# Efficient use of phase 1 information in two-phase case-control studies based on administrative databases

Dem Fachbereich 03 Mathematik und Informatik der
Universität Bremen
zur Erlangung des akademischen Grades eines
Dr. rer. nat.

eingereichte Dissertation

von
Frau Dipl.-Math. Sigrid Behr
aus
Hamm (Westf.)

Datum der Einreichung: 2013/03/28

Erstgutachter: Prof. Dr. Iris Pigeot-Kübler
Zweitgutachter: Pascal Wild, PhD

Tag der mündlichen Prüfung: 2013/05/16

# Danksagung

Auf diesem Wege möchte ich mich bei allen ganz herzlich bedanken, die mich bei meiner Dissertation unterstützt haben. Der größte Dank gebührt meinem Betreuer Walter Schill, der mein Interesse für das Dissertationsthema geweckt hat und dessen Ideen und Anregungen den Verlauf der Arbeit geprägt haben. Vielen Dank, Walter, für die fachliche und moralische Unterstützung in den letzten Jahren!
Besonders möchte ich mich auch bei meiner Doktormutter und Erstgutachterin Iris Pigeot bedanken, die jederzeit(!) für mich da war und einen absolut reibungslosen Abschluss der Arbeit, insbesondere in der Endphase, ermöglicht hat. Zudem möchte ich auch meinem Zweitgutachter Pascal Wild, der gesamten Prüfungskommission und meinen Kolleginnen und Kollegen danken.
Alle Personen zu nennen, denen ich für ihr Verständnis und ihre Unterstützung in den letzten Jahren dankbar bin, ist unmöglich. Stellvertretend für alle, die verstanden haben, wenn ich keine Zeit hatte, die mir den Rücken freigehalten haben, meine Launen ertragen haben und mich mit ihrer Freundschaft unterstützt haben, möchte ich mich bei Ann, Andy, Andreas, Dieter, Gräfin Emma, Michael, Sandra, Ulla und meiner Familie bedanken.

# Funding

Most parts of the presented work were funded by the research grants PI 345/4-1 and PI 345/5-1 of the German research foundation (DFG).

# Abstract

Population-based administrative databases are valuable data sources for scientific research. A common problem of studies based on secondary data is, however, missing confounder information. For instance in pharmacoepidemiological studies based on claims data of statutory health insurances, information on the potentially important confounders body mass index (BMI) and smoking behaviour is lacking. Obtaining additional information from another data source, e.g. from survey data, for at least a subsample of the study population resolves this problem. Two-phase designs can be employed for the combined analysis of both data sources. In logistic two-phase studies a dichotomous outcome and some covariate information are available for the whole population in phase 1 whereas the full vector of covariates is only observed for a subsample in phase 2. Information of phase 1 is utilised in the analysis of the phase 2 sample via stratification by phase 1 covariates. This approach allows for an unbiased parameter estimation and can improve efficiency of the estimation of covariates included in the stratification. An additional gain in efficiency can be achieved if the stratification is also used for sampling of the phase 2 data set and the most informative subjects are sampled with a high probability.

Two-phase designs have been developed for field studies. Field studies usually comprise very few covariates in phase 1 because ascertainment of each additional covariate for the whole population is associated with additional costs. In these traditional two-phase studies, the stratification is simply defined by cross-classification of all available phase 1 covariates. Two-phase database studies include a multitude of phase 1 covariates. Cross-classification of all available covariates would therefore lead to a tremendously high number of strata and, due to the restricted size of the phase 2 sample, to empty cells. Since two-phase methods cannot be applied in presence of empty cells, new stratification strategies are needed which account for all relevant phase 1 covariates but do not result in empty cells.

The aim of this thesis is the development of strategies for the efficient use of phase 1 information in logistic two-phase studies comprising a multitude of phase 1 covariates. An alternative stratification strategy is proposed, which is based on percentiles of a disease score. In this context, the disease score includes as a summary measure information on several phase 1 covariates in the stratification. The core of the thesis is the development of a design criterion which allows the identification of efficient stratifications at the planning stage of the study when only phase

1 information is available. Both the new stratification strategy and the design criterion are applied to an empirical two-phase database study investigating the risk of serious bleedings associated with phenprocoumon exposure. Additionally, the novel approaches are evaluated in a simulation study mimicking the empirical study. It is shown that those stratifications assessed as efficient by the design criterion result in the smallest standard errors for the estimates of most phase 1 covariates in the simulation study. The best stratifications regarding unbiased and efficient parameter estimation are defined by cross-classification of variables used for sampling of the phase 2 data and percentiles of the disease score. Moreover, empirically based recommendations for the planning of future two-phase studies are deduced from the results of the simulation study.

An outlook on survey methodological approaches for the analysis of partially missing data completes the thesis. The multiple imputation approach is applied to the empirical study and compared to results of the two-phase analyses. It becomes evident that multiple imputation is much more efficient with respect to the estimation of coefficients for phase 1 covariates than two-phase methods. However, biased estimates are observed in a simulation study, when the imputation model does not sufficiently account for the selective sampling of the phase 2 data. A final assessment of the comparison between two-phase methods and survey methodological approaches is still missing.

# Zusammenfassung

Administrative Datenbanken sind wertvolle Datenquellen für populationsbasierte Studien. Ein großes Problem dieser Sekundärdaten ist jedoch, dass oft nicht alle für die Studie notwendigen Informationen in den Daten enthalten sind. In pharmakoepidemiologischen Studien, die auf Basis von Abrechnungsdaten der gesetzlichen Krankenversicherungen durchgeführt werden, fehlen z.B. Informationen über wichtige verzerrende Faktoren (Confounder) wie den Body Mass Index (BMI) und das Rauchverhalten. Daten über fehlende Confounder müssen separat erhoben werden oder aus anderen Datenquellen zugespielt werden. Da es häufig nicht möglich ist, die zusätzlichen Daten für die gesamte Studienpopulation zu erheben, muss die Auswertungsmethode mit partiell fehlender Information umgehen können. Zwei-Phasen-Designs ermöglichen eine effiziente gemeinsame Auswertung beider Datenquellen. Das Design lässt sich sinnvoll anwenden, wenn gewisse Information, z.B. über den Krankheitsstatus und einige Kovariablen, für eine breite Studiengesamtheit verfügbar ist (Phase 1), während zusätzliche oder genauere Information nur für eine Teilmenge (Phase 2) vorliegt. In der Analyse der Phase-2-Stichprobe wird Information über die Kovariablenverteilung in den Phase-1-Daten genutzt, wodurch die Standardfehler der Parameterschätzungen der Phase-1-Variablen und korrelierter Phase-2-Variablen reduziert werden. Diese Nutzung von Phase-1-Information erfolgt über eine Stratifizierung der Datensätze nach Phase-1-Variablen. Ein zusätzlicher Effizienzgewinn kann dadurch erreicht werden, dass die Stratifizierung nicht nur zur Auswertung, sondern zuvor auch zur Selektion der Phase-2-Stichprobe verwendet wird, sodass ‘informative’ Personen mit einer höheren Wahrscheinlichkeit ausgewählt werden.

Zwei-Phasen-Designs wurden ursprünglich für Feldstudien entwickelt. Feldstudien enthalten in Phase 1 nur sehr wenige Variablen, da die Erhebung jeder zusätzlichen Variable mit Kosten verbunden ist. Die Stratifizierung wird in traditionellen Zwei-Phasen-Studien einfach als Kreuzklassifikation aller Phase-1-Variablen definiert. In Datenbankstudien steht hingegen eine Vielzahl an Phase-1-Variablen zur Verfügung, die in die Stratifizierung eingeschlossen werden können. Kreuzklassifikation aller Phase-1-Variablen führt in dieser Situation zu einer sehr großen Anzahl von Straten und, aufgrund des begrenzten Umfangs der Phase-2-Stichprobe, zu unbesetzten Zellen. Da Zwei-Phasen-Methoden nicht für Stratifizierungen mit leeren Straten angewendet werden können, wird eine Stratifizierungsstrategie benötigt, die

zwar einerseits viele Phase-1-Variablen einschließt, andererseits aber nicht zu leeren Zellen führt.

Das Ziel dieser Arbeit ist die Entwicklung von Strategien zur effizienten Nutzung von Phase-1-Information in logistischen Zwei-Phasen-Studien, die eine Vielzahl von Phase-1-Variablen enthalten. Dazu wird zunächst ein zur Kreuzklassifikation alternatives Stratifizierungsverfahren vorgeschlagen, das auf der Verwendung von Disease Scores beruht. Die Disease Scores werden dabei als zusammenfassendes Maß vieler Phase-1-Variablen genutzt. Im Hauptteil der Arbeit wird ein Designkriterium entwickelt, das den Vergleich verschiedener Stratifizierungen in Hinblick auf eine effiziente Parameterschätzung ermöglicht. Sowohl das neue Stratifizierungsverfahren als auch das Designkriterium werden in einer empirischen Zwei-Phasen-Datenbankstudie, die das Risiko schwerer Blutungen nach Phenprocoumoneinnahme untersucht, angewendet und in einer auf der empirischen Studie basierenden Simulationsstudie überprüft. Es zeigt sich, dass die mit dem Designkriterium als effizient eingestuften Stratifizierungen in der Simulationsstudie tatsächlich für viele Phase-1-Variablen zu den kleinsten Standardfehlern der geschätzten Parameter führen. Bezüglich der Effizienz und der gleichzeitigen Vermeidung von Verzerrungen in der Parameterschätzung sind Stratifizierungen, die aus einer Kreuzklassifikation von Perzentilen des Disease Scores und von Variablen bestehen, die zur Selektion der Phase-2-Stichprobe genutzt wurden, in dieser Datenkonstellation am besten geeignet. Aus den Ergebnissen der Simulationsstudie lassen sich außerdem Empfehlungen für die Planung zukünftiger Zwei-Phasen-Studien ableiten.

Die Arbeit schließt mit einem Ausblick auf surveymethodologische Ansätze, die ebenfalls für die Analyse unvollständiger Datensätze eingesetzt werden können. Insbesondere wird der Multiple-Imputation-Ansatz auf die empirische Studie angewandt und mit den Ergebnissen der Zwei-Phasen-Analyse verglichen. Die Schätzung mittels Multiple Imputation resultiert in dieser Studie im Vergleich zu Zwei-Phasen-Analysen in wesentlich kleineren Standardfehlern für die geschätzten Effekte der Phase-1-Variablen. In einer Simulationsstudie wird jedoch gezeigt, dass die Parameterschätzer bei Verwendung falscher Imputationsmodelle verzerrt sind. Eine abschließende Beurteilung der surveymethodologischen Verfahren im Vergleich zur Zwei-Phasen-Analyse ist noch nicht erfolgt.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| AOK | Regional health insurance (German: Allgemeine Ortskrankenkasse) |
| ARE | Asymptotic relative efficiency |
| ATC | Anatomical Therapeutic Chemical |
| ASA | Acetylsalicylic acid |
| BA | German Federal Employment Agency (German: Bundesagentur für Arbeit) |
| BIPS | Leibniz Institute for Prevention Research and Epidemiology GmbH - BIPS |
| BMI | Body mass index |
| CATI | Computer-assisted telephone interview |
| CI | Confidence interval |
| COPD | Chronic obstructive pulmonary disease |
| CRAN | Comprehensive R Archive Network |
| DDD | Defined Daily Dose |
| DMP | Disease Management Programme |
| DSC | Disease score |
| EM | Expectation-Maximisation |
| GePaRD | German Pharmacoepidemiological Research Database |
| GI | Gastrointestinal |
| ICD-10-GM | International Statistical Classification of Diseases, version 10, German modification |
| ID | Identification number |
| IM | Imputation model |
| lidA | German cohort study on work, age and health (German: lidA- leben in der Arbeit) |

| | |
|---|---|
| MAR | Missing-at-Random |
| MCMC | Markov chain Monte Carlo |
| ML | Maximum likelihood |
| MLE | Maximum likelihood estimator |
| NSAID | Non-steroidal anti-inflammatory drug |
| OPS | Classification system for medical procedures (German: Operationen- und Prozedurenschlüssel) |
| OR | Odds ratio |
| OTC | Over-the-counter |
| PASS | Panel study 'The Labour Market and Social Security' (German: Panel Arbeitsmarkt und Soziale Sicherung) |
| PL | Pseudo likelihood |
| PT | Penalty term |
| PZN | Central pharmaceutical reference number (German: Pharmazentralnummer) |
| SE | Standard error |
| SGB | Code of Social Law (German: Sozialgesetzbuch) |
| SHI | Statutory health insurance |
| SSRI | Selective serotonin reuptake inhibitor |
| STD | Standard deviation |
| WL | Weighted likelihood |
| WLE | Weighted likelihood estimator |

# List of Symbols

| | |
|---|---|
| $D$ | Disease indicator, p.22 |
| $\delta_m$ | Probability of covariate value $\tilde{\mathbf{x}}_m$, p.25 |
| $EFF_{\mathscr{S},k}$ | Relative efficiency for stratification $\mathscr{S}$ and covariate $k$, p.42 |
| $G$ | Marginal distribution function of $\mathbf{X}$, p.24 |
| $g$ | Density of the marginal distribution of $\mathbf{X}$, p.24 |
| $I_I$ | Phase 1 information obtained by ordinary logistic regression, p.52 |
| $I_P$ | Information of the profile log-likelihood, p.30 |
| $I_R$ | Phase 1 information according to Reilly and Pepe, 1995, p.36 |
| $I^*(\mathbf{\Phi})$ | Observed information of the pseudo model, p.26 |
| $J^*(\mathbf{\Phi})$ | Expected information of the pseudo model, p.26 |
| $L_p$ | Profile likelihood, p.24 |
| $l_P$ | Profile log-likelihood, p.25 |
| $L_W$ | Weighted likelihood, p.35 |
| $l^*$ | Log-likelihood of the pseudo model, p.25 |
| $N$, $N_i$, $N_{ij}$, $N_{ijk}$ | Phase 1 frequencies, p.22, p.34 |
| $n$, $n_i$, $n_{ij}$, $n_{ijk}$ | Phase 2 frequencies, p.22 |
| $\widehat{PT}_I(\mathscr{S})$ | Approximation of the penalty term for stratification $\mathscr{S}$, p.53 |
| $\pi_1$ | Marginal disease probability $Pr(D = 1)$, p.22 |
| $Q_{ij}$ | Marginal stratum probability, p.28 |
| $Q^*_{ij}(\tilde{\mathbf{x}}_m)$ | Stratum probability for a given covariate value $\tilde{\mathbf{x}}_m$, p.28 |
| $R^I$ | Indicator for selection into phase 1, p.27 |
| $R^{II}$ | Indicator for selection into phase 2, p.27 |
| $\mathscr{S}$ | Stratification, p.22 |
| $S_{ij}$ | Stratum $j$ for cases $(i = 1)$ or controls $(i = 0)$, p.22 |
| $\hat{\boldsymbol{\theta}}_{IM}$ | Multiple imputation estimator, p.79 |
| $\hat{\boldsymbol{\theta}}_P$ | EM estimator, p.34 |

| | |
|---|---|
| $\mathbf{X}$ | Complete covariate vector, p.22 |
| $\mathbf{Y}$ | Vector of phase 1 covariates, p.22 |
| $\mathbf{Z}$ | Vector of phase 2 covariates, p.22 |
| $\hat{\Omega}_I$ | Weighted score variability based on phase 1 data, p.52 |

# Chapter 1

# Introduction

Population-based administrative databases are a valuable data source for scientific research. Examples for such databases in Germany are the micro data of the German Federal Employment Agency (BA) and the claims data of the statutory health insurances (SHI). These data sources are often used for empirical research, either as a data source for database studies, as a sampling frame for survey studies, or as a source of additional information to be linked with survey data. A common problem in database studies based on secondary data is, however, missing confounder information, which can be resolved by obtaining additional information for at least a subset of the study sample. In this context, the combined use of administrative data and other data sources is reasonable. Currently, there are some examples of studies combining population-based administrative data with survey data or other secondary data. For instance in the panel study 'The Labour Market and Social Security' (PASS), the register on unemployment benefit II recipients was used as the sampling frame for one of the two samples included in the panel (Rudolph and Trappmann, 2007). In PASS, information on postal code, unemployment benefit receipt and the number of so-called 'Bedarfsgemeinschaften' was extracted from the administrative data for each household and used for the selection of the study sample. Another example is the study 'lidA - leben in der Arbeit. German Cohort Study on Work, Age and Health' in which interview data is planned to be linked to administrative data of the BA and to health insurance data to investigate the influence of work-related factors on the health of workers of advanced age (March et al., 2012).

In both examples, the available information from the administrative data source is not fully considered in the design and analysis of the studies. In the main analysis of PASS, information from the administrative data is only utilised to generate sampling weights which means that only that part of information is used on which the sample selection was based. Further analyses investigating non-response and memory error incorporate additional variables from the administrative data (Kreuter et al., 2010). However, these analyses are based on subjects included in both data sources and ignore information for subjects not included in the intersection of the data sources. In the lidA study, it is also planned to base the study exclusively on subjects included in the survey who agreed to the data linkage.

To use the available data sources more efficiently in such studies, two-phase designs can be employed. Two-phase designs were first suggested by Neyman, 1938 who proposed a two-step approach of sampling in a field survey with the aim to investigate a characteristic which is expensive to obtain but for which an easily collectable surrogate exists. In the first step of sampling, the surrogate is obtained for a large population. In the second step of sampling, a stratified random sample is drawn from the large population where the surrogate is used to define the strata. The characteristic of interest is only obtained for the stratified subsample but the analysis of the characteristic also utilises the distribution of the surrogate in the large population, thereby enhancing the efficiency of the estimator. More than 40 years later, the two-phase design was introduced into the field of epidemiology by Walker, 1982 and White, 1982. During the late 1980s and the 1990s methodological work has been published regarding the estimation in two-phase studies with binary outcome (e.g., Breslow and Cain, 1988; Flanders and Greenland, 1991; Schill et al., 1993), though the design was rarely used in practice. Also in the late 1990s, Schaubel et al., 1997 suggested the use of the two-phase design for epidemiological studies in which the population is obtained from an administrative database and confounder information is collected from surveys or other data sources. The few published examples of two-phase studies based on administrative or registry data employ simple two-phase designs using the main exposure of interest as a single stratification variable (Collet et al., 1998; Sharpe et al., 2000; Martel et al., 2009). These studies ignore information, e.g. on age and sex, which is also available in the administrative data for the full population.

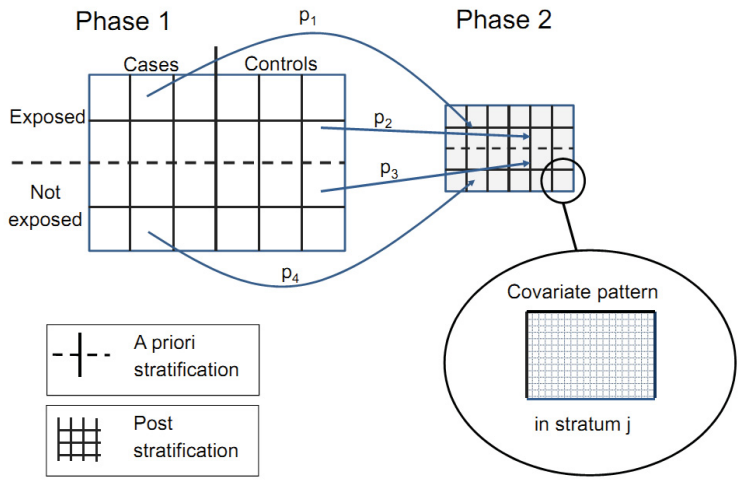



**Figure 1.1:** Concept of a two-phase case-control design

The aim of this thesis is to investigate two-phase designs for case-control studies based on administrative databases which use the maximum amount of available phase 1 information for the efficient estimation of the adjusted exposure effect of the exposure under study. Incorporation of the phase 1 information in the two-phase analysis is accomplished by the performed stratification which is therefore the key element of a two-phase design. To further illustrate the role of stratification, the concept of a two-phase case-control design is depicted in Figure 1.1. Two types of stratification can be distinguished: A stratification defined before the phase 2 sample is drawn is called *a priori stratification*; if the stratification is defined to analyse an already existing phase 2 sample it is called *post stratification*. The post stratification can be based on any set of variables available in phase 1. As will be seen in Chapter 3, the estimation methods do not differentiate between a priori and post stratification. For design considerations, however, the distinction is worthwhile.

Figure 1.1 shows a two-phase case-control study in which an a priori stratification is defined based on a binary covariate, which is in most applications the exposure of main interest. The a priori stratification classifies the phase 1 data set into four strata: exposed cases, unexposed cases, exposed controls and unexposed controls. From each of the four strata, a random sample is selected with a specified probability,

here denoted by $p_1, \ldots, p_4$. The resulting phase 2 data set is further stratified by additional phase 1 variables, i.e. by those variables that are available for everyone in the phase 1 sample. The phase 2 data set is then analysed with respect to the complete covariate pattern which is only known for subjects included in the phase 2 sample and taking into account the stratum distribution in the phase 1 data set. The choice of the stratification and the choice of the selection probabilities have a substantial impact on the efficiency of the parameter estimates obtained from the two-phase analysis.

In this thesis, an empirical study on the risk of bleeding associated with phenprocoumon exposure is used as a recurring theme to illustrate the impact of different analysis approaches in two-phase database studies. Chapter 2 starts with a description of two ordinary case-control studies on phenprocoumon exposure and bleedings which are published in Behr et al., 2010 (Appendix B) and Garbe et al., 2013 (Appendix C) and which serve as a motivation for the empirical two-phase study. The chapter also introduces the two underlying data sources: claims data included in the German Pharmacoepidemiological Research Database (GePaRD) and interview data obtained in a health survey. In anticipation of the theoretical discussion of two-phase database studies in the following chapters, first results of the empirical two-phase study are reported based on the publication of Behr et al., 2012 (Appendix D) in *Pharmacoepidemiology and Drug Safety*. Chapter 3 provides details on the parameter estimation in two-phase studies. After defining the notation and the study design, likelihood-based estimation methods for two-phase studies are introduced. The main part of the chapter is devoted to the derivation of maximum likelihood estimators in case-control and two-phase case-control studies using a profile likelihood approach. The focus of Chapter 4 lies on the identification of stratifications which use the available phase 1 information efficiently. The chapter starts with a discussion on the limitations of cross-classification, i.e. of the stratification strategy used in traditional two-phase field studies, which can also be observed in the empirical study (Behr et al., 2012; Appendix D). An alternative stratification strategy based on percentiles of a disease score is proposed by Behr and Schill, 2013 (Appendix E) in a paper that will be submitted to a methodological epidemiology journal. The core of the chapter as well as of the thesis is the development of a design criterion for the identification of the most efficient stratifications with respect to small standard errors of the parameter estimates. This part of the thesis has not yet been sum-

marised in a manuscript and is therefore described in detail. Both novel approaches, the new stratification strategy and the design criterion, are applied in the empirical study as well as in a simulation study based on this study. Chapter 5 gives a first impression of approaches beyond two-phase methods. In particular, results of the two-phase analysis using the most efficient stratification identified in Chapter 4 are compared with results of the survey methodological approach multiple imputation. Chapter 6 concludes with a discussion of the presented work with regard to design considerations of future two-phase studies comprising a rich phase 1 data set.

# Chapter 2

# The empirical study: Two-phase case-control study on the risk of serious bleedings associated with phenprocoumon treatment

## 2.1 Motivation

The Leibniz Institute for Prevention Research and Epidemiology - BIPS established a research database which consists in its current form of administrative data from four German statutory health insurances to study the use, effectiveness and safety of pharmaceutics and vaccines (for a detailed description of the German Pharmacoepidemiological Research Database, in short GePaRD, see Section 2.2.1). One of the first research projects that was initiated based on this database investigated the risk of serious bleedings associated with phenprocoumon exposure. The research project was intended to fulfil two purposes. First, the knowledge gap concerning the magnitude of the bleeding risk related to phenprocoumon treatment should be closed. Second, with serious bleedings being a well-known complication of oral anticoagulants the suitability of the database for pharmacoepidemiological studies should be investigated.

A series of nested case-control studies was conducted to assess the overall bleeding risk and the risk of bleeding in specific locations (gastrointestinal, intracerebral, subarachnoid and urogenital bleedings) for phenprocoumon treatment. Except for the outcome definition and the selection of potential confounders, the designs of the case-control studies were very similar. The study on intracerebral bleedings has been published in *Pharmacoepidemiology and Drug Safety* (Behr et al., 2010; Appendix B). Details on the design of the case-control studies are described in that publication. The study on subarachnoid bleedings has been submitted to *Stroke* (Garbe et al., 2013; Appendix C). In addition, the risk of bleeding was assessed for drug interactions with phenprocoumon (Jobski et al., 2011). In the next section, the focus is on an unpublished study investigating the overall risk of serious bleeding which motivated the conduct of the two-phase case-control study. After giving a brief summary of the background and the study design, the results are described in more detail.

### 2.1.1 Case-control study on the overall bleeding risk

Before the new classes of oral anticoagulants were launched to the market in 2008, coumarins were the most widely used oral anticoagulants. Phenprocoumon is a coumarin which accounts for more than 99% of the coumarin prescriptions in Germany whereas outside of Germany warfarin is used instead (Hein and Schwabe, 2007). For this reason, there have been no epidemiological studies investigating the bleeding risk for phenprocoumon alone and only a few studies investigating the combined use of warfarin and phenprocoumon (e.g., Johnsen et al., 2003; Groenbaek et al., 2008). To estimate the overall risk of serious bleeding associated with phenprocoumon exposure a case-control study was conducted nested in a cohort of more than 265,000 subjects who were insured at the regional healthcare provider AOK Bremen/Bremerhaven between 2004 and 2006. Cohort entry was defined as the first day after six months of continuous insurance time in the study period and cohort exit was determined as the end of the insurance period, the date of hospitalisation due to serious bleedings, the date of death, or the end of the study period whichever came first. Claims data from the GePaRD (see Section 2.2.1) was used to identify cases by hospital discharge diagnoses of serious bleedings and to obtain covariate information for cases and their respective controls which were matched by

age and sex with a ratio of ten controls per case using a risk-set sampling approach. Phenprocoumon exposure was assessed on the index day, i.e. the day of hospitalisation for cases and the day resulting in the same duration of follow-up for controls. To be more precise, exposure on the index day was defined as a phenprocoumon prescription overlapping the index day where the duration of a prescription was estimated by the amount of prescribed substance divided by the estimated average daily dose. Adjusted bleeding odds ratios (OR) and 95% confidence intervals were estimated for phenprocoumon exposure in a multivariable conditional logistic regression model including comorbid diseases and concomitant medications as covariates as well as interaction terms between phenprocoumon exposure and the covariates, age and sex, respectively. A list of all covariates and details on their definition are given in a paper recently published in *Pharmacoepidemiology and Drug Safety* (Behr et al., 2012; Appendix D).

A total of 2,113 cases of serious bleedings and 21,128 matched controls were identified in the cohort. The mean age in the case-control sample was 68 years (standard deviation (STD)=17.3 years) and slightly more women (54%) than men were included. Most cases were hospitalised due to gastrointestinal bleedings (56%) and cerebral bleedings (17%). Regarding phenprocoumon exposure, more cases (10%) than controls (3%) were exposed on the index day. The prevalence of comorbid conditions and concurrent medications was generally higher among cases than among controls, e.g., diabetes was prevalent in 17% of the cases and 11% of the controls (Table 2.1).

The results of the multivariable regression model are presented in Table 2.2. The multivariable analysis revealed a significantly increased risk of serious bleedings associated with phenprocoumon exposure. As indicated by the significant interaction terms, the magnitude of the risk depended on age and sex and was 4.6-fold increased for a 68 years old male phenprocoumon user compared to a respective male not taking phenprocoumon. The risk associated with phenprocoumon was approximately two times higher for females and decreased for older age. Because of the matched design of the study age and sex could not be evaluated as independent risk factors for serious bleedings. Most of the considered covariates were identified as significant risk factors for serious bleedings. Among these, medications affecting the coagulation, such as platelet aggregation inhibitors, heparin, and analgesics, were associated

**Table 2.1:** Characteristics of cases and controls for selected comorbid conditions and concurrent medications

| | **Cases** N=2,113 | **Controls** N=21,128 |
|---|---|---|
| **Comorbid conditions** | | |
| Alcohol dependence | 164 (7.76%) | 407 (1.93%) |
| Bleeding history | 58 (2.74%) | 93 (0.44%) |
| Diabetes mellitus | 360 (17.04%) | 2,365 (11.19%) |
| Diverticular disease | 219 (10.36%) | 1,163 (5.50%) |
| Hypertension | 1,355 (64.13%) | 11,168 (52.86%) |
| Liver failure | 324 (15.33%) | 2,252 (10.66%) |
| Renal failure | 324 (15.33%) | 1,388 (6.57%) |
| **Concurrent medications** | | |
| Analgesics, antirheumatics, ASA | 477 (22.57%) | 2,239 (10.60%) |
| Diuretics | 726 (34.36%) | 3,834 (18.15%) |
| H2-receptor antagonists | 69 (3.27%) | 318 (1.51%) |
| Platelet aggregation inhibitors, heparin | 334 (15.33%) | 1,154 (5.46%) |
| Proton pump inhibitors | 372 (17.61%) | 1,164 (5.51%) |
| SSRIs | 36 (1.70%) | 118 (0.56%) |

ASA: acetylsalicylic acid, SSRI: selective serotonin reuptake inhibitor

with an approximately two-fold increased risk.

