.84-1.45)
Kidney, Renal Pelvis, Ureter	13 166	2.42 (1.50-3.92)[3]	1.04 (0.78-1.37)	1.07 (0.79-1.46)
Eye and Orbit	728	0.74 (0.30-1.79)	0.79 (0.42-1.52)	1.03 (0.53-1.98)
Brain and Other Nervous System	4 824	1.08 (0.58-2.00)	0.95 (0.66-1.37)	0.83 (0.55-1.24)
Endocrine System	3 236	1.24 (0.67-2.32)	0.70 (0.46-1.06)	0.85 (0.55-1.33)
Thyroid	2 871	1.12 (0.63-1.97)	0.68 (0.45-1.01)	0.81 (0.53-1.25)
Lymphoma	29 991	2.41 (1.43-4.08)	0.96 (0.71-1.29)	1.09 (0.79-1.49)
Hodgkin Lymphoma	909	3.14 (1.75-5.62)[3]	1.05 (0.63-1.74)	0.96 (0.54-1.73)
NHL, including CLL[5]	29 082	2.38 (1.51-3.76)[3]	0.95 (0.74-1.22)	1.09 (0.82-1.43)
Myeloma	7 792	4.71 (2.72-8.16)[3]	1.15 (0.81-1.64)	1.08 (0.74-1.59)
Leukemia	8 863	6.51 (3.81-11.13)[3]	1.66 (1.19-2.32)	1.38 (0.95-1.99)
Mesothelioma	1 703	1.83 (0.95-3.52)	0.87 (0.54-1.39)	1.08 (0.66-1.78)
Kaposi Sarcoma	293	3.22 (1.51-6.86)	1.60 (0.83-3.09)	0.64 (0.21-1.88)

[1] The OR for overall cancer risk is adjusted for sex, age, race, selection year, and any condition as categorical variables. The ORs for the site-specific cancer groups like digestive system are adjusted for sex, age, race, and any condition as categorical variables and selection year as continuous variable. The ORs for subtypes of the site-specific cancer groups are adjusted for sex and race as categorical variables and age, selection year, and any condition as continuous variables.

[2] The number of cases may not add up to the totals for cancer specific sites and for cancer overall because miscellaneous cancers were excluded.

[3] Association is significant at $p<0.0004$ (Bonferroni correction for all 129 comparisons, 43 sites for 3 time periods).

[4] Women only (537 males are excluded from analysis).

[5] CLL = chronic lymphocytic leukemia

Table A.4: Associations of blood transfusions with cancer risk, overall and for specific sites and subtypes in men based on covariate matched controls.

		OR[1] (95% CI)		
	Total[2]	0-12 months	13-30 months	31-48 months
All cancers	280 034	1.92 (1.79-2.07)[3]	1.08 (1.00-1.16)	1.06 (0.98-1.15)
Oral Cavity and Pharynx	5 025	1.27 (0.44-3.64)	1.64 (0.87-3.09)	1.15 (0.59-2.27)
Lip	766	0.56 (0.17-1.86)	1.59 (0.79-3.18)	1.11 (0.50-2.49)
Tongue	1 075	1.26 (0.46-3.46)	1.85 (0.97-3.52)	1.38 (0.67-2.85)
Salivary Gland	777	1.17 (0.41-3.34)	1.58 (0.79-3.14)	0.65 (0.27-1.58)
Mouth	1 041	1.88 (0.70-5.00)	1.70 (0.88-3.28)	1.26 (0.60-2.65)
Pharynx	1 366	1.27 (0.47-3.43)	1.49 (0.78-2.84)	1.25 (0.61-2.56)
Digestive System	56 391	2.94 (1.22-7.07)	1.12 (0.66-1.91)	1.08 (0.62-1.88)
Esophagus	3 735	1.68 (0.67-4.22)	1.26 (0.70-2.26)	1.44 (0.78-2.69)
Stomach	5 917	3.85 (1.64-9.04)	1.25 (0.71-2.21)	1.37 (0.75-2.51)
Colon	24 262	3.47 (1.57-7.64)	1.11 (0.68-1.83)	0.92 (0.53-1.60)
Rectum	8 591	1.92 (0.80-4.62)	0.73 (0.41-1.29)	0.91 (0.50-1.68)
Anus	390	2.70 (0.95-7.68)	0.78 (0.29-2.14)	1.01 (0.36-2.79)
Liver	2 972	3.23 (1.32-7.89)	1.68 (0.93-3.00)	1.63 (0.86-3.07)
Pancreas	6 931	2.02 (0.83-4.90)	1.01 (0.57-1.77)	1.02 (0.55-1.87)
Respiratory System	48 452	1.76 (0.71-4.37)	1.29 (0.75-2.20)	1.32 (0.76-2.29)
Lung	45 246	1.79 (0.81-3.94)	1.27 (0.80-2.03)	1.33 (0.81-2.19)
Bones and Joints	184	1.36 (0.31-5.86)	0.82 (0.22-3.03)	1.00 (0.27-3.70)
Soft Tissue	1 267	1.87 (0.62-5.64)	1.12 (0.53-2.35)	0.84 (0.37-1.90)
Skin excluding Basal and Squamous	9 852	1.02 (0.36-2.88)	1.01 (0.54-1.87)	1.02 (0.53-1.93)
Melanoma	8 547	0.98 (0.39-2.42)	0.94 (0.54-1.62)	1.03 (0.57-1.86)
Male Genital System	91 147	0.89 (0.36-2.22)	0.80 (0.48-1.36)	0.81 (0.47-1.40)
Prostate	90 546	0.89 (0.40-1.96)	0.81 (0.51-1.27)	0.81 (0.50-1.32)
Urinary System	29 951	1.62 (0.63-4.17)	1.10 (0.63-1.94)	1.11 (0.62-1.99)
Urinary Bladder	22 692	1.39 (0.60-3.23)	1.12 (0.68-1.84)	1.10 (0.64-1.89)
Kidney, Renal Pelvis, Ureter	7 013	2.37 (0.99-5.71)	1.03 (0.58-1.81)	1.09 (0.60-2.00)
Eye and Orbit	376	0.84 (0.21-3.36)	1.00 (0.38-2.67)	1.21 (0.45-3.28)
Brain and Other Nervous System	2 271	0.83 (0.27-2.52)	1.09 (0.55-2.17)	0.83 (0.39-1.75)
Endocrine System	1 014	1.33 (0.43-4.15)	0.78 (0.34-1.76)	1.25 (0.56-2.78)
Thyroid	827	1.18 (0.41-3.37)	0.70 (0.31-1.59)	1.30 (0.60-2.82)
Lymphoma	14 009	2.48 (0.94-6.51)	0.95 (0.51-1.76)	1.13 (0.60-2.12)
Hodgkin Lymphoma	425	2.51 (0.89-7.11)	1.14 (0.49-2.67)	1.13 (0.45-2.88)
NHL, including CLL[4]	13 584	2.47 (1.06-5.71)	0.95 (0.55-1.62)	1.13 (0.64-1.99)
Myeloma	3 744	4.65 (1.70-12.71)	1.23 (0.62-2.46)	1.02 (0.49-2.12)
Leukemia	4 532	6.71 (2.53-17.82)[3]	1.78 (0.93-3.41)	1.36 (0.68-2.72)
Mesothelioma	1 340	1.63 (0.54-4.92)	0.91 (0.42-1.93)	1.12 (0.52-2.45)
Kaposi Sarcoma	183	3.76 (1.12-12.64)	0.90 (0.27-3.00)	0.56 (0.11-2.75)

[1] The OR for overall cancer risk is adjusted for age, race, selection year, and any condition as categorical variables. The ORs for the site-specific cancer groups like digestive system are adjusted for age, race, and any condition as categorical variables and selection year as continuous variable. The ORs for subtypes of the site-specific cancer groups are adjusted for race as categorical variable and age, selection year, and any condition as continuous variables.

[2] The number of cases may not add up to the totals for cancer specific sites and for cancer overall because miscellaneous cancers were excluded.

[3] Association is significant at $p<0.0004$ (Bonferroni correction for all 111 comparisons, 37 sites for 3 time periods).

[4] CLL = chronic lymphocytic leukemia

Table A.5: Associations of blood transfusions with cancer risk, overall and for specific sites and subtypes in women based on covariate matched controls.