The observed interaction between phenprocoumon exposure and sex has not been described in the literature previously. Whether the elevation of risk is causally related to female sex cannot be concluded from this study because the higher bleeding risk can also be related to the usually lower body mass index (BMI) in women. Since information on BMI is not included in the administrative data, the risk of bleeding could not be adjusted for it which is a relevant limitation of the study. Furthermore, claims data does only include information on prescriptions which are reimbursed by the health care provider. Especially painkillers like acetylsalicylic acid (ASA) are often bought over-the-counter (OTC) leading to an underestimation of exposure to analgesics and potentially diluting the appropriate estimation of the bleeding risk

**Table 2.2:** Results of the multivariable conditional logistic regression analysis for serious bleedings

| **Multivariable model** | **Odds Ratio**[b] | **95% Confidence Interval** | **p-value** |
|---|---|---|---|
| Phenprocoumon exposure | 4.60 | 2.93-7.23 | <.001 |
| *Interaction:* phen. * sex | 1.86 | 1.27-2.72 | 0.001 |
| *Interaction:* phen. * age[a] | 0.98 | 0.96-1.00 | 0.029 |
| **Comorbid conditions** | | | |
| Alcohol dependence | 3.67 | 2.96-4.55 | <.001 |
| Bleeding history | 3.91 | 2.68-5.69 | <.001 |
| Diabetes mellitus | 1.13 | 0.98-1.29 | 0.082 |
| Diverticular disease | 1.63 | 1.36-1.94 | <.001 |
| *Interaction:* phen. * diverticular disease | 0.58 | 0.32-1.05 | 0.070 |
| Hypertension | 1.16 | 1.03-1.31 | 0.017 |
| *Interaction:* phen. * hypertension | 0.82 | 0.52-1.30 | 0.391 |
| Liver failure | 1.39 | 1.22-1.58 | <.001 |
| Renal failure | 1.67 | 1.42-1.96 | <.001 |
| *Interaction:* phen. * renal failure | 0.60 | 0.37-0.96 | 0.035 |
| **Concurrent medications** | | | |
| Analgesics, antirheumatics, ASA | 1.88 | 1.66-2.12 | <.001 |
| Diuretics | 1.60 | 1.42-1.80 | <.001 |
| H2-receptor antagonists | 1.38 | 1.03-1.85 | 0.029 |
| Platelet aggregation inhibitors, heparin | 2.22 | 1.92-2.58 | <.001 |
| Proton pump inhibitors | 2.16 | 1.87-2.49 | <.001 |
| SSRIs | 1.93 | 1.26-2.96 | 0.002 |

[a] Age centred at 68 years.

[b] Also adjusted for corticosteroid use, cancer, and chronic obstructive pulmonary disease (COPD).
ASA: acetylsalicylic acid, SSRI: selective serotonin reuptake inhibitor

for these substances.

To overcome these limitations, additional information was obtained from a subsample of insurance members in a health survey described in Section 2.2.2 where a

two-phase design was used for the combined analysis of both data sources. The design and the main results of the two-phase study are briefly described in Section 2.3. More details of the study are published in Behr et al., 2012 (Appendix D). Since this two-phase study is used to illustrate design aspects of two-phase studies throughout the thesis, the underlying data sources are described in the next sections.

## 2.2 Data sources

### 2.2.1 The administrative claims database

The German Pharmacoepidemiological Research Database (GePaRD) currently comprises claims data of the years 2004-2009 from four statutory health insurances with about 17 million members during this time period. The content and structure of the insurance claims data is regulated by the Code of Social Law V which allows pooling of the data from different health care providers according to the structure presented in Figure 2.1. The data can be subdivided into four blocks corresponding to distinct data-generating processes which are regulated by different paragraphs of the Code of Social Law. The first block consists of the basic claims data comprising demographic characteristics of the insurance member and information on insurance periods. The second block of data arises from hospital stays and includes inpatient diagnoses and inpatient medical procedures. Since inpatient service is reimbursed by lump sums according to the main discharge diagnosis, inpatient pharmaceutical treatment is not recorded in the claims data. The third block contains outpatient treatment data such as information about the treating physician, outpatient diagnoses and outpatient medical procedures. The fourth block consists of outpatient prescription data including all prescriptions which were filled by the patient in a pharmacy and were reimbursed by the health care provider. Thus, no information on OTC medication and on non-refundable medication is available in the claims data. The four blocks are linked by pseudonymous subject identifiers.

All diagnoses are coded according to the German modification of the International Statistical Classification of Diseases version 10 (ICD-10-GM). Medical procedures are coded using the OPS classification system for surgeries and medical procedures.

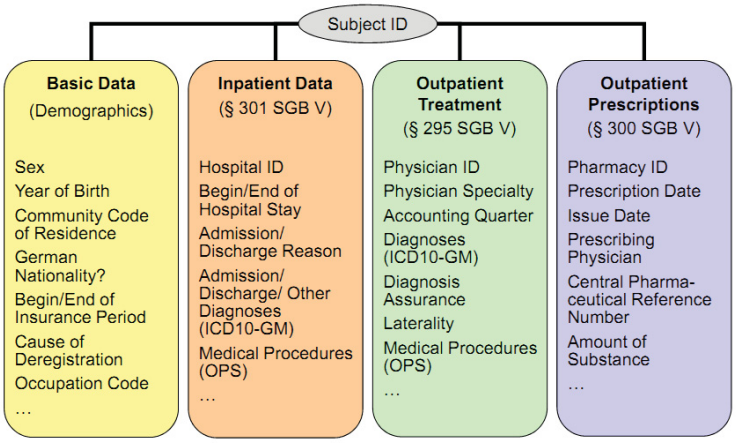



**Figure 2.1:** Structure of GePaRD

The prescription data can be linked by using the central pharmaceutical reference number (PZN) to information included in a pharmaceutical reference database maintained at the BIPS that comprises the ATC code, trade name, generic name, and the Defined Daily Dose (DDD) for each substance.

Pigeot and Ahrens, 2008 investigated the usefulness of German SHI data for the conduct of pharmacoepidemiological studies in a pilot database covering 3.6 million subjects. They compared the age and sex distribution and hospital admission rates for specific diseases obtained from the claims data with the respective values of the official statistics and found good agreement. They concluded that claims data can be used to define confounder information on medical conditions and on concomitant medication but also alluded to the limitation of the data to identify lifestyle-related confounders such as smoking.

### 2.2.2 Additional information obtained in a health survey

For the empirical study on the overall bleeding risk related to phenprocoumon use additional information on lifestyle- and health-related factors was obtained in computer-assisted telephone interviews (CATI) for a subsample of insurants included in the case-control study described in Section 2.1. The health survey was conducted in collaboration with the regional statutory health insurance AOK Bre-

men/Bremerhaven who contacted randomly selected insurance members (see Section 2.3) by mail and asked for their willingness to participate in the survey. Members who agreed to participate had to send their telephone number together with a signed informed consent to the BIPS and were afterwards called by the field work unit of the BIPS. The data collection and data handling process was described in a data protection concept which was reviewed and accepted by the data protection officers of the SHI and of the University of Bremen. In addition, the ethical committee of the University of Bremen did not raise ethical concerns regarding the conduct of the study.

**Table 2.3:** Response proportion in the health survey

| | Batch 1 | Batch 2 | Total |
|---|---|---|---|
| Letters sent | 1,619 | 1,661 | 3,280 |
| Response | 382 (23.59%) | 229 (13.78%) | 611 (18.62%) |
| Interviews conducted | 312 (19.27%) | 193 (11.5%) | 505 (15.35%) |

Since claims data of the year 2007 became available shortly before the survey started, the selection of insurants was based on claims data from 2004 until 2006 and preliminary data from 2007. Several exclusion criteria were applied to ensure that all persons were able to participate in the telephone interview. Persons were excluded if they were younger than 18 years or older than 75 years, if they quit membership in the SHI, if they had a diagnosis of dementia between 2004 and 2007, or if they died. Two batches of 2,000 insurants each were selected and contacted consecutively. In the initial batch of 2,000 persons, all persons with serious bleedings and all persons with phenprocoumon exposure were included to achieve a balanced design with respect to phenprocoumon exposure and serious bleedings. The remaining persons were selected randomly from the set of unexposed persons without bleedings. After insurants had been selected for the two batches, the SHI removed also those persons who lived in nursing homes, who had left the SHI after 2007 or died after 2007. All interviews were conducted between October 2009 and January 2010. A total of 505 interviews could be completed which corresponds to a response proportion of 15% (Table 2.3). The response proportions for the two batches are summarised in Table 2.3.

The following information was requested in the telephone interviews:

- Personal information about the date of birth, height, weight and the general health status,
- hospital stays since 2004 including the reason for hospitalisation and the length of stay,
- administration and indication of specific drug substances since 2006, i.e., phenprocoumon, acetylsalicylic acid (ASA), diclofenac, ibuprofen and St. John's wort which is an antidote for phenprocoumon,
- treated and untreated gastric disorders since 2006, e.g., gastrointestinal bleedings, gastric ulcer, reflux,
- other non-traumatic bleedings since 2004 which needed medical treatment,
- smoking history and current smoking behaviour.

The survey data was checked for internal plausibility and for external plausibility by comparing the survey information with claims data for those items which were available in both data sources. Very few inconsistencies were detected in the internal plausibility checks. The comparison with the claims data revealed that bleedings were reported with a low sensitivity whereas the sensitivity and the specificity were high for phenprocoumon exposure. In contrast, prescriptions of the well-known substance ASA were overreported and prescriptions of the less known pain killers ibuprofen and diclofenac were underreported. As bleeding information can be reliably obtained from the claims data the low reporting quality of bleedings imposed no restrictions on the two-phase study described in the next section. However, information on OTC use of ASA and St. John's wort and on other pain killers was deemed to be not reliable due to the observed response errors for the prescribed drugs. As a consequence, this information was not included in the two-phase study.

## 2.3 The two-phase study

A two-phase study was conducted to consider the additional confounder information on BMI and smoking from the survey data for the estimation of the adjusted bleeding

risk associated with phenprocoumon use based on the case-control database study. This section describes the design of the two-phase study and gives a short overview about the results. Details on the results are presented in Behr et al., 2012 (Appendix D).

### 2.3.1 Study design

The design of the two-phase case-control study was very similar to the database study presented in Section 2.1 using the same case definition and the same rules for the identification of phenprocoumon exposure as well as the derivation of covariates. The underlying cohort was defined in an extended data set consisting of claims data of the AOK Bremen/Bremerhaven from 2004 until 2007. In addition to the inclusion criterion of continuous membership in the SHI for at least six months, the same exclusion criteria were applied as for the sample selection of the health survey. Cohort entry and cohort exit were defined as in the database study. However, in contrast to this study, controls were not matched by age, sex and follow-up time but were randomly selected from the set of cohort members who did not become a case during the study with a case:control ratio of 1:20. Since the phase 2 subsample, i.e. the insurants who were asked to participate in the health survey, had been selected from preliminary data before the phase 1 case-control sample was drawn, it had to be ensured that all survey participants were also part of the case-control sample. Therefore, all subjects in the set of potential controls who completed the telephone interview were included in the case-control sample. For each control, a random index date was chosen from the period between cohort entry and cohort exit. Additional information on BMI and smoking behaviour was available from the health survey for a subset of 502 subjects. The remaining three subjects who participated in the survey had to be excluded because two of them were not entitled to receive benefits from the SHI during the study period. Interview data of the third subject was of very low quality (i.e., most items were missing, in particular no dates were available).

The setup of the two-phase study is depicted in Figure 2.2. The phase 2 sample was planned as a stratified random sample with approximately equal numbers of exposed cases, unexposed cases, exposed controls and unexposed controls. Aiming

at a balanced phase 2 data set of 1,000 subjects and assuming a response proportion of 25%, 4,000 subjects were selected with sampling fractions of one for cases and exposed controls and of 0.08 for unexposed controls. The actual phase 2 sample size was only half of the planned sample size and the actual sampling fractions differed widely from the planned sampling fractions (Figure 2.2). Because of the unequal sampling fractions, the phase 2 sample was not representative for the phase 1 sample.

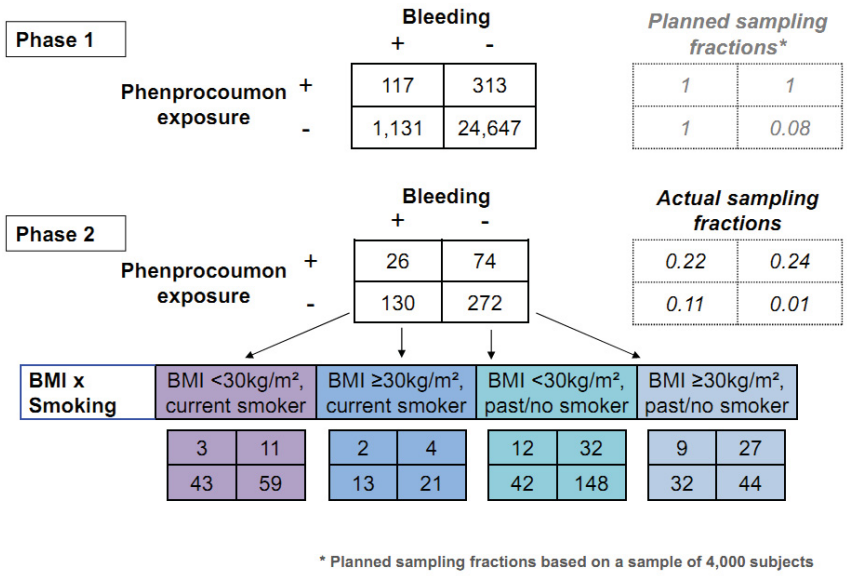



**Figure 2.2:** Setup of the two-phase study on serious bleedings associated with phenprocoumon use

As will be explained in Chapter 3, overall representativeness of the phase 2 sample is not required for adopting the two-phase approach but representativeness of each stratum has to be ensured to fulfil the Missing-at-Random (MAR) assumption. Since unintended selection processes might have led to distorted phase 2 strata, a non-response analysis was conducted to identify prognostic factors for participation in the survey. The logistic regression analysis modelling the probability of selection into the phase 2 sample revealed case status and phenprocoumon exposure as the most relevant prognostic factors which was expected due to the stratified sampling scheme. Furthermore, age at cohort entry was identified as an important predictor with a higher selection probability for older age. Significant but weak associations

were found for sex, use of statins and diverticular disease. All other covariates were removed from the model by the applied backward selection procedure due to non-significance ($p \geq 0.05$). The results of the non-response analysis are given in Table 2.4. Taking these results into account, a stratification was constructed by cross-classifying age ($< 50$ years, $50- < 65$ years, and $\geq 65$ years), sex, and phenprocoumon exposure into 12 strata. Use of statins and diverticular disease were not considered in the stratification because of the relatively low prevalence of both factors (3.6% for statins, 0.7% for diverticular disease) and the only weak association.

**Table 2.4:** Results of the logistic regression analysis modelling the probability of selection into phase 2

| **Backward selection model** | **OR (95% CI)** | **p-value** |
|---|---|---|
| Case-control status | 5.35 (4.31 - 6.64) | $< .001$ |
| Phenprocoumon exposure | 7.33 (5.57 - 9.64) | $< .001$ |
| Sex (reference: male) | 1.25 (1.03 - 1.50) | 0.021 |
| Age (reference: $< 50$ years) | | |
| $\geq 50- < 65$ years | 3.64 (2.87 - 4.63) | $< .001$ |
| $\geq 65$ years | 4.20 (3.21 - 5.50) | $< .001$ |
| Diverticular disease | 1.82 (1.01 - 3.25) | 0.045 |
| Statin use | 1.56 (1.16 - 2.10) | 0.003 |

A two-phase logistic model was fit to estimate the risk of bleeding associated with phenprocoumon use adjusted for the phase 1 covariates age, sex, hypertension, and diabetes mellitus as well as for the phase 2 covariates BMI and smoking. Odds ratios and corresponding 95% confidence intervals (CI) were estimated by using the maximum likelihood (ML) approach described by Breslow and Holubkov, 1997a and taking into account the stratification specified above. The theoretical background of ML estimation in two-phase studies is given in Chapter 3. Further details on the statistical analysis are specified in Behr et al., 2012 (Appendix D). The results of the two-phase analysis were compared to risk estimates from a logistic regression analysis based on the full phase 1 data set adjusted for the same phase 1 covariates but without consideration of BMI and smoking. In addition, a full phase 1 model

was used to estimate the risk for phenprocoumon adjusted for all available phase 1 variables.

### 2.3.2 Results

For the phase 1 case-control sample 1,248 cases and 24,960 controls were selected from 186,438 subjects included in the study cohort. The phase 2 sample consisted of 155 cases and 343 controls with complete information on BMI and smoking[1]. Similar estimators for the risk of bleeding associated with phenprocoumon use were observed in the two-phase analysis (OR=4.96, 95% CI: 2.91-8.45), the corresponding phase 1 analysis (OR=4.14, 95% CI: 2.95-5.81), and in the full phase 1 model (OR=3.93, 95% CI: 2.75-5.61). Regression coefficients for age and the estimated interaction between age and phenprocoumon exposure were also of similar size in all models (cf. Table 3 and Table 5 in Appendix D). The interaction between sex and phenprocoumon exposure was not statistically significant ($p < 0.05$) in the full phase 1 model (OR=1.33, 95% CI: 0.83-2.15) but in the reduced phase 1 model (OR=1.59, 95% CI: 1.01-2.49) and in the two-phase analysis (OR=1.75, 95% CI: 1.07-2.86). Nevertheless, the odds ratios related to this interaction consistently indicated an increased risk of bleeding for women taking phenprocoumon compared to men taking phenprocoumon. Moreover, the two-phase model revealed an increased risk of bleeding for BMI $\geq$ 30 kg/m$^2$ and smoking with odds ratios of 1.57 (95% CI: 1.00-2.47) and 2.30 (95% CI: 1.50-3.53), respectively. There was no evidence for an interaction between BMI and phenprocoumon exposure. The effects of hypertension and diabetes mellitus could not be estimated with sufficient precision in the two-phase analysis.

### 2.3.3 Discussion

Additional adjustment for BMI and smoking in the two-phase analysis did not alter the estimated bleeding risk associated with phenprocoumon exposure. This study confirmed the increased risk for female phenprocoumon users which was observed

[1]One control without information on BMI and one case and two controls without information on smoking were excluded from the two-phase analysis.

in the case-control study described in Section 2.1.1 although the risk was less pronounced in the two-phase study.
On the one hand the two-phase analysis yielded accurate and precise risk estimators for phenprocoumon exposure and age, on the other hand two-phase estimators for covariates like hypertension and diabetes mellitus were inconclusive. Since inclusion of phase 1 information on these covariates in the stratification may improve the precision of the estimators, subsequent analyses were conducted to explore the performance of alternative stratifications in this study. The results of these analyses are presented and discussed in Chapter 4.

# Chapter 3

# Two-phase methodology for case-control studies

Several likelihood-based methods have been developed for the parameter estimation in logistic two-phase studies. Three of these approaches, the full maximum likelihood (ML), the pseudo likelihood (PL) and the weighted likelihood (WL) estimation, are outlined in this chapter. After introducing the notation and the two-phase setting considered in this thesis, the ML estimator is derived for simple case-control studies and for two-phase case-control studies by using a profile likelihood approach. Moreover, the relation between the ML estimator and the PL estimator is briefly discussed. As an alternative to the profile likelihood approach, the EM-algorithm is described as a computational approach for ML estimation in two-phase studies. Finally, the WL method is established which is known to be less efficient than ML estimation. The WL approach is introduced here because its easily interpretable covariance formula allows for the definition of a design criterion that is derived in Chapter 4. The chapter concludes with some remarks on the implementation of these methods in SAS and R software.

## 3.1 Notation

Throughout the thesis, a binary disease model with a linear-logistic form is considered. Let $D$ denote a binary disease indicator that equals one for diseased subjects (cases) and zero for subjects free of disease (controls) and let $\mathbf{X}$ be a $p$-dimensional covariate vector. The model for contracting the disease given the covariate vector $\mathbf{X}$ is therefore:

$$Pr(D = 1|\mathbf{X} = \mathbf{x}) = (1 + e^{-(\alpha+\boldsymbol{\beta}^T\mathbf{x})})^{-1} \tag{3.1}$$

where $Pr()$ stands for probability. It is well known that the $\boldsymbol{\beta}$-part of the parameter vector $\boldsymbol{\theta}^T = (\alpha, \boldsymbol{\beta}^T)$ can be estimated from data collected in retrospective or prospective studies. The intercept $\alpha$, which indicates the baseline disease probability, can only be estimated in retrospective studies if additional information on the marginal disease probability is available. The focus here lies on the retrospective case-control design, a situation in which the covariate distribution is observed conditional on the disease status, i.e. data of the form $(\mathbf{X}|D)$ is observed. To be more precise, suppose a case-control study is conducted with $N$ subjects, $N_1$ cases and $N_0$ controls, $N = N_1 + N_0$, in which covariate values $\mathbf{x}$ are sampled from the conditional distribution of $\mathbf{X}$ given $D$. The likelihood contribution of a subject with covariate vector $\mathbf{x}$ and disease status $d$ is then given by $Pr(\mathbf{X} = \mathbf{x}|D = d)$.

In the setting considered in this thesis, the case-control study is nested within a cohort using data from a population-based administrative database which allows for the estimation of the marginal disease probabilities $\pi_1 = Pr(D = 1)$ and $\pi_0 = Pr(D = 0)$. Although the claims data regarded here comprises a multitude of covariates the complete covariate vector $\mathbf{X}$ might not be entirely available. Then, a two-phase study is reasonable, in which the claims data constitutes the phase 1 data set and phase 2 data comprises additional information from another data source for a subset of $n$ subjects. Suppose that the phase 1 data includes information about the disease status $D$ and about covariates $\mathbf{Y}$. With $\mathbf{Z}$ denoting the additional covariate information collected only for the phase 2 sample, the complete vector of explanatory variables is given by $\mathbf{X}^T = (\mathbf{Y}^T, \mathbf{Z}^T)$. A stratification $\mathscr{S}$ can be constructed as a function of $\mathbf{Y}$, which is then used to stratify the case-control sample in $2 \times J$ strata, where $S_{ij}$ is the $j$th stratum of cases (if $i = 1$) or of controls (if $i = 0$) for $j = 1, \ldots, J$. Assume that the phase 2 data set is a stratified random sample,

for which $n_{ij}$ individuals have been sampled randomly from the $N_{ij}$ individuals in stratum $S_{ij}$ for $i \in \{0,1\}$ and $j = 1, \ldots, J$. In this situation, the MAR-assumption holds in each stratum.

In the following, the subscript $i$ is used to denote the disease status ($i \in \{0,1\}$), $j$ is used to denote the stratum ($j \in \{1, \ldots, J\}$) and $k$ denotes a single subject ($k \in \{1, \ldots, N\}$). Furthermore, $\bullet_+$ means summation over the respective index.

Before introducing the methodology for a two-phase study, the concept of parameter estimation via a profile likelihood is explained for the simple case-control study in the next section.

## 3.2 Estimation via the profile likelihood

### 3.2.1 Estimation in case-control studies

Although the likelihood for the retrospective case-control study contains terms of the form $Pr(\mathbf{X}|D)$, Prentice and Pyke, 1979 have shown that the semiparametric maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ can be obtained from the ordinary (prospective) logistic regression model (see Equation (3.1)). Arguments are mainly based on Bayes' theorem (i.e. $Pr(D|\mathbf{X}) = Pr(\mathbf{X}|D)Pr(D)Pr(\mathbf{X})^{-1}$) leaving the covariate distribution $Pr(\mathbf{X})$ unspecified.

An alternative approach to prove the results of Prentice and Pyke was given by Scott and Wild, 2001 who used the profile likelihood to obtain a semiparametric estimator of $\boldsymbol{\beta}$. For this purpose, they had to consider the estimator of $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^T)^T$ which can only be derived from prospective sampling. In fact, the argument of prospective sampling is only needed for the estimation of the intercept parameter $\alpha$ which will be discussed later. Thus, the semiparametric MLE of $\boldsymbol{\theta}$ will be derived in the following, where the MLE of $\alpha$ cannot formally be derived in a case-control sample and will therefore be discarded. Suppose that the $N$ observations in the case-control sample are realisations of $(D_1, \mathbf{X}_1), \ldots, (D_N, \mathbf{X}_N)$ i.i.d. random variables

with density $f(d|\bar{\mathbf{x}};\boldsymbol{\theta})g(\mathbf{x})$ where

$$f(d|\bar{\mathbf{x}};\boldsymbol{\theta}) = \left(\frac{\exp(\boldsymbol{\theta}^T\bar{\mathbf{x}})}{1+\exp(\boldsymbol{\theta}^T\bar{\mathbf{x}})}\right)^d \left(\frac{1}{1+\exp(\boldsymbol{\theta}^T\bar{\mathbf{x}})}\right)^{1-d}, \tag{3.2}$$

$d \in \{0,1\}, \bar{\mathbf{x}}^T = (1,\mathbf{x}^T) \in \mathbb{R}^{p+1}, \boldsymbol{\theta} \in \mathbb{R}^{p+1}$, and $g(\mathbf{x})$ is the density of the marginal distribution of $\mathbf{X}$ which is left unspecified. In the following, $(D,\mathbf{X})$ will be written instead of $(D_1,\mathbf{X}_1),\ldots,(D_N,\mathbf{X}_N)$ and $\mathbf{x}$ and $\bar{\mathbf{x}}$ will be used interchangeably. The first component of $\mathbf{x}$ is one whenever $\mathbf{x}$ is used in combination with $\boldsymbol{\theta}$. Then, the likelihood for the case-control study is given by

$$L(\boldsymbol{\theta},g) = \prod_{k=1}^{N} f(d_k|\mathbf{x}_k;\boldsymbol{\theta})g(\mathbf{x}_k)\pi_1^{-d_k}\pi_0^{-(1-d_k)}. \tag{3.3}$$

The probabilities $\pi_0 = Pr(D=0)$ and $\pi_1 = Pr(D=1)$ in (3.3) cannot be omitted from the likelihood function because they are related to $f$ and $g$ by:

$$\pi_i = \int f(D=i|\mathbf{x};\boldsymbol{\theta})dG(\mathbf{x}), \quad i=0,1, \tag{3.4}$$

with $G$ denoting the marginal distribution function of $\mathbf{X}$.

Since only the estimation of $\boldsymbol{\theta}$ is of interest, the profile likelihood, that treats $g$ as nuisance parameter, simplifies the parameter estimation. The profile likelihood of $L$ is defined as

$$L_P(\boldsymbol{\theta}) = \max_g L(\boldsymbol{\theta},g).$$

To maximise the profile likelihood $L_P$ or the profile log-likelihood $l_P$, the density $g$ is assumed to be discrete where the support of $G$ consists of the $M$ observed distinct values of $\mathbf{X}$, $\{\tilde{\mathbf{x}}_1,\ldots,\tilde{\mathbf{x}}_M\}$. Let $n_{im}$ denote the number of subjects with disease status $i$ and covariate value $\tilde{\mathbf{x}}_m$ and let $\delta_m$ denote the probability of $\tilde{\mathbf{x}}_m$. Using (3.4) and summing over cases and controls as well as over the distinct covariate values the log-likelihood evolves from (3.3) as follows:

$$l(\boldsymbol{\theta},\boldsymbol{\delta}) =$$

$$\sum_{k=1}^{N}\log f(d_k|\tilde{\mathbf{x}}_k;\boldsymbol{\theta}) + \sum_{k=1}^{N}\log g(\mathbf{x}_k) - \sum_{k=1}^{N} d_k\log(\pi_1) - \sum_{k=1}^{N}(1-d_k)\log(\pi_0) = \tag{3.5}$$

$$\sum_{i=0}^{1}\sum_{m=1}^{M} n_{im}\log f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta}) + \sum_{m=1}^{M} n_{+m}\log\delta_m - \sum_{i=0}^{1} N_i\log\left(\sum_{l=1}^{M} f(D=i|\tilde{\mathbf{x}}_l;\boldsymbol{\theta})\delta_l\right).$$

The profile likelihood is obtained (for fixed $\boldsymbol{\theta}$) by maximising (3.5) over $\boldsymbol{\delta}$ taking into consideration the constraint $\sum_m \delta_m = 1$ by using a Lagrange multiplier. Scott and

Wild showed (cf. Appendix A.1) that the Lagrange multiplier $\lambda$ for this constrained maximisation problem is $\lambda = 0$ and that the estimator $\hat{\delta}_m$ fulfils the equation

$$\delta_m = \tau_m(\boldsymbol{\theta}) = \frac{n_{+m}}{\sum_i \frac{N_i f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})}{\pi_i}}, \quad m = 1, \ldots, M. \tag{3.6}$$

Since the Lagrange multiplier vanishes from the optimisation problem, it is obvious that the solution of $l(\boldsymbol{\theta}, \boldsymbol{\delta}) = 0$ automatically fulfils the constraint, which has also been noted by Prentice and Pyke.

Using (3.5) and (3.6), the profile log-likelihood $l_P(\boldsymbol{\theta})$ can be transformed to

$$l_P(\boldsymbol{\theta}) = l^*(\boldsymbol{\theta}) = \sum_{k=1}^{N} \log f^*(d_k|\mathbf{x}_k;\boldsymbol{\theta}) \tag{3.7}$$

$$\text{with} \quad f^*(D = i|\mathbf{x};\boldsymbol{\theta}) = \frac{\mu_i f(D = i|\mathbf{x};\boldsymbol{\theta})}{\sum_h \mu_h f(D = h|\mathbf{x};\boldsymbol{\theta})}, \quad \mu_i = \frac{N_i}{\pi_i}, \quad i \in \{0, 1\}.$$

The function $l^*$ corresponds to the log-likelihood of a prospective *pseudo model* that would arise from data which is sampled from the joint distribution $(D, \mathbf{X})$ but is only recorded with probability $\frac{\mu_i}{N} = \frac{N_i}{N\pi_i}$ for $d = i$. It follows for the pseudo model that

$$\begin{aligned} \text{logit} f^*(D = 1|\mathbf{x};\boldsymbol{\theta}) &= \text{logit} f(D = 1|\mathbf{x};\boldsymbol{\theta}) + \log(\kappa) \\ &= \alpha + \log(\kappa) + \boldsymbol{\beta}^T\mathbf{x}, \quad \kappa = \frac{\frac{N_1}{\pi_1}}{\frac{N_0}{\pi_0}}. \end{aligned} \tag{3.8}$$

Hence, the same parameter vector $\boldsymbol{\beta}$ and an intercept $\alpha^*$ which is related to the intercept $\alpha$ in $f$ by $\alpha^* = \alpha + \log(\kappa)$ can be estimated by using the prospective model $f^*$. In practice this can be done by applying an ordinary logistic regression model. Of note, the intercept $\alpha$ can only be estimated with additional information on $\kappa$. For details on the derivation of (3.7) and (3.8) see Appendix A.2.

The MLE $\hat{\boldsymbol{\theta}}$ maximising the profile likelihood $L_P$ is the $(p+1)$-dimensional component of the solution of

$$\frac{\partial l_P(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial l^*(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}} = 0,$$

where $\boldsymbol{\Phi}^T = (\boldsymbol{\theta}^T, \kappa)$. This follows from (3.7) and because of

$$\frac{\partial l^*(\boldsymbol{\Phi})}{\partial \kappa} = 0 \quad \text{(see Appendix A.3).}$$

Remarkably, the dimension of the nuisance parameters could be reduced from $M$ for $(\delta_1, \ldots, \delta_M)$ to 1 for $\kappa$.

Using results of Gill et al., 1988, it can be shown that the above results which have been developed for a discrete distribution of $\mathbf{X}$ are also valid for continuous covariates. A detailed argumentation is given in Holubkov, 1995. The strong consistency of $\hat{\boldsymbol{\beta}}$ has been proven by Prentice and Pyke, 1979. Furthermore, Scott and Wild, 2001 demonstrated that the usual large sample results for maximum likelihood estimates also apply for $\hat{\boldsymbol{\beta}}$ by showing that $\hat{\boldsymbol{\Phi}}^T = (\hat{\boldsymbol{\theta}}^T, \hat{\kappa})$ is approximately normally distributed with expectation $\boldsymbol{\Phi}_0$ and covariance matrix

$$\mathrm{Cov}(\hat{\boldsymbol{\Phi}}) = J^*(\boldsymbol{\Phi}_0)^{-1}\mathrm{Cov}(S^*(\boldsymbol{\Phi}_0))J^*(\boldsymbol{\Phi}_0)^{-1} = J^*(\boldsymbol{\Phi}_0)^{-1} - k\mathbf{e}\mathbf{e}^T$$

where $\boldsymbol{\Phi}_0$ is the true parameter value, $k = \kappa^2\left(\frac{1}{N_1} + \frac{1}{N_0}\right)$, $\mathbf{e} = (0, \ldots, 0, 1)^T$, $S^*(\boldsymbol{\Phi}) = \frac{\partial l^*(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}}$ is the score function and $J^*(\boldsymbol{\Phi}) = -E\left(\frac{\partial^2 l^*}{\partial \boldsymbol{\Phi}\partial \boldsymbol{\Phi}^T}\right)$ is the expected pseudo information matrix. Since the expected pseudo information matrix coincides with the observed information matrix for the logistic model, the observed information matrix $I^*(\boldsymbol{\Phi}) = \frac{\partial^2 l^*}{\partial \boldsymbol{\Phi}\partial \boldsymbol{\Phi}^T}$ can be used as well.

### 3.2.2 Estimation in two-phase studies

In this section, the semiparametric maximum likelihood estimation introduced for simple case-control studies in the last section is extended to the more complex two-phase design. This will provide a theoretical basis for the parameter estimation in database studies with a case-control design augmented by additional information. In a first step, the sampling scheme from which the data arises is described and the corresponding likelihood function is derived. Based on this likelihood function, a semiparametric MLE is derived following the approach described by Scott and Wild, 2001. At the end of this section the relation between the semiparametric MLE and the pseudo likelihood estimator, which has been developed by Breslow and Cain, 1988, is established following the arguments of Scott and Wild, 1997.

**Likelihood function for the two-phase database study:** Consider a two-phase case-control study in which the phase 1 data set is obtained from a population-based claims database and the phase 2 data comprises additional information orig-

inating from another data source for a subset of subjects. The phase 1 case-control sample can then be considered as a random sample of $N_0$ controls and $N_1$ cases from a finite population of size $\mathscr{N}$ with $\mathscr{N}_0$ controls and $\mathscr{N}_1$ cases. From data of the form $(D, \mathbf{X})$, the phase 1 data set is sampled from the conditional distribution $(\mathbf{X}|D)$. Cases and controls are sampled with probability $p_i^I = Pr(R^I = 1|D = i)$ with $R^I$ being the indicator for selection into phase 1. Furthermore, assume that the stratum, $s_k$, is known for each individual $k$ with $s_k = (i, j)$ if $(d_k, \mathbf{x}_k) \in S_{ij}$ and that $(D, \mathbf{X})$ is observed for all individuals in phase 2. Let $(D, \mathbf{X})$ have the density $f(d|\mathbf{x}; \boldsymbol{\theta})g(\mathbf{x})$ as specified in the previous section. Moreover, assume that the probability of being selected for the phase 2 sample depends only on the stratum, i.e., the MAR-assumption is fulfilled with $Pr(R^{II} = 1|s_k, \mathbf{x_k}) = Pr(R^{II} = 1|s_k), \quad k = 1, \ldots, N$, where $R^{II}$ is the indicator for selection into phase 2. Let $Pr(R^{II} = 1|s_k = (i, j)) = p_{ij}^{II}$ be the selection probability for an individual $k$ in stratum $S_{ij}$. Then, it follows for the density of the observed and unobserved data for subject $k$ that

$$\begin{aligned} Pr(d_k, \mathbf{x}_k, s_k, r_k^I, r_k^{II}) &= Pr(r_k^I|d_k, \mathbf{x}_k, s_k, r_k^{II}) Pr(d_k, \mathbf{x}_k, s_k, r_k^{II}) \\ &= Pr(r_k^I|d_k) Pr(r_k^{II}|d_k, \mathbf{x}_k, s_k) Pr(d_k, \mathbf{x}_k, s_k) \\ &= Pr(r_k^I|d_k) Pr(r_k^{II}|s_k) Pr(d_k, \mathbf{x}_k, s_k), \end{aligned} \tag{3.9}$$

where $r_k^I$ and $r_k^{II}$ denote the subject-specific values of $R^I$ and $R^{II}$. Suppose that neither $p_i^I$ nor $p_{ij}^{II}$ contain any information on the parameter vector $\boldsymbol{\theta}$. Then, $p_i^I$ and $p_{ij}^{II}$ can be omitted from the likelihood function for estimating $\boldsymbol{\theta}$. The contribution to the likelihood for the $\mathscr{N}_i - N_i$ individuals not selected into phase 1 (i.e., $r_k^I = 0$) is $Pr(D = i) = \pi_i$ for $i \in \{0, 1\}$. The contribution of the $N_{ij} - n_{ij}$ subjects included in the phase 1 sample but not selected into the phase 2 sample (i.e., $r_k^{II} = 0$) is $Pr((d_k, \mathbf{x}_k) \in S_{ij})$ for $i \in \{0, 1\}$ and $j = 1, \ldots, J$. The contribution to the likelihood function for individuals in phase 2 with full covariate information is $Pr(d_k, \mathbf{x}_k) = f(d_k|\mathbf{x}_k; \boldsymbol{\theta})g(\mathbf{x}_k)$. According to (3.9) and the above, the likelihood can be written as

$$L(\boldsymbol{\theta}, g) = \prod_{i=0}^{1} \pi_i^{\mathscr{N}_i - N_i} \prod_{j=1}^{J} Pr((D, \mathbf{X}) \in S_{ij})^{N_{ij} - n_{ij}} \prod_{k=1}^{n_{ij}} f(d_k|\mathbf{x}_k; \boldsymbol{\theta})g(\mathbf{x}_k). \tag{3.10}$$

**Semiparametric likelihood estimation:** The semiparametric MLE is derived by applying the profile likelihood approach described by Scott and Wild, 2001 to the likelihood function (3.10). Compared with the likelihood function used by Scott and

Wild, the likelihood function given in (3.10) includes an additional term to account for sampling phase 1 data from a finite population.

To obtain the profile log-likelihood $l_P(\boldsymbol{\theta})$ it is again assumed that the marginal distribution function $G$ of $\mathbf{X}$ has finite support $\{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_M\}$ and that $g(\tilde{\mathbf{x}}_m) = \delta_m$. Let $n_{++m}$ denote the number of subjects with covariate value $\tilde{\mathbf{x}}_m$ in the phase 2 sample and $n_{ijm}$ the respective number of subjects in stratum $S_{ij}$. The first step is then to solve the following constrained maximisation problem for fixed $\boldsymbol{\theta}$:

$$\begin{aligned}\max_g l(\boldsymbol{\theta}, g) = \max_g \Bigg[ & \sum_{i=0}^{1} (\mathscr{N}_i - N_i) \log\Big(\sum_{m=1}^{M} f(D = i|\tilde{\mathbf{x}}_m; \boldsymbol{\theta})\delta_m\Big) \\ & + \sum_{i=0}^{1}\sum_{j=1}^{J} (N_{ij} - n_{ij}) \log\left(\sum_{m=1}^{M} Q_{ij}^*(\tilde{\mathbf{x}}_m)\delta_m\right) \\ & + \sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{m=1}^{M} n_{ijm} \log f(D = i|\tilde{\mathbf{x}}_m; \boldsymbol{\theta}) + \sum_{m=1}^{M} n_{++m} \log \delta_m \Bigg] \end{aligned} \tag{3.11}$$

with respect to the constraint $\sum_m \delta_m = 1$. The first term in (3.11) results from Equation (3.4). For simplification of the notation, the stratum probability $Q_{ij}^*$ for a given covariate vector $\tilde{\mathbf{x}}_m$ is introduced in (3.11) with

$$Q_{ij}^*(\tilde{\mathbf{x}}_m) = Pr((D, \mathbf{X}) \in S_{ij}|\tilde{\mathbf{x}}_m).$$

Note that $Q_{ij} = \sum_m Q_{ij}^*(\tilde{\mathbf{x}}_m)\delta_m$ is the marginal stratum probability for stratum $S_{ij}$. The solution of (3.11) is derived as in the previous section by introducing a Lagrange multiplier $\lambda$. It is shown in Appendix (A.4) that $\lambda = -\mathscr{N}$ and that the estimator $\hat{\boldsymbol{\delta}}$ fulfils the following system of equations

$$\begin{aligned}\delta_m = \tau_m(\boldsymbol{\theta}) = & \frac{n_{++m}}{\mathscr{N} - \sum_i \frac{\mathscr{N}_i - N_i}{\pi_i} f(D = i|\tilde{\mathbf{x}}_m; \boldsymbol{\theta}) - \sum_i \sum_j (N_{ij} - n_{ij}) \frac{Q_{ij}^*(\tilde{\mathbf{x}}_m)}{Q_{ij}}}, \\ & m = 1, \ldots, M. \end{aligned} \tag{3.12}$$

After inserting (3.12) in (3.11) and omitting additive constants the profile log-

likelihood is given by

$$
\begin{aligned}
& l_P(\boldsymbol{\theta}) \\
=& l^*(\boldsymbol{\theta}, \boldsymbol{\tau}(\boldsymbol{\theta})) \qquad (3.13) \\
=& \sum_{i=0}^{1} (\mathscr{N}_i - N_i) \log \pi_i + \sum_{i=0}^{1} \sum_{j=1}^{J} (N_{ij} - n_{ij}) \log Q_{ij} \\
&+ \sum_{i=0}^{1} \sum_{j=1}^{J} \sum_{m=1}^{M} n_{ijm} \log f(D = i | \tilde{\mathbf{x}}_m; \boldsymbol{\theta}) \\
&- \sum_{i=0}^{1} \sum_{j=1}^{J} \sum_{m=1}^{M} n_{ijm} \log \left( \mathscr{N} - \sum_{i=0}^{1} \frac{\mathscr{N}_i - N_i}{\pi_i} f(D = i | \tilde{\mathbf{x}}_m; \boldsymbol{\theta}) - \sum_{i=0}^{1} \sum_{j=1}^{J} \frac{N_{ij} - n_{ij}}{Q_{ij}} Q^*_{ij}(\tilde{\mathbf{x}}_m) \right)
\end{aligned}
$$

with $Q_{ij}$ and $\pi_i$ such that they maximise $l^*$ and fulfil

$$0 < Q_{ij} < 1, \quad \sum_{i=0}^{1} \sum_{j=1}^{J} Q_{ij} = 1$$

and

$$0 < \pi_i < 1, \quad \pi_1 + \pi_0 = 1,$$

respectively. Setting

$$\rho_{ij} = \log \left( \frac{Q_{ij}}{Q_{1J}} \right), ij \in \{0,1\} \times \{1, \ldots, J\}, ij \neq 1J, \quad \rho_{1J} = 0,$$

and solving

$$\frac{\partial l^*}{\partial \boldsymbol{\rho}} = 0$$

with respect to the $2J$-dimensional vector

$$\boldsymbol{\rho} = (\rho_{01}, \ldots, \rho_{1J-1}, \text{logit}(\pi_1))^T$$

yields an estimator that fulfils the constraints on $Q_{ij}$ and $\pi_i$ (see Appendix (A.5)). Obviously, $\boldsymbol{\rho}$ depends on $\boldsymbol{\theta}$ via $Q_{ij}$.

The second step to obtain the semiparametric MLE is to find the MLE $\hat{\boldsymbol{\theta}}$ which maximises $l^*(\boldsymbol{\theta}, \boldsymbol{\tau}(\boldsymbol{\theta}))$ or equivalently $l^*(\boldsymbol{\theta}, \boldsymbol{\rho}(\boldsymbol{\theta}))$. Seber and Wild (1989, Theorem 2.2, pp. 39) have shown that the MLE $\hat{\boldsymbol{\theta}}$ can be obtained as solution $\hat{\boldsymbol{\Phi}} = (\hat{\boldsymbol{\theta}}^T, \hat{\boldsymbol{\rho}}(\boldsymbol{\theta})^T)^T$ of

$$\frac{\partial l^*(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}} = 0,$$

if $\hat{\boldsymbol{\rho}}(\boldsymbol{\theta})$ solves

$$\frac{\partial l^*(\boldsymbol{\theta}, \boldsymbol{\rho}(\boldsymbol{\theta}))}{\partial \boldsymbol{\rho}} = 0$$

for any fixed $\boldsymbol{\theta}$.

Moreover, the information of the profile log-likelihood $I_P$ is given by

$$I_P = -\left(\frac{\partial^2 l_P}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right) = -\left(\frac{\partial^2 l^*(\boldsymbol{\theta}, \boldsymbol{\rho})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} - \frac{\partial^2 l^*(\boldsymbol{\theta}, \boldsymbol{\rho})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\rho}^T}\left(\frac{\partial^2 l^*(\boldsymbol{\theta}, \boldsymbol{\rho})}{\partial \boldsymbol{\rho} \partial \boldsymbol{\rho}^T}\right)^{-1} \frac{\partial^2 l^*(\boldsymbol{\theta}, \boldsymbol{\rho})}{\partial \boldsymbol{\rho} \partial \boldsymbol{\theta}^T}\right).$$

Using the standard rule for the inverse of a partitioned matrix, the inverse profile information $I_P^{-1}$ can be derived from the $(p+1) \times (p+1)$ submatrix of $I(\boldsymbol{\Phi})^{-1}$ where

$$\begin{aligned} I(\boldsymbol{\Phi}) &= -\frac{\partial^2 l^*}{\partial \boldsymbol{\Phi} \partial \boldsymbol{\Phi}^T} \\ &= \begin{pmatrix} I_{\boldsymbol{\theta}\boldsymbol{\theta}} & I_{\boldsymbol{\theta}\boldsymbol{\rho}} \\ I_{\boldsymbol{\rho}\boldsymbol{\theta}} & I_{\boldsymbol{\rho}\boldsymbol{\rho}} \end{pmatrix}. \end{aligned}$$

If $I(\boldsymbol{\Phi})$ is positive definite at $\hat{\boldsymbol{\Phi}}$, $l^*$ has a local maximum at $\hat{\boldsymbol{\Phi}}$. However, Scott and Wild, 2001 noted that this assumption is often violated in practice. As a solution they proposed to maximise the profile likelihood function directly by applying the Newton-Raphson algorithm to solve the profile likelihood for $\boldsymbol{\theta}$ with an inner Newton-Raphson loop that maximises the profile likelihood with respect to $\boldsymbol{\rho}(\boldsymbol{\theta})$.

**Relation to the pseudo likelihood:** A pseudo model can be established for the two-phase case-control situation analogously to the simple case-control situation. For simplification, the first term of the likelihood function (3.10) will be ignored in the following. This will only affect the estimation of the intercept parameter because the $\mathscr{N}_i - N_i$ subjects not selected for the case-control sample only contribute disease information but no information about the association between disease and covariates. Calculations along the lines of Appendix A.4 yield $\lambda = N$. As a consequence, $\hat{\boldsymbol{\delta}}(\boldsymbol{\theta})$ solves the following system of equations:

$$\begin{aligned} \delta_m = \tau_m(\boldsymbol{\theta}) &= \frac{n_{++m}}{N \sum_i \sum_j \mu_{ij} Q_{ij}^*(\tilde{\mathbf{x}}_m)} \\ \text{with } \mu_{ij} &= 1 - \frac{N_{ij} - n_{ij}}{N Q_{ij}}, \quad m = 1, \ldots, M. \end{aligned} \tag{3.14}$$

If $\mathbf{X}$ fully determines $S_{ij}$ for $j = 1, \ldots, J$, which is true for the considered study situation, $Q_{ij}^*(x)$ in (3.14) can be substituted by $f(D = i|\mathbf{x}; \boldsymbol{\theta})$ for $(D, \mathbf{X}) \in S_{ij}$.

With this substitution and following the arguments given in Scott and Wild, 1997 a pseudo model $f^*$ can be defined as

$$f^*(D=i|\mathbf{x};\boldsymbol{\theta}) = \frac{\mu_{ij} f(D=i|\mathbf{x};\boldsymbol{\theta})}{\sum_i \sum_j \mu_{ij} f(D=i|\mathbf{x};\boldsymbol{\theta})}. \tag{3.15}$$

Furthermore, it follows that

$$\begin{aligned}
\frac{\mu_{ij}}{1-\mu_{ij}} &= \frac{\mu_{ij}}{\frac{N_{ij}-n_{ij}}{NQ_{ij}}} = \mu_{ij}\left(\frac{N_{ij}-n_{ij}}{N\sum_m Q^*_{ij}(\tilde{\mathbf{x}}_m)\delta_m}\right)^{-1} \\
&= \frac{\mu_{ij}}{N_{ij}-n_{ij}}\left(N\sum_m Q^*_{ij}(\tilde{\mathbf{x}}_m)\frac{n_{++m}}{N\sum_i\sum_j \mu_{ij}Q^*_{ij}(\tilde{\mathbf{x}}_m)}\right) \\
&= \frac{\sum_l n_{++l} f^*(D=i|\tilde{\mathbf{x}}_l;\boldsymbol{\theta})}{N_{ij}-n_{ij}}
\end{aligned} \tag{3.16}$$

where the index $l$ is such that $\tilde{\mathbf{x}}_l \in S_{ij}$. Then, $\mu_{ij}$ can be rewritten as

$$\mu_{ij} = \frac{n_{ij}-\gamma_{ij}}{N_{ij}-\gamma_{ij}} \text{ with } \gamma_{ij} = n_{ij} - \sum_l n_{++l} f^*(D=i|\tilde{\mathbf{x}}_l;\boldsymbol{\theta}). \tag{3.17}$$

Since the parameters $\gamma_{ij}$ are the differences between the observed stratum counts and the expected stratum counts under the pseudo model, they can be interpreted as measures for the model fit of the pseudo model.

Using (3.14), (3.15) and (3.16), the profile log-likelihood can be written as (see Appendix A.6)

$$\begin{aligned}
\tilde{l}_P(\boldsymbol{\theta}) &= \tilde{l}^*(\boldsymbol{\theta}, \boldsymbol{\gamma}(\boldsymbol{\theta})) \\
&= \sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{m=1}^{M} n_{ijm} \log f^*(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta}) - \sum_{i=0}^{1}\sum_{j=1}^{J} n_{ij}\log(n_{ij}-\gamma_{ij}) \\
&+ \sum_{i=0}^{1}\sum_{j=1}^{J} N_{ij}\log(N_{ij}-\gamma_{ij}).
\end{aligned} \tag{3.18}$$

Note, that $\tilde{l}_P$ in (3.18) and $l_P$ in (3.13) differ by the term which is related to the disease probability in the finite population.

Scott and Wild, 1997 have shown that when solving

$$\frac{\partial \tilde{l}^*(\boldsymbol{\theta}, \boldsymbol{\gamma}(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = 0$$

all parts including the partial derivatives of $\boldsymbol{\gamma}$ vanish. Thus, the maximum likelihood estimator of $\tilde{l}_P$ corresponds to the maximum likelihood estimator of the pseudo model, i.e. it solves

$$\frac{\partial}{\partial \boldsymbol{\theta}} \sum_{i=0}^{1} \sum_{j=1}^{J} \sum_{m=1}^{M} \log f^*(D = i|\tilde{\mathbf{x}}_m; \boldsymbol{\theta}) = 0.$$

Furthermore, $f^*$ has a logistic form with

$$f^*(d|\mathbf{x}; \boldsymbol{\theta}) = \left( \frac{\exp\left(\log\left(\frac{\mu_{1j}}{\mu_{0j}}\right) + \boldsymbol{\theta}^T \mathbf{x}\right)}{1 + \exp\left(\log\left(\frac{\mu_{1j}}{\mu_{0j}}\right) + \boldsymbol{\theta}^T \mathbf{x}\right)} \right)^d \left( \frac{1}{1 + \exp\left(\log\left(\frac{\mu_{1j}}{\mu_{0j}}\right) + \boldsymbol{\theta}^T \mathbf{x}\right)} \right)^{1-d}.$$

To calculate the maximum likelihood estimator in practice, Scott and Wild proposed the following algorithm: Starting with fixed values $\hat{\mu}_{ij}^0$ the pseudo model is fitted with respect to $\boldsymbol{\theta}$. Using the estimator $\hat{\boldsymbol{\theta}}^a$, updated values $\hat{\mu}_{ij}^a$ are calculated by the inner loop and the loop starts with fitting the pseudo model with the updated values to estimate $\hat{\boldsymbol{\theta}}^{a+1}$. If $\hat{\mu}_{ij}^0 = \frac{n_{ij}}{N_{ij}}$ are used as starting values, the estimator of the first iteration yields the pseudo likelihood estimator which has been proposed by Breslow and Cain, 1988. Details on the implementation of methods for the parameter estimation in two-phase studies in current software packages are given in Section 3.4.

**Large sample theory:** The estimation procedure described in this section for covariates with discrete distributions can be extended to continuous covariates with the same arguments that were raised for simple case-control studies. Since the standard constrained ML theory (Aitchison and Silvey, 1958) cannot be applied in this setting, the asymptotic properties of $\hat{\boldsymbol{\theta}}$ have to be derived specifically for continuous covariates. As already mentioned for the simple case-control study, the intercept $\alpha$ cannot be estimated without bias when using the usual parametrisation. Since the intercept is of minor interest, only the large sample results for the parameter vector $\boldsymbol{\beta}$ are considered here. The strong consistency of $\hat{\boldsymbol{\beta}}$ has been proven by Vaart and Wellner, 2001. Breslow et al., 2003 showed for two-phase data arising from Bernoulli sampling that $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is asymptotically normal with mean $\mathbf{0}$ and variance $I_P^{-1}$, where $\boldsymbol{\beta}_0$ is the true parameter value. Moreover, the semiparametric MLE $\hat{\boldsymbol{\beta}}$ was demonstrated to be fully efficient because its variance achieves the asymptotic information bound for the two-phase case-control sampling scheme. Lee, 2007 extended

these results to more general sampling schemes allowing samples also to be drawn from different populations.

**Remarks on the assumptions for two-phase methods:** Two assumptions are usually imposed in the literature concerning two-phase methods: the missing-at-random (MAR) assumption and the conditional independence assumption. The MAR assumption supposes that the selection of subjects into the phase 2 sample only depends on known covariates. This is also postulated for the above described scenario because subjects are selected randomly from each stratum for phase 2 and therefore $Pr(\mathbf{x}_k|d_k, s_k, R_k^{II} = 1) = Pr(\mathbf{x}_k|d_k, s_k, R_k^{II} = 0)$. The conditional independence assumption, which is given by $Pr(d_k|s_k, \mathbf{x}_k) = Pr(d_k|\mathbf{x}_k)$, is always fulfilled in the considered setting because it is assumed that the covariate vector $\mathbf{x}_k$ fully determines the stratum membership. This assumption is fulfilled if at least those covariates which have been used to build the stratification are included in the regression model. The functional relation used for modelling $\mathbf{x}_k$ in $f$ does not have an impact on the validity of the assumption, e.g., age may be included as continuous variable in $f$ although only age groups are considered in the stratification.

## 3.3 Further estimation procedures

Several estimation procedures have been proposed for the parameter estimation in two-phase studies, one of which, the profile likelihood approach, has been described in the previous section. Another semiparametric MLE was developed by Breslow and Holubkov, 1997a who derived the score equation directly by solving a maximisation problem with several constraints. Since both semiparametric MLEs have been shown to be equivalent (Scott et al., 2007), the Breslow-Holubkov estimator is not described in detail here. A computational approach to obtain the MLE from two-phase case-control data is based on the application of an Expectation-Maximisation(EM)-algorithm to a Poisson likelihood which has been suggested by Schill and Drescher, 1997. This approach will be briefly outlined in this section. Furthermore, the WL approach will be introduced which is less efficient than the MLE but will nevertheless be used for the derivation of design criteria in Chapter 4 due to the simplicity of its covariance formula.

### 3.3.1 ML estimation via the EM-algorithm

The EM-algorithm is an iterative numerical approach that can be used to determine maximum likelihood estimators from data which has not been completely observed. Its application to the incomplete phase 1 data of a two-phase study as suggested by Schill and Drescher, 1997 is described in the following.

Consider that the vector of covariates $\mathbf{X}$ assumes a finite number of values, $\mathbf{x}_{j1}, \ldots, \mathbf{x}_{jM_j}, \quad j = 1, \ldots, J$, in each stratum of the two-phase study. Then, phase 1 and phase 2 data can be classified into cells $(i, j, m)$, $i = 0, 1$, $j = 1, \ldots, J$, $m = 1, \ldots, M_j$, where $i$ denotes the disease status, $j$ denotes the stratum, and $m$ denotes the covariate value. The number $N_{ijm}$ of subjects in cell $(i, j, m)$ of the phase 1 data set is unobserved whereas the respective number $n_{ijm}$ in the phase 2 data set as well as the stratum counts $N_{ij}$ in phase 1 are observed. We assume that $N_{ijm}$ is Poisson distributed with expectation

$$\mu_{ijm} = \begin{cases} \exp(\zeta_{jm} + \alpha + x_{jm}^T \boldsymbol{\beta}) & \text{if } i = 1 \\ \exp(\zeta_{jm}) & \text{if } i = 0 \end{cases} . \tag{3.19}$$

Then, the EM-algorithm can be applied to estimate the parameter vector $\boldsymbol{\theta}_P = (\boldsymbol{\zeta}^T, \alpha, \boldsymbol{\beta}^T)^T$, $\boldsymbol{\zeta} = (\zeta_{11}, \ldots, \zeta_{JM_J})^T$. In the E-step of the algorithm, the unobserved cell counts $N_{ijm}$ are replaced by their expected values given the observed data, i.e., by $\hat{N}_{ijm} = \mathrm{E}[N_{ijm} | n_{ijm}, N_{ij}]$. The M-step then estimates $\boldsymbol{\theta}_P$ by maximising the likelihood of model (3.19) on the completed data.

Using that $\tilde{N}_{ijm} = N_{ijm} - n_{ijm}$ given the stratum counts $(N_{ij}, n_{ij})$ is multinomially distributed, i.e., $\tilde{N}_{ijm} | (N_{ij}, n_{ij}) \sim Mult(N_{ij} - n_{ij}, (\frac{\mu_{ij1}}{\mu_{ij+}}, \ldots, \frac{\mu_{ijM_j}}{\mu_{ij+}}))$, the expected cell counts can be written as

$$\hat{N}_{ijm} = n_{ijm} + (N_{ij} - n_{ij}) \frac{\hat{\mu}_{ijm}}{\hat{\mu}_{ij+}},$$

where $\hat{\mu}_{ijm}$ is calculated by inserting $\hat{\boldsymbol{\theta}}_P$ in (3.19). Iteration of the E- and M-step yields the MLE.

Scott and Wild, 1991 have shown that the MLE $\hat{\boldsymbol{\beta}}$ resulting from the Poisson model (3.19) coincides with the MLE of a logistic model. Hence, the EM estimator is equal to the MLE obtained from the profile likelihood approach discussed in Section 3.2.2.

### 3.3.2 WL estimation

The weighted likelihood method has been proposed for the parameter estimation in two-phase studies by Flanders and Greenland, 1991. An equivalent approach, the mean-score method, has been derived a few years later by Reilly and Pepe, 1995. Here, the description of the method is based on the results of Reilly and Pepe.

Since the intercept $\alpha$ is not of much interest, only the estimation procedure for the parameter vector $\boldsymbol{\beta}$ is considered in this section. Reilly and Pepe suggested to solve the following estimating equation to obtain estimates for $\boldsymbol{\beta}$:

$$\sum_{q\in\{k:r_k^{II}=1\}} \frac{\partial}{\partial\boldsymbol{\beta}} \log f(d_q|\mathbf{x}_q;\boldsymbol{\beta}) + \sum_{q'\in\{k:r_k^{II}=0\}} \hat{E}\left(\frac{\partial}{\partial\boldsymbol{\beta}} \log f(d_{q'}|\mathbf{X}) \;\middle|\; d_{q'}, s_{q'}\right), \qquad (3.20)$$

where $r_k^{II}$ is the indicator for selection into phase 2 as specified in Section 3.2.2 and $\hat{E}\left(\frac{\partial}{\partial\boldsymbol{\beta}} \log f(d_{q'}|\mathbf{X}) \;\middle|\; d_{q'}, s_{q'}\right)$ is calculated as

$$\hat{E}\left(\frac{\partial}{\partial\boldsymbol{\beta}} \log f(d_{q'}|\mathbf{X}) \;\middle|\; d_{q'}, s_{q'}\right) = \sum_{r\in\{k:r_k^{II}=1, s_k=s_{q'}\in S_{ij}\}} \frac{1}{n_{ij}} \frac{\partial}{\partial\boldsymbol{\beta}} \log f(d_{q'}|\mathbf{x}_r;\boldsymbol{\beta}).$$

Since complete covariate information is only available for subjects included in the phase 2 sample, the estimation equation is divided into two parts. One part corresponds to the usual score contribution of subjects with complete covariate information and the other part corresponds to the contribution of subjects without complete covariate information. For subjects without complete covariate information, the score contribution is estimated from subjects in the same stratum who have complete covariate information, i.e., the mean score contribution of phase 2 subjects in the specific stratum is considered for these subjects. Equation (3.20) corresponds to the respective estimating equation arising from the weighted likelihood

$$L_W = \prod_{i=0}^{1} \prod_{j=1}^{J} \prod_{k=1}^{n_{ij}} f(d_k|\mathbf{x}_k;\boldsymbol{\beta})^{\frac{N_{ij}}{n_{ij}}}.$$

Reilly and Pepe have proven that the estimator $\hat{\boldsymbol{\beta}}$ solving (3.20) is strongly consistent for the true parameter value $\boldsymbol{\beta}_0$ and that $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is asymptotically normal with mean $\mathbf{0}$ and variance

$$I_R^{-1} + I_R^{-1}\Omega I_R^{-1}, \qquad (3.21)$$

where

$$I_R = E\left(\frac{-\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\log f(D|\mathbf{X};\boldsymbol{\beta})\right)$$

and

$$\Omega = \sum_{i=0}^{1}\sum_{j=1}^{J}\frac{Q_{ij}(1-p_{ij}^{II})}{p_{ij}^{II}}\,\mathrm{Cov}\left(\frac{\partial}{\partial\boldsymbol{\beta}}\log f(D|\mathbf{X};\boldsymbol{\beta})\;\Big|\;(D,\mathbf{X})\in S_{ij}, R^{II}=1\right). \quad (3.22)$$

The variance formula is further discussed in Section 4.