		OR[1] (95% CI)		
	Total[2]	0-12 months	13-30 months	31-48 months
All cancers	272 917	2.18 (2.03-2.33)[3]	1.01 (0.95-1.08)	1.04 (0.97-1.12)
Oral Cavity and Pharynx	3 872	1.50 (0.51-4.44)	1.25 (0.70-2.22)	1.21 (0.62-2.34)
Lip	353	1.96 (0.62-6.20)	0.96 (0.42-2.24)	1.17 (0.47-2.93)
Tongue	967	1.31 (0.45-3.84)	1.35 (0.73-2.49)	1.25 (0.61-2.56)
Salivary Gland	485	2.07 (0.69-6.23)	1.13 (0.54-2.38)	1.09 (0.46-2.60)
Mouth	1 337	1.35 (0.47-3.87)	1.17 (0.65-2.12)	1.27 (0.64-2.51)
Pharynx	690	1.44 (0.48-4.33)	1.52 (0.80-2.91)	1.11 (0.50-2.50)
Digestive System	68 995	3.27 (1.33-8.04)	1.04 (0.66-1.64)	1.00 (0.59-1.69)
Esophagus	1 846	1.90 (0.69-5.26)	1.39 (0.79-2.44)	1.35 (0.70-2.60)
Stomach	5 044	3.96 (1.55-10.1)	1.07 (0.63-1.80)	1.20 (0.65-2.20)
Colon	34 218	3.90 (1.67-9.14)	1.08 (0.70-1.67)	0.95 (0.56-1.60)
Rectum	9 017	2.05 (0.79-5.32)	0.71 (0.42-1.18)	0.81 (0.44-1.46)
Anus	891	1.48 (0.50-4.34)	0.72 (0.35-1.48)	1.45 (0.70-2.98)
Liver	2 067	3.35 (1.26-8.90)	1.76 (1.02-3.03)	1.67 (0.88-3.16)
Pancreas	10 183	2.25 (0.88-5.75)	0.90 (0.55-1.48)	1.03 (0.58-1.83)
Respiratory System	44 187	1.66 (0.64-4.32)	1.10 (0.68-1.78)	1.16 (0.67-2.02)
Lung	43 029	1.68 (0.69-4.05)	1.10 (0.71-1.69)	1.16 (0.70-1.93)
Bones and Joints	212	1.19 (0.29-4.93)	1.60 (0.64-4.00)	1.01 (0.30-3.35)
Soft Tissue	1 237	2.09 (0.68-6.39)	0.79 (0.39-1.58)	1.13 (0.53-2.42)
Skin excluding Basal and Squamous	6 668	1.07 (0.36-3.15)	0.92 (0.52-1.61)	1.09 (0.58-2.06)
Melanoma	5 712	1.10 (0.40-3.00)	0.90 (0.54-1.50)	1.04 (0.57-1.90)
Breast	64 701	0.94 (0.36-2.46)	0.83 (0.52-1.33)	0.93 (0.54-1.59)
Female Genital System	24 812	1.53 (0.57-4.14)	0.89 (0.53-1.49)	0.97 (0.54-1.75)
Cervix	1 727	1.69 (0.61-4.68)	1.29 (0.73-2.28)	1.30 (0.66-2.53)
Uterus	12 210	1.15 (0.44-3.02)	0.81 (0.49-1.32)	0.88 (0.50-1.56)
Ovary	7 940	1.96 (0.75-5.14)	0.83 (0.49-1.38)	0.96 (0.53-1.74)
Vagina / Vulva	2 221	1.52 (0.55-4.24)	1.17 (0.67-2.03)	1.28 (0.67-2.43)
Urinary System	15 392	2.03 (0.74-5.54)	1.01 (0.59-1.71)	1.10 (0.60-2.01)
Urinary Bladder	9 107	1.76 (0.68-4.60)	0.95 (0.58-1.57)	1.14 (0.64-2.04)
Kidney, Renal Pelvis, Ureter	6 153	2.47 (0.94-6.45)	1.05 (0.62-1.75)	1.05 (0.58-1.93)
Eye and Orbit	352	0.64 (0.15-2.81)	0.62 (0.22-1.76)	0.85 (0.29-2.53)
Brain and Other Nervous System	2 553	1.28 (0.42-3.90)	0.86 (0.46-1.59)	0.83 (0.41-1.69)
Endocrine System	2 222	1.23 (0.41-3.72)	0.66 (0.35-1.26)	0.69 (0.33-1.44)
Thyroid	2 044	1.11 (0.39-3.15)	0.66 (0.36-1.22)	0.63 (0.31-1.29)
Lymphoma	15 982	2.37 (0.88-6.39)	0.95 (0.56-1.62)	1.05 (0.57-1.92)
Hodgkin Lymphoma	484	3.67 (1.27-10.63)	0.99 (0.45-2.15)	0.83 (0.33-2.14)
NHL, including CLL[4]	15 498	2.33 (0.93-5.83)	0.95 (0.59-1.53)	1.05 (0.60-1.84)
Myeloma	4 048	4.76 (1.70-13.3)	1.07 (0.58-1.97)	1.12 (0.56-2.23)
Leukemia	4 331	6.29 (2.29-17.32)[3]	1.55 (0.86-2.78)	1.38 (0.70-2.72)
Mesothelioma	363	2.43 (0.73-8.05)	0.84 (0.34-2.08)	1.02 (0.37-2.77)
Kaposi Sarcoma	110	2.39 (0.62-9.12)	2.74 (1.12-6.68)	0.74 (0.15-3.75)

[1] The OR for overall cancer risk is adjusted for age, race, selection year, and any condition as categorical variables. The ORs for the site-specific cancer groups like digestive system are adjusted for age, race, and any condition as categorical variables and selection year as continuous variable. The ORs for subtypes of the site-specific cancer groups are adjusted for race as categorical variable and age, selection year, and any condition as continuous variables.

[2] The number of cases may not add up to the totals for cancer specific sites and for cancer overall because miscellaneous cancers were excluded.

[3] Association is significant at $p<0.0004$ (Bonferroni correction for all 123 comparisons, 41 sites for 3 time periods).

[4] CLL = chronic lymphocytic leukemia

Table A.6: Propensity score analysis for associations of blood transfusions with cancer risk, overall and for specific sites and subtypes.

		OR[1] (95% CI)		
	Total[2]	0-12 months	13-30 months	31-48 months
All cancers	552 951	1.84 (1.76-1.92)[3]	0.94 (0.90-0.98)	1.04 (0.99-1.10)
Oral Cavity and Pharynx	8 897	1.21 (0.79-1.85)	1.23 (0.95-1.59)	1.06 (0.78-1.44)
Lip	1 119	0.97 (0.55-1.73)	1.27 (0.82-1.95)	1.09 (0.64-1.86)
Tongue	2 042	1.10 (0.69-1.74)	1.31 (0.95-1.82)	1.14 (0.77-1.69)
Salivary Gland	1 262	1.36 (0.83-2.22)	1.16 (0.77-1.74)	0.73 (0.43-1.22)
Mouth	2 418	1.42 (0.92-2.18)	1.30 (0.95-1.78)	1.14 (0.78-1.66)
Pharynx	2 056	1.09 (0.69-1.74)	1.12 (0.80-1.57)	1.06 (0.70-1.62)
Digestive System	125 386	2.74 (1.99-3.78)[3]	0.92 (0.77-1.11)	1.02 (0.82-1.26)
Esophagus	5 581	1.51 (1.03-2.22)	1.05 (0.82-1.36)	1.23 (0.91-1.65)
Stomach	10 961	3.35 (2.40-4.68)[3]	0.94 (0.75-1.19)	1.19 (0.91-1.56)
Colon	58 480	3.30 (2.47-4.39)[3]	0.96 (0.81-1.14)	0.98 (0.79-1.21)
Rectum	17 608	1.84 (1.31-2.59)	0.69 (0.56-0.87)	0.91 (0.70-1.18)
Anus	1 281	1.59 (0.98-2.57)	0.67 (0.41-1.09)	1.23 (0.78-1.93)
Liver	5 039	2.52 (1.76-3.62)[3]	1.09 (0.84-1.41)	1.22 (0.91-1.66)
Pancreas	17 114	1.92 (1.37-2.69)[3]	0.85 (0.69-1.05)	1.00 (0.78-1.29)
Respiratory System	92 639	1.45 (1.02-2.05)	0.95 (0.79-1.15)	1.08 (0.86-1.34)
Lung	88 275	1.46 (1.09-1.97)	0.94 (0.80-1.10)	1.08 (0.89-1.31)
Bones and Joints	396	1.22 (0.55-2.72)	1.41 (0.74-2.69)	1.08 (0.46-2.55)
Soft Tissue	2 504	1.81 (1.14-2.89)	0.87 (0.60-1.24)	1.05 (0.69-1.59)
Skin excluding Basal and Squamous	16 520	0.97 (0.64-1.48)	0.94 (0.74-1.19)	1.09 (0.82-1.45)
Melanoma	14 259	0.98 (0.68-1.41)	0.93 (0.75-1.15)	1.11 (0.86-1.43)
Breast[4]	64 701	0.89 (0.46-1.74)	0.95 (0.67-1.35)	1.01 (0.65-1.56)
Female Genital System	24 812	1.43 (0.72-2.86)	0.96 (0.65-1.42)	1.05 (0.65-1.70)
Cervix	1 727	1.65 (0.79-3.44)	1.24 (0.76-2.01)	1.21 (0.68-2.17)
Uterus	12 210	1.23 (0.63-2.41)	1.04 (0.71-1.53)	1.06 (0.66-1.71)
Ovary	7 940	1.95 (1.00-3.81)	0.93 (0.62-1.39)	1.02 (0.62-1.67)
Vagina / Vulva	2 221	1.32 (0.64-2.72)	1.06 (0.67-1.67)	1.11 (0.64-1.92)
Male Genital System	91 147	0.93 (0.48-1.78)	0.97 (0.67-1.40)	0.95 (0.61-1.47)
Prostate	90 546	0.93 (0.53-1.64)	0.96 (0.70-1.33)	0.95 (0.64-1.41)
Urinary System	45 343	1.59 (1.10-2.30)	0.93 (0.75-1.15)	1.06 (0.83-1.36)
Urinary Bladder	31 799	1.37 (0.98-1.91)	0.94 (0.78-1.13)	1.06 (0.84-1.33)
Kidney, Renal Pelvis, Ureter	13 166	2.13 (1.51-3.00)[3]	0.89 (0.71-1.11)	1.06 (0.81-1.38)
Eye and Orbit	728	0.71 (0.32-1.55)	0.85 (0.44-1.65)	1.27 (0.68-2.40)
Brain and Other Nervous System	4 824	1.05 (0.66-1.67)	1.06 (0.78-1.44)	0.91 (0.63-1.32)
Endocrine System	3 236	1.20 (0.74-1.92)	0.76 (0.53-1.10)	0.88 (0.58-1.32)
Thyroid	2 871	1.07 (0.69-1.67)	0.74 (0.51-1.07)	0.81 (0.54-1.23)
Lymphoma	29 991	2.16 (1.48-3.14)[3]	0.86 (0.69-1.08)	1.12 (0.86-1.46)
Hodgkin Lymphoma	909	2.91 (1.83-4.63)[3]	1.02 (0.62-1.67)	0.97 (0.56-1.67)
NHL, including CLL[5]	29 082	2.14 (1.55-2.95)[3]	0.86 (0.71-1.04)	1.13 (0.89-1.42)
Myeloma	7 792	3.90 (2.63-5.80)[3]	0.85 (0.64-1.14)	1.10 (0.79-1.54)
Leukemia	8 863	4.84 (3.29-7.12)[3]	0.95 (0.72-1.24)	1.21 (0.88-1.67)
Mesothelioma	1 703	1.68 (1.01-2.78)	0.81 (0.53-1.24)	1.15 (0.72-1.84)
Kaposi Sarcoma	293	2.44 (1.27-4.68)	0.95 (0.47-1.92)	0.48 (0.17-1.39)

[1] The OR for overall cancer risk is adjusted for sex, age, race and selection year as categorical variables and for the estimated propensity scores as continuous variables. The ORs for the site specific cancer groups like digestive system are adjusted for sex, age and race as categorical variables and selection year and the estimated propensity scores as continuous variables. The ORs for subtypes of the site specific cancer groups are adjusted for sex and race as categorical variables and age, selection year and the estimated propensity scores as continuous variables.