The WL estimator has been shown to be less efficient than the MLE (see e.g., Schill and Drescher, 1997; Breslow and Holubkov, 1997b; Breslow and Chatterjee, 1999). The loss in efficiency is particularly high if the sampling fractions vary substantially between the strata.

## 3.4 Implementation in R and SAS software

As mentioned in the introduction, applications of the two-phase design to case-control data are rare. One reason for the reluctance to conduct logistic two-phase studies might be that two-phase methods have not been implemented in standard statistical software like R and SAS until recently. During the last years, however, the following well-documented and easy-to-use R and SAS packages became available for the analysis of two-phase studies.

**The `survey` package:** Lumley, 2004 released the R package `survey` which provides functions for the analysis of complex survey samples. Among these, the function `twophase` allows the handling of two-phase designs. The Horvitz-Thompson estimator (Horvitz and Thompson, 1952) which is equivalent to the WL estimator is implemented in this package. The covariance delivered by the `survey` package differs slightly from that of the WL method by a correction factor that accounts for sampling from a finite population.

**The `osDesign` package:** More recently, Haneuse et al., 2011 published the R package `osDesign` that is designated to the analysis of two-phase case-control studies. WL, PL, and ML methods are included in this package. The ML approach is based on Breslow and Holubkov, 1997a who derived an estimation equation for a

constrained ML problem, where the constraints are imposed for each phase of retrospective sampling, i.e. there are phase 1 and phase 2 constraints. Unfortunately, in specific data constellations the implementation of the ML method reveals estimators which do not fulfil all of the constraints. For instance, when analysing the example data set "Ohio" used by Haneuse et al., 2011 with a model including only sex and race as dependent variables a warning is produced instead of ML estimators. A further disadvantage of `osDesign` is that only empirical but no model-based standard errors are provided for the WL estimators. Nevertheless, in many data constellations the package provides valid estimators in short computing time, hence being suitable for simulation studies. The `osDesign` package was used for parameter estimation in the two phase analyses described in Sections 2.3 and 4.1.

**The `sas-twophase-package`:** The `sas-twophase-package` by Schill et al., 2013 is a SAS-based software package for the analysis of two-phase studies that is available from the BIPS homepage (`http://www.bips.uni-bremen.de/sastwophase`). It is a collection of SAS-macros providing WL, PL, and ML estimators and corresponding model-based standard errors. ML estimators are calculated by using the EM-algorithm of Section 3.3.1. The EM-algorithm is unfortunately associated with a long computing time and may also fail due to lack of memory for large phase 2 data sets (e.g., n=10,000). A detailed discussion on the performance of the EM-algorithm is given in Schill et al., 2013. All two-phase analyses presented in Sections 4.2 and 4.5 were conducted with an earlier version of the `sas-twophase-package` which also included a SAS-implementation of the Breslow-Holubkov estimator as programmed in the `osDesign` package.

Although the currently available software is sufficient to analyse single two-phase studies there is still a lack of software which is suitable for ML estimation in simulation studies. More recent methodological development concentrated on efficient computing of semiparametric estimators in two-phase studies based on the profile likelihood approach introduced in Section 3.2 (see Scott and Wild, 2006; Hirose and Lee, 2012). The R-based package `missreg`, in which *Description of the "missreg" library* implemented these methods, is unfortunately not available at the Comprehensive R Archive Network (CRAN, `http://cran.r-project.org`) and has not the appropriate format for packages in current versions of R. Nevertheless, the package and a detailed documentation can be downloaded from Wild's homepage

(`http://www.stat.auckland.ac.nz/~wild/software.html`). Since the `missreg` package has been designed more generally for the analysis of data arising from selective sampling or data with partially missing information using arbitrary regression models, this software requires comprehensive statistical knowledge of the underlying methods.

# Chapter 4

# Stratification strategies for the efficient use of phase 1 information

Stratification according to phase 1 variables is the key element of a two-phase design. This follows immediately from the estimation methods described in the previous chapter. In two-phase analyses, stratification fulfils two purposes: First, stratification incorporates information on the distribution of phase 1 covariates, estimated in the large phase 1 data set, in the two-phase analysis. This leads to reduced standard errors of the corresponding parameter estimators. Second, stratification is able to account for the intended and unintended selection processes which then eliminates the bias introduced by the selective sampling of the phase 2 sample. A selection process is considered to be intended if it is controlled by the study investigator who assigns selection probabilities for subjects with specific characteristics. In this thesis, intended selection processes are equivalent to the concept of a priori stratification which has been introduced in Chapter 1. Unintended selection processes are not under control of the study investigator but may occur in many applications. Examples for unintended selection processes are non-response in surveys or self-selection regarding the participation in special programmes such as Disease Management Programmes (DMPs) offered by the German statutory health insurances. An example of an unintended selection process has been described for the empirical study in Chapter 2.3. In this study, age and sex were identified from the set of phase 1 variables as predictors for the selection into phase 2. These predictors were included

in a post stratification in order to reduce the bias introduced by the selective sampling. Using post stratifications (see Chapter 1 for an explanation of this concept) is reasonable to account for unintended selection processes and can also be useful to consider additional covariate information from phase 1 to improve efficiency of the parameter estimation. The focus of this chapter is on the use of stratifications in the sense of their first purpose, namely for the reduction of standard errors in two-phase database studies.

A special feature of two-phase database studies is that the phase 1 data set comprises information on a multitude of covariates which could be included in the stratification. This is in contrast to traditional two-phase studies where the data originates from field studies and the ascertainment of each phase 1 variable is associated with additional costs. Hence, only few phase 1 covariates are available in these studies and the stratification is usually simply defined by a cross-classification of the phase 1 covariates. Applying the same approach to two-phase database studies is impossible because cross-classification of all available phase 1 covariates would lead to a tremendously high number of strata. Since two-phase methods require non-empty strata in phase 2, the number of strata is limited by the size and composition of the phase 2 data set. As a consequence, the stratification needs to be defined in view of the following dilemma: On the one hand, consideration of each additional covariate in the cross-classification increases the number of strata and therefore the risk of empty cells. On the other hand, ignoring information on phase 1 covariates results in unnecessarily large standard errors of the corresponding parameter estimators.

This chapter addresses potential solutions of this dilemma by illustrating it in the empirical study where cross-classification by additional covariates results in empty cells but ignoring these covariates precludes the estimation of precise covariate effects. In the second section, an alternative stratification strategy is proposed which uses percentiles of a disease score. The alternative strategy is applied to the empirical study and is further evaluated in a simulation study based on the empirical study. The remainder of the chapter deals with the planning of an efficient stratification. For this purpose, a criterion for the comparison of stratifications with respect to their efficiency is derived from the covariance formula of the WL approach. This criterion is then applied to the empirical study to identify candidates for efficient stratifications which are subsequently evaluated in a simulation study.

## 4.1 Limitations of cross-classification in the empirical study

The two-phase study of Chapter 2.3 on the one hand revealed a precise estimator for the parameter of main interest, i.e., for phenprocoumon exposure, but on the other hand failed to provide meaningful risk estimators for the comorbidities hypertension and diabetes mellitus. The post stratification which was used in the two-phase analysis, in the following denoted by stratification A, did neither include phase 1 information on hypertension nor on diabetes mellitus but was defined by cross-classification of age, sex, and phenprocoumon exposure only (cf. Chapter 2.3). Therefore, the two-phase analysis is now repeated with two alternative post stratifications, one including additional information on hypertension (stratification B) and the other one including additional information on diabetes mellitus (stratification C). In this section, results of these additional two-phase analyses are presented to clarify the aforementioned dilemma which exists for the application of cross-classification if many phase 1 covariates are available. Hence, the focus is on the problem of empty strata and on the efficiency of parameter estimation for hypertension and diabetes. Further results are given in Behr et al., 2012 (Appendix D).

Table 4 in Appendix D illustrates the problem of sparsely populated strata by showing the number of subjects in each cell of the phase 1 and phase 2 data set stratified by post stratifications A, B, and C. Overall, only few cases are exposed to phenprocoumon. Thus, the six subgroups of exposed cases defined by age and sex in stratification A have cell counts ranging from 2 to 7 in the phase 2 data set. Further subclassification of these sparsely populated cells would certainly result in empty cells as can be seen for the example of hypertension in Figure 4.1. Four cells in the youngest age group are empty in phase 2 and need to be collapsed with other cells before two-phase methods can be applied. Stratification B is therefore constructed by combining the cells of exposed male subjects in the youngest age group for cases and controls as well as the cells of exposed female subjects in the same age group (see Table 4 in Appendix D). That means that information on hypertension is ignored in these cells. Furthermore, the distinction in male and female is neglected in the youngest age group of cases and controls with a diagnosis of hypertension but without exposure to phenprocoumon. There are many alternative

ways to collapse the four cells. The rationale for the chosen one is to combine the smallest strata. A similar approach is used to define stratification C, where six strata have to be collapsed to avoid empty cells. Since hypertension and diabetes mellitus are the most prevalent comorbidities (see Table 2 in Appendix D) even more empty cells are expected when subdividing the cells in stratification A by any other covariate. Inclusion of two, three or more phase 1 covariates in the stratification would only add information to the most frequent cells, e.g., to controls who are not exposed to phenprocoumon, because sparsely populated cells cannot be subclassified any further.

**Number of cases in phase 1 (phase 2)**

| Exposed | Male | | Female | |
|---|---|---|---|---|
| | Hyp - | Hyp + | Hyp + | Hyp - |
| < 50 years | 7 (2) | 2 (0) | 1 (0) | 3 (2) |
| 50 - < 65 years | 20 (5) | 8 (2) | 10 (3) | 6 (1) |
| >= 65 years | 12 (2) | 16 (3) | 20 (3) | 13 (3) |

| Not Exposed | Male | | Female | |
|---|---|---|---|---|
| | Hyp - | Hyp + | Hyp + | Hyp - |
| < 50 years | 196 (12) | 23 (2) | 17 (0) | 123 (8) |
| 50 - < 65 years | 183 (16) | 76 (8) | 92 (16) | 135 (18) |
| >= 65 years | 84 (13) | 64 (6) | 68 (20) | 82 (11) |

**Number of controls in phase 1 (phase 2)**

| Exposed | Male | | Female | |
|---|---|---|---|---|
| | Hyp - | Hyp + | Hyp + | Hyp - |
| < 50 years | 11 (2) | 5 (1) | 1 (0) | 21 (5) |
| 50 - < 65 years | 44 (11) | 30 (9) | 19 (5) | 22 (8) |
| >= 65 years | 41 (9) | 44 (15) | 25 (9) | 23 (8) |

| Not Exposed | Male | | Female | |
|---|---|---|---|---|
| | Hyp - | Hyp + | Hyp + | Hyp - |
| < 50 years | 8,045 (27) | 289 (3) | 313 (1) | 7,143 (42) |
| 50 - < 65 years | 2,614 (45) | 717 (19) | 795 (18) | 2,204 (47) |
| >= 65 years | 852 (19) | 443 (11) | 632 (17) | 887 (17) |

**Figure 4.1:** Cross-classification by age, sex, and hypertension (Hyp+/Hyp-) according to post stratification B.

For a more theoretical comparison of the stratifications, the efficiency of parameter estimation is measured by the square root of the asymptotic relative efficiency (ARE) which is defined for stratification $\mathscr{S}$ and covariate $k$ by

$$EFF_{\mathscr{S},k} = \sqrt{ARE} = \frac{SE(\hat{\beta}_{1,k})}{SE(\hat{\beta}_{II\mathscr{S},k})},$$

where $SE(\hat{\beta}_{1,k})$ and $SE(\hat{\beta}_{II\mathscr{S},k})$ denote the standard errors of parameter estimates based on the phase 1 analysis and the two-phase analysis, respectively. The estimated efficiency for each estimated parameter is compared for post stratifications A, B, and C in Table 4.1. The corresponding estimated ORs and 95% confidence intervals are given in Table 6 in Appendix D. There is no notable difference between the stratifications with respect to the efficiency of parameter estimates for

phenprocoumon, age, sex, and the two-way interactions between phenprocoumon exposure and age and sex. The efficiency of the parameter estimates for hypertension and diabetes is more than doubled if the respective covariate is considered in the stratification. The precision gained with post stratification B for the estimation of the bleeding risk associated with hypertension and with post stratification C for diabetes mellitus leads to statistically significant results for these parameters.

**Table 4.1:** Efficiency of two-phase estimators for different post stratifications

| **Multivariable model**[a] | **Efficiency** | | |
|---|---|---|---|
| | Stratification A[b] $\frac{\widehat{SE}(\hat{\beta}_{1,k})}{\widehat{SE}(\hat{\beta}_{II\mathscr{S}_A,k})}$ | Stratification B[c] $\frac{\widehat{SE}(\hat{\beta}_{1,k})}{\widehat{SE}(\hat{\beta}_{II\mathscr{S}_B,k})}$ | Stratification C[d] $\frac{\widehat{SE}(\hat{\beta}_{1,k})}{\widehat{SE}(\hat{\beta}_{II\mathscr{S}_C,k})}$ |
| Phenprocoumon exposure | 0.64 | 0.69 | 0.68 |
| Age (centred at 55 years) | 0.33 | 0.33 | 0.33 |
| *Interaction:* phen. * age | 0.81 | 0.81 | 0.81 |
| Female sex | 0.72 | 0.72 | 0.75 |
| *Interaction:* phen. * sex | 0.92 | 0.93 | 0.90 |
| **Comorbid conditions:** | | | |
| Diabetes mellitus | 0.34 | 0.35 | **0.75** |
| Hypertension | 0.33 | **0.70** | 0.34 |

[a] The multivariable model includes also the variables smoking, BMI, and the interaction between BMI and phenprocoumon for which no efficiency can be calculated since these have only been collected in phase 2.

[b] Stratification A includes information on age, sex, and phenprocoumon exposure.

[c] Stratification B includes information on age, sex, phenprocoumon exposure, and hypertension.

[d] Stratification C includes information on age, sex, phenprocoumon exposure, and diabetes mellitus.

These results demonstrate that the choice of variables of which phase 1 information is considered in the two-phase analysis has a substantial impact on the efficiency of parameter estimates. However, also the problem of empty strata, that occurs with cross-classification for covariate combinations of low prevalence in small phase 2 samples, becomes obvious. Defining stratifications by cross-classification of covariates only works well for a small number of covariates if the covariate distribution is balanced or if the phase 2 sample is large. Collapsing strata of low prevalence is an option to avoid empty strata. A disadvantage of this option is that the process

of choosing strata to be combined is laborious, data-driven, and somehow arbitrary.

The stated dilemma for cross-classification is the motivation for the development of alternative stratification strategies which include the maximum amount of phase 1 information without increasing simultaneously the number of strata with each additional covariate. Such an alternative strategy is proposed in the next section.

## 4.2 Alternative stratification strategy: Using a disease score for stratification

A new stratification strategy is proposed and evaluated in a paper which will be submitted for publication to a methodological epidemiology journal (Behr and Schill, 2013; Appendix E). The concept of stratifying on disease scores is introduced and results of the paper are briefly summarised in this section.

### 4.2.1 Background

The requirement for the new stratification approach is that information on all available phase 1 covariates can be included without inducing a large number of strata. A similar problem exists for the estimation of adjusted risks for rare events if a multivariable model including a large number of covariates has to be fit to data with only few events. Scores like the Charlson index (Charlson et al., 1987) or the Chronic Disease Score (Korff et al., 1992), which summarise information on several covariates in one value, have been used for confounder adjustment in this situation. The idea of using a score to adjust for imbalances of the covariate distribution between treatment groups has also been employed in the propensity score approach (Rosenbaum and Rubin, 1983). Based on this rationale, the idea of constructing a score as a summary measure of several covariates will be exploited to solve the stratification problem in two-phase studies.

### 4.2.2 Definition of stratifications based on disease scores

Let $\mathbf{Y} = (Y_1, \ldots, Y_{p'})^T$ denote the vector of all available phase 1 covariates and let $\mathbf{Y}_I$ denote a regression vector consisting of components of $\mathbf{Y}$ as well as of product terms between components of $\mathbf{Y}$, which represent main effects and interactions in a regression model. A disease score is defined as the probability of disease given the covariate vector, i.e., $Pr(D = 1|\mathbf{Y}_I)$. The disease score (DSC) is calculated for each subject $h$ in the phase 1 sample by

$$DSC_h = \frac{\exp(\hat{\alpha}_I + \hat{\boldsymbol{\beta}}_I^T \mathbf{y}_{Ih})}{1 + \exp(\hat{\alpha}_I + \hat{\boldsymbol{\beta}}_I^T \mathbf{y}_{Ih})}, \tag{4.1}$$

where the parameters $\alpha_I$ and $\boldsymbol{\beta}_I$ are estimated in a logistic regression model based on the full phase 1 sample. To define a stratification, the range of disease scores is split into subclasses. Choice of the number and placement of the cut-points influences the size of the strata and the amount of information on phase 1 covariates which is included in the stratification. A stratification with a small number of wide strata will include less information than a stratification with a large number of narrow strata. Narrow strata, however, can lead to empty strata for cases or for controls because cut-points are defined by percentiles of the marginal distribution of disease scores but the conditional distribution in cases $(DSC|D = 1)$ may differ from the conditional distribution in controls $(DSC|D = 0)$. As an example, consider the distribution of disease scores estimated in the empirical study depicted in Figure E.1 in Appendix E. The marginal distribution is dominated by the conditional distribution of controls due to a case:control ratio of 1:20. It is thus possible that values below the 50th percentile of the marginal distribution are less frequent in cases. Another reason for empty strata in cases or controls might be a lack of overlap in disease scores. The overlap in the empirical example is large. In studies where the phase 1 covariates $\mathbf{Y}$ discriminate well between cases and controls the overlap may be very small. Although the same problem arises with the propensity score approach, arguments raised for instance by Myers and Louis, 2007 cannot be applied to stratification in two-phase designs because the aim for using the score is different. Whereas the propensity score is used as a balancing score to allow for an unbiased comparison of two groups, the disease score is used as a stratification variable which is correlated with each relevant phase 1 variable. In this context, relevance is measured by the estimated effect in the disease model. Since determination of an optimal partition

of the range of disease scores requires a performance criterion for stratifications in two-phase studies, this discussion is postponed to Section 4.4.

### 4.2.3 Application to the empirical study

The proposed stratification strategy is applied to the empirical two-phase study. In contrast to the two-phase model considered in Sections 2.3 and 4.1 the model in this application includes eight additional phase 1 variables listed in Table E.2 in Appendix E. The disease score is estimated by using the same set of phase 1 covariates as well as interactions between phenprocoumon exposure and age and sex, respectively. Three post stratifications (D, E*, F) are defined by percentiles of the disease score based on five (D, F) or two (E*) cut-points. To account for the stratified sampling by phenprocoumon exposure, post stratifications D, E*, and F are additionally classified by phenprocoumon exposure and E* as well as F also include information on age group and sex (see Table E.1 in Appendix E). These post stratifications are compared with post stratifications A, B, and C of Section 4.1 regarding potential bias and standard errors of parameters estimates in the two-phase model.

The comparison of the stratifications reveals that stratifying on percentiles of a disease score leads to a reduction of standard errors of estimated parameters for the phase 1 covariates. However, the reduction is negligible. In particular, the standard errors of parameter estimates for hypertension and diabetes are much more reduced by using stratifications B and C which include phase 1 information of these covariates directly (see Table E.2 in Appendix E). Potential bias, assessed by the difference between phase 1 and two-phase estimators, is observed for post stratification D in estimators related to phenprocoumon exposure and sex. This bias can be explained by two arguments: First, the response analysis of Section 2.3 demonstrated that the phase 2 sample is not random with respect to age which introduces bias in a complete-case analysis. Second, stratification D includes information on sex only via the disease score. This is not sufficient to correct the aforementioned bias. Furthermore, some two-phase estimators related to comorbidities and concomitant medications differ from the respective phase 1 estimators indicating bias. Estimators obtained from a stratification based on disease scores are closer to the phase 1

estimators which might be interpreted as bias reduction. Since the true parameter values are not known in this study, it cannot be concluded which estimators are biased. The performance of the stratifications regarding bias is therefore evaluated in the simulation study where the true parameter values are specified in the design.

### 4.2.4 Performance in the simulation study

The simulation study investigating the performance of stratifications based on disease scores is part of a more complex simulation study. The other parts as well as technical details of the set-up are described in Section 4.5. In this section, stratifications A, B, C, D, and E* of the previous section as well as a simpler variant of E*, called stratification E, are evaluated regarding bias and efficiency of the two-phase estimators. Bias is defined by the difference between the two-phase estimator and the true parameter value and efficiency is defined as specified in Section 4.1. Furthermore, failure proportions are considered, where an analysis is counted as a failure in case of empty strata or convergence problems of the estimation procedure. The set-up of this part of the simulation study is depicted in Figure E.2 in Appendix E.

The simulation study confirms the results of Section 4.2.3: Direct inclusion of information on hypertension and diabetes by cross-classification is more efficient than including this information via a disease score (see Figure E.4 in Appendix E). Furthermore, the bias that occurs with post stratification D in the empirical study due to insufficient consideration of the sampling procedure is also apparent in the simulated two-phase studies. Whereas in the empirical study parameters related to sex are estimated with bias when using stratification D, the same problem is observed for age in the simulation study when analysing age-stratified phase 2 samples with stratification D. The reason for the bias is quite obvious because in this situation the a priori stratification is finer than the post stratification. Since age is correlated with many covariates, ignoring the full age information in the post stratification also leads to biased estimates for other covariates (see Figure E.3 in Appendix E).

Another finding of the simulation study is related to the proportion of failures, which are mostly caused by empty strata ($> 95\%$ of failures). Only a small number of strata is feasible for the evaluation of small phase 2 samples because stratifications with a larger number of strata are associated with high failure proportions.

Stratifications based on percentiles of disease scores are more likely to result in empty strata than stratifications with the same number of strata defined by cross-classification. However, for a given number of strata, stratifications using a disease score include information on much more covariates than stratifications defined by cross-classification. This observation indicates that some of the strata defined by disease scores are sparsely populated in the phase 1 sample. This most probably results from the different conditional distributions of disease scores in cases and controls. The problem of empty strata vanishes if the sampling stratification and the evaluation stratification coincide.

Among all tested combinations of a priori and post stratifications, using stratification E* for sampling and analysis delivers the best performance with respect to low failure proportions, small bias and high efficiency. As can be seen from Table E.3 in Appendix E, efficiencies are considerably improved for most phase 1 covariates. Therefore, the same inferences on phase 1 covariates can be drawn from the two-phase analysis using stratification E* and the phase 1 analysis which is based on a data set that is ten times larger. Further results of the simulation study are described in Behr and Schill, 2013 and in Section 4.5.

### 4.2.5 Discussion

The first applications of using disease scores for stratification in two-phase analyses yield promising results. Advantages of this stratification approach are:

1. Phase 1 information on all relevant covariates can be included in the analysis without increasing the number of strata with each additional covariate.
2. Standard errors of these covariates are notably reduced.

A considerable gain in efficiency is only achieved if the stratification is also used for sampling the phase 2 data. If the phase 2 data set is sampled with respect to a cross-classification of variables, the post stratification has to include these variables directly and cannot be solely defined on percentiles of the disease score. In this case, the post stratification is defined by cross-classification of the a priori stratification and subclasses of the disease score, where the number and placement of the cut-

points is restricted by the requirement of non-empty strata. Choice of number and placement of the cut-points is also influenced by the difference between the conditional distributions of disease scores in cases and controls. It remains unclear how subclasses of the disease score should be defined to yield efficient a priori and post stratifications.

## 4.3 Criteria for planning the a priori stratification

As has been demonstrated in Sections 4.1 and 4.2 it is difficult to define an efficient stratification using information on a multitude of phase 1 covariates. There is no general rule available stating that one stratification strategy is better than the other. Additionally, for each strategy many possibilities exist to build stratifications. In particular, it is an open question how efficient stratifications can be identified at the planning stage of a two-phase study when only phase 1 information is available. Thus, a criterion is desirable which allows for the comparison of a priori stratifications with respect to efficiency based on phase 1 data alone. In this section, such a design criterion is derived from the variance formula of the weighted likelihood estimator (WLE).

### 4.3.1 Variance of the WLE

Before the design criterion is derived, the variance formula (3.21) of the WLE is discussed in more detail based on results and arguments given in Reilly and Pepe, 1995. In the following, only the variance of the estimated parameter vector $\hat{\boldsymbol{\beta}}$ is considered whereas rows and columns of the covariance matrix referring to the estimated intercept $\hat{\alpha}$ are ignored.

The covariance matrix of the WLE $\hat{\boldsymbol{\beta}}$ is given by

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \frac{1}{N} I_R^{-1} + \frac{1}{N} I_R^{-1} \Omega I_R^{-1}. \tag{4.2}$$

Since $N^{-1} I_R^{-1}$ corresponds to the variance that would have been obtained if complete data was available for everyone in phase 1, the second part of the formula, $N^{-1} I_R^{-1} \Omega I_R^{-1}$, can be interpreted as penalty term which is a measure for the loss in

efficiency due to the partially missing covariate information. Reilly and Pepe have shown that the information $I_R$ can be consistently estimated by

$$\hat{I}_R = \frac{1}{N}\left[\sum_{q\in\{k:r_k^{II}=1\}}\left(-\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\log f(d_q|\mathbf{x}_q;\boldsymbol{\beta})\right)\right.$$
$$\left.+\sum_{q'\in\{k:r_k^{II}=0\}}\hat{E}\left(-\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\log f(d_{q'}|\mathbf{X};\boldsymbol{\beta})\;\Big|\;d_{q'},s_{q'}\right)\right]_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}, \qquad (4.3)$$

where

$$\hat{E}\left(-\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\log f(d_{q'}|\mathbf{X};\boldsymbol{\beta})\;\Big|\;d_{q'},s_{q'}\right)=\sum_t\frac{1}{n_{ij}}\left(-\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\log f(d_t|\mathbf{x}_t;\boldsymbol{\beta})\right)$$

and $\sum_t$ is the summation over phase 2 subjects in the respective stratum determined by $s_{q'}$, i.e., $t\in\{k: r_k^{II}=1, s_k=s_{q'}\in S_{ij}\}$.

A consistent estimator of $\Omega$ (see (3.22)) is given by

$$\hat{\Omega}=\sum_{i=0}^{1}\sum_{j=1}^{J}\frac{N_{ij}}{N}\frac{N_{ij}-n_{ij}}{n_{ij}}\widehat{\mathrm{Cov}}\left(SC(\hat{\boldsymbol{\beta}})\;\Big|\;(D,\mathbf{X})\in S_{ij},R^{II}=1\right), \qquad (4.4)$$

where $\widehat{\mathrm{Cov}}$ denotes the empirical covariance matrix in stratum $(i,j)$ of the phase 2 data set and $SC(\hat{\boldsymbol{\beta}})$ denotes the score vector evaluated at $\hat{\boldsymbol{\beta}}$, i.e.,

$$SC(\hat{\boldsymbol{\beta}})=\left[\frac{\partial}{\partial\boldsymbol{\beta}}\log f(D|\mathbf{X};\boldsymbol{\beta})\right]_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}.$$

In (4.4), $\frac{N_{ij}}{N}$ and $\frac{n_{ij}}{N_{ij}}$ are used as consistent estimators of the marginal stratum probabilities $Q_{ij}$ and the selection probabilities $p_{ij}^{II}$.

Let $\hat{\Pi}_k$ denote the estimated probability of contracting the disease for subject $k$ which is defined by

$$\hat{\Pi}_k=\frac{\exp(\hat{\alpha}+\hat{\boldsymbol{\beta}}^T\mathbf{x}_k)}{1+\exp(\hat{\alpha}+\hat{\boldsymbol{\beta}}^T\mathbf{x}_k)}$$

with $\hat{\boldsymbol{\beta}}$ being the WLE. The subject-specific score vector $(SC_k)$ evaluated at $\hat{\boldsymbol{\beta}}$ is then given by

$$SC_k(\hat{\boldsymbol{\beta}})=(d_k-\hat{\Pi}_k)\mathbf{x}_k.$$

Accordingly, the information $\hat{I}_k(\hat{\boldsymbol{\beta}})$ of subject $k$ has the form

$$\hat{I}_k(\hat{\boldsymbol{\beta}})=\hat{\Pi}_k(1-\hat{\Pi}_k)\mathbf{x}_k\mathbf{x}_k^T.$$

Using the above expressions, $\hat{\Omega}$ and $\hat{I}_R$ can be written as

$$\hat{\Omega} = \sum_{i=0}^{1} \sum_{j=1}^{J} \frac{N_{ij}}{N} \frac{N_{ij} - n_{ij}}{n_{ij}} \widehat{\mathrm{Cov}} \left( (D - \hat{\Pi})\mathbf{X} \;\middle|\; (D, \mathbf{X}) \in S_{ij}, R^{II} = 1 \right) \tag{4.5}$$

and

$$\begin{aligned} \hat{I}_R &= \frac{1}{N} \sum_{i=0}^{1} \sum_{j=1}^{J} \left[ \sum_{k=1}^{n_{ij}} \hat{\Pi}_{ijk}(1 - \hat{\Pi}_{ijk}) \mathbf{x}_{ijk} \mathbf{x}_{ijk}^T + \frac{N_{ij} - n_{ij}}{n_{ij}} \sum_{k=1}^{n_{ij}} \hat{\Pi}_{ijk}(1 - \hat{\Pi}_{ijk}) \mathbf{x}_{ijk} \mathbf{x}_{ijk}^T \right] \\ &= \frac{1}{N} \sum_{i=0}^{1} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} \frac{N_{ij}}{n_{ij}} \hat{\Pi}_{ijk}(1 - \hat{\Pi}_{ijk}) \mathbf{x}_{ijk} \mathbf{x}_{ijk}^T. \end{aligned} \tag{4.6}$$

Only the second part of (4.2), i.e. the penalty term, depends on the stratification $\mathscr{S}$ through $\Omega$. Therefore, a design criterion considering only the penalty term should be sufficient for the identification of efficient stratifications. Since both estimators $\hat{I}_R$ and $\hat{\Omega}$ depend on $\mathscr{S}$, this is only 'asymptotically' correct in the sense that the estimators converge to $I_R$ and $\Omega$, respectively, as the sample size tends to infinity. Nevertheless, the first part of (4.2) is assumed to be neglectable for the definition of a design criterion because

1. for large phase 1 and relatively small phase 2 samples the phase 1 sample variance $N^{-1} I_R^{-1}$ is considered to be much smaller than the penalty term;
2. only the weights of phase 2 contributions change for different stratifications in a fixed phase 2 sample which is expected to result in major differences only in presence of extreme weights.

### 4.3.2 Constructing a design criterion based on phase 1 data

The aim is to define a design criterion which identifies the most efficient stratifications from a set of stratifications by using only phase 1 data. The idea is to predict for each stratification the size of the penalty term and therefore the loss in efficiency that would occur in a two-phase analysis using the respective stratification. It is assumed that phase 2 data is not available for the prediction, so that only the efficiency of parameter estimation for phase 1 covariates can be assessed. Furthermore, the subject-specific score vector and information matrix cannot be obtained from a

two-phase analysis but have to be taken from a usual logistic regression analysis of the phase 1 data.

Recall from Chapter 3 that the vector of complete covariate information $\mathbf{X}^T = (\mathbf{Y}^T, \mathbf{Z}^T)$ is partitioned into phase 1 information $\mathbf{Y}$ and phase 2 variables $\mathbf{Z}$. Let $\hat{\boldsymbol{\beta}}_I$ denote the MLE of the phase 1 model $f(D|\mathbf{Y};\boldsymbol{\beta}_I)$ and $I_I$ the corresponding information matrix obtained from usual logistic regression. For simplicity of notation, in the following no distinction is made between $\mathbf{Y}$ and the regression vector $\mathbf{Y}_I$ based on $\mathbf{Y}$ as well as between $\mathbf{X}$ and the respective regression vector.

The penalty term related to phase 1 covariates can be approximated based on phase 1 data by

1. replacing $N\hat{I}_R$ with the estimated phase 1 information $\hat{I}_I$,
2. estimating the score variability in each phase 1 stratum by the empirical covariance
$$\widehat{\mathrm{Cov}}\left((D-\hat{\Pi}_I)\mathbf{Y} \;\middle|\; (D,\mathbf{X}) \in S_{ij}\right),$$
where $\hat{\Pi}_I$ is the estimated disease probability
$$\hat{\Pi}_I = \frac{\exp(\hat{\alpha}_I + \hat{\boldsymbol{\beta}}_I^T\mathbf{y})}{1+\exp(\hat{\alpha}_I + \hat{\boldsymbol{\beta}}_I^T\mathbf{y})}, \tag{4.7}$$
and
3. setting the phase 2 stratum size $n_{ij}$ to
$$n_{ij} = \min\left\{\frac{n}{2J}, N_{ij}\right\},$$
where $n$ is the phase 2 sample size and $J$ is the number of strata in order to achieve a balanced phase 2 sample with an equal number of subjects in each stratum.

Then, an approximation of the penalty term (PT) is obtained as

$$\widehat{PT}_I(\mathscr{S}) = N\hat{I}_I^{-1}\left(\sum_{i=0}^{1}\sum_{j=1}^{J}\frac{N_{ij}}{N}\frac{N_{ij}-n_{ij}}{n_{ij}}\widehat{\mathrm{Cov}}\left((D-\hat{\Pi}_I)\mathbf{Y} \;\middle|\; (D,\mathbf{X}) \in S_{ij}\right)\right)\hat{I}_I^{-1} \tag{4.8}$$

$$=: N\hat{I}_I^{-1}\hat{\Omega}_I\hat{I}_I^{-1}.$$

The rationale for using the estimated phase 1 covariance $\hat{I}_I^{-1}$ follows immediately from the previous section because $N^{-1}I_R^{-1}$ corresponds to the covariance that would have been obtained from the full phase 1 sample if complete covariate information was available for everyone in phase 1. It has to be noted that $\hat{I}_I^{-1}$ is obtained from a model including only phase 1 covariates whereas $N^{-1}\hat{I}_R^{-1}$ is the estimated covariance of parameter estimators from a model that includes also phase 2 variables. Therefore, $\widehat{PT}_I$ is a $p' \times p'$-matrix instead of a $p \times p$-matrix, where $p' = \dim(\mathbf{Y})$ and $p = \dim(\mathbf{X})$. This means that no information on the penalties associated with the estimation of phase 2 parameters is available. Also the variability of the score is estimated by a model including only phase 1 variables instead of phase 1 and phase 2 variables. The consequences of ignoring phase 2 variables in the approximation of the penalty term is discussed in the next section. A further difference is that the score variability is estimated in the phase 1 strata instead of phase 2 strata as is defined in (4.4). However, if the MAR assumption is fulfilled for the phase 2 sample, strata with a large variability in phase 1 will also have a large variability in phase 2. In particular, if stratifications are ranked with respect to the score variability, ranks will be equal or at least similar in phase 1 and phase 2 strata. Setting $n_{ij}$ as in 3. is based on the assumption of a balanced design taking into account the restriction imposed by the phase 1 stratum sizes $N_{ij}$. This choice is arbitrary and could be replaced depending on another design, e.g., by assuming random sampling of the phase 2 data set.

Although the approximation differs from the penalty term calculated for the two-phase analysis, it may be used to predict the performance of stratifications with respect to an efficient estimation of parameters related to phase 1 covariates. For the comparison of different stratifications, the diagonal elements of $\widehat{PT}_I$, which represent the variances of the respective parameter estimates, are plotted as shown in Section 4.4, Figure 4.2, where each column refers to a single stratification. Stratifications with small penalty terms for all or the most important phase 1 covariates are then considered as the most efficient stratifications. The choice of the most important covariates is based on qualitative criteria which take into account for instance prior knowledge and the specific study question. Thus, the design criterion is defined as follows:

**Design Criterion.** *Consider a two-phase study based on data* $(D, \mathbf{X})$*, where* $D$ *is*

*the disease indicator and* $\mathbf{X}^T = (\mathbf{Y}^T, \mathbf{Z}^T)$ *is the complete covariate vector available in phase 2 with* $\mathbf{Y}$ *being the vector of covariates available in phase 1 and* $\mathbf{Z}$ *being the additional information available in phase 2. Then, efficient stratifications are identified by the following steps:*

1. *Define a ranking of covariates* $X_1, \ldots, X_p$ *according to their importance for the study.*
2. *Define a set of candidate stratifications* $\{\mathscr{S}_1, \ldots, \mathscr{S}_L\}$.
3. *Estimate* $\widehat{PT}_I(\mathscr{S}_l)$, $l = 1, \ldots, L$, *and select the stratifications with the smallest diagonal elements of these matrices for all or the most important covariates.*

The application of the proposed design criterion is exemplified for the empirical study in Section 4.4. Its properties and limitations are discussed in the next section.

### 4.3.3 Properties and limitations of the design criterion

The suggested design criterion relies on an approximation of the penalty term that is based on the full phase 1 data and a model including only phase 1 covariates. Since no information on the efficiency of parameter estimation for phase 2 variables is provided by the criterion, stratifications are only selected with respect to efficient estimation of parameters for phase 1 variables. This is a clear limitation. However, at the planning stage, when only phase 1 data is available, this limitation cannot be avoided. If known proxies for phase 2 variables are available in phase 1, this limitation can be at least partly compensated by including these variables in the selected stratification.

**Approximation error** Ignoring phase 2 variables in the model also leads to an approximation error of the penalties related to the estimation of parameters for phase 1 covariates because $N\hat{I}_R \neq \hat{I}_I$ and $\hat{\Omega} \neq \hat{\Omega}_I$. To examine this error in more detail, consider the penalty term $\widehat{PT}_k$ of $\hat{\beta}_k$ related to a phase 1 covariate $X_k$ and

its approximation $\widehat{PT}_{I,k}$:

$$\widehat{PT}_k = \frac{1}{N} \sum_{u=1}^{p} \sum_{v=1}^{p} \left(i_R^{-1}\right)_{kv} (\omega)_{vu} \left(i_R^{-1}\right)_{uk}, \tag{4.9}$$

$$\widehat{PT}_{I,k} = N \sum_{u=1}^{p'} \sum_{v=1}^{p'} \left(i_I^{-1}\right)_{kv} (\omega_I)_{vu} \left(i_I^{-1}\right)_{uk}, \tag{4.10}$$

where $\left(i_R^{-1}\right)_{ab}$, $(\omega)_{ab}$, $\left(i_I^{-1}\right)_{ab}$, and $(\omega_I)_{ab}$ are the elements of the $a$th row and $b$th column of the matrices $\hat{I}_R^{-1}$, $\hat{\Omega}$, $\hat{I}_I^{-1}$, and $\hat{\Omega}_I$, respectively. Apart from the differences between $N^{-1}\left(i_R^{-1}\right)_{ab}$ and $\left(i_I^{-1}\right)_{ab}$ as well as between $(\omega)_{ab}$ and $(\omega_I)_{ab}$ for $1 \leq a, b \leq p'$, the two terms differ by the summands related to phase 2 variables, i.e., by

$$\frac{1}{N} \sum_{u=p'+1}^{p} \sum_{v=p'+1}^{p} \left(i_R^{-1}\right)_{kv} (\omega)_{vu} \left(i_R^{-1}\right)_{uk}.$$

The size of this sum depends on the covariance of $X_k$ and the phase 2 variables as well as on the empirical covariance of the score in each stratum for components related to phase 2 variables. Whereas the first dependency is not influenced by the stratification, the second dependency is. Stratifications including information on $X_k$ and on proxies for the phase 2 variables are likely to reduce the score variability of the respective components, though it cannot be calculated based on phase 1 data. This is another argument for including proxies for phase 2 variables in the stratification.

The difference between $N^{-1}\left(i_R^{-1}\right)_{ab}$ and $\left(i_I^{-1}\right)_{ab}$ for $1 \leq a, b \leq p'$ is not neglectable because it has been shown that at least the variance $\left(i_I^{-1}\right)_{kk}$ is inflated by including additional covariates in the model (see e.g., Whittemore, 1981; Hsieh et al., 1998). However, since the size of the difference is independent of the stratification, it does not need to be considered in the design criterion.

The terms $(\omega)_{ab}$ and $(\omega_I)_{ab}$ for $1 \leq a, b \leq p'$ differ with respect to the score variability which is estimated for $(\omega)_{ab}$ by the empirical covariance of the score $SC = (D - \hat{\Pi})\mathbf{Y}$ in each phase 2 stratum and for $(\omega_I)_{ab}$ by the empirical covariance of the score $SC_I = (D - \hat{\Pi}_I)\mathbf{Y}$ in each phase 1 stratum. As has been discussed in the previous section, the issue of using phase 1 strata instead of phase 2 strata is not relevant for the design criterion. The difference between

$$\hat{\Pi}_h = \frac{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_h)}{1 + \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_h)} \quad \text{and} \quad \hat{\Pi}_{I,h} = \frac{\exp(\hat{\alpha}_I + \hat{\boldsymbol{\beta}}_I^T \mathbf{y}_h)}{1 + \exp(\hat{\alpha}_I + \hat{\boldsymbol{\beta}}_I^T \mathbf{y}_h)}$$

for each subject $h$ might, however, lead to divergent conclusions on the performance of the investigated stratifications. If the phase 1 data includes most of the information on the disease such that $\hat{\Pi} \approx \hat{\Pi}_I$, the design criterion results in valid predictions of the penalties. As two-phase database studies comprise a multitude of phase 1 covariates, this condition is assumed to be fulfilled. Even if $\hat{\Pi} \not\approx \hat{\Pi}_I$, the approximation error will most likely only lead to an underestimation of the efficiency of stratifications including proxy information on phase 2 covariates. This limitation of the design criterion is already known.

**Relation to the disease score** As can be seen from (4.7), the estimated disease probability $\hat{\Pi}_I$ which is used in the approximation of the penalty term coincides with the disease score DSC defined in (4.1). Since the empirical covariance of the score in stratum $S_{ij}$ only depends on the variabilities of $\hat{\Pi}_I$ and $\mathbf{Y}$ in this stratum, it is obviously reduced when the stratification is based on $\hat{\Pi}_I$. Thus, stratifications defined by percentiles of the disease score are likely to have small penalty terms.

**Transferability to ML estimation** The design criterion is based on the variance of the WLE because the variance formula allows for a relatively easy approximation of the variance by using only phase 1 data. For the variance of the MLE (cf. Section 3.2.2) no decomposition in the phase 1 variance and a penalty term exists. ML is, however, the preferred estimation approach because it is more efficient than the WL approach (cf. Section 3.3.2). The design criterion might underestimate the efficiency of ML estimators especially for stratifications with a high variability of sampling fractions. If the variability of sampling fractions is similar for the compared stratifications, it is assumed that efficient stratifications according to the design criterion are also efficient when using ML estimation.

How the theoretical properties of the design criterion translate into practice is investigated in Sections 4.4 and 4.5 where the criterion is used to identify efficient stratifications for the empirical study which are then evaluated in a simulation study.

## 4.4 Planning an efficient stratification for the empirical study

The design criterion is applied to the empirical two-phase study to identify efficient stratifications for the extended two-phase model considered in Section 4.2.3. Hence, penalties are to be predicted for the parameter estimation of the phase 1 covariates phenprocoumon exposure, age, sex, hypertension, diabetes mellitus, and eight binary covariates for comorbid diseases and concomitant medications. According to Section 4.3.2, the search is conducted in three steps: First, the covariates are ranked according to their importance for the study. Second, a set of candidate stratifications is defined taking into account the ranking of covariates. Third, penalties are predicted for these stratifications and evaluated graphically to identify the best stratifications with respect to small penalties. The three steps are detailed in the following.

### Step 1: Ranking of covariates

The ranking of covariates according to their importance for the study is based on the study objective, subject matter knowledge, and prior knowledge on the covariates. Since the study aims at assessing the risk of bleeding associated with phenprocoumon use, the most important variable is phenprocoumon exposure. Age and sex are chosen to be the second most important covariates because older age and male sex are known risk factors for bleedings (see for instance Hernandez-Diaz and Rodriguez, 2002). The variable age is of particular importance because it is the only continuous phase 1 variable. Moreover, age is associated with most of the considered comorbid diseases and concomitant medications. All other covariates are defined to be of secondary importance. As hypertension and diabetes are the most prevalent diseases, they are considered to be more important than diseases with a lower prevalence. Consequently, ranks are assigned to the phase 1 covariates as shown in Table 4.2.

**Table 4.2:** Ranking of phase 1 covariates

| Rank | Covariate |
|---|---|
| 1 | Phenprocoumon exposure |
| 2 | Age |
| 3 | Sex |
| 4 | Hypertension, diabetes mellitus |
| 5 | Ischemic heart disease, liver disease, GI disease, use of NSAIDs, ASA, diuretics, statins, gastroprotective drugs |

ASA: acetylsalicylic acid, GI: gastrointestinal, NSAID: non-steroidal anti-inflammatory drug

### Step 2: Definition of candidate stratifications

Taking into account the ranking of covariates, 31 stratifications are constructed by cross-classification of covariates. Starting with a stratification based on the most important covariate, this covariate is then combined with the second most important covariate and so forth. Since cross-classification of phenprocoumon exposure, age (in three categories), sex and two further covariates results in up to 48 strata, no more than two covariates are added to the variables of most and second most importance. Moreover, to avoid empty cells, only combinations with the most prevalent covariates hypertension and diabetes are considered.

In addition, five stratifications are defined by percentiles of two different disease scores. While the first disease score (DSC1) corresponds to the one used in Section 4.2.3, the second disease score (DSC2) is obtained from a logistic regression model including all phase 1 covariates except phenprocoumon, age, and sex. For both disease scores, stratifications are defined by cross-classification of phenprocoumon exposure, age, sex, and either deciles or the 90th and 95th percentiles of the disease score. The fifth stratification is based on deciles of DSC1 alone. Hence, the set of candidate stratifications consists of the 36 stratifications listed in Table 4.3.

**Table 4.3:** Definition of stratifications

| ID No. | Stratified by... |
|---|---|
| 1 | ...phenprocoumon exposure (phen.) |
| 2a | ... phen., age[a] |
| 2b | ... phen., sex |
| 2ab | ... phen., age[a], sex |
| 3a | ... phen., age[a], sex, hypertension |
| 3b | ... phen., age[a], sex, diabetes mellitus |
| 3ab | ... phen., age[a], sex, hypertension, diabetes mellitus |
| 3ac | ... phen., age[a], sex, hypertension, liver disease |
| 3ad | ... phen., age[a], sex, hypertension, ischemic heart disease |
| 3ae | ... phen., age[a], sex, hypertension, GI disease |
| 3af | ... phen., age[a], sex, hypertension, use of ASA |
| 3ag | ... phen., age[a], sex, hypertension, use of NSAIDs |
| 3ah | ... phen., age[a], sex, hypertension, use of statins |
| 3ai | ... phen., age[a], sex, hypertension, use of diuretics |
| 3aj | ... phen., age[a], sex, hypertension, use of gastroprotective drugs |
| 3bc | ... phen., age[a], sex, diabetes mellitus, liver disease |
| 3bd | ... phen., age[a], sex, diabetes mellitus, ischemic heart disease |
| 3be | ... phen., age[a], sex, diabetes mellitus, GI disease |
| 3bf | ... phen., age[a], sex, diabetes mellitus, use of ASA |
| 3bg | ... phen., age[a], sex, diabetes mellitus, use of NSAIDs |
| 3bh | ... phen., age[a], sex, diabetes mellitus, use of statins |
| 3bi | ... phen., age[a], sex, diabetes mellitus, use of diuretics |
| 3bj | ... phen., age[a], sex, diabetes mellitus, use of gastroprotective drugs |
| 4_1a | ... phen., age[a], sex, deciles of DSC1 |
| 4_1b | ... phen., age[a], sex, 90th/95th %-ile of DSC1 |
| 4_1c | ... deciles of DSC1 |
| 4_2a | ... phen., age[a], sex, deciles of DSC2 |
| 4_2b | ... phen., age[a], sex, 90th/95th %-ile of DSC2 |

[a] Age is considered in categories <50 years, 50-<65 years, and ≥65 years.

ASA: acetylsalicylic acid, DSC1: disease score 1, DSC2: disease score 2, GI: gastrointestinal, NSAID: non-steroidal anti-inflammatory drug, phen.: phenprocoumon

## Step 3: Prediction of penalties and choosing the best stratification

The penalty terms $\widehat{PT}_I(\mathscr{S})$ are calculated for the 36 candidate stratifications and each phase 1 covariate assuming balanced phase 2 samples of sizes 500, 1,000, 2,000, and 10,000. The size of $\widehat{PT}_I(\mathscr{S})$ is assessed graphically as shown in Figure 4.2, where all covariates except phenprocoumon exposure and age can be included in the same graph because their penalties have the same order of magnitude. Since the same picture arises for all phase 2 sample sizes, results are only presented for the sample size 2,000.

Regarding phenprocoumon exposure, all stratifications have the same low values of $\widehat{PT}_I(\mathscr{S})$ except for the stratification that does not include phenprocoumon exposure explicitly (4_1c) (see Figure 4.3). For age, the smallest penalty occurs for the stratification defined by phenprocoumon exposure, age, sex, and deciles of DSC1 (4_1a) and the highest values are observed for those stratifications ignoring age information (1, 2b) and for some of the stratifications including phenprocoumon exposure, age, sex, diabetes mellitus, and one further covariate (see Figure 4.3). Concerning the other covariates, the best results are obtained for stratifications based on a disease score as can be seen in Figure 4.2. Interestingly, most stratifications based on cross-classification of phenprocoumon exposure, age, sex, and at least one comorbid disease or concomitant medication are less efficient for the estimation of the age effect compared to other stratifications including age information. This can only be explained by different sampling fractions in the age groups. They apparently lead to sampling of less informative subjects with respect to age. The gap that can be seen in Figure 4.2 between dots related to variables included in the stratification and those not included indicates that the loss of age information also results in less efficient estimation for variables not included in the stratification. Thus, including additional covariates in a stratification defined by cross-classification does not only lead to a gain in efficiency for the additional covariate but may also result in efficiency loss for the other covariates.

Considering only the size of $\widehat{PT}_I(\mathscr{S})$, stratifications based on phenprocoumon exposure, age, sex, and deciles of a disease score (4_1a, 4_2a) would be chosen as the most promising stratifications for an efficient study design. A disadvantage of both

stratifications is that the strata referring to the lower percentiles are very sparsely populated. The risk of empty strata is then increased for real phase 2 samples in which phase 2 information presumably cannot be obtained for all selected persons. Therefore, the coarser stratifications (4_1b, 4_2b) are to be preferred in practice. To evaluate how these stratifications perform in a realistic two-phase analysis a simulation study is conducted mimicking the empirical data situation.

Among the 36 stratifications compared in this section are also stratifications based on two partitions of the disease score: one includes ten subclasses with equidistant cut-points and the second one consists of three subclasses with cut-points that ensure a sufficient number of cases in each subclass. The estimated penalties indicate that the stratification based on the ten narrow subclasses is more efficient than that based on the three broad subclasses. Although not further investigated here, the design criterion seems to be useful for determining good partitions of the disease score.

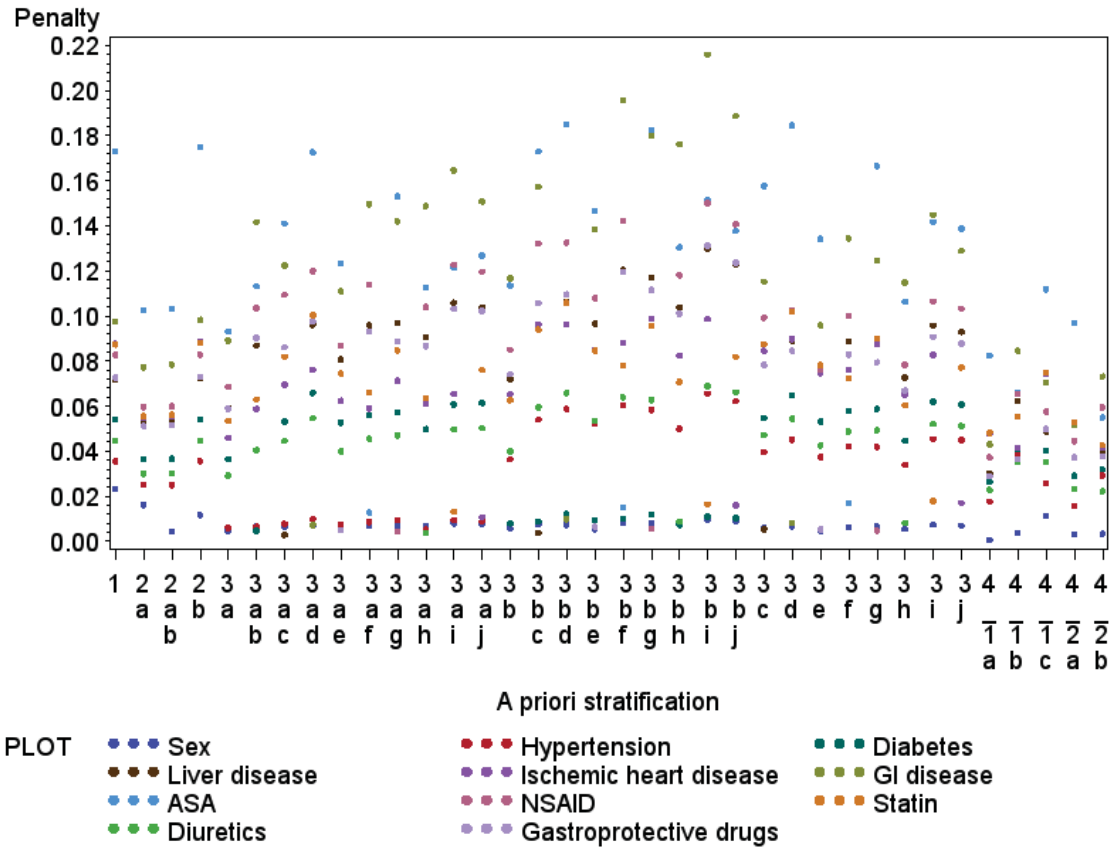
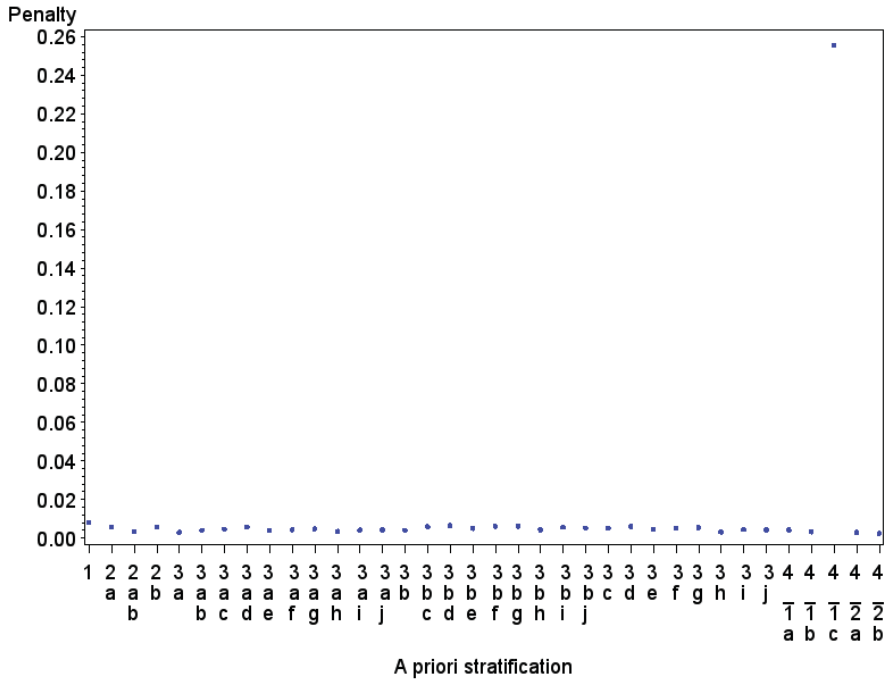
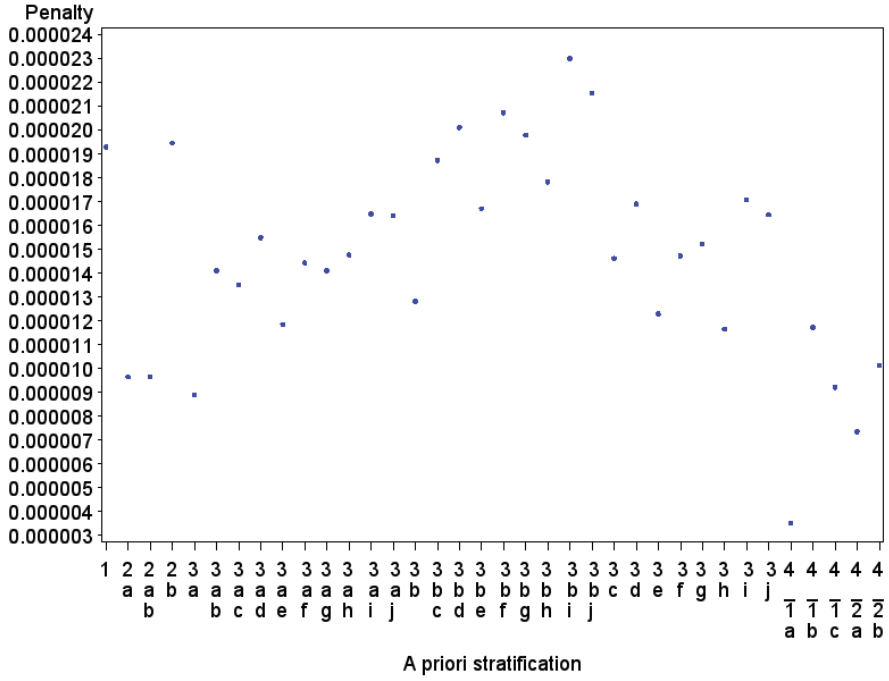



**Figure 4.2:** Penalty terms estimated from phase 1 data for a phase 2 sample of size 2,000




(a) Phenprocoumon exposure




(b) Age

**Figure 4.3:** Penalty terms for phenprocoumon exposure and age estimated from phase 1 data for a phase 2 sample of size 2,000

## 4.5 Simulation study to assess the performance of selected stratifications

Complementary to the empirical study, a simulation study is conducted to compare different stratification strategies in two-phase analyses with a rich phase 1 data set. The simulation study mimics the empirical study but allows the investigation of larger phase 2 samples and the examination of bias in the parameter estimation. The objectives are specifically

1. to assess the performance of stratifications defined by percentiles of a disease score in comparison to cross-classification,
2. to assess the performance of the a priori stratifications selected by the design criterion, and
3. to assess the interplay of a priori and post stratifications.

This section starts with a description of the set-up of the complete simulation study whereas results are only reported for the second and third objectives. Results according to the first objective have already been described in Section 4.2.4.

### 4.5.1 Set-up of the simulation study

**Simulation of phase 1 and phase 2 data sets** At first, 1,000 phase 1 data sets are simulated which have the same size and the same covariate structure as the empirical study. Technically, the covariates are simulated in three steps: The first step relates to the simulation of 12 binary covariates according to the estimated covariate distribution of the empirical study for 26,208 subjects. Then, age is simulated based on the conditional distribution of age given the binary covariates. In the third step, the phase 2 variables BMI and smoking are generated for each subject as categorical variables with four and three categories using the distribution estimated from the phase 2 data set of the empirical study. The disease status is then derived according to the estimated logistic disease model of the empirical study (see Table E.3 in Appendix E). From each phase 1 data set, 44 phase 2 samples of 500, 1,000, 2,000, and 10,000 subjects, respectively, are drawn with respect to

11 sampling schemes including random sampling and stratified sampling by the a priori stratifications listed in Table 4.4. Of the ten a priori stratifications, the simplest stratification includes only two strata defined by phenprocoumon exposure, the next six stratifications are derived by cross-classification of phenprocoumon exposure and several covariates, and the last three are based on the disease scores DSC1 and DSC2. In particular, stratifications IX and X are those selected in Section 4.4 by the design criterion. Moreover, stratification IX corresponds to stratification E* used in Section 4.2. The number of strata ranges from two until 40 strata (see Table 4.4). To avoid empty cells for stratifications with more than 20 strata, those strata occurring only in cases or only in controls are collapsed by ignoring information on the least important covariate. Consider for example a stratification defined by cross-classification of phenprocoumon exposure, age, sex, and hypertension and suppose that no male cases exposed to phenprocoumon in the youngest age group are diagnosed with hypertension. Then, this stratum is combined with the respective stratum of persons without hypertension. This approach ensures that information on the more important covariates such as phenprocoumon and age is always included in the stratification.