[2] The number of cases may not add up to the totals for cancer specific sites and for cancer overall because miscellaneous cancers were excluded.

[3] Association is significant at $p<0.0004$ (Bonferroni correction for all 129 comparisons, 43 sites for 3 time periods).

[4] Women only (537 males are excluded from analysis).

[5] CLL = chronic lymphocytic leukemia

Table A.7: Disease risk score analysis for associations of blood transfusions with cancer risk, overall and for specific sites and subtypes.

		OR[1] (95% CI)		
	Total[2]	0-12 months	13-30 months	31-48 months
All cancers	552 951	1.87 (1.79-1.96)[3]	0.96 (0.91-1.00)	1.04 (0.99-1.10)
Oral Cavity and Pharynx	8 897	1.27 (0.81-1.98)	1.33 (1.04-1.71)	1.16 (0.86-1.57)
Lip	1 119	0.99 (0.55-1.79)	1.27 (0.84-1.92)	1.11 (0.68-1.82)
Tongue	2 042	1.17 (0.73-1.88)	1.46 (1.07-1.99)	1.32 (0.90-1.92)
Salivary Gland	1 262	1.42 (0.86-2.35)	1.27 (0.87-1.86)	0.84 (0.50-1.39)
Mouth	2 418	1.44 (0.93-2.25)	1.30 (0.96-1.77)	1.27 (0.89-1.83)
Pharynx	2 056	1.22 (0.76-1.96)	1.34 (0.97-1.86)	1.13 (0.76-1.70)
Digestive System	125 386	2.89 (2.05-4.06)[3]	1.00 (0.84-1.21)	1.03 (0.83-1.28)
Esophagus	5 581	1.64 (1.11-2.44)	1.22 (0.95-1.56)	1.40 (1.05-1.87)
Stomach	10 961	3.62 (2.55-5.15)[3]	1.08 (0.86-1.36)	1.29 (0.99-1.69)
Colon	58 480	3.45 (2.55-4.68)[3]	1.03 (0.87-1.22)	0.95 (0.77-1.17)
Rectum	17 608	1.85 (1.29-2.65)	0.67 (0.54-0.84)	0.84 (0.64-1.09)
Anus	1 281	1.63 (1.00-2.67)	0.68 (0.42-1.09)	1.31 (0.84-2.05)
Liver	5 039	2.77 (1.89-4.05)[3]	1.34 (1.05-1.72)	1.40 (1.04-1.89)
Pancreas	17 114	2.00 (1.40-2.86)[3]	0.90 (0.73-1.11)	1.05 (0.81-1.35)
Respiratory System	92 639	1.56 (1.08-2.25)	1.09 (0.90-1.31)	1.20 (0.96-1.51)
Lung	88 275	1.58 (1.15-2.16)	1.08 (0.92-1.26)	1.21 (0.99-1.46)
Bones and Joints	396	1.17 (0.53-2.62)	1.21 (0.63-2.32)	1.03 (0.47-2.30)
Soft Tissue	2 504	1.86 (1.15-3.02)	0.89 (0.62-1.28)	0.99 (0.66-1.50)
Skin excluding Basal and Squamous	16 520	0.97 (0.62-1.50)	0.90 (0.71-1.15)	1.06 (0.80-1.40)
Melanoma	14 259	0.96 (0.66-1.41)	0.86 (0.70-1.07)	1.05 (0.81-1.36)
Breast[4]	64 701	0.82 (0.41-1.64)	0.76 (0.53-1.07)	0.92 (0.59-1.43)
Female Genital System	24 812	1.38 (0.68-2.81)	0.84 (.57-1.24)	0.97 (0.60-1.57)
Cervix	1 727	1.50 (0.70-3.19)	1.14 (0.71-1.82)	1.19 (0.67-2.13)
Uterus	12 210	1.04 (0.52-2.07)	0.78 (0.53-1.12)	0.89 (0.56-1.43)
Ovary	7 940	1.76 (0.89-3.50)	0.77 (0.52-1.14)	0.95 (0.58-1.55)
Vagina / Vulva	2 221	1.36 (0.65-2.86)	1.08 (0.70-1.67)	1.26 (0.74-2.17)
Male Genital System	91 147	0.85 (0.41-1.73)	0.74 (0.50-1.09)	0.82 (0.52-1.29)
Prostate	90 546	0.85 (0.46-1.57)	0.74 (0.53-1.03)	0.82 (0.55-1.23)
Urinary System	45 343	1.66 (1.12-2.45)	0.99 (0.80-1.23)	1.11 (0.87-1.43)
Urinary Bladder	31 799	1.43 (1.01-2.02)	0.99 (0.82-1.20)	1.12 (0.89-1.41)
Kidney, Renal Pelvis, Ureter	13 166	2.24 (1.56-3.22)[3]	0.97 (0.78-1.21)	1.10 (0.84-1.42)
Eye and Orbit	728	0.69 (0.31-1.54)	0.77 (0.41-1.43)	1.04 (0.55-1.96)
Brain and Other Nervous System	4 824	1.01 (0.62-1.64)	0.92 (0.68-1.24)	0.85 (0.59-1.22)
Endocrine System	3 236	1.15 (0.70-1.88)	0.66 (0.46-0.95)	0.87 (0.58-1.30)
Thyroid	2 871	1.03 (0.65-1.63)	0.63 (0.44-0.91)	0.82 (0.55-1.23)
Lymphoma	29 991	2.17 (1.46-3.23)[3]	0.85 (0.68-1.08)	1.08 (0.82-1.40)
Hodgkin Lymphoma	909	2.81 (1.75-4.53)[3]	0.93 (0.58-1.49)	0.97 (0.55-1.69)
NHL, including CLL[5]	29 082	2.16 (1.54-3.03)[3]	0.85 (0.70-1.04)	1.08 (0.86-1.36)
Myeloma	7 792	4.19 (2.76-6.35)[3]	0.99 (0.75-1.32)	1.06 (0.76-1.48)
Leukemia	8 863	5.31 (3.54-7.96)[3]	1.22 (0.94-1.59)	1.25 (0.91-1.72)
Mesothelioma	1 703	1.68 (0.99-2.84)	0.79 (0.51-1.21)	1.07 (0.68-1.69)
Kaposi Sarcoma	293	2.93 (1.53-5.62)	1.47 (0.78-2.76)	0.61 (0.21-1.77)

[1] The OR for overall cancer risk is adjusted for sex, age, race and selection year as categorical variables and for the estimated disease risk scores as continuous variables. The ORs for the site specific cancer groups like digestive system are adjusted for sex, age and race as categorical variables and selection year and the estimated disease risk scores as continuous variables. The ORs for subtypes of the site specific cancer groups are adjusted for sex and race as categorical variables and age, selection year and the estimated disease risk scores as continuous variables.

[2] The number of cases may not add up to the totals for cancer specific sites and for cancer overall because miscellaneous cancers were excluded.

[3] Association is significant at $p<0.0004$ (Bonferroni correction for all 129 comparisons, 43 sites for 3 time periods).

[4] Women only (537 males are excluded from analysis).

[5] CLL = chronic lymphocytic leukemia

Table A.8: Sensitivity analysis for significant cancers with and without additional adjustment for specific medical conditions. Odds ratios (ORs) are adjusted for sex, age, race, selection year, any condition and for the 96 conditions separately. The minimal and maximal odds ratios from the 96 logistic regressions are reported.

		0-12 months			13-30 months			31-48 months		
	Total	OR[1]	OR min[2]	OR max[2]	OR[1]	OR min[2]	OR max[2]	OR[1]	OR min[2]	OR max[2]
All cancers	552 951	2.05	2.01	2.07	1.04	0.99	1.06	1.05	1.03	1.07
Oral Cavity and Pharynx	8 897	1.38	1.31	1.37	1.44	1.31	1.44	1.18	1.12	1.18
Tongue	2 042	1.29	1.21	1.29	1.58	1.45	1.61	1.32	1.2	1.37
Pharynx	2 056	1.34	1.27	1.37	1.49	1.28	1.52	1.19	1.12	1.24
Digestive System	125 386	3.14	2.98	3.12	1.08	0.98	1.1	1.04	1.02	1.05
Esophagus	5 581	1.78	1.68	1.77	1.3	1.19	1.32	1.4	1.3	1.41
Stomach	10 961	3.92	3.62	3.95	1.16	0.97	1.19	1.29	1.21	1.32
Colon	58 480	3.73	3.53	3.71	1.11	0.98	1.12	0.95	0.92	0.97
Rectum	17 608	2	1.95	2.04	0.72	0.67	0.75	0.86	0.83	0.9
Liver	5 039	3.29	2.94	3.36	1.72	1.32	1.75	1.65	1.27	1.66
Pancreas	17 114	2.16	2.12	2.18	0.95	0.89	0.94	1.03	0.99	1.03
Respiratory System	92 639	1.71	1.58	1.7	1.18	1.05	1.18	1.24	1.15	1.25
Lung	88 275	1.72	1.6	1.72	1.17	1.04	1.17	1.24	1.15	1.25
Soft Tissue	2 504	1.99	1.92	1.99	0.94	0.87	0.95	0.99	0.93	1.01
Urinary System	45 343	1.79	1.68	1.8	1.05	0.97	1.08	1.09	1.05	1.12
Kidney, Renal Pelvis, Ureter	13 166	2.42	2.19	2.43	1.04	0.91	1.06	1.07	1	1.09
Lymphoma	29 991	2.41	2.26	2.42	0.96	0.84	0.99	1.09	1.03	1.12
Hodgkin Lymphoma	909	3.14	2.79	3.19	1.05	0.84	1.13	0.96	0.88	0.99
NHL, including CLL[3]	29 082	2.38	2.24	2.39	0.95	0.84	0.98	1.09	1.04	1.12
Myeloma	7 792	4.71	4.06	4.81	1.15	0.82	1.17	1.08	0.92	1.07
Leukemia	8 863	6.51	5.34	6.45	1.66	1.14	1.7	1.38	1.11	1.37
Kaposi Sarcoma	293	3.22	2.71	3.16	1.6	1.21	1.67	0.64	0.55	0.68

[1] Odds ratios adjusted for sex, age, race, selection year and any condition.