The phase 1 data sets are simulated by using the SAS call routine `rantbl` for generation of the covariates and `ranbin` for the disease status. Random selection of the phase 2 samples is achieved by invoking the SAS call routine `ranuni` and selecting subjects with the 500-10,000 smallest random numbers. Two seed streams are deployed, one for simulation of the phase 1 data and one for drawing the phase 2 samples. Independence of the simulations is guaranteed because only 5% and 60% of the period length of the random number generator is exhausted.

**Analysis of the data sets** Two-phase analysis of the 44,000 two-phase studies are conducted employing ML and WL methods which utilise the respective a priori stratification. Phase 2 samples drawn according to four selected a priori stratifications (0, I, II, VIII) are additionally analysed with respect to six post stratifications (A-F, see Table 4.5). For each parameter estimator $\hat{\beta}_{ki}$ for variable $k$ in simulation $i$ within-simulation standard errors $SE(\hat{\beta}_{ki})$ are calculated as well as bias and efficiency as defined in Section 4.2.4. Besides the within-simulation standard errors, also the empirical standard error is determined representing the uncertainty of $\hat{\beta}_k$

between the simulations. The empirical standard error is calculated as

$$SE_{emp} = \sqrt{\frac{1}{n_{sim} - 1} \sum_{i=1}^{n_{sim}} (\hat{\beta}_{ki} - \bar{\hat{\beta}}_k)^2},$$

where $\bar{\hat{\beta}}_k$ is the average parameter estimate. The stratifications are compared regarding failure proportions and the average bias as well as the average efficiency of the two-phase estimators, where failures are defined as in Section 4.2.4. For Objective 2, also the average within-simulation standard errors of the estimated coefficients of the phase 2 variables BMI and smoking are compared.

**Table 4.4:** Definition of a priori stratifications

| ID in simulation | Stratified by... | No of strata[b] | ID in Section4.4 |
|---|---|---|---|
| 0 | Random sample | 0 | – |
| I | ...phenprocoumon exposure (phen.) | 2 | (1) |
| II | ...phen., age[a] | 6 | (2a) |
| III | ...phen., sex | 4 | (2b) |
| IV | ...phen., age[a], sex | 12 | (2ab) |
| V | ...phen., age[a], sex, hypertension | 20 | (3a) |
| VI | ...phen., age[a], sex, diabetes mellitus (DM) | 20 | (3b) |
| VII | ...phen., age[a], sex, hypertension, DM | 40 (39-40) | (3ab) |
| VIII | ...deciles of DSC1 | 10 | (4_1c) |
| IX (=E*) | ...phen., age[a], sex, 90th/95th %-ile of DSC1 | 21 (20-22) | (4_1b) |
| X | ...phen., age[a], sex, 90th/95th %-ile of DSC2 | 30 (30-30) | (4_2b) |

[a] Age is considered in categories <50 years, 50-<65 years, and ≥65 years.

[b] If the number of strata varies, the median number of strata is presented with the quartiles Q1 and Q3 in parentheses.

**Table 4.5:** Definition of post stratifications

| ID | Stratified by... | No of strata |
|---|---|---|
| A | ...phenprocoumon exposure (phen.), age, sex | 10 |
| B | ...phen., age[a], sex, hypertension | 20 |
| C | ...phen., age[a], sex, diabetes mellitus | 20 |
| D | ...phen., 50th/75th/90th/95th/99th %-ile of DSC1 | 12 |
| E | ...phen., age[a], sex, 90th %-ile of DSC1 | 20 |
| F | ...phen., age[b], sex, hypertension | 16 |

[a] Age is considered in categories <50 years, 50-<65 years, and ≥65 years.

[b] Age is considered in categories <65 years and ≥65 years.

### 4.5.2 Results

#### Performance of a priori stratifications

For the comparison of a priori stratifications, only two-phase analyses are considered in which the a priori stratification is used for sampling and analysis of the phase 2 data. Results with respect to other post stratifications are described in the next section. The simulation study confirms the good performance of stratifications IX and X, which have been identified as most efficient by the design criterion in Section 4.4. Regarding the estimated coefficients for phenprocoumon exposure, age, and sex, both are among the most efficient stratifications. Concerning the parameter estimation for hypertension and diabetes, efficiencies are lower compared to stratifications including these covariates directly but higher compared to other stratifications. Both stratifications are more efficient for the parameter estimation of all other phase 1 covariates. Moreover, the parameter estimation is unbiased and no failures are observed with these stratifications. The simulation study also demonstrates that stratification VIII is worse than other stratifications because it is associated with high failure proportions and inefficient parameter estimation for most phase 1 covariates including the parameter of main interest (phenprocoumon). However, effects of the phase 2 covariates smoking and BMI are estimated with the smallest standard errors when using this stratification. Furthermore, the study reveals a problem

for cross-classification when the WL approach is used: Since cross-classification of several covariates leads to a high variability of sampling fractions, the WL estimators are biased and inefficient. More detailed results of the simulation study are described below.

**Failure proportion** The failure proportions are approximately zero for all a priori stratifications except for stratification VIII where proportions of up to 70% failures are observed for small phase 2 samples (see Table 4.7).

**Bias** The stratified sampling of most phase 2 samples introduces bias in complete-case analyses (i.e., in logistic regression analyses using only phase 2 data) for all parameter estimates related to variables included in the a priori stratification. In two-phase analysis using the ML approach, the bias is removed for all parameter estimates and all stratifications (results not shown). When applying the WL approach to phase 2 samples of sizes 500 and 1,000, residual bias occurs in parameter estimates of hypertension and diabetes mellitus for stratifications including these variables (i.e., V, VI, VII; see Table 4.6). This might be explained by the high variability of sampling fractions for these stratifications which is particularly high for small samples. Sampling fractions vary more for these stratifications because the number of strata is large and strata of subjects with diabetes mellitus and hypertension and/or phenprocoumon exposure are very sparsely populated whereas strata with healthy subjects not taking phenprocoumon contain a large number of subjects. For small phase 2 samples, the proportion of subjects selected from these large strata is very small thereby leading to very high weights in the weighted analysis. For instance, the smallest average sampling fraction in samples of size 500 for a priori stratification VII is 0.0009 leading to a weight of 1111.11.

**Efficiency** With respect to efficiency, the performance of the a priori stratifications differs widely subject to the covariates. Regarding phenprocoumon exposure, all stratifications have similar efficiencies except for stratification VIII which is inefficient (see Figure 4.5(a)). The coefficient for age can only be estimated efficiently if age is included in the a priori stratification either directly or indirectly via a disease score (see Figure 4.5(b)). High efficiencies are also observed if the covariate infor-

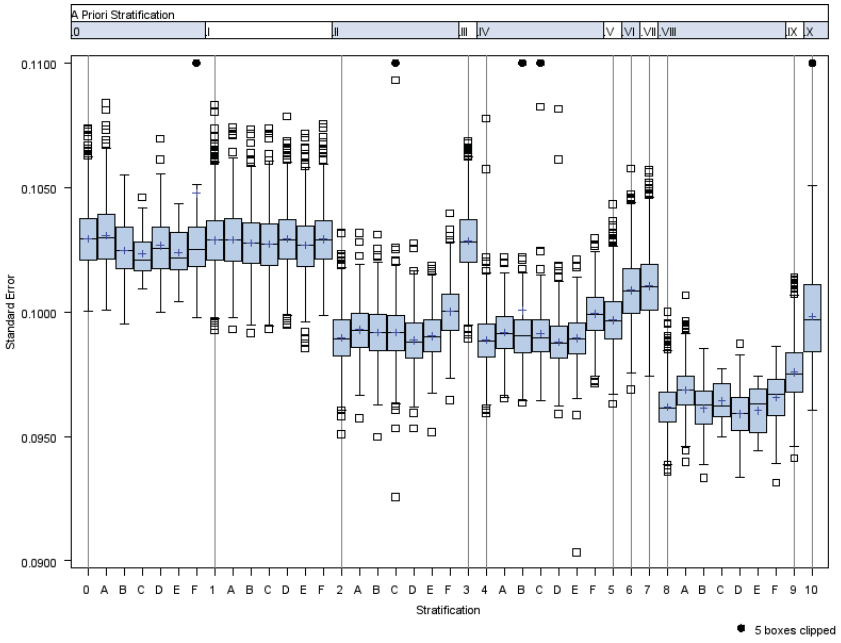
**Table 4.6:** Results of ML and WL estimation using a priori stratification VII based on phase 2 samples of size $n = 500$

| **Estimation approach** | | **ML** | **WL** |
|---|---|---|---|
| *Number of simulations:* | | $n_{sim} = 999$ | $n_{sim} = 1,000$ |
| Multivariable model | True $\beta^a$ | $\overline{\hat{\beta}}^b$ (SE) | $\overline{\hat{\beta}}^b$ (SE) |
| Phenprocoumon exposure | 1.37 | 1.37 (0.25) | 1.30 (0.33) |
| Age (centred at 55 years) | 0.038 | 0.038 (0.005) | 0.038 (0.009) |
| *Interaction:* phen. * age | -0.034 | -0.034 (0.019) | -0.031 (0.022) |
| Female sex | -0.12 | -0.12 (0.11) | -0.13 (0.17) |
| *Interaction:* phen. * sex | 0.29 | 0.29 (0.27) | 0.30 (0.34) |
| BMI $\geq$30kg/m$^2$ | 0.45 | 0.47 (0.23) | 0.49 (0.40) |
| *Interaction:* phen. * BMI | -0.39 | -0.40 (0.41) | -0.41 (0.58) |
| Current smoker | 0.83 | 0.87 (0.19) | 0.97 (0.37) |
| **Comorbid conditions:** | | | |
| Hypertension | 0.12 | 0.11 (0.12) | **-0.02 (0.20)** |
| Diabetes mellitus | 0.24 | 0.23 (0.12) | **0.08 (0.19)** |
| Ischemic heart disease | 0.05 | 0.05 (0.27) | 0.06 (0.47) |
| Liver disease | 0.27 | 0.27 (0.27) | 0.36 (0.51) |
| GI disease | 0.32 | 0.35 (0.35) | 0.47 (0.65) |
| Use of NSAIDs | 0.37 | 0.39 (0.31) | 0.49 (0.57) |
| Use of ASA | 0.56 | 0.58 (0.39) | 0.67 (0.59) |
| Use of diuretics | 0.21 | 0.22 (0.21) | 0.27 (0.38) |
| Use of statins | -0.11 | -0.12 (0.27) | -0.13 (0.48) |
| Use of gastroprotective drugs | 0.57 | 0.59 (0.28) | 0.69 (0.52) |

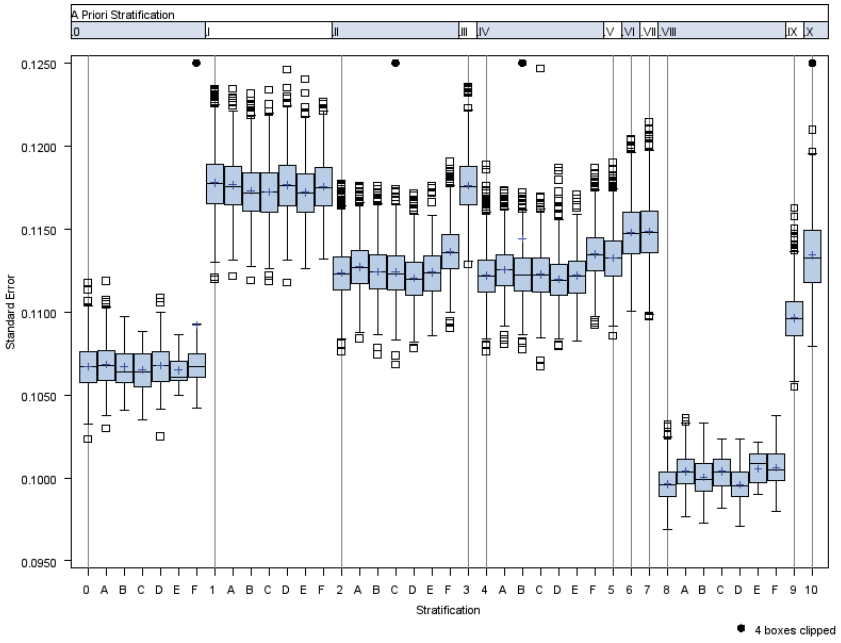
[a] Parameter that has been used for the simulation of the disease status.

[b] Parameter estimate and standard error (SE) are averaged over all simulations.

ASA: acetylsalicylic acid, GI: gastrointestinal, NSAID: non-steroidal anti-inflammatory drug, phen.: phenprocoumon




(a) Smoking




(b) BMI

**Figure 4.4:** Standard errors of ML estimators for phase 2 variables in samples of size 2,000

A priori stratifications used in the analysis are denoted by the numbers 0 - 10, post stratifications are denoted by the capitals A-F.

mation is directly included in the stratification as is exemplified for the covariate diabetes in Figure 4.5(c). A priori stratifications IX and X, which are based on a disease score, are also efficient for all other phase 1 covariates not included directly in any a priori stratification (e.g., for gastroprotective drugs see Figure 4.5(d)).

**Standard error of phase 2 variables** The standard errors (SE) for the main effects of smoking and BMI are depicted in Figure 4.4. The SE for the interaction between phenprocoumon exposure and BMI is dominated by the SE for the effect of phenprocoumon. Hence, the performance of stratifications is similar to that shown in Figure 4.5(a). Interestingly, stratification VIII results in the smallest errors for both parameter estimates. Stratifications IX and X, which are also based on a disease score, reveal smaller errors than the other a priori stratifications. The effect of BMI, however, is estimated with a better precision in the random samples (0) (see Figure 4.4(b)). These results indicate that there is no strong proxy for smoking or BMI in the phase 1 data set. In particular, inclusion of sex information does not reduce the SE of the estimated BMI parameter.

### Interplay of a priori and post stratifications

The results described in the previous section are only relevant for two-phase studies where the actually derived phase 2 sample coincides with the planned phase 2 sample. In practice, the samples may differ due to unintended selection processes such as non-response of subjects with specific characteristics. If the stratification used in the two-phase analysis does not account for all variables influencing the selection probability, parameter estimates are biased. This is one reason for employing a post stratification that differs from the a priori stratification. The issue of bias occurring when the post stratification does not account for the selective sampling of phase 2 has already been addressed in Section 4.2.4. Potential non-response is also an argument for not defining very fine a priori stratifications because non-response can lead to empty cells for sparsely populated strata. A possible solution would be to use a relatively coarse a priori stratification for sampling and a fine post stratification for the evaluation. Whether this approach is feasible and leads to efficient parameter estimation can be seen from the results described in this section.

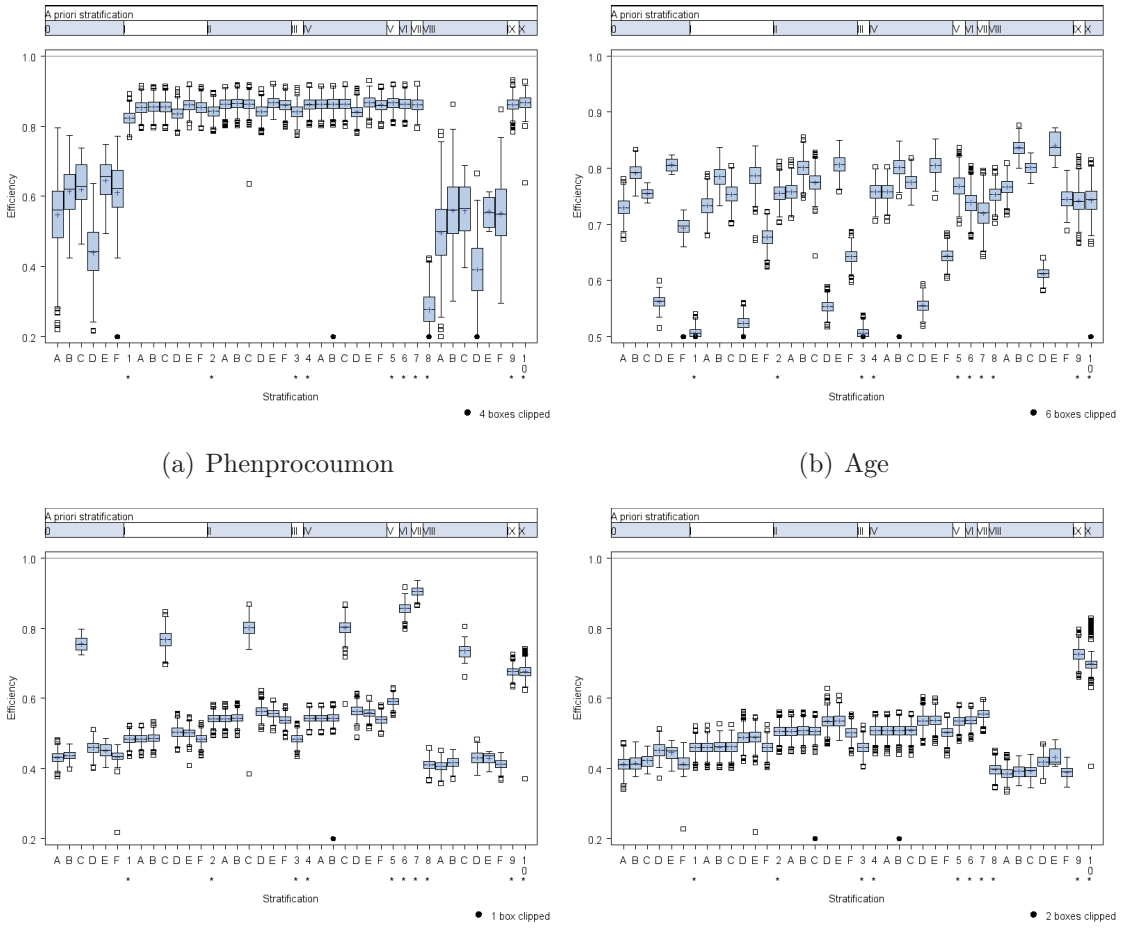
The simulation study shows that random samples and samples in which subjects with rare exposures of interest are not oversampled are associated with high failure proportions and therefore with a high risk of empty strata. Regarding efficiency, the impact of the sampling and the evaluation stratification depends on the covariate. Sampling according to rare exposures, e.g., phenprocoumon, is more efficient than including this exposure only in the post stratification. For more prevalent covariates, the impact of the a priori stratification is less important. Thus, it is sufficient to include covariates such as sex only in the post stratification. For all other phase 1 covariates in between these extremes inclusion in the a priori stratification, at least via a disease score, results in small efficiency gains. More details on the results are presented below. As results regarding failure proportions and efficiency are similar for ML and WL estimation (except for the range of efficiency), they are only presented for ML estimation but apply for WL estimation as well.

**Failure proportion** Table 4.7 shows the failure proportions observed for combinations of a priori and post stratifications. The worst failure proportions are obtained for random samples and samples drawn according to deciles of a disease score (VIII). These samples only have acceptable failure proportions if they comprise at least 2,000 subjects when analysed with stratification A or if they comprise at least 10,000 subjects when analysed with stratifications other than stratification A. The high failure proportions are caused by empty strata. They occur because sparsely populated strata are not necessarily oversampled when the phase 2 samples are not drawn according to the post stratifications. The finding that post stratifications with few strata have much lower failure proportions than stratifications with more strata has already been discussed in Section 4.2.4.

**Efficiency** Due to the high failure proportions observed for some stratifications in samples of size 500 and 1,000, the comparison of efficiency is exemplified for samples of size 2,000. The interplay of a priori and post stratifications with regard to efficiency of parameter estimation is illustrated for phenprocoumon exposure, age, diabetes mellitus, and use of gastroprotective drugs in Figure 4.5. The efficiency of parameter estimates for phenprocoumon is mainly driven by the a priori stratification. The inefficiency obtained for random samples (0) and samples stratified

**Table 4.7:** Failure proportion of ML-estimation for different a priori/post stratification combinations

| Sample size | A priori strat. | Post stratification | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | A | B | C | D | E | F | a priori |
| 500 | 0 | 94.50% | 100% | 100% | 100.00% | 100% | 100% | – |
| | I | 1.80% | 73.60% | 98.60% | 46.80% | 99.90% | 28.30% | 0.50% |
| | II | 0.10% | 79.80% | 99.00% | 20.70% | 99.60% | 3.00% | 0.10% |
| | IV | 0.40% | 78.30% | 96.50% | 19.60% | 99.90% | 2.10% | 0.40% |
| | VIII | 97.40% | 100% | 100% | 100% | 100% | 100% | 70.40% |
| 1,000 | 0 | 67.50% | 99.70% | 99.80% | 97.50% | 100% | 99.70% | – |
| | I | 0.00% | 17.80% | 66.60% | 16.70% | 90.90% | 3.50% | 0.00% |
| | II | 0.00% | 30.60% | 76.30% | 2.20% | 93.70% | 0.00% | 0.00% |
| | IV | 0.30% | 30.50% | 75.30% | 1.80% | 94.10% | 0.30% | 0.30% |
| | VIII | 66.60% | 99.90% | 100% | 98.80% | 100% | 99.90% | 20.60% |
| 2,000 | 0 | 24.30% | 92.20% | 97.40% | 83.60% | 98.60% | 92.40% | – |
| | I | 0.00% | 0.20% | 6.20% | 1.80% | 37.90% | 0.00% | 0.00% |
| | II | 0.00% | 3.30% | 21.10% | 0.00% | 64.50% | 0.00% | 0.00% |
| | IV | 0.30% | 3.30% | 20.10% | 0.00% | 63.60% | 0.30% | 0.30% |
| | VIII | 13.50% | 88.90% | 98.00% | 78.90% | 98.90% | 88.50% | 0.80% |
| 10,000 | 0 | 0.00% | 0.20% | 5.00% | 2.40% | 6.50% | 0.20% | – |
| | I | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| | II | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| | IV | 0.30% | 0.30% | 0.30% | 0.00% | 0.30% | 0.30% | 0.30% |
| | VIII | 0.00% | 0.20% | 2.30% | 0.80% | 3.30% | 0.20% | 0.00% |




(a) Phenprocoumon

(b) Age

(c) Diabetes mellitus

(d) Gastroprotective drugs

**Figure 4.5:** Efficiency of ML estimators in phase 2 samples of size 2,000

A priori stratifications used in the analysis are denoted by the numbers 1* - 10*, post stratifications are denoted by the capitals A-F.

by the disease score (VIII) is not improved by any post stratification (see Figure 4.5(a)). For an efficient estimation of the age parameter, it is only important that age is included in the post stratification (see Figure 4.5(b)). To be more precise, stratifications D, I, and III, which do not include age, have the worst efficiencies. Moreover, including age in the sampling stratification does not lead to a gain in efficiency for the estimation of the age parameter as can be seen for stratifications II and IV. However, sampling on age slightly improves the efficiency of parameter estimates for covariates related to age (see Figure 4.5(c) and Figure 4.5(d). Regarding all other phase 1 covariates, an efficient parameter estimation depends on both, the sampling and the evaluation stratification. This is best illustrated by the example of diabetes mellitus: It is apparent from Figure 4.5(c) that inclusion of diabetes mellitus in the post stratification (C) improves the efficiency. An even higher efficiency is achieved if the a priori stratification includes diabetes mellitus (VI, VII). Although the direct inclusion of a covariate is more efficient than the inclusion via a disease score, considerable gains in efficiency are observed for most phase 1 covariates when the sampling and evaluation stratification include percentiles of a disease score (see stratifications IX and X in Figure 4.5(d)). In particular, these stratifications are more efficient for the parameter estimation for most phase 1 covariates then a priori stratifications including age information (II, IV).

### 4.5.3 Discussion

A summary of the simulation results regarding all three objectives of the study is at the same time also a summary of the whole chapter. Using a disease score for the definition of the stratification permits the inclusion of phase 1 information on many covariates thereby improving the efficiency of parameter estimation. Since covariate information is considered only indirectly, stratifications based on disease scores are coarser with respect to the specific covariate than stratifications defined by cross-classification. Therefore, a combination of both strategies, i.e. cross-classification by important variables and percentiles of a disease score, results in the best performance with respect to bias and efficiency. To avoid bias, important variables mean at least all variables used for sampling.

In the light of these results it is reassuring that these stratifications are chosen

by the design criterion. The simulation study confirms the choice suggested by the criterion also for ML estimation, although the criterion is derived from the WL approach. In the investigated data scenario, the selected stratifications are also efficient for the estimation of parameters related to phase 2 variables. This result probably cannot be transferred to two-phase studies which include proxy information for phase 2 variables in the phase 1 data set.

The simulation study has also demonstrated that cross-classification of several phase 1 covariates is not only associated with the problem of empty strata. When using the WL approach for parameter estimation, this stratification strategy may also lead to unstable results caused by the high variability of sampling weights.

In conclusion, the following recommendation can be derived from the simulation study for the definition of a priori and post stratifications for two-phase studies with a rich phase 1 data set. At the planning stage of the study, when only phase 1 data is available, the a priori stratification should be defined based on prior knowledge and the design criterion described in Section 4.4. For step 2 of the planning process, i.e., for the definition of candidate stratifications, it should be taken into account that variables like sex, which are very common, do not need to be considered in the a priori stratification. Instead of including information on age in the a priori stratification the inclusion of percentiles of a disease score yields more efficient estimates. Very rare exposures of interest should be included in any case. For the analysis of the phase 2 sample, it is crucial that the post stratification is finer than the a priori stratification and includes all variables influencing the selection probability of the phase 2 sample. Therefore, a response analysis should be conducted to identify all relevant phase 1 variables. Furthermore, age should be included in the post stratification for two reasons: (1) age itself is a risk factor for the disease; (2) age is related to most comorbidities and use of comedications. It has to be noted that this recommendation may only be valid for similar two-phase studies investigating other pharmacoepidemiological research questions or for studies with a similar data structure. This data structure is characterised by a rare exposure of interest and mostly binary covariates. The impact of age is likely to be the same in other pharmacoepidemiological studies.

# Chapter 5

# Beyond two-phase methods: Approaches for using the full phase 1 information

Two-phase methodology is not the only approach to incorporate phase 1 information into the analysis of the phase 2 data set. Calibration and estimation of weights in a weighted analysis as well as multiple imputation are survey methodological approaches which can be used in the setting of a two-phase study.

Calibration of weights was proposed by Deville et al., 1993 for the inclusion of information on auxiliary variables that are available for the whole study population. In context of a two-phase study, the phase 1 variables can be understood as auxiliary variables. The weights, defined by the inverse sampling fractions for the WL approach (see Section 3.3.2), are adjusted according to the condition that the weighted phase 2 totals of the phase 1 variables equal the known phase 1 totals, i.e.

$$\sum_{i=1}^{n} w_i \mathbf{y}_i = \sum_{i=1}^{N} \mathbf{y}_i,$$

where $w_i$, $i = 1, \ldots, n$, are the adjusted weights and $\mathbf{y}_i$ is the vector of phase 1 covariates for subject $i$.

Robins et al., 1994 suggested to estimate the weights by a model that predicts the sampling probability based on phase 1 variables. To achieve unbiased estimates

in a stratified two-phase study, the prediction model should also include the a priori stratification, i.e., the model should be defined by $Pr(R^{II} = 1|\mathbf{Y}, \mathscr{S})$. This has been noted by Breslow et al., 2013 who compared calibration and estimation of weights with WL, PL and ML estimation in a two-phase stratified case-control study. Compared to WL estimation, they found reduced standard errors of the estimated coefficients for the phase 1 variables but also differences in the parameter estimates for phase 1 as well as phase 2 variables. Standard errors obtained with adjusted weights were not smaller than those obtained with ML estimation. A detailed discussion of using calibrated and estimated weights for the analysis of the particular two-phase studies considered in this thesis is beyond the scope of this chapter.

Multiple imputation was introduced by Rubin, 1978 to handle non-response in sample surveys. It is a "filling-in" method, because the missing values are replaced by plausible estimates. Instead of imputing only one value for each missing observation, the imputation is repeated several times to reflect the uncertainty related to the imputation process. Details of the method are described in the next section. Marti and Chavance, 2011 compared multiple imputation to weighted methods using inverse probability, calibrated, and estimated weights in a case-cohort study which can also be interpreted as a two-phase study. They found that the multiple imputation estimator was slightly more efficient than the weighted estimators for the phase 1 variables but it was also biased in some situations.

A simple multiple imputation method called "hot-deck multiple imputation" has been shown to be equivalent to WL estimation (Reilly and Pepe, 1997). For this approach, missing values are replaced by randomly taking values from subjects with the same observed variables instead of estimating these values. Since WL estimation is less efficient than ML estimation, hot-deck imputation is not further considered. Rubin, 1987 proved that the variance estimator developed for multiple imputation is not appropriate for estimating the variance of the hot-deck multiple imputation estimator because it is biased. Hence, the variance of the multiple imputation estimator differs from the WLE variance and might especially be more efficient. How multiple imputation performs in two-phase studies is investigated in this chapter. After a brief introduction of the theory, the approach is applied to the empirical two-phase study. Furthermore, a simulation study similar to the study described in

Section 4.5 is conducted to assess bias in multiple imputation estimates.

## 5.1 Introduction to multiple imputation

Multiple imputation has been widely used for the handling of missing data in various fields of application including epidemiology (see e.g. Rubin, 1996, for a selection of examples). A detailed description of the technique, its requirements and its properties is given for instance in Rubin, 1987, Schafer, 1999, and Little and Rubin, 2002. The concept is briefly summarised in the following.

Multiple imputation is conducted in three steps: First, $M \geq 2$ complete data sets are generated by replacing each missing value by a $M$-dimensional vector of plausible values. These values are simulated with respect to an imputation model estimating the predictive distribution of the missing data. Second, the $M$ complete data sets are analysed with any adequate method for the analysis of complete data sets to obtain the desired estimate $\hat{\boldsymbol{\theta}}_m$, $m = 1, \ldots, M$. Third, the $M$ estimates $\hat{\boldsymbol{\theta}}_m$ are combined to constitute the multiple imputation estimate and the respective variance for statistical inference. The multiple imputation estimate is the mean of the $M$ estimates

$$\hat{\boldsymbol{\theta}}_{IM} = \frac{1}{M} \sum_{m=1}^{M} \hat{\boldsymbol{\theta}}_m.$$

The total variance $\mathrm{Var}(\hat{\boldsymbol{\theta}}_{IM})$ is composed of the estimated between-imputation variance $\hat{B}$ and the average within-imputation variance estimate $\hat{W}$

$$\begin{aligned}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_{IM}) &= \frac{M+1}{M}\hat{B} + \hat{W} \\ &= \frac{M+1}{M}\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_{IM}\right)\left(\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_{IM}\right)^T + \frac{1}{M}\sum_{m=1}^{M}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_m).\end{aligned}$$

Two requirements have to be fulfilled to achieve unbiased estimates $\hat{\boldsymbol{\theta}}_{IM}$: (1) the missing-at-random (MAR) assumption has to be fulfilled; (2) the imputation model has to give reasonable predictions for the missing observations. Particularly, the imputation model and the analysis model applied in the second step have to coincide with respect to the underlying assumptions. If the imputation model for instance ignores interaction terms, these interactions are estimated with bias in the analysis model (see Schafer, 1999, for a discussion of this issue).

Although multiple imputation has been developed for missing data that arises from mechanisms like non-response or loss to follow-up, two-phase database studies can be perceived as a missing-data situation in which multiple imputation can be applied. The first requirement, the MAR assumption, also needs to be fulfilled for the application of two-phase methods. Hence, no additional assumption is imposed. The second requirement depends on the availability of information about the missing observations. In two-phase database studies, a multitude of phase 1 covariates is available to define the imputation model. Thus, for sufficiently large phase 2 samples the second requirement is most probably satisfied unless the imputation model is misspecified.

Besides the imputation model, the method used for imputation is important for the validity of results. Several imputation methods have been implemented in SAS software (see SAS/STAT 9.2 User's Guide). The most commonly applied method is Markov chain Monte Carlo (MCMC) which can be used for continuous, categorical and mixed multivariate data. If the data has a monotone missing pattern, the logistic regression method (Rubin, 1987) and the propensity score method (Lavori et al., 1995) can be applied to impute ordinal and continuous covariates, respectively. Variables $Z_1, \ldots, Z_p$ have a monotone missing pattern if a missing value for $Z_j$ implies that values for $Z_k$, $k > j$, are also missing. Since values of phase 2 covariates are missing for all subjects outside of the phase 2 sample and are observed for all subjects in the phase 2 sample, two-phase studies have a monotone missing pattern. Both methods are easier to apply than MCMC because convergence of MCMC is often difficult to verify. These methods are therefore employed in the application to the empirical study as well as in the simulation study. A sketch of the methods is given below.

**Logistic regression method** To impute values of an ordinal variable $Z$ a logistic regression model is fit to the observed values of $Z$ and explanatory variables $Y_1, \ldots, Y_k$:

$$\text{logit}(Pr(Z \leq i|\mathbf{y};\boldsymbol{\theta}_i)) = \alpha_i + \boldsymbol{\beta}^T\mathbf{y}, \quad i = 1, \ldots, I.$$

The resulting parameter estimates $\hat{\boldsymbol{\theta}}_i^T = (\hat{\alpha}_i, \hat{\beta}_1, \ldots, \hat{\beta}_k)$, $i = 1, \ldots, I$, and the corresponding covariance matrix are used to define the posterior predictive distribution of $\boldsymbol{\theta}_i$ given by $\boldsymbol{\theta}_i|Z, \mathbf{Y} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_i, \widehat{\text{Cov}}(\hat{\boldsymbol{\theta}}_i))$. For the $m$th imputation of $Z$,

new parameters $\boldsymbol{\theta}_i^m$ are drawn from the posterior distribution and the probabilities $p_i^m = Pr(Z \leq i|\mathbf{y}; \boldsymbol{\theta}_i^m)$ are calculated for a subject with missing $Z$ and covariate values $\mathbf{y}$. These probabilities are compared to a random number $u^m$ generated from a uniform distribution over the interval $[0, 1]$. Values of $Z$ are then imputed according to the following rule: If $u^m \leq p_1^m$, $Z$ is set to 1; if $p_1^m < u^m \leq p_2^m$, $Z$ is set to 2; and so forth.

The performance of the logistic regression method for the imputation of categorical variables has been investigated by Allison, 2005 who concludes that this method is preferable to linear imputation methods such as MCMC.

**Propensity score method** To impute values of a continuous variable $Z$, an approximate Bayesian bootstrap imputation is conducted that draws observations from groups defined by quintiles of the propensity score representing the probability of a missing value in $Z$ given the covariates $\mathbf{Y}$. To be more precise, the imputation consists of the following steps: Let $R$ denote the indicator for missing observations in $Z$ with $R = 0$ if $Z$ is missing and $R = 1$ if $Z$ is observed. The first step is then to estimate the propensity score $p = Pr(R = 0|\mathbf{y})$ for each observation by logistic regression. In the second step, the $n_0$ observations with missing values for $Z$, $Z_{mis}$, and the $n_1$ observations with observed values for $Z$, $Z_{obs}$, are divided into five groups according to their propensity score resulting in groups of sizes $n_{0k}$ and $n_{1k}$, $k = 1, \ldots, 5$. In the third step, $n_{1k}$ observations are randomly drawn with replacement from the $k$th group of $Z_{obs}$ to generate a new set $Z^*_{obs,k}$. Finally, to impute values for $Z_{mis,k}$, $n_{0k}$ observations are drawn randomly with replacement from $Z^*_{obs,k}$.

It has been mentioned by Schafer, 1999 that the propensity score method may not be appropriate for estimating the relationship between $Z$ and a covariate which is unrelated to the indicator of missing information $R$.

In contrast to most applications of multiple imputation, the proportion of missing information is usually much higher in two-phase studies. For example in the empirical two-phase study more than 98% of the information on BMI and smoking is missing. A further difference between two-phase studies and common missing-data situations is that missing data in two-phase studies results from the design

instead of from an unintended (by the investigator) or random process. In the next section, multiple imputation is applied to the empirical two-phase study to assess if the technique is also reasonable for relatively small phase 2 data sets obtained by stratified sampling and if it is superior to two-phase methodology concerning bias and standard errors of the parameter estimates.

## 5.2 Multiple imputation in the empirical study

In the empirical study on phenprocoumon exposure and serious bleedings, the variables BMI and smoking have to be imputed based on the phase 2 data set to obtain a complete data set. Both variables have been included as binary covariates in all previous analyses, although smoking information is available in three categories and BMI is observed as continuous covariate. Five imputation models are used to impute values for smoking and BMI. While smoking information is always imputed in three categories, BMI is considered as binary, categorical (four categories) or continuous variable. The main focus is on the imputation of categorical values for BMI to be consistent with the simulation study described in Section 4.5. To evaluate the performance of multiple imputation utilising the maximum and minimum amount of information on BMI, respectively, additional imputation models are applied to impute continuous and binary values for BMI. Two further models are employed that account for the stratified sampling of the phase 2 data set. Details on each model are specified in the following:

**IM1** **Imputation Model 1** employs the propensity score method to generate continuous values for BMI and the logistic regression method to derive categorical values of smoking. For both methods, disease status and all phase 1 variables are included as explanatory variables in the model.

**IM2** **Imputation Model 2** imputes categorical values for BMI and smoking by using the logistic regression method with the same set of explanatory variables as in IM1.

**IM3** **Imputation Model 3** generates binary values for BMI and smoking by using the logistic regression method with the same set of explanatory variables as in IM1.

**IM4** **Imputation Model 4** includes separate imputations in the four strata defined by disease and phenprocoumon exposure. In each stratum, categorical values of BMI and smoking are imputed by a logistic regression model including all phase 1 variables except phenprocoumon exposure.

**IM5** **Imputation Model 5** also uses the logistic regression method to generate categorical values for BMI and smoking. Stratum indicators for disease status and phenprocoumon exposure are explicitly included as covariates in the model in addition to all phase 1 variables except phenprocoumon exposure.

All imputations are conducted using the SAS procedure PROC MI.

Twentyfive complete data sets are imputed with each imputation model. Usually, only five to ten imputations are recommended (see e.g. Schafer, 1999) because the relative efficiency of an estimate based on $M$ imputations compared to an infinite number of imputations is $(1+\frac{\lambda}{M})^{-1}$, where $\lambda$ is the proportion of missing information. The rationale for choosing $M = 25$ is the high proportion of missing information in this two-phase study. Each complete data set is analysed with two logistic regression models. The first analysis model includes all phase 1 variables, two binary covariates for BMI $\geq 30$ kg/m$^2$ and current smoking as well as interaction terms for phenprocoumon exposure and age and sex, respectively. The second analysis model additionally includes an interaction term for BMI and phenprocoumon exposure. The SAS procedure PROC MIANALYZE is used to combine the results according to the rules described in the previous section.

Estimates obtained from the four multiple imputation models are compared to phase 1 and two-phase estimates for both analysis models in Table 5.1 and Table 5.2. The estimated coefficients differ between the imputation models especially for phenprocoumon exposure, smoking and the interaction between phenprocoumon and BMI. In particular, the parameter estimate for phenprocoumon based on data sets generated according to IM5 is much smaller than those based on other data sets. Except for the standard error of the phenprocoumon coefficient for IM5, the standard errors for multiple imputation estimates of phase 1 variables are comparable to those obtained from the phase 1 analysis. When comparing imputation models IM1, IM2, and IM3 with respect to the estimation of coefficients for the phase 1 variables, it can be seen that IM1 is more efficient than IM2 and IM2 is more efficient than

IM3. This indicates that categorising as well as dichotomising BMI information is associated with a loss in efficiency. The stratified imputation implemented for IM4 is also associated with a slight loss in efficiency for these variables. Regarding estimation of the phase 2 variables in the model without BMI interaction, IM1 and IM4 are more efficient than IM2, IM3 and IM5. The largest standard errors in both analysis models are observed for IM5.

The results suggest that multiple imputation is much more efficient for the estimation of the phase 1 variables than two-phase methods whereas phase 2 variables can be estimated with a comparable precision. However, the estimated coefficients differ between the models especially for the variable of main interest. As the true parameter value is not known, it cannot be concluded which estimate is biased. To investigate the potential for bias in the four imputation models, a simulation study is conducted and described in the next section.

**Table 5.1:** Comparison of phase 1, two-phase, and multiple imputation analysis of the empirical study (analysis model without BMI interaction)

| Multivariable model | Log OR (SE) | | | | | | |
|---|---|---|---|---|---|---|---|
| | Phase 1 | Two-phase | Multiple imputation | | | | |
| N=26 208 | | Strat. E* | IM1 | IM2 | IM3 | IM4 | IM5 |
| Phenprocoumon use | 1.37 (0.18) | 1.44 (0.22) | 1.39 (0.18) | 1.39 (0.18) | 1.39 (0.18) | 1.33 (0.19) | 0.97 (0.34) |
| Age[a] | 0.04 (<.01) | 0.06 (0.01) | 0.04 (<.01) | 0.04 (<.01) | 0.04 (<.01) | 0.04 (<.01) | 0.04 (0.01) |
| IA: phen.×age | -0.04 (0.01) | -0.04 (0.02) | -0.03 (0.01) | -0.04 (0.01) | -0.03 (0.01) | -0.02 (0.01) | -0.04 (0.01) |
| Female sex | -0.18 (0.06) | -0.14 (0.10) | -0.08 (0.08) | -0.04 (0.09) | -0.14 (0.07) | -0.05 (0.09) | -0.09 (0.09) |
| IA: phen.×sex | 0.39 (0.24) | 0.55 (0.28) | 0.38 (0.24) | 0.38 (0.24) | 0.39 (0.24) | 0.42 (0.25) | 0.30 (0.26) |
| *Comorbid conditions:* | | | | | | | |
| Diabetes mellitus | 0.29 (0.09) | 0.14 (0.25) | 0.28 (0.09) | 0.26 (0.10) | 0.25 (0.11) | 0.27 (0.10) | 0.27 (0.10) |
| Hypertension | 0.16 (0.08) | 0.50 (0.23) | 0.15 (0.08) | 0.14 (0.11) | 0.11 (0.10) | 0.15 (0.10) | 0.16 (0.10) |
| Ischemic heart disease | 0.10 (0.11) | 0.21 (0.28) | 0.06 (0.12) | 0.09 (0.14) | 0.09 (0.14) | 0.02 (0.14) | 0.05 (0.12) |
| Liver disease | 0.56 (0.10) | 0.67 (0.29) | 0.59 (0.10) | 0.57 (0.11) | 0.55 (0.11) | 0.61 (0.13) | 0.57 (0.11) |
| Gastrointestinal disease | 0.44 (0.12) | 0.61 (0.35) | 0.44 (0.12) | 0.45 (0.12) | 0.40 (0.14) | 0.42 (0.15) | 0.46 (0.12) |
| *Current use of:* | | | | | | | |
| NSAIDs | 0.38 (0.11) | 0.40 (0.31) | 0.37 (0.11) | 0.37 (0.12) | 0.38 (0.13) | 0.43 (0.14) | 0.35 (0.12) |
| ASA | 0.60 (0.13) | 0.51 (0.40) | 0.61 (0.14) | 0.59 (0.15) | 0.57 (0.16) | 0.49 (0.18) | 0.58 (0.15) |
| Diuretics | 0.21 (0.08) | -0.56 (0.25) | 0.20 (0.09) | 0.18 (0.09) | 0.16 (0.11) | 0.18 (0.11) | 0.17 (0.10) |
| Statins | -0.05 (0.12) | -0.28 (0.34) | -0.04 (0.13) | -0.03 (0.13) | 0.04 (0.13) | 0.02 (0.14) | -0.02 (0.13) |
| Gastroprotective drugs | 0.73 (0.09) | 0.65 (0.31) | 0.74 (0.10) | 0.74 (0.10) | 0.80 (0.11) | 0.74 (0.12) | 0.74 (0.10) |
| *Phase 2 variables:* | | | | | | | |
| BMI ≥30kg/m$^2$ | — | 0.41 (0.21) | 0.26 (0.15) | 0.19 (0.24) | 0.37 (0.24) | 0.21 (0.21) | 0.24 (0.94) |
| Current smoker | — | 0.72 (0.22) | 0.59 (0.21) | 0.65 (0.26) | 0.77 (0.21) | 0.99 (0.20) | 0.91 (0.91) |

[a] Age is centred around 55 years.

SE: standard error, IA: interaction, NSAID: non-steroidal anti-inflammatory drug, ASA: acetylsalicylic acid, BMI: body mass index

**Table 5.2:** Comparison of phase 1, two-phase, and multiple imputation analysis of the empirical study (analysis model with BMI interaction)

| Multivariable model | Log OR (SE) | | | | | | |
|---|---|---|---|---|---|---|---|
| | Phase 1 | Two-phase | Multiple imputation | | | | |
| N=26 208 | | Strat. E* | IM1 | IM2 | IM3 | IM4 | IM5 |
| Phenprocoumon use | 1.37 (0.18) | 1.44 (0.22) | 1.53 (0.23) | 1.46 (0.22) | 1.46 (0.23) | 1.61 (0.30) | 0.83 (1.33) |
| Age[a] | 0.04 (<.01) | 0.06 (0.01) | 0.04 (<.01) | 0.04 (<.01) | 0.05 (<.01) | 0.04 (<.01) | 0.04 (<.01) |
| IA: phen.×age | -0.04 (0.01) | -0.04 (0.02) | -0.03 (0.01) | -0.03 (0.01) | -0.03 (0.01) | -0.02 (0.02) | -0.04 (0.02) |
| Female sex | -0.18 (0.06) | -0.14 (0.10) | -0.08 (0.08) | -0.04 (0.09) | -0.14 (0.07) | -0.04 (0.09) | -0.09 (0.09) |
| IA: phen.×sex | 0.39 (0.24) | 0.55 (0.28) | 0.38 (0.24) | 0.38 (0.24) | 0.40 (0.24) | 0.44 (0.26) | 0.21 (0.33) |
| *Comorbid conditions:* | | | | | | | |
| Diabetes mellitus | 0.29 (0.09) | 0.14 (0.25) | 0.28 (0.09) | 0.26 (0.10) | 0.25 (0.11) | 0.26 (0.10) | 0.29 (0.10) |
| Hypertension | 0.16 (0.08) | 0.50 (0.23) | 0.16 (0.08) | 0.14 (0.11) | 0.11 (0.10) | 0.15 (0.10) | 0.17 (0.09) |
| Ischemic heart disease | 0.10 (0.11) | 0.21 (0.28) | 0.06 (0.12) | 0.09 (0.14) | 0.09 (0.14) | 0.06 (0.16) | 0.04 (0.12) |
| Liver disease | 0.56 (0.10) | 0.67 (0.29) | 0.59 (0.10) | 0.57 (0.11) | 0.55 (0.11) | 0.61 (0.13) | 0.57 (0.11) |
| Gastrointestinal disease | 0.44 (0.12) | 0.61 (0.35) | 0.44 (0.12) | 0.45 (0.12) | 0.40 (0.14) | 0.41 (0.15) | 0.47 (0.12) |
| *Current use of:* | | | | | | | |
| NSAIDs | 0.38 (0.11) | 0.40 (0.31) | 0.37 (0.11) | 0.37 (0.12) | 0.38 (0.13) | 0.43 (0.13) | 0.36 (0.12) |
| ASA | 0.60 (0.13) | 0.51 (0.40) | 0.60 (0.14) | 0.59 (0.15) | 0.57 (0.16) | 0.49 (0.18) | 0.56 (0.15) |
| Diuretics | 0.21 (0.08) | -0.56 (0.25) | 0.20 (0.09) | 0.18 (0.09) | 0.16 (0.11) | 0.17 (0.11) | 0.17 (0.10) |
| Statins | -0.05 (0.12) | -0.28 (0.34) | -0.05 (0.13) | -0.03 (0.13) | 0.04 (0.13) | 0.00 (0.14) | 0.00 (0.13) |
| Gastroprotective drugs | 0.73 (0.09) | 0.65 (0.31) | 0.74 (0.10) | 0.74 (0.10) | 0.80 (0.11) | 0.73 (0.12) | 0.75 (0.10) |
| *Phase 2 variables:* | | | | | | | |
| BMI $\geq$30kg/m$^2$ | — | 0.41 (0.21) | 0.29 (0.17) | 0.20 (0.25) | 0.38 (0.26) | 0.27 (0.23) | 0.15 (0.64) |
| IA: phen.×BMI | — | -0.59 (0.54) | -0.37 (0.39) | -0.18 (0.32) | -0.17 (0.33) | -0.70 (0.69) | 0.05 (2.73) |
| Current smoker | — | 0.72 (0.22) | 0.59 (0.21) | 0.65 (0.26) | 0.77 (0.21) | 0.99 (0.20) | 0.89 (0.85) |

[a] Age is centred around 55 years.

SE: standard error, IA: interaction, NSAID: non-steroidal anti-inflammatory drug, ASA: acetylsalicylic acid, BMI: body mass index

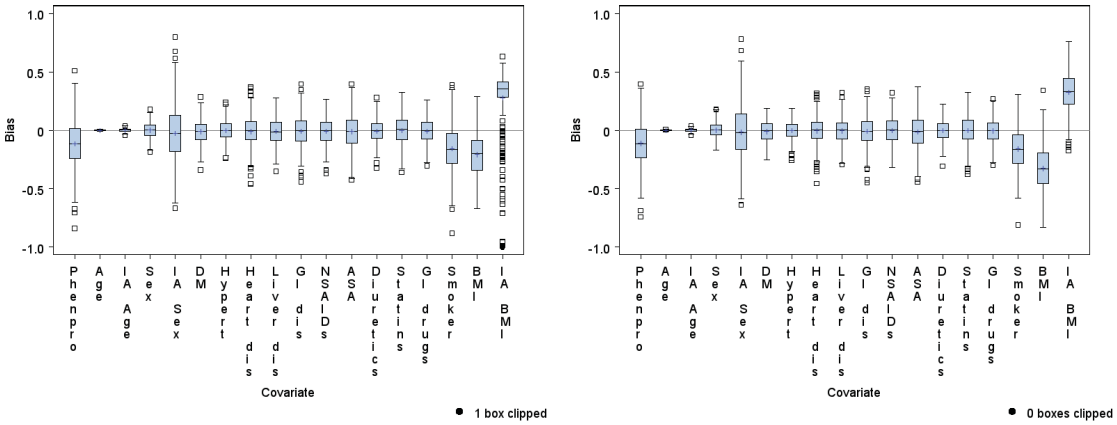
## 5.3 Simulation study to assess bias in multiple imputation analyses

To reduce complexity of the simulation study, only imputation models IM2 and IM4 are investigated. Four scenarios are simulated with 500 repetitions each, of which three address bias occurring with IM2 and one evaluates bias occurring with IM4. For Scenarios 1, 2 and 4, phase 1 data sets are simulated as described in Chapter 4.5. For simulation of Scenario 3, the phase 1 data set is generated according to a disease model omitting the interaction between phenprocoumon exposure and BMI. Phase 2 samples of 500, 1,000, and 2,000 subjects are drawn from each phase 1 data set by simple random sampling (Scenarios 1 and 3) or by stratified sampling according to a priori stratification IV, i.e. with respect to phenprocoumon exposure, age, and sex (Scenarios 2 and 4). Further a priori stratifications and phase 2 sample sizes considered in the simulation study of Section 4.5 are not implemented, again to reduce complexity of this study. Values of the phase 2 variables BMI and smoking are first deleted for all subjects not included in the phase 2 subsamples and afterwards imputed according to IM2 or IM4. The characteristics of the scenarios are summarised in Table 5.3.

**Table 5.3:** Simulation scenarios for investigation of bias

| | Scenario | | | |
|---|---|---|---|---|
| Characteristic | 1 | 2 | 3 | 4 |
| Disease model includes interaction with BMI | × | × | | × |
| Random sample in phase 2 | × | | × | |
| Stratified sample in phase 2 | | × | | × |
| Imputation Model 2 | × | × | × | |
| Imputation Model 4 | | | | × |

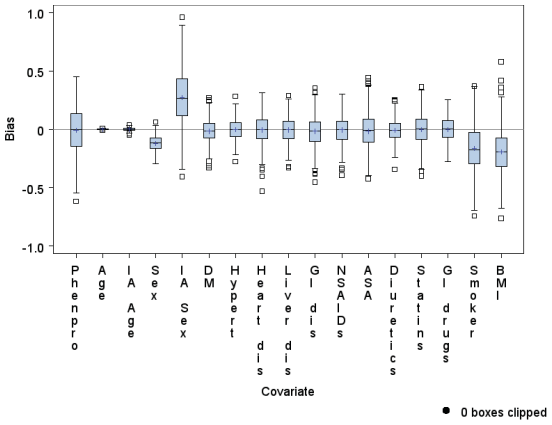
The bias observed in each of the scenarios is depicted for phase 2 samples of 500 subjects in Figure 5.1. Coefficients for BMI and smoking are estimated with bias in scenarios 1, 2 and 3. Scenarios 1 and 2, which are based on a phase 1 data set simulated according to a disease model including BMI interaction, also the parameter estimate for the interaction between phenprocoumon exposure and BMI as



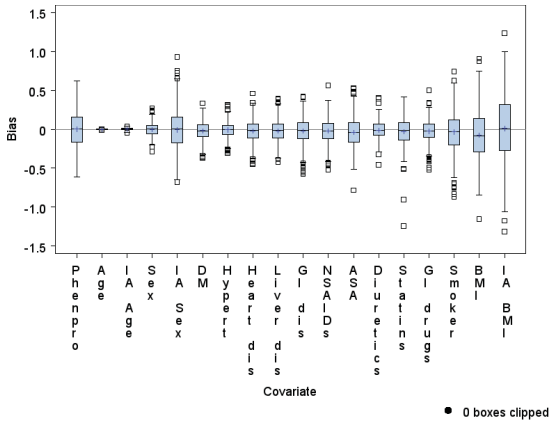

(a) Scenario 1




(b) Scenario 2




(c) Scenario 3




(d) Scenario 4

**Figure 5.1:** Bias in multiple imputation analysis in phase 2 samples of size 500

dis: disease, DM: diabetes mellitus, hypert: hypertension, IA: interaction, Phenpro: phenprocoumon

well as the parameter estimate for phenprocoumon exposure are biased. The bias is slightly larger in Scenario 2 which differs from Scenario 1 only by the selection of the phase 2 sample. Whereas Scenario 1 involves random phase 2 samples, stratified sampling has been conducted to sample the phase 2 data sets in Scenario 2. When comparing Scenario 1 with Scenario 3 it becomes apparent, that the bias in Scenario 1 is not only caused by the interaction between phenprocoumon exposure and BMI because the bias of parameter estimates of BMI and smoking is similar in both scenarios. Additional bias can be seen in Scenario 3 for sex and the interaction between phenprocoumon exposure and sex. Biased sampling of the phase 2 samples can be excluded as reason for the bias because both scenarios are based on random phase 2 samples. The only plausible explanation for the bias is that the simulation of BMI and smoking for the phase 1 data set is based on the empirical distribution of these variables in the phase 2 sample of the empirical study which is a stratified sample. If the stratified sampling is accounted for by imputing the missing values separately in the four strata defined by phenprocoumon exposure and disease status, as has been done in Scenario 4, the parameter estimation is unbiased thereby confirming the explanation. These results are based on phase 2 samples of size 500 meaning that 98% of the information on BMI and smoking has been imputed. Slightly less bias is observed for phase 2 samples of 1,000 and 2,000 subjects for which 96% and 92% of the data has been imputed.

The average parameter estimates and standard errors are summarised in Table 5.4 for IM4 based on phase 2 samples of 500 and 2,000 subjects. For comparison, the table also shows the true parameter vector and results of the phase 1 analysis as well as for IM2 based on phase 2 samples of size 2,000. As has been seen before, the parameter estimates obtained with IM4 are unbiased. Furthermore, for samples of size 2,000 the standard error for coefficients of phase 1 covariates is nearly as small as the respective error in the phase 1 analysis. Standard errors for coefficients of the phase 2 variables are of the same magnitude than those obtained from two-phase analyses (see Table E.3 in Appendix E). Compared to standard errors obtained from IM2, the loss in efficiency for IM4 is very small when using phase 2 samples of 2,000 subjects.

**Table 5.4:** Simulation results: parameter estimates and standard errors in multiple imputation analyses

| **Analysis** | | **Phase 1** | **IM4** | **IM4** | **IM2** |
|---|---|---|---|---|---|
| *Size of phase 2 sample:* | | | $n = 500$ | $n = 2,000$ | $n = 2,000$ |
| Multivariable model | True $\beta^a$ | $\overline{\beta}^b$ (SE) | $\overline{\beta}^b$ (SE) | $\overline{\beta}^b$ (SE) | $\overline{\beta}^b$ (SE) |
| Phenprocoumon exposure | 1.37 | 1.22 (0.17) | 1.38 (0.24) | 1.35 (0.19) | 1.31 (0.19) |
| Age (centred at 55 years) | 0.04 | 0.04 (<0.01) | 0.04 (<0.01) | 0.04 (<0.01) | 0.04 (<0.01) |
| *Interaction:* phen. * age | -0.03 | -0.03 (0.01) | -0.03 (0.01) | -0.03 (0.01) | -0.03 (0.01) |
| Female sex | -0.12 | -0.12 (0.05) | -0.12 (0.09) | -0.12 (0.06) | -0.12 (0.05) |
| *Interaction:* phen. * sex | 0.29 | 0.27 (0.23) | 0.28 (0.25) | 0.27 (0.23) | 0.27 (0.23) |
| BMI $\geq$30kg/m$^2$ | 0.45 | – | 0.38 (0.28) | 0.38 (0.11) | 0.23 (0.10) |
| *Interaction:* phen. * BMI | -0.39 | – | -0.38 (0.44) | -0.32 (0.26) | -0.18 (0.26) |
| Current smoker | 0.83 | – | 0.80 (0.28) | 0.81 (0.11) | 0.70 (0.09) |
| **Comorbid conditions:** | | | | | |
| Hypertension | 0.12 | 0.11 (0.07) | 0.11 (0.10) | 0.11 (0.07) | 0.12 (0.07) |
| Diabetes mellitus | 0.24 | 0.23 (0.08) | 0.22 (0.13) | 0.23 (0.09) | 0.23 (0.08) |
| Ischemic heart disease | 0.05 | 0.05 (0.10) | 0.03 (0.15) | 0.04 (0.11) | 0.05 (0.11) |
| Liver disease | 0.27 | 0.26 (0.09) | 0.25 (0.15) | 0.27 (0.10) | 0.27 (0.10) |
| GI disease | 0.32 | 0.31 (0.11) | 0.30 (0.18) | 0.31 (0.12) | 0.31 (0.11) |
| Use of NSAIDs | 0.37 | 0.36 (0.09) | 0.35 (0.15) | 0.37 (0.11) | 0.37 (0.10) |
| Use of ASA | 0.56 | 0.54 (0.12) | 0.52 (0.20) | 0.55 (0.14) | 0.55 (0.13) |
| Use of diuretics | 0.21 | 0.20 (0.07) | 0.20 (0.12) | 0.21 (0.08) | 0.21 (0.08) |
| Use of statins | -0.11 | -0.11 (0.11) | -0.14 (0.16) | -0.11 (0.12) | -0.11 (0.11) |
| Use of gastroprotective drugs | 0.57 | 0.56 (0.09) | 0.55 (0.14) | 0.56 (0.10) | 0.57 (0.09) |

[a] Parameter that has been used for the simulation of the disease status.

[b] Parameter estimate and standard error (SE) are averaged over all simulations.

ASA: acetylsalicylic acid, GI: gastrointestinal, NSAID: non-steroidal anti-inflammatory drug, phen.: phenprocoumon

## 5.4 Discussion

The results of multiple imputation in the empirical study as well as the simulation study have shown that coefficients of phase 1 covariates can be estimated more precisely compared to two-phase methods. Regarding phase 2 variables and related interactions, efficiency is comparable for multiple imputation and two-phase methods. The problem of using multiple imputation for the analysis of two-phase studies is the potential for bias. Especially if the phase 2 data set is not a random sample and is relatively small, e.g. more than 90% of the information has to be imputed, parameter estimates may be biased. With the implemented imputation models, the only way to avoid biased estimates is the imputation in strata defined by variables used for sampling of the empirical phase 2 data set. If it is assumed that the stratified imputation in IM4 also leads to unbiased estimates in the empirical study, differences in the parameter estimates obtained with other imputation models as well as from the phase 1 and the two-phase analysis can be interpreted as bias. However, due to the large standard errors in the empirical study, the observed differences are not relevant.