[2] Odds ratios are adjusted for sex, age, race, selection year, any condition and for the 96 conditions separately. The minimal and maximal odds ratios from the 96 logistic regressions are reported.

[3] CLL = chronic lymphocytic leukemia

Appendix B

R-code

There are two different functions for Poisson and logistic regression models in cohort studies and two different functions for conditional and unconditional analysis in pair matched case-control studies (all functions called 'AnalyticBias_rsolve') which calculate the asymptotic bias as described in Chapter 6. The package 'rootSolve' is needed for the calculations. An explanation of the input parameters for the functions and examples for the function call are given below.

```
#Function:
#################################################################
AnalyticBias_rsolve<-function(
setting,   #can be a vector or a matrix including: intercept of exposure model, intercept of outcome model,
          #confounder prevalence, exposure effect and number of confounders
false_p.Y, #Misspecified outcome models: 1=PS, 2=DRS, 3=Crude
Scale,     #0 for probability based, 1 for logit (log) based
beta_est  #initial values for optimization e.g. c(0,0,0) for PS and DRS and c(0,0) for Crude
)
```

Examples:

```
#Example for cohort study:
#################################################################
alphas=c(0.22,1.1)
betas=c(1.1,0.22)
setting=c(-3.10,-3.36,0.1,1,10)
out.PS=AnalyticBias_rsolve(setting,1,1,c(0,0,0))
out.DRS=AnalyticBias_rsolve(setting,2,1,c(0,0,0))
out.Crude=AnalyticBias_rsolve(setting,3,1,c(0,0))

#Example for matched case-control study:
#################################################################
alphas=c(1.6,1.6)
betas=c(2,-0.6)
setting=c(-3.10,-3.36,0.1,1,2)
#matched analysis
out.PS.m=AnalyticBias_rsolve(setting,1,1,c(0))
out.DRS.m=AnalyticBias_rsolve(setting,2,1,c(0))
#unmatched analysis
out.PS.un=AnalyticBias_rsolve(setting,1,1,c(0,0,0))
out.DRS.un=AnalyticBias_rsolve(setting,2,1,c(0,0,0))
```

B.1 Analytic bias - cohort study

B.1.1 Poisson regression

```
AnalyticBias_rsolve<-function(setting,false_p.Y,Scale,beta_est)
{
alpha0_count=setting[1]
beta0_count=setting[2]
prev=setting[3]
beta_T=setting[4]
n.confounder=setting[5]
p=list(c(prev,1-prev))
l.p=rep(p,n.confounder)
kombis.p=expand.grid(l.p)
X=list(c(1,0))
l.X=rep(X,n.confounder)
kombis.X=expand.grid(l.X)
p.x=apply(kombis.p,1,prod)
T=c(1,0)
alpha_cond=rep(alphas,n.confounder/2)
beta_cond=rep(betas,n.confounder/2)
alpha=c(alpha0_count,alpha_cond)
beta=c(beta0_count,beta_T,beta_cond)

#-----------------------------------------------------------------------
AnalyticBias<-function(beta_est,false_p.Y,Scale)
{
#################################################################
#Exposure model: P(T=1|X) (=True PS)
#################################################################

p.T<-function(z) {
(1+exp(-(alpha0_count+alpha[2:length(alpha)]%*%t(z))))^(-1)
}

#################################################################
#True Outcome model
#################################################################

p.Y_true<-function(t,z){
exp(beta0_count+beta[2]*t+beta[3:length(beta)]%*%t(z))
}

#################################################################
#False Outcome models
#################################################################
#-------------------------
#      PS
#-------------------------
if (false_p.Y==1){
        if (Scale==0){
        ####false outcome model PS on probability scale
                p.Y_estPS<-function(t,z,beta_est){
                exp(beta_est[1]+beta_est[2]*t+beta_est[3]*p.T(z))
                }
        }
        if (Scale==1){
        ####false outcome model PS on logit scale
                p.Y_estPS<-function(t,z,beta_est){
```

```
                exp(beta_est[1]+beta_est[2]*t+beta_est[3]*log(p.T(z)/(1-p.T(z))))
                }
        }

        ##########################################################
        #Solve expectation for outcome Y
        ##########################################################
        Exp.Y<-function(t,z,beta_est) {
        r <- rep(NA, 3)
        lambda=p.Y_true(t,z)
        lambda.star=p.Y_estPS(t,z,beta_est)
        r[1]<-lambda-lambda.star
        r[2]<-t*r[1]
        if (Scale==0){
                r[3]<-p.T(z)*r[1]       #PS on probability scale
        }
        if (Scale==1){
                r[3]<-log(p.T(z)/(1-p.T(z)))*r[1] #PS on logit scale
        }
        r
        }
        ##########################################################
}
#-------------------------
#      DRS
#-------------------------
if (false_p.Y==2){
        ####DRS
        p.DRS<-function(z) {
        exp(beta0_count+beta[3:length(beta)]%*%t(z))
        }
        if (Scale==0){
        ####false outcome model DRS
                p.Y_estDRS<-function(t,z,beta_est){
                exp(beta_est[1]+beta_est[2]*t+beta_est[3]*p.DRS(z))
                }
        }
        if (Scale==1){
        ####false outcome model DRS on log scale
                p.Y_estDRS<-function(t,z,beta_est){
                exp(beta_est[1]+beta_est[2]*t+beta_est[3]*log(p.DRS(z)))
        }
        }

        ##########################################################
        #Solve expectation for outcome Y
        ##########################################################
        Exp.Y<-function(t,z,beta_est) {
        r <- rep(NA, 3)
        lambda=p.Y_true(t,z)
        lambda.star=p.Y_estDRS(t,z,beta_est)
        r[1]<-lambda-lambda.star
        r[2]<-t*r[1]
        if (Scale==0){
                r[3]<-p.DRS(z)*r[1]    #DRS
        }
        if (Scale==1){
                r[3]<-log(p.DRS(z))*r[1]    #DRS on log scale
        }
```

```
        r
        }
        ##############################################################
}
#------------------------
#      Crude
#------------------------
if (false_p.Y==3){
        p.Y_crude<-function(t,beta_est){
        exp(beta_est[1]+beta_est[2]*t)
        }
        ##############################################################
        #Solve expectation for outcome Y
        ##############################################################
        Exp.Y<-function(t,z,beta_est) {
        r <- rep(NA, 2)
        lambda=p.Y_true(t,z)
        lambda.star=p.Y_crude(t,beta_est)
        r[1]<-lambda-lambda.star
        r[2]<-t*r[1]
        r
        }
        ##############################################################
}

##############################################################
#Solve expectation for Exposure T
##############################################################

Exp.T<-function(z,beta_est){
p.T(z)*Exp.Y(T[1],z,beta_est)+(1-p.T(z))*Exp.Y(T[2],z,beta_est)
}

##############################################################
#Solve expectation for confounders Z
##############################################################

Exp.Z<-function(beta_est) {
nn=dim(kombis.X)[1]
A=matrix(NA,nn,length(beta_est))
for (i in 1:nn) {
A[i,]=prod(kombis.p[i,])*Exp.T(kombis.X[i,],beta_est)
}
colSums(A)
}

##############################################################

F.A<-function(beta_est) {
Exp.Z(beta_est)
}
##############################################################
}
#-------------------------------------------------------------
##############################################################
#optimization
##############################################################
solution.1=multiroot(AnalyticBias(beta_est,false_p.Y,Scale), beta_est)$root
}
```

B.1.2 Logistic regression

```
AnalyticBias_rsolve<-function(setting,false_p.Y,Scale,beta_est)
{
alpha0_count=setting[1]
beta0_count=setting[2]
prev=setting[3]
beta_T=setting[4]
n.confounder=setting[5]

p=list(c(prev,1-prev))
l.p=rep(p,n.confounder)
kombis.p=expand.grid(l.p)
X=list(c(1,0))
l.X=rep(X,n.confounder)
kombis.X=expand.grid(l.X)
p.x=apply(kombis.p,1,prod)
T=c(1,0)

alpha_cond=rep(alphas,n.confounder/2)
beta_cond=rep(betas,n.confounder/2)
alpha=c(alpha0_count,alpha_cond)
beta=c(beta0_count,beta_T,beta_cond)

#-------------------------------------------------------------------
AnalyticBias<-function(beta_est,false_p.Y,Scale)
{

#################################################################
#Exposure model: P(T=1|X) (=True PS)
#################################################################

p.T<-function(z) {
(1+exp(-(alpha0_count+alpha[2:length(alpha)]%*%t(z))))^(-1)
}

#################################################################
#True Outcome model
#################################################################

p.Y_true<-function(t,z){
(1+exp(-(beta0_count+beta[2]*t+beta[3:length(beta)]%*%t(z))))^(-1)
}

#################################################################
#False Outcome models
#################################################################
#-------------------------
#      PS
#-------------------------
if (false_p.Y==1){

if (Scale==0){
####false outcome model PS on probability scale
p.Y_estPS<-function(t,z,beta_est){
(1+exp(-(beta_est[1]+beta_est[2]*t+beta_est[3]*p.T(z))))^(-1)
}
}
if (Scale==1){
```

```
####false outcome model PS on logit scale
p.Y_estPS<-function(t,z,beta_est){
(1+exp(-(beta_est[1]+beta_est[2]*t+beta_est[3]*log(p.T(z)/(1-p.T(z))))))^(-1)
}
}

#############################################################
#Solve expectation for outcome Y
#############################################################
Exp.Y<-function(t,z,beta_est) {
r <- rep(NA, 3)
r[1]<-p.Y_estPS(t,z,beta_est)-p.Y_true(t,z)
r[2]<-t*r[1]
if (Scale==0){
r[3]<-p.T(z)*r[1]  #PS on probability scale
}
if (Scale==1){
r[3]<-log(p.T(z)/(1-p.T(z)))*r[1] #PS on logit scale
}
r
}
#############################################################
}
#-------------------------
#      DRS
#-------------------------
if (false_p.Y==2){
####DRS
p.DRS<-function(z) {
(1+exp(-(beta0_count+beta[3:length(beta)]%*%t(z))))^(-1)
}

if (Scale==0){
####false outcome model DRS on probability scale
p.Y_estDRS<-function(t,z,beta_est){
(1+exp(-(beta_est[1]+beta_est[2]*t+beta_est[3]*p.DRS(z))))^(-1)
}
}
if (Scale==1){
####false outcome model DRS on log scale
p.Y_estDRS<-function(t,z,beta_est){
(1+exp(-(beta_est[1]+beta_est[2]*t+beta_est[3]*log(p.DRS(z)/(1-p.DRS(z))))))^(-1)
}
}

#############################################################
#Solve expectation for outcome Y
#############################################################
Exp.Y<-function(t,z,beta_est) {
r <- rep(NA, 3)
r[1]<-p.Y_estDRS(t,z,beta_est)-p.Y_true(t,z)
r[2]<-t*r[1]
if (Scale==0){
r[3]<-p.DRS(z)*r[1]    #DRS on probability scale
}
if (Scale==1){
r[3]<-log(p.DRS(z)/(1-p.DRS(z)))*r[1]  #DRS on logit scale
}
r
```

```
}
##########################################################
}
#-------------------------
#      Crude
#-------------------------
if (false_p.Y==3){
p.Y_crude<-function(t,beta_est){
(1+exp(-(beta_est[1]+beta_est[2]*t)))^(-1)
}

##########################################################
#Solve expectation for outcome Y
##########################################################
Exp.Y<-function(t,z,beta_est) {
r <- rep(NA, 2)
r[1]<-p.Y_crude(t,beta_est)-p.Y_true(t,z)
r[2]<-t*r[1]
r
}
##########################################################
}

##########################################################
#Solve expectation for Exposure T
##########################################################

Exp.T<-function(z,beta_est) {
p.T(z)*Exp.Y(T[1],z,beta_est)+(1-p.T(z))*Exp.Y(T[2],z,beta_est)
}

##########################################################
#Solve expectation for confounders Z
##########################################################
Exp.Z<-function(beta_est) {
nn=dim(kombis.X)[1]
A=matrix(NA,nn,length(beta_est))
for (i in 1:nn) {
A[i,]=prod(kombis.p[i,])*Exp.T(kombis.X[i,],beta_est)
}
colSums(A)
}

##########################################################

F.A<-function(beta_est) {
Exp.Z(beta_est)
}

##########################################################
}
#-----------------------------------------------------------

#Optimierung
solution.1=multiroot(AnalyticBias(beta_est,false_p.Y,Scale), beta_est)$root
}
```

B.2 Analytic bias - matched case-control study

B.2.1 Matched analysis, conditional logistic regression

```
AnalyticBias_rsolve<-function(setting,false_p.Y,Scale,beta_est)
{
alpha_0=setting[1]
beta_0=setting[2]
prev=setting[3]
beta_T=setting[4]
n.confounder=setting[5]

p=list(c(prev,1-prev))
l.p=rep(p,n.confounder)
kombis.p=expand.grid(l.p)
X=list(c(1,0))
l.X=rep(X,n.confounder)
kombis.X=expand.grid(l.X)

alpha_cond=rep(alphas,n.confounder/2)
beta_cond=rep(betas,n.confounder/2)
alpha=c(alpha_0,alpha_cond)
beta=c(beta_0,beta_T,beta_cond)

#-----------------------------------------------------------------------
AnalyticBias<-function(beta_est,false_p.Y,Scale) {

#################################################################
#Exposure model: P(T=1|X) (=True PS)
#################################################################

p.T<-function(z) {
(1+exp(-(alpha[1]+alpha[2:length(alpha)]%*%t(z))))^(-1)
}

#################################################################
#Solve expectation for outcome Y
#################################################################

Exp.Y<-function(ti,tj,zi,zj,beta_est) {
(ti- (ti*exp(ti*beta_est) + tj*exp(tj*beta_est)) / (exp(ti*beta_est) + exp(tj*beta_est)))*
(1+exp(-beta[1] - beta[2]*ti-beta[3:length(beta)]%*%t(zi)))^(-1)*
(1+exp(beta[1] + beta[2]*tj+beta[3:length(beta)]%*%t(zj)))^(-1)

+(tj- (ti*exp(ti*beta_est) + tj*exp(tj*beta_est)) / (exp(ti*beta_est) + exp(tj*beta_est)))*
(1+exp(-beta[1] - beta[2]*tj-beta[3:length(beta)]%*%t(zj)))^(-1)*
(1+exp(beta[1] + beta[2]*ti+beta[3:length(beta)]%*%t(zi)))^(-1)
}

#################################################################
#Solve expectation for exposure T
#################################################################
#T=(1,0) #Ti=1 and Tj=0 or Ti=0 and Tj=1; else 0

Exp.T<-function(zi,zj,beta_est) {
p.T(zi)*(1-p.T(zj))*Exp.Y(1,0,zi,zj,beta_est)+
p.T(zi)*p.T(zj)*Exp.Y(1,1,zi,zj,beta_est)+
(1-p.T(zi))*(1-p.T(zj))*Exp.Y(0,0,zi,zj,beta_est)+
(1-p.T(zi))*(p.T(zj))*Exp.Y(0,1,zi,zj,beta_est)
```

```
}
#############################################################################
#--------------------------
#      PS
#--------------------------
if (false_p.Y==1){

####Generate all possible matching combinations: kombis.XPS
if (Scale==0){
p.T_all<-(1+exp(-(alpha[1]+alpha[2:length(alpha)]%*%t(kombis.X))))^(-1)
kombis.XPS=data.frame(kombis.X,as.vector(p.T_all))
}

if (Scale==1){
p.T_all<-alpha[1]+alpha[2:length(alpha)]%*%t(kombis.X)
kombis.XPS=data.frame(kombis.X,as.vector(p.T_all))
}

#############################################################################
#Solve expectation for confounders Z
#############################################################################

Exp.Z<-function(beta_est) {
i=1
A=matrix(NA,dim(kombis.XPS)[1],1)
for (i in 1:dim(kombis.XPS)[1]) {
        zi=kombis.X[i,1:dim(kombis.X)[2]]
        pzi=prod(kombis.p[i,])

        j=1
        B=matrix(NA,dim(kombis.XPS)[1],1)
        g.zi=matrix(NA,dim(kombis.XPS)[1],1)
        for (j in 1:dim(kombis.XPS)[1]) {
                zj=kombis.X[j,1:dim(kombis.X)[2]]
                pzj=prod(kombis.p[j,])

                #Indicator: equal scores I=1 else I=0
                if (kombis.XPS[i,dim(kombis.XPS)[2]]==kombis.XPS[j,dim(kombis.XPS)[2]]){
                        g.zi[j,]=(1+exp(beta[1] + beta[2]*1+beta[3:length(beta)]%*%t(zj)))^(-1)*p.T(zj)*pzj +
                        (1+exp(beta[1] + beta[2]*0+beta[3:length(beta)]%*%t(zj)))^(-1)*(1-p.T(zj))*pzj
                        B[j,]=pzi*pzj*Exp.T(zi,zj,beta_est)
                }
                if (kombis.XPS[i,dim(kombis.XPS)[2]]!=kombis.XPS[j,dim(kombis.XPS)[2]]){
                        g.zi[j,]=0
                        B[j,]=0
                }
        }
        A[i,]=colSums(B)/colSums(g.zi)
}
colSums(A)
}
}
#--------------------------
#      DRS
#--------------------------
if (false_p.Y==2){

#############Generate all possible matching combinations: kombis.XDRS

```

```
if (Scale==0){
        p.DRS_all<-(1+exp(-(beta[1]+beta[3:length(beta)]%*%t(kombis.X))))^(-1)
        kombis.XDRS=data.frame(kombis.X,as.vector(p.DRS_all))
}
if (Scale==1){
        p.DRS_all<-beta[1]+beta[3:length(beta)]%*%t(kombis.X)
        kombis.XDRS=data.frame(kombis.X,as.vector(p.DRS_all))
}

#########################################################
#Solve expectation for confounders Z
#########################################################

Exp.Z<-function(beta_est) {
i=1
A=matrix(NA,dim(kombis.XDRS)[1],1)
for (i in 1:dim(kombis.XDRS)[1]) {
        zi=kombis.X[i,1:dim(kombis.X)[2]]
        pzi=prod(kombis.p[i,])

        j=1
        B=matrix(NA,dim(kombis.XDRS)[1],1)
        g.zi=matrix(NA,dim(kombis.XDRS)[1],1)
        for (j in 1:dim(kombis.XDRS)[1]) {
                zj=kombis.X[j,1:dim(kombis.X)[2]]
                pzj=prod(kombis.p[j,])

                #Indicator: equal scores I=1 else I=0
                if (kombis.XDRS[i,dim(kombis.XDRS)[2]]==kombis.XDRS[j,dim(kombis.XDRS)[2]]){
                        g.zi[j,]=(1+exp(beta[1] + beta[2]*1+beta[3:length(beta)]%*%t(zj)))^(-1)*p.T(zj)*pzj +
                        (1+exp(beta[1] + beta[2]*0+beta[3:length(beta)]%*%t(zj)))^(-1)*(1-p.T(zj))*pzj
                        B[j,]=pzi*pzj*Exp.T(zi,zj,beta_est)
                }
                if (kombis.XDRS[i,dim(kombis.XDRS)[2]]!=kombis.XDRS[j,dim(kombis.XDRS)[2]]){
                        g.zi[j,]=0
                        B[j,]=0
                }
        }
        A[i,]=colSums(B)/colSums(g.zi)
}
colSums(A)
}
}

#########################################################

F.A<-function(beta_est) {
Exp.Z(beta_est)
}

#########################################################
}
#-----------------------------------------------------------

#########################################################
#optimization
#########################################################
solution.1=multiroot(AnalyticBias(beta_est,false_p.Y,Scale), beta_est)$root
}
```

B.2.2 Unmatched analysis, unconditional logistic regression

```
1
2
3  AnalyticBias_rsolve<-function(setting,false_p.Y,Scale,beta_est)
4  {
5  alpha_0=setting[1]
6  beta_0=setting[2]
7  prev=setting[3]
8  beta_T=setting[4]
9  n.confounder=setting[5]
10
11 p=list(c(prev,1-prev))
12 l.p=rep(p,n.confounder)
13 kombis.p=expand.grid(l.p)
14 X=list(c(1,0))
15 l.X=rep(X,n.confounder)
16 kombis.X=expand.grid(l.X)
17
18 alpha_cond=rep(alphas,n.confounder/2)
19 beta_cond=rep(betas,n.confounder/2)
20 alpha=c(alpha_0,alpha_cond)
21 beta=c(beta_0,beta_T,beta_cond)
22
23 #-------------------------------------------------------------------
24 AnalyticBias<-function(beta_est,false_p.Y,Scale) {
25
26 ###################################################################
27 #Exposure model: P(T=1|X) (=True PS)
28 ###################################################################
29
30 p.T<-function(z) {
31 (1+exp(-(alpha[1]+alpha[2:length(alpha)]%*%t(z))))^(-1)
32 }
33
34 ###################################################################
35 #True Outcome model
36 ###################################################################
37
38 p.Y_true<-function(t,z){
39 (1+exp(-(beta[1]+beta[2]*t+beta[3:length(beta)]%*%t(z))))^(-1)
40 }
41
42 ###################################################################
43 #False Outcome models
44 ###################################################################
45 #-------------------------
46 #      PS
47 #-------------------------
48 if (false_p.Y==1){
49
50 if (Scale==0){
51 ####false outcome model PS on probability scale
52 p.Y_estPS<-function(t,z,beta_est){
53 (1+exp(-(beta_est[1]+beta_est[2]*t+beta_est[3]*p.T(z))))^(-1)
54 }
55 }
56 if (Scale==1){
57 ####false outcome model PS on logit scale
```

```
p.Y_estPS<-function(t,z,beta_est){
(1+exp(-(beta_est[1]+beta_est[2]*t+beta_est[3]*log(p.T(z)/(1-p.T(z))))))^(-1)
}
}

#############################################################
#Solve expectation for outcome Y
#############################################################
Exp.Y<-function(ti,tj,zi,zj,beta_est) {

###P(Yi=1,Yj=0) and P(Yi=0,Yj=1)
p.Yi1.Yj0=p.Y_true(ti,zi)*(1-p.Y_true(tj,zj))
p.Yi0.Yj1=p.Y_true(tj,zj)*(1-p.Y_true(ti,zi))

###False outcome models for Yi and Yj
p.Yi_false=p.Y_estPS(ti,zi,beta_est)
p.Yj_false=p.Y_estPS(tj,zj,beta_est)

#--------------------------
r <- rep(NA, 3)
r[1]<-(1-p.Yi_false-p.Yj_false)*p.Yi1.Yj0 + (1-p.Yi_false-p.Yj_false)*p.Yi0.Yj1
r[2]<-(ti-ti*p.Yi_false-tj*p.Yj_false)*p.Yi1.Yj0 + (tj-ti*p.Yi_false-tj*p.Yj_false)*p.Yi0.Yj1

if (Scale==0){  #PS on probability scale
r[3]<-(p.T(zi)-p.T(zi)*p.Yi_false-p.T(zj)*p.Yj_false)*p.Yi1.Yj0 +
(p.T(zj)-p.T(zi)*p.Yi_false-p.T(zj)*p.Yj_false)*p.Yi0.Yj1
}

if (Scale==1){   #PS on logit scale
lPS1=log(p.T(zi)/(1-p.T(zi)))
lPS2=log(p.T(zj)/(1-p.T(zj)))
r[3]<-(lPS1-lPS1*p.Yi_false-lPS2*p.Yj_false)*p.Yi1.Yj0 + (lPS2-lPS1*p.Yi_false-lPS2*p.Yj_false)*p.Yi0.Yj1
}
r
#--------------------------
}
#############################################################
}
#--------------------------
#      DRS
#--------------------------
if (false_p.Y==2){

####DRS
p.DRS<-function(z) {
(1+exp(-(beta[1]+beta[3:length(beta)]%*%t(z))))^(-1)
}

if (Scale==0){
####false outcome model DRS on probability scale
p.Y_estDRS<-function(t,z,beta_est){
(1+exp(-(beta_est[1]+beta_est[2]*t+beta_est[3]*p.DRS(z))))^(-1)
}
}

if (Scale==1){
####false outcome model DRS on log scale
p.Y_estDRS<-function(t,z,beta_est){
(1+exp(-(beta_est[1]+beta_est[2]*t+beta_est[3]*log(p.DRS(z)/(1-p.DRS(z))))))^(-1)
```

```
}
}


#############################################################
#Solve expectation for outcome Y
#############################################################

Exp.Y<-function(ti,tj,zi,zj,beta_est) {

###P(Yi=1,Yj=0) and P(Yi=0,Yj=1)
p.Yi1.Yj0=p.Y_true(ti,zi)*(1-p.Y_true(tj,zj))
p.Yi0.Yj1=p.Y_true(tj,zj)*(1-p.Y_true(ti,zi))

###False outcome models for Yi and Yj
p.Yi_false=p.Y_estDRS(ti,zi,beta_est)
p.Yj_false=p.Y_estDRS(tj,zj,beta_est)

#--------------------------
r <- rep(NA, 3)
r[1]<-(1-p.Yi_false-p.Yj_false)*p.Yi1.Yj0 + (1-p.Yi_false-p.Yj_false)*p.Yi0.Yj1
r[2]<-(ti-ti*p.Yi_false-tj*p.Yj_false)*p.Yi1.Yj0 + (tj-ti*p.Yi_false-tj*p.Yj_false)*p.Yi0.Yj1

if (Scale==0){  #DRS on probability scale
r[3]<-(p.DRS(zi,beta)-p.DRS(zi)*p.Yi_false-p.DRS(zj)*p.Yj_false)*p.Yi1.Yj0+
(p.DRS(zj,beta)-p.DRS(zi)*p.Yi_false-p.DRS(zj)*p.Yj_false)*p.Yi0.Yj1
}

if (Scale==1){   #DRS on logit scale
lDRS1=log(p.DRS(zi)/(1-p.DRS(zi)))
lDRS2=log(p.DRS(zj)/(1-p.DRS(zj)))
r[3]<-(lDRS1-lDRS1*p.Yi_false-lDRS2*p.Yj_false)*p.Yi1.Yj0+
(lDRS2-lDRS1*p.Yi_false-lDRS2*p.Yj_false)*p.Yi0.Yj1
}
r
#--------------------------
}
#############################################################
}


#############################################################################
#Solve expectation for Exposure T
#############################################################################
#T=(1,0) #Ti=1 and Tj=0 or Ti=0 and Tj=1; else 0

Exp.T<-function(zi,zj,beta_est) {
p.T(zi)*(1-p.T(zj))*Exp.Y(1,0,zi,zj,beta_est)+
p.T(zi)*p.T(zj)*Exp.Y(1,1,zi,zj,beta_est)+
(1-p.T(zi))*(1-p.T(zj))*Exp.Y(0,0,zi,zj,beta_est)+
(1-p.T(zi))*(p.T(zj))*Exp.Y(0,1,zi,zj,beta_est)
}

#############################################################################
#--------------------------
#      PS
#--------------------------
if (false_p.Y==1){

```

```
####Generate all possible matching combinations: kombis.XPS
if (Scale==0){
p.T_all<-(1+exp(-(alpha[1]+alpha[2:length(alpha)]%*%t(kombis.X))))^(-1)
kombis.XPS=data.frame(kombis.X,as.vector(p.T_all))
}

if (Scale==1){
p.T_all<-alpha[1]+alpha[2:length(alpha)]%*%t(kombis.X)
kombis.XPS=data.frame(kombis.X,as.vector(p.T_all))
}

#########################################################
#Solve expectation for confounders Z
#########################################################

Exp.Z<-function(beta_est){
i=1
nn=length(beta_est)
A=matrix(NA,dim(kombis.XPS)[1],nn)
for (i in 1:dim(kombis.XPS)[1]) {
        zi=kombis.X[i,1:dim(kombis.X)[2]]
        pzi=prod(kombis.p[i,])

        j=1
        B=matrix(NA,dim(kombis.XPS)[1],nn)
        g.zi=matrix(NA,dim(kombis.XPS)[1],1)
        for (j in 1:dim(kombis.XPS)[1]) {
                zj=kombis.X[j,1:dim(kombis.X)[2]]
                pzj=prod(kombis.p[j,])

                #Indicator: equal scores I=1 else I=0
                if (kombis.XPS[i,dim(kombis.XPS)[2]]==kombis.XPS[j,dim(kombis.XPS)[2]]){
                        g.zi[j,]=(1+exp(beta[1] + beta[2]*1+beta[3:length(beta)]%*%t(zj)))^(-1)*p.T(zj)*pzj +
                        (1+exp(beta[1] + beta[2]*0+beta[3:length(beta)]%*%t(zj)))^(-1)*(1-p.T(zj))*pzj
                        B[j,]=pzi*pzj*Exp.T(zi,zj,beta_est)
                }
                if (kombis.XPS[i,dim(kombis.XPS)[2]]!=kombis.XPS[j,dim(kombis.XPS)[2]]){
                        g.zi[j,]=0
                        B[j,]=0
                }
        }
        A[i,]=colSums(B)/colSums(g.zi)
}
colSums(A)
}
}
#--------------------------
#      DRS
#--------------------------
if (false_p.Y==2){

####Generate all possible matching combinations: kombis.XDRS

if (Scale==0){
        p.DRS_all<-(1+exp(-(beta[1]+beta[3:length(beta)]%*%t(kombis.X))))^(-1)
        kombis.XDRS=data.frame(kombis.X,as.vector(p.DRS_all))
}
if (Scale==1){
        p.DRS_all<-beta[1]+beta[3:length(beta)]%*%t(kombis.X)
```

```
235         kombis.XDRS=data.frame(kombis.X,as.vector(p.DRS_all))
236 }
237
238 ##########################################################
239 #Solve expectation for confounders Z
240 ##########################################################
241
242 Exp.Z<-function(beta_est) {
243 i=1
244 nn=length(beta_est)
245 A=matrix(NA,dim(kombis.XDRS)[1],nn)
246 for (i in 1:dim(kombis.XDRS)[1]) {
247         zi=kombis.X[i,1:dim(kombis.X)[2]]
248         pzi=prod(kombis.p[i,])
249
250         j=1
251         B=matrix(NA,dim(kombis.XDRS)[1],nn)
252         g.zi=matrix(NA,dim(kombis.XDRS)[1],1)
253         for (j in 1:dim(kombis.XDRS)[1]) {
254                 zj=kombis.X[j,1:dim(kombis.X)[2]]
255                 pzj=prod(kombis.p[j,])
256
257                 #Indicator: equal scores I=1 else I=0
258                 if (kombis.XDRS[i,dim(kombis.XDRS)[2]]==kombis.XDRS[j,dim(kombis.XDRS)[2]]){
259                         g.zi[j,]=(1+exp(beta[1] + beta[2]*1+beta[3:length(beta)]%*%t(zj)))^(-1)*p.T(zj)*pzj +
260                         (1+exp(beta[1] + beta[2]*0+beta[3:length(beta)]%*%t(zj)))^(-1)*(1-p.T(zj))*pzj
261                         B[j,]=pzi*pzj*Exp.T(zi,zj,beta_est)
262                 }
263                 if (kombis.XDRS[i,dim(kombis.XDRS)[2]]!=kombis.XDRS[j,dim(kombis.XDRS)[2]]){
264                         g.zi[j,]=0
265                         B[j,]=0
266                 }
267         }
268         A[i,]=colSums(B)/colSums(g.zi)
269 }
270 colSums(A)
271 }
272 }
273
274 ##########################################################
275
276 F.A<-function(beta_est) {
277 Exp.Z(beta_est)
278 }
279
280 ##########################################################
281 }
282 #---------------------------------------------------------
283
284 ##########################################################
285 #optimization
286 ##########################################################
287
288 solution.1=multiroot(AnalyticBias(beta_est,false_p.Y,Scale), beta_est)$root
289 }
```

Bibliography

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.

Alarcón, G. S., McGwin, G., Bertoli, A. M., Fessler, B. J., Calvo-Alén, J., Bastian, H. M., Vilà, L. M., and Reveille, J. D. (2007). Effect of hydroxychloroquine on the survival of patients with systemic lupus erythematosus: Data from lumina, a multiethnic us cohort (lumina l). *Annals of the Rheumatic Diseases*, 66(9):1168–1172.

Allen, A. S. and Satten, G. A. (2011). Control for confounding in case-control studies using the stratification score, a retrospective balancing score. *American Journal of Epidemiology*, 173(7):752–760.

Anderson, S. A., Menis, M., O'Connell, K., and Burwen, D. R. (2007). Blood use by inpatient elderly population in the united states. *Transfusion*, 47(4):582–592.

Arbogast, P. G., Kaltenbach, L., Ding, H., and Ray, W. A. (2008). Adjustment for multiple cardiovascular risk factors using a summary risk score. *Epidemiology*, 19(1):30–37.

Arbogast, P. G. and Ray, W. A. (2009). Use of disease risk scores in pharmacoepidemiologic studies. *Statistical Methods in Medical Research*, 18(1):67–80.

Arbogast, P. G. and Ray, W. A. (2011). Performance of disease risk scores, propensity scores, and traditional multivariable outcome regression in the presence of multiple confounders. *American Journal of Epidemiology*, 174(5):613–620.

Arbogast, P. G. and Seeger, J. D. (2012). Summary variables in observational research: Propensity scores and disease risk scores. *Effective Health Care Program Research Report*, No. 33. (Prepared by DEcIDE Methods Center under Contract No. HHSA 290-2005-0016-I, Task Order 10.) AHRQ Publication No. 11(12) -EHC055-EF. Rockville, MD: Agency for Healthcare Research and Quality. Available at: http://effectivehealthcare.ahrq.gov/reports/final.cfm (accessed 13.12.2012).

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

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

Austin, P. C. (2010). Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *American Journal of Epidemiology*, 172(9):1092–1097.

Austin, P. C. (2011a). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424.

Austin, P. C. (2011b). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, 10(2):150–161.

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, 33(6):1057–1069.

Austin, P. C., Grootendorst, P., and Anderson, G. M. (2007a). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A monte carlo study. *Statistics in Medicine*, 26(4):734–753.

Austin, P. C., Grootendorst, P., Normand, S. L., and Anderson, G. M. (2007b). Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A monte carlo study. *Statistics in Medicine*, 26(4):754–768.

Barlow, W. E., Ichikawa, L., Rosner, D., and Izumi, S. (1999). Analysis of case-cohort designs. *Journal of Clinical Epidemiology*, 52(12):1165–1172.

Braitman, L. E. and Rosenbaum, P. R. (2002). Rare outcomes, common treatments: analytic strategies using propensity scores. *Annals of Internal Medicine*, 137(8):693–695.

Breslow, N. E. (2005). Case-control studies. In Ahrens, W. and Pigeot, I., editors, *Handbook of Epidemiology*, pages 287–319. Berlin: Springer-Verlag.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research. Volume I - The Analysis of Case-Control Studies*. Lyon: International Agency for Research on Cancer (IARC Scientific Publications No. 32).

Breslow, N. E. and Day, N. E. (1987). *Statistical Methods in Cancer Research. Volume II - The Design and Analysis of Cohort Studies*. Lyon: International Agency for Research on Cancer (IARC Scientific Publications No. 82).

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

Brookhart, M. A., Stürmer, T., Glynn, R. J., Rassen, J., and Schneeweiss, S. (2010). Confounding control in healthcare database research: challenges and potential approaches. *Medical Care*, 48(6 Suppl):114–120.

Bross, M. H., Soch, K., and Smith-Knuppel, T. (2010). Anemia in older persons. *American Family Physician*, 82(5):480–487.

Buddeberg, F., Schimmer, B. B., and Spahn, D. R. (2008). Transfusion-transmissible infections and transfusion-related immunomodulation. *Best Practice & Research Clinical Anaesthesiology*, 22(3):503–517.

Cadarette, S. M., Gagne, J. J., Solomon, D. H., Katz, J. N., and Stürmer, T. (2010). Confounder summary scores when comparing the effects of multiple drug exposures. *Pharmacoepidemiology and Drug Safety*, 19(1):2–9.

Caliendo, M. and Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1):31–72.

Castillo, J. J., Dalia, S., and Pascual, S. K. (2010). Association between red blood cell transfusions and development of non-hodgkin lymphoma: A meta-analysis of observational studies. *Blood*, 116(16):2897–2907.

Cazzola, M. and Malcovati, L. (2005). Myelodysplastic syndromes - coping with ineffective hematopoiesis. *New England Journal of Medicine*, 352(6):536–538.

Cepeda, M. S., Boston, R., Farrar, J. T., and Strom, B. L. (2003). Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology*, 158(3):280–287.

Chang, C. M., Quinlan, S. C., Warren, J. L., and Engels, E. A. (2010). Blood transfusions and the subsequent risk of hematologic malignancies. *Transfusion*, 50(10):2249–2257.

Cook, E. F. and Goldman, L. (1989). Performance of tests of significance based on stratification by a multivariate confounder score or by a propensity score. *Journal of Clinical Epidemiology*, 42(4):317–324.

Cornfield, J. (1951). A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute*, 11(6):1269–1275.

Costanza, M. C. (1995). Matching. *Preventive Medicine*, 24(5):425–433.

Coussens, L. M. and Werb, Z. (2002). Inflammation and cancer. *Nature*, 420(6917):860–867.

D'Agostino, Jr., R. B. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17(19):2265–2281.

Davenport, J. (1996). Macrocytic anemia. *American Family Physician*, 53(1):155–162.

De Martel, C., Ferlay, J., Franceschi, S., Vignat, J., Bray, F., Forman, D., and Plummer, M. (2012). Global burden of cancers attributable to infections in 2008: A review and synthetic analysis. *The Lancet Oncology*, 13(6):607–615.

Dehejia, R. H. and Wahba, S. (2002). Propensity score matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84(1):151–161.

Dwyre, D. M., Fernando, L. P., and Holland, P. V. (2011). Hepatitis b, hepatitis c and hiv transfusion-transmitted infections in the 21st century. *Vox Sanguinis*, 100(1):92–98.

Edgren, G., Bagnardi, V., Bellocco, R., Hjalgrim, H., Rostgaard, K., Melbye, M., Reilly, M., Adami, H. O., Hall, P., and Nyrén, O. (2010). Pattern of declining hemoglobin concentration before cancer diagnosis. *International Journal of Cancer*, 127(6):1429–1436.

El-Serag, H. B., Johnson, M. L., Hachem, C., and Morgana, R. O. (2009). Statins are associated with a reduced risk of hepatocellular carcinoma in a large cohort of patients with diabetes. *Gastroenterology*, 136(5):1601–1608.

Engels, E. A., Pfeiffer, R. M., Ricker, W., Wheeler, W., Parsons, R., and Warren, J. L. (2011). Use of surveillance, epidemiology, and end results-medicare data to conduct case-control studies of cancer among the us elderly. *American Journal of Epidemiology*, 174(7):860–870.

Etminan, M., Samii, A., and Brophy, J. M. (2010). Statin use and risk of epilepsy: A nested case-control study. *Neurology*, 75(17):1496–1500.

Franklin, J. M., Rassen, J. A., Ackermann, D., Bartels, D. B., and Schneeweiss, S. (2014). Metrics for covariate balance in cohort studies of causal effects. *Statistics in Medicine*, 33(10):1685–1699.

Fritz, A., Percy, C., Jack, A., Shanmugarathnam, K., Sobin, L., Parkin, D. M., and Whelan, S. (2000). *International Classification of Diseases for Oncology. 3rd ed.* Geneva: World Health Organization 2000.

Gail, M. H., Wieand, S., and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71(3):431–444.

Glynn, R. J., Gagne, J. J., and Schneeweiss, S. (2012). Role of disease risk scores in comparative effectiveness research with emerging therapies. *Pharmacoepidemiology*, 21(S2):138–147.

Glynn, R. J., Schneeweiss, S., and Stürmer, T. (2006). Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic & Clinical Pharmacology & Toxicology*, 98(3):253–259.

Grulich, A. E., van Leeuwen, M. T., Falster, M. O., and Vajdic, C. M. (2007). Incidence of cancers in people with hiv/aids compared with immunosuppressed transplant recipients: a meta-analysis. *Lancet*, 370(9581):59–67.

Gu, X. S. and Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4):405–420.

Hade, E. M. and Lu, B. (2014). Bias associated with using the estimated propensity score as a regression covariate. *Statistics in Medicine*, 33(1):74–87.

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488.

Heinze, G. and Jüni, P. (2011). An overview of the objectives of and the approaches to propensity score analyses. *European Heart Journal*, 32(14):1704–1708.

Hernán, M. and Robins, J. (2015). *Causal inference: Part I, Chapters 1-10 (updated 14 May 2014)*. Chapman & Hall/CRC. Available at: http://www.hsph.harvard.edu/miguel-hernan/causal-inference-book (accessed 24.06.2014).

Hirano, K. and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2(3-4):259–278.

Hjalgrim, H., Edgren, G., Rostgaard, K., Reilly, M., Tran, T. N., Titlestad, K. E., Shanwell, A., Jersild, C., Adami, J., Wikman, A., Gridley, G., Wideroff, L., Nyrén, O., and Melbye, M. (2007). Cancer incidence in blood transfusion recipients. *Journal of the National Cancer Institute*, 99(24):1864–1874.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236.

Holland, P. (1986). Statistics and causal inference. *American Statistical Association*, 81(396):945–960.

Howell, M. D., Novack, V., Grgurich, P., Soulliard, D., Novack, L., Pencina, M., and Talmor, D. (2010). Iatrogenic gastric acid suppression and the risk of nosocomial clostridium difficile infection. *Archives of Internal Medicine*, 170(9):784–790.

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 171(2):481–502.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

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

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86.

Jadad, A. (1998). *Randomized controlled trials: a user's guide*. London: BMJ Books.

Joffe, M. M. and Rosenbaum, P. R. (1999). Invited commentary: propensity scores. *American Journal of Epidemiology*, 150(4):327–333.

Kass, P. H. and Gold, E. B. (1999). Modern epidemiologic study designs. In Ahrens, W. and Pigeot, I., editors, *Handbook of Epidemiology*, pages 321–344. Berlin: Springer-Verlag.

Kleinbaum, D. G., Kupper, L. L., and Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. New York: Van Nostrand Reinhold.

Langholz, B. (2005). Case-control study, nested. In Armitage, P. and Colton, T., editors, *Encyclopedia of Biostatistics, 2nd Edition.*, pages 646–655. Chichester: Wiley & Sons.

Leacy, F. P. and Stuart, E. A. (2013). On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Statistics in Medicine*. aop.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23(19):2937–2960.

Lunt, M. (2014). Selecting an appropriate caliper can be essential for achieving good balance with propensity score matching. *American Journal of Epidemiology*, 179(2):226–235.

Mantel, N. (1973). Synthetic retrospective studies and related topics. *Biometrics*, 29(3):479–486.

Martens, E. P., Pestman, W. R., de Boer, A., Belitser, S. V., and Klungel, O. H. (2008). Systematic differences in treatment effect estimates between propensity score methods and logistic regression. *International Journal of Epidemiology*, 37(5):1142–1147.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models (2nd ed)*. London: Chapman and Hall.

Michalia, M., Kompoti, M., Panagiotakopoulou, A., Kallitsi, G., Charitidi, M., Trikka-Graphakos, E., and Clouva-Molyvdas, P. M. (2012). Impact of red blood cells transfusion on icu-acquired bloodstream infections: A case-control study. *Journal of Critical Care*, 27(6):655–661.

Miettinen, O. (1976a). Estimability and estimation in case-referent studies. *American Journal of Epidemiology*, 103(2):226–235.

Miettinen, O. S. (1976b). Stratification by a multivariate confounder score. *American Journal of Epidemiology*, 104(6):609–620.

Ming, K. and Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics*, 56(1):118–124.

Månsson, R., Joffe, M. M., Sun, W., and Hennessy, S. (2007). On the estimation and use of propensity scores in case-control and case-cohort studies. *American Journal of Epidemiology*, 166(3):332–339.

Modén, B., Merlo, J., Ohlsson, H., and Rosvall, M. (2010). Psychotropic drugs and falling accidents among the elderly: A nested case control study in the whole population of scania, sweden. *Journal of Epidemiology and Community Health*, 64(5):440–446.

Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472.

Patorno, E., Grotta, A., Bellocco, R., and Schneeweiss, S. (2013). Propensity score methodology for confounding control in health care utilization databases. *Epidemiology Biostatistics and Public Health*, 10(3):e8940.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., and Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379.

Pfeiffer, R. M. and Riedl, R. (2014). Bias for score adjusted analysis. Manuscript submitted for publication.

Pike, M. C., Anderson, J., and Day, N. (1979). Some insights into miettinen's multivariate confounder score approach to case-control study analysis. *Journal of Epidemiology and Community Health*, 33(1):104–106.

Pocock, S. J. (1983). *Clinical Trials: A practical approach*. Chichester: Wiley & Sons.

Potosky, A. L., Riley, G. F., Lubitz, J. D., Mentnech, R. M., and Kessler, L. G. (1993). Potential for cancer related health services research using a linked medicare-tumor registry database. *Medical Care*, 31(8):732–748.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73(1):1–11.

Rassen, J. A., Shelat, A. A., Myers, J.and Glynn, R. J., Rothman, K. J., and Schneeweiss, S. (2012). One-to-many propensity score matching in cohort studies. *Pharmacoepidemiology and Drug Safety*, 21(Suppl 2):69–80.

Riedl, R., Engels, E. A., Warren, J. L., Berghold, A., Ricker, W., and Pfeiffer, R. M. (2013). Blood transfusions and the subsequent risk of cancers in the united states elderly. *Transfusion*, 53(10):2198–2206.

Robins, J. M., Gail, M. H., and Lubin, J. H. (1986). More on "biased selection of controls for case-control analyses of cohort studies". *Biometrics*, 42(2):293–199.

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

Robins, J. M., Mark, S. D., and Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48(2):479–495.

Rockey, D. C. (2010). Occult and obscure gastrointestinal bleeding: Causes and clinical management. *Nature Reviews Gastroenterology and Hepatology*, 7(5):265–279.

Rogers, M. A. M., Blumberg, N., Heal, J. M., and Langa, K. M. (2011). Utilization of blood transfusion among older adults in the united states. *Transfusion*, 51(4):710–718.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394.

Rosenbaum, P. R. (2002). *Observational studies, 2nd Edition*. New York: Springer-Verlag.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. New York: Springer-Verlag.

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

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524.

Rosenbaum, P. R. and Rubin, D. B. (1985a). The bias due to incomplete matching. *Biometrics*, 41(1):103–116.

Rosenbaum, P. R. and Rubin, D. B. (1985b). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Rosenstock, S., Jørgensen, T., Bonnevie, O., and Andersen, L. (2003). Risk factors for peptic ulcer disease: A population based prospective cohort study comprising 2416 danish adults. *Gut*, 52(2):186–193.

Rothman, K. J., Greenland, S., and L., L. T. (2008). *Modern Epidemiology, 3rd Edition*. Philadelphia: Lippincott Williams & Wilkins.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58.

Rubin, D. B. (1980). Bias reduction using mahalanobis-metric matching. *Biometrics*, 36(2):293–298.

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge University Press.

Sackett, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases*, 32(1-2):51–68.

Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H., and A., B. M. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20(4):512–522.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for r. *Journal of Statistical Software*, 42(7):1–52.

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

Shah, B. R., Laupacis, A., Hux, J. E., and Austin, P. C. (2005). Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *Journal of Clinical Epidemiology*, 58(6):550–559.

Strauss, D. (1998). On miettinen's multivariate confounder score. *Journal of Clinical Epidemiology*, 51(3):233–236.

Stürmer, T., Joshi, M., Glynn, R. J., Avorn, J., Rothman, K. J., and Schneeweiss, S. (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology*, 59(5):437–447.

Stürmer, T., Rothman, K. J., Avorn, J., and Glynn, R. J. (2010). Treatment effects in the presence of unmeasured confounding: Dealing with observations in the tails of the propensity score distribution-a simulation study. *American Journal of Epidemiology*, 172(7):843–854.

Stürmer, T., Schneeweiss, S., Brookhart, M. A., Rothman, K. J., Avorn, J., and Glynn, R. J. (2005). Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: Nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. *American Journal of Epidemiology*, 161(9):891–898.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1–21.

Tadrous, M., Gagne, J. J., Stürmer, T., and Cadarette, S. M. (2013). Disease risk score as a confounder summary method: Systematic review and recommendations. *Pharmacoepidemiology and Drug Safety*, 22(2):122–129.

Thillemann, T. M., Pedersen, A. B., Mehnert, F., Johnsen, S. P., and Søballe, K. (2010). The risk of revision after primary total hip arthroplasty among statin users: A nationwide population-based nested case-control study. *Journal of Bone and Joint Surgery - Series A*, 92(5):1063–1072.

Vamvakas, E. C. and Blajchman, M. A. (2007). Transfusion-related immunomodulation (trim): An update. *Blood Reviews*, 21(6):327–348.

Vamvakas, E. C. and Taswell, H. F. (1994). Epidemiology of blood transfusion. *Transfusion*, 34(6):464–470.

Wacholder, S. (1991). Practical considerations in choosing between the case-cohort and nested case-control designs. *Epidemiology*, 2(2):155–158.

Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., and Mor, V. (2004). Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and Drug Safety*, 13(12):841–853.

Westreich, D., Cole, S. R., Funk, M. J., Brookhart, M. A., and Stürmer, T. (2011). The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiology and Drug Safety*, 20(3):317–320.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

Williamson, E., Morley, R., Lucas, A., and Carpenter, J. (2012). Propensity scores: from naive enthusiasm to intuitive understanding. *Statistical Methods in Medical Research*, 21(3):273–293.

Woodward, M. (1999). *Epidemiology: study design and data analysis*. Boca Raton: Chapman & Hall/CRC.

Wyss, R., Girman, C. J., LoCasale, R. J., Brookhart, A. M., and Stürmer, T. (2013). Variable selection for propensity score models when estimating treatment effects on multiple outcomes: a simulation study. *Pharmacoepidemiology and Drug Safety*, 22(1):77–85.

Wyss, R., Lunt, M., Alan Brookhart, M. A., Glynn, R. J., and Stürmer, T. (2014). Reducing bias amplification in the presence of unmeasured confounding through out-of-sample estimation strategies for the disease risk score. *Journal of Causal Inference*. aop.

Zippin, C., Lum, D., and Hankey, B. F. (1995). Completeness of hospital cancer case reporting from the seer program of the national cancer institute. *Cancer*, 76(11):2343–2350.