Compared to the study conducted by Marti and Chavance, 2011, the simulation study described here reveals substantially higher bias for some imputation models. Marti and Chavance observed only about 2% bias in a simulated case-cohort study when using a misspecified imputation model. In the simulation using IM2 49% bias occurs for the parameter estimate of BMI. Furthermore, with a standard error of 0.1 the biased estimate ($\bar{\beta} = 0.23$) is significantly different from the true parameter value ($\beta = 0.45$). The simulated case-cohort study is only comparable with respect to the size of the phase 1 (N=25,000) and the phase 2 (n=1,000) data sets. Otherwise it is based on completely simulated data consisting of two Gaussian and one binary variable, whereas this simulation study is based on the covariate distribution observed in the empirical study including 15 variables. Another important difference might be that the subcohort in the simulated case-cohort study is a random sample of controls and not a stratified sample.

In conclusion, multiple imputation in two-phase studies with a rich phase 1 data set leads on the one hand to efficient parameter estimates of phase 1 as well as phase 2 variables by utilising the complete phase 1 data set. On the other hand

the risk of bias is large, particularly if the phase 2 data set is not a random sample. For the described data constellation, the study design has been accounted for in the imputation model by imputing the data separately in strata. This approach will only be feasible if each stratum includes a sufficient number of subjects. The analyses described in this chapter should only be viewed as a first impression of applying multiple imputation to two-phase studies, because the plurality of techniques available for multiple imputation has not been entirely investigated.

# Chapter 6

# Conclusion

Aim of this thesis was to explore two-phase designs for case-control studies based on administrative databases that make use of all available phase 1 data to estimate adjusted exposure effects efficiently. Phase 1 information is utilised in the two-phase analysis by stratification which leads to a reduction of standard errors compared to an analysis including only phase 2 data. This approach also allows for an unbiased estimation of regression coefficients if all variables influencing the composition of the phase 2 data set are included in the stratification. A specific feature of two-phase database studies is the availability of a rich phase 1 data set providing information on most variables of interest for the whole study population. Hence, information on many phase 1 variables can potentially be included in the analysis. This also establishes the possibility to identify variables influencing the selection into the phase 2 subsample by modelling the probability of selection in a non-response analysis (Behr et al., 2012; Appendix D; Behr and Schill, 2013). As an example of a two-phase database study, a study investigating the risk of bleeding associated with phenprocoumon exposure was conducted in which phase 1 data was extracted from claims data of one statutory health insurance and additional information on body mass index and smoking in phase 2 was ascertained in a health survey. This empirical study was exploited as a vehicle to (1) illustrate the dilemma occurring with stratification on multiple variables, (2) propose a new stratification strategy resolving the dilemma and (3) develop a design criterion for the selection of the most efficient stratifications. The dilemma was identified for cross-classification of phase 1 covariates which is the stratification strategy used in traditional two-phase field studies (cf.

Section 4.1): On the one hand, inclusion of each additional covariate in the stratification increases the number of strata and therefore the risk of empty cells. On the other hand, ignoring information on phase 1 covariates results in unnecessarily large standard errors of the respective parameter estimates. Defining the stratification by percentiles of a disease score (DSC), which is a summary measure of several phase 1 covariates, solves the problem (cf. Section 4.2). With this stratification strategy, the number of strata does not depend on the number of phase 1 variables but on the number of cut-points chosen to define the stratification. In the empirical study as well as in simulations based on this study, this stratification strategy led to a reduction of the standard errors of parameter estimates for most phase 1 covariates. A limitation of the stratification based on disease scores is that it does not fully account for the bias introduced by stratified sampling of the phase 2 sample. This limitation can be compensated by defining the stratification by cross-classification of variables used for sampling and subclasses of the disease score (Behr and Schill, 2013). However, if several variables have been used for sampling, the choice of subclasses of the disease score is restricted by the requirement of non-empty strata. The choice of the number and placement of cut-points is essential for the efficiency of the stratification. It should consider the overlap of the conditional score distributions in cases ($DSC|D = 1$) and controls ($DSC|D = 0$) to avoid empty cells. In this thesis, stratifications based on one to five cut-points showed a good performance with respect to the design criterion proposed in Section 4.3 and in the empirical study as well as the simulation study. The optimal choice of number and placement of the cut-points has not been investigated yet in a systematic way. The design criterion might be used in future research to determine rules for good stratifications based on disease scores.

The motivation behind the development of the design criterion was the lack of guidance for choosing an efficient stratification from the variety of possible stratifications. The criterion combines qualitative and quantitative aspects. The qualitative part involves the ranking of covariates by their importance based on prior knowledge and the formation of a set of candidate stratifications. The quantitative part was derived from the variance of the weighted likelihood estimator (WLE) which can be expressed as the sum of the phase 1 variance and a penalty term. The penalty term is approximated for each stratification and each variable on the basis of phase 1 data. The resulting penalty terms are then compared graphically. The design

criterion was tested in the empirical study and successfully validated for the specific data constellation in a simulation study. The qualitative aspect of the design criterion can be interpreted as both a strength and a weakness at the same time. Its strength lies in the inclusion of prior and expert knowledge and its weakness in the subjectivity. A clear limitation of the criterion is that it is unable to determine efficient stratifications for the parameter estimation of phase 2 variables. If known proxies of the phase 2 variables are available in the phase 1 data, inclusion of these variables in the stratification may compensate this limitation. Schill and Wild, 2006 were confronted with a similar problem, when they tried to find optimal sampling fractions for a given stratification based on phase 1 data to efficiently estimate a parameter vector. They suggested to define scenarios in which the stratumwise covariate distributions among controls are specified by the conditional probabilities $Pr(\mathbf{X}|\mathscr{S}, D = 0)$ and the parameter vector of the logistic disease model is assigned. In contrast to the situation considered by Schill and Wild, in the present situation separate scenarios would have to be specified for each stratification out of the set of candidate stratifications which would increase the complexity of the comparison.

The simulation study conducted to evaluate the performance of the new stratification strategy and the design criterion also provided guidance for the planning of future two-phase studies. At the planning stage of a two-phase study, when only phase 1 data is available, it has to be decided which variables will be used for sampling the phase 2 data, i.e. which variables compose the a priori stratification. When employing an a priori stratification for sampling and a different post stratification for the analysis, the post stratification always has to be finer than the a priori stratification. To improve efficiency it is not worthwhile to include covariates like sex in the a priori stratification which solely have frequently occurring categories. Large gains in efficiency can be achieved when covariates related to rare exposures such as phenprocoumon exposure are included in the a priori stratification. Furthermore, the specific role of age became obvious in the study. Although all age categories are frequent and therefore do not need to be considered in the a priori stratification to increase efficiency of the estimation of the age effect, inclusion of age in the a priori stratification is recommendable because this leads to smaller standard errors of parameter estimates for rare diseases and exposures which are related to age. Even more efficient for the estimation of these effects is the inclusion of disease scores in the a priori stratification.

For the empirical study, a stratification defined by cross-classification of phenprocoumon exposure, age, sex and three subclasses of a disease score was identified as the most efficient stratification to estimate the effects of most phase 1 covariates. The results of the two-phase analysis using this stratification were compared to the results of multiple imputation analyses (cf. Section 5). Multiple imputation, which is a survey methodological approach for the analysis of incomplete data sets, has been tested in this thesis for its applicability to two-phase data sets. In comparison to the two-phase analysis in the empirical study, multiple imputation was more efficient for the parameter estimation of most phase 1 covariates and at least equally efficient for the parameter estimation of the other variables. In the simulation study it became apparent that multiple imputation leads to biased parameter estimates if the imputation model does not sufficiently account for the sampling design of the phase 2 data set. To be more precise, the inclusion of stratum indicators in the imputation model was not sufficient to correct for the bias introduced by stratified sampling. Applying stratified imputation, i.e. separate imputation in the four strata used for sampling, revealed unbiased parameter estimates. In conclusion, the first application of multiple imputation in a two-phase database study was on the one hand promising with respect to small standard errors of parameter estimates for phase 1 covariates. On the other hand, using an incorrect imputation model led to biased estimates of coefficients for phase 2 variables as well as for variables associated with the phase 2 variables. In that regard, two-phase methods were more robust against bias. It has to be noted that the adequacy of multiple imputation for two-phase studies comprising a rich phase 1 data set has only been investigated for this specific study situation and might not be transferable to studies with for instance a smaller phase 2 data set or a more complex sampling strategy.

This aspect has also to be taken into account in the assessment of the results achieved in this thesis since they all rely on the empirical study and simulations based on it. The data constellation in the empirical study is characterised by the following facts:

1. The study investigates the association between a rare disease and a rare exposure.
2. The phase 2 data set comprises the most informative subjects with respect to disease and exposure.

3. Most variables which are relevant for the disease model are available in phase 1.
4. Lack of adjustment for phase 2 variables does not confound the estimated effect of any phase 1 covariate.
5. Except for age, all variables are included as binary covariates in the logistic model.

While the first and the fifth characteristic are typical for pharmacoepidemiological database studies, the second, third and fourth characteristic depend on the study design and the specific study situation. Both the performance of the new stratification strategy and the design criterion may be different in other data constellations. The performance of stratifications based on a disease score is likely to depend on the specific disease model and on the overlap of the conditional distributions of the score for cases and controls. Furthermore, it will be related to the number of variables influencing the composition of the phase 2 sample. The performance of the design criterion is most probably connected to the accuracy of the approximation of the penalty term. The approximation error depends for instance on the difference between the disease probability estimated by using phase 1 data only and the disease probability estimated by using the complete covariate information in phase 2. It has been assumed that these probabilities do not differ strongly because of the rich phase 1 data set. In studies for which an important risk factor for the disease is only available in phase 2 this assumption will most likely be violated. Therefore, both novel approaches should be evaluated in further data constellations.

It is envisaged to use the design criterion for the planning of a two-phase study investigating the risk of diabetic complications in patients treated with different antidiabetic drugs. In this study, phase 1 data will comprise claims data of one German statutory health insurance and phase 2 data will be derived from the documentation of the Disease Management Programme (DMP) for type 2 diabetes which provides information on body mass index, blood pressure as well as duration, intensity and control of diabetes. A challenge will be to model the selectivity of participation in the DMP, which may be influenced by several variables. Also the new stratification strategy will be tested in this study. This study will differ from the present empirical study in that the phase 2 data set will not include the most informative subjects for answering the study question because subjects are not selected into the phase 2

sample by the investigator but are self-selected. Furthermore, phase 2 data in this study may comprise important confounders.

Methods for the combined analysis of several data sources will certainly be relevant in context of the National Cohort which is a long-term German population study with the objective to investigate the causes, risk factors and prevention of widespread diseases (`http:\www.nationale-kohorte.de`). To achieve this objective it is planned to link secondary data, e.g. health insurance data, to the data collected in the study for all participants who agreed to the linkage of the data. Although this data constellation will differ from the two-phase database studies considered in this thesis, there are important parallels: (1) the data from the National Cohort can be viewed as phase 1 data comprising a multitude of variables; (2) the linked secondary data will only provide information for a subsample of participants because secondary data will not be available for each participant and not all participants will agree to the linkage. Therefore, the secondary data can be interpreted as the phase 2 data set comprising additional information, whereas the most important information is included in the phase 1 data. If the proportion of missing information is high, application of the two-phase methods considered in this thesis may be reasonable. Otherwise, it is a typical missing-data situation where multiple imputation is likely to be the better approach. Moreover, these two-phase methods are restricted to logistic regression whereas multiple imputation can be applied more general. Further research is needed to improve the methods for the analysis of such data and to better understand their strengths and limitations. Besides multiple imputation and the two-phase methods described in this thesis, weighted analyses using calibrated or estimated weights may be a promising approach. Recently, pseudo likelihood methods have been extended to include estimated weights (Scott and Wild, 2011). A first application of this approach showed good results in comparison to the ordinary two-phase analysis (Breslow et al., 2013). A comparison of these approaches for diverse data constellations is still missing.

# Appendix A

# Mathematical details for Chapter 3

## A.1 Proof of (3.6)

To prove (3.6) the following constrained maximisation problem has to be solved:

$$\text{Maximise } l(\boldsymbol{\theta}, \boldsymbol{\delta}) \quad \text{subject to } \sum_{m=1}^{M} \delta_m - 1 = 0,$$

where $l(\boldsymbol{\theta}, \boldsymbol{\delta})$ is defined according to (3.5). If $\hat{\boldsymbol{\delta}}$ solves the maximisation problem, then there exists a Lagrange multiplier $\lambda \in \mathbb{R}$ such that

$$\left[ \frac{\partial}{\partial \boldsymbol{\delta}} l(\boldsymbol{\theta}, \boldsymbol{\delta}) + \lambda \frac{\partial}{\partial \boldsymbol{\delta}} \left( \sum_{m=1}^{M} \delta_m - 1 \right) \right]_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}} = 0.$$

For fixed $m$ it follows that

$$\left[ \frac{\partial l(\boldsymbol{\theta}, \boldsymbol{\delta})}{\partial \delta_m} + \lambda \right]_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}} = 0 \tag{A.1}$$

$$\Leftrightarrow \quad \left[ n_{+m} \frac{1}{\delta_m} - \sum_{i=0}^{1} N_i \frac{f(D = i|\tilde{\mathbf{x}}_m; \boldsymbol{\theta})}{\sum_{l=1}^{M} f(D = i|\tilde{\mathbf{x}}_l; \boldsymbol{\theta}) \delta_l} + \lambda \right]_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}} = 0.$$

Multiplication with $\delta_m$ and summation over $m$ leads to:

$$\left[ \sum_{m=1}^{M} n_{+m} - \sum_{i=0}^{1} N_i \frac{\sum_{m=1}^{M} f(D = i|\tilde{\mathbf{x}}_m; \boldsymbol{\theta}) \delta_m}{\sum_{l=1}^{M} f(D = i|\tilde{\mathbf{x}}_l; \boldsymbol{\theta}) \delta_l} + \lambda \sum_{m=1}^{M} \delta_m \right]_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}} = 0$$

$$\Leftrightarrow \quad \left[ N - N + \lambda \sum_{m=1}^{M} \delta_m \right]_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}} = 0.$$

Since $\sum_{m=1}^{M} \delta_m = 1$ it follows that $\lambda = 0$ and from (A.1) and (3.4) it follows immediately that $\hat{\boldsymbol{\delta}}$ fulfils the system of equations

$$\delta_m = \tau_m(\boldsymbol{\theta}) = \frac{n_{+m}}{\sum_{i=0}^{1} \frac{N_i f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})}{\pi_i}}, \quad m = 1, \ldots, M.$$

□

## A.2 Proof of (3.7) and (3.8)

The profile likelihood as given in (3.7) can be derived from (3.5) and (3.6) as follows:

$$\begin{aligned}
& l_P(\boldsymbol{\theta}) = l(\boldsymbol{\theta}, \tau(\boldsymbol{\theta})) \\
&= \sum_{i=0}^{1}\sum_{m=1}^{M} n_{im} \log f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta}) + \sum_{m=1}^{M} n_{+m} \log\left(\frac{n_{+m}}{\sum_h \frac{N_h f(D=h|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})}{\pi_h}}\right) - N_1 \log \pi_1 - N_0 \log \pi_0 \\
&= \sum_{i=0}^{1}\sum_{m=1}^{M} n_{im} \log f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta}) - \sum_{m=1}^{M} n_{+m} \log \sum_h \frac{N_h f(D=h|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})}{\pi_h} + \sum_{m=1}^{M} n_{+m} \log(n_{+m}) \\
&\quad - N_1 \log \pi_1 - N_0 \log \pi_0 \\
&= \sum_{i=0}^{1}\sum_{m=1}^{M} n_{im} \log\left(\frac{f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})}{\sum_h f(D=h|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})\frac{N_h}{\pi_h}}\right) + \sum_{m=1}^{M} n_{+m} \log(n_{+m}) - N_1 \log \pi_1 - N_0 \log \pi_0 \\
&= \sum_{i=0}^{1}\sum_{m=1}^{M} n_{im} \log\left(\frac{f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})\frac{N_i}{\pi_i N}}{\sum_h f(D=h|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})\frac{N_h}{\pi_h N}}\right) - \sum_{i=0}^{1} N_i \log\left(\frac{N_i}{\pi_i}\right) \\
&+ \sum_{m=1}^{M} n_{+m} \log(n_{+m}) - N_1 \log \pi_1 - N_0 \log \pi_0 \\
&= \sum_{i=0}^{1}\sum_{m=1}^{M} n_{im} \log\left(\frac{f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})\mu_i}{\sum_h f(D=h|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})\mu_h}\right) + \sum_{m=1}^{M} n_{+m} \log(n_{+m}) - N_1 \log N_1 - N_0 \log N_0.
\end{aligned}$$

Since the last three terms are independent of $\boldsymbol{\theta}$, (3.7) holds.

□

Using (3.7) it follows immediately that

$$\begin{aligned}\log\left(\frac{f^*(D=1|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})}{f^*(D=0|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})}\right) &= \log(f^*(D=1|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})) - \log(f^*(D=0|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})) \\ &= \log(f(D=1|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})\frac{N_1}{\pi_1}) - \log(f(D=0|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})\frac{N_0}{\pi_0}) \\ &= \log\left(\frac{f(D=1|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})}{f(D=0|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})}\right) + \log\left(\frac{\frac{N_1}{\pi_1}}{\frac{N_0}{\pi_0}}\right)\end{aligned}$$

and therefore (3.8).

□

## A.3 Proof of $\frac{\partial l^*(\boldsymbol{\Phi})}{\partial \kappa} = 0$

Using (3.2) and (3.8) the pseudo log-likelihood $l^*$ is given in terms of $(\boldsymbol{\theta}, \kappa)$ by

$$\begin{aligned}l^*(\boldsymbol{\theta},\kappa) &= \sum_{k=1}^{N} \log f^*(d_k|\mathbf{x}_k;\boldsymbol{\theta},\kappa) \\ &= \sum_{k=1}^{N} \log\left(\left(\frac{\exp(\boldsymbol{\theta}^T\mathbf{x}_k + \log(\kappa))}{1+\exp(\boldsymbol{\theta}^T\mathbf{x}_k + \log(\kappa))}\right)^{d_k}\left(\frac{1}{1+\exp(\boldsymbol{\theta}^T\mathbf{x}_k + \log(\kappa))}\right)^{1-d_k}\right) \\ &= \sum_{k=1}^{N} d_k \log\left(\frac{\exp(\boldsymbol{\theta}^T\mathbf{x}_k + \log(\kappa))}{1+\exp(\boldsymbol{\theta}^T\mathbf{x}_k + \log(\kappa))}\right) + \sum_{k=1}^{N}(1-d_k)\log\left(\frac{1}{1+\exp(\boldsymbol{\theta}^T\mathbf{x}_k + \log(\kappa))}\right) \\ &= \sum_{k=1}^{N} d_k(\boldsymbol{\theta}^T\mathbf{x}_k + \log(\kappa)) - \sum_{k=1}^{N} d_k \log(1+\exp(\boldsymbol{\theta}^T\mathbf{x}_k + \log(\kappa)) \\ &\quad - \sum_{k=1}^{N}(1-d_k)\log(1+\exp(\boldsymbol{\theta}^T\mathbf{x}_k + \log(\kappa))) \\ &= \sum_{k=1}^{N} d_k(\boldsymbol{\theta}^T\mathbf{x}_k + \log(\kappa)) - \sum_{k=1}^{N}\log(1+\exp(\boldsymbol{\theta}^T\mathbf{x}_k + \log(\kappa))).\end{aligned}$$

Then it follows on the one hand for $\kappa$ that

$$\begin{aligned}\frac{\partial l^*(\boldsymbol{\theta},\kappa)}{\partial\kappa} &= \sum_{k=1}^{N}\frac{d_k}{\kappa} - \sum_{k=1}^{N}\frac{\exp(\boldsymbol{\theta}^T\mathbf{x}_k + \log(\kappa))\frac{1}{\kappa}}{1+\exp(\boldsymbol{\theta}^T\mathbf{x}_k + \log(\kappa))} \\ &= \frac{1}{\kappa}\left(N_1 - \sum_{m=1}^{M} n_{+m} f^*(D=1|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})\right).\end{aligned}$$

On the other hand it can be deduced from (3.6) and (3.7) that

$$\begin{aligned}
\mu_1 = \frac{N_1}{\pi_1} &= \frac{N_1}{\sum_{m=1}^{M} f(D=1|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})\delta_m} = \frac{N_1}{\sum_{m=1}^{M} f(D=1|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})\frac{n_{+m}}{\sum_i \mu_i f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})}} \\
&\Leftrightarrow \sum_{m=1}^{M} \frac{\mu_1 f(D=1|\tilde{\mathbf{x}}_m;\boldsymbol{\theta}) n_{+m}}{\sum_i \mu_i f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})} = N_1 \\
&\Leftrightarrow \sum_{m=1}^{M} f^*(D=1|\tilde{\mathbf{x}}_m;\boldsymbol{\theta}) n_{+m} = N_1.
\end{aligned}$$

Combining both arguments completes the proof.

□

## A.4 Proof of $\lambda = \mathscr{N}$ and of (3.12)

If $\hat{\boldsymbol{\delta}}$ solves the maximisation problem (3.11) then there exists a Lagrange multiplier $\lambda$ such that

$$\left[\frac{\partial l(\boldsymbol{\theta},\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} + \lambda\right]_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}} = 0.$$

It follows that $\lambda = -\mathscr{N}$ because

$$\begin{aligned}
&\left[\frac{\partial l(\boldsymbol{\theta},\boldsymbol{\delta})}{\partial \delta_m} + \lambda\right]_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}} = 0 \\
\Leftrightarrow &\left[\sum_{i=0}^{1}(\mathscr{N}_i - N_i)\frac{f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})}{\sum_l f(D=i|\tilde{\mathbf{x}}_l;\boldsymbol{\theta})\delta_l} + \sum_{i=0}^{1}\sum_{j=1}^{J}(N_{ij}-n_{ij})\frac{Q_{ij}^*(\tilde{\mathbf{x}}_m)}{\sum_l Q_{ij}^*(\tilde{\mathbf{x}}_l)\delta_l}\right. \\
&\left. + \frac{n_{++m}}{\delta_m} + \lambda\right]_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}} = 0 \\
\Leftrightarrow &\left[\sum_{i=0}^{1}(\mathscr{N}_i - N_i)\frac{f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})\delta_m}{\sum_l f(D=i|\tilde{\mathbf{x}}_l;\boldsymbol{\theta})\delta_l} + \sum_{i=0}^{1}\sum_{j=1}^{J}(N_{ij}-n_{ij})\frac{Q_{ij}^*(\tilde{\mathbf{x}}_m)\delta_m}{\sum_l Q_{ij}^*(\tilde{\mathbf{x}}_l)\delta_l}\right. \\
&\left. + n_{++m} + \lambda\delta_m\right]_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}} = 0 \qquad \text{(A.2)}
\end{aligned}$$

and with summation over $m$:

$$\left[ \sum_{i=0}^{1} (\mathscr{N}_i - N_i) \frac{\sum_m f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})\delta_m}{\sum_l f(D=i|\tilde{\mathbf{x}}_l;\boldsymbol{\theta})\delta_l} + \sum_{i=0}^{1}\sum_{j=1}^{J} (N_{ij} - n_{ij}) \frac{\sum_m Q^*_{ij}(\tilde{\mathbf{x}}_m)\delta_m}{\sum_l Q^*_{ij}(\tilde{\mathbf{x}}_l)\delta_l} \right.$$
$$\left. + \sum_{m=1}^{M} n_{++m} + \lambda \sum_{m=1}^{M} \delta_m \right]_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}} = 0$$
$$\Leftrightarrow \sum_{i=0}^{1} (\mathscr{N}_i - N_i) + \sum_{i=0}^{1}\sum_{j=1}^{J} (N_{ij} - n_{ij}) + n + \lambda = 0$$
$$\Leftrightarrow -\mathscr{N} = \lambda.$$

(A.2) yields that $\boldsymbol{\delta}$ fulfils the system of equations

$$\delta_m = \tau_m(\boldsymbol{\theta}) = \frac{n_{++m}}{\mathscr{N} - \sum_i \frac{\mathscr{N}_i - N_i}{\pi_i} f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta}) - \sum_i \sum_j (N_{ij} - n_{ij}) \frac{Q^*_{ij}(\tilde{\mathbf{x}}_m)}{Q_{ij}}}.$$

□

## A.5 Parametrisation of $Q_{ij}$

With

$$Q_{ij} = \frac{\exp(\rho_{ij})}{\sum_i \sum_j \exp(\rho_{ij})}$$

it follows immediately that $0 < Q_{ij} < 1$ and $\sum_i \sum_j Q_{ij} = 1$. This is also true if $\rho_{1J}$ is set to 0. Then it follows from

$$\log(Q_{ij}) = \log\left(\frac{\exp(\rho_{ij})}{\sum_i \sum_j \exp(\rho_{ij})}\right) = \rho_{ij} - \log\left(\sum_i \sum_j \exp(\rho_{ij})\right)$$

and

$$\log(Q_{1J}) = \log\left(\frac{\exp(\rho_{1J})}{\sum_i \sum_j \exp(\rho_{ij})}\right) = \rho_{1J} - \log\left(\sum_i \sum_j \exp(\rho_{ij})\right)$$

that

$$\log(Q_{ij}) - \log(Q_{1J}) = \rho_{ij} - \rho_{1J} = \rho_{ij}$$

and thus

$$\log\left(\frac{Q_{ij}}{Q_{1J}}\right) = \rho_{ij}, \quad \text{for } ij \neq 1J.$$

Hence, both parameterisations are equivalent.

□

## A.6 Proof of (3.18)

Omitting the first term from (3.11) leads to the profile log-likelihood for the pseudo model:

$$\begin{aligned}\tilde{l}_P(\boldsymbol{\theta}) = &\sum_{i=0}^{1}\sum_{j=1}^{J}(N_{ij}-n_{ij})\log\left(\sum_{m=1}^{M}Q_{ij}^*(\tilde{\mathbf{x}}_m)\delta_m\right)\\ &+\sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{m=1}^{M}n_{ijm}\log f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})+\sum_{m=1}^{M}n_{++m}\log\delta_m.\end{aligned}$$

With (3.14) and substituting $Q_{ij}^*$ by $f(D=i|\tilde{\mathbf{x}};\boldsymbol{\theta})$ it follows

$$\begin{aligned}\tilde{l}_P(\boldsymbol{\theta}) =&\sum_{i=0}^{1}\sum_{j=1}^{J}(N_{ij}-n_{ij})\log\left(\sum_{m=1}^{M}Q_{ij}^*(\tilde{\mathbf{x}}_m)\frac{n_{++m}}{N\sum_i\sum_j\mu_{ij}Q_{ij}^*(\tilde{\mathbf{x}}_m)}\right)\\ &+\sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{m=1}^{M}n_{ijm}\log f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})+\sum_{m=1}^{M}n_{++m}\log\left(\frac{n_{++m}}{N\sum_i\sum_j\mu_{ij}Q_{ij}^*(\tilde{\mathbf{x}}_m)}\right)\\ =&\sum_{i=0}^{1}\sum_{j=1}^{J}(N_{ij}-n_{ij})\log\left(\sum_{l:\tilde{\mathbf{x}}_l\in S_{ij}}f(D=i|\tilde{\mathbf{x}}_l;\boldsymbol{\theta})\frac{n_{++l}}{N\sum_i\sum_j\mu_{ij}f(D=i|\tilde{\mathbf{x}}_l;\boldsymbol{\theta})}\right)\\ &+\sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{m=1}^{M}n_{ijm}\log f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})\\ &+\sum_{m=1}^{M}n_{++m}\log\left(\frac{n_{++m}}{N\sum_i\sum_j\mu_{ij}f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})}\right).\end{aligned}$$

The constant $N$ can be omitted from the denominator of the first and third term because it can be written as a single summand in the profile log-likelihood. Inserting

(3.15) in the following equation yields

$$\begin{aligned}
\tilde{l}_P(\boldsymbol{\theta}) =& \sum_{i=0}^{1}\sum_{j=1}^{J}(N_{ij}-n_{ij})\log\left(\sum_{l:\tilde{\mathbf{x}}_l\in S_{ij}}\frac{n_{++l}f^*(D=i|\tilde{\mathbf{x}}_l;\boldsymbol{\theta})}{\mu_{ij}}\right)\\
&+\sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{m=1}^{M}n_{ijm}\log f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})\\
&+\sum_{m=1}^{M}n_{++m}\log(n_{++m})-\sum_{m=1}^{M}\sum_{i=0}^{1}\sum_{j=1}^{J}n_{ijm}\log\left(\sum_i\sum_j\mu_{ij}f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})\right)\\
=&\sum_{i=0}^{1}\sum_{j=1}^{J}(N_{ij}-n_{ij})\log\left(\sum_{l:\tilde{\mathbf{x}}_l\in S_{ij}}n_{++l}f^*(D=i|\tilde{\mathbf{x}}_l;\boldsymbol{\theta})\right)-\sum_{i=0}^{1}\sum_{j=1}^{J}(N_{ij}-n_{ij})\log(\mu_{ij})\\
&+\sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{m=1}^{M}n_{ijm}\log\left(\frac{f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})}{\sum_i\sum_j\mu_{ij}f(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta}}\right)\\
&+\sum_{m=1}^{M}n_{++m}\log(n_{++m}).
\end{aligned}$$

This equation is simplified as follows by inserting $\gamma_{ij}$ according to (3.17) and by ignoring the last term:

$$\begin{aligned}
\tilde{l}_P(\boldsymbol{\theta}) =& \sum_{i=0}^{1}\sum_{j=1}^{J}(N_{ij}-n_{ij})\log(n_{ij}-\gamma_{ij})-\sum_{i=0}^{1}\sum_{j=1}^{J}(N_{ij}-n_{ij})\log\left(\frac{n_{ij}-\gamma_{ij}}{N_{ij}-\gamma_{ij}}\right)\\
&+\sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{m=1}^{M}n_{ijm}\log\left(f^*(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})\right)-\sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{m=1}^{M}n_{ijm}\log(\mu_{ij})\\
=&\sum_{i=0}^{1}\sum_{j=1}^{J}(N_{ij}-n_{ij})\log(N_{ij}-\gamma_{ij})+\sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{m=1}^{M}n_{ijm}\log\left(f^*(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})\right)\\
&-\sum_{i=0}^{1}\sum_{j=1}^{J}n_{ij}\log(n_{ij}-\gamma_{ij})+\sum_{i=0}^{1}\sum_{j=1}^{J}n_{ij}\log(N_{ij}-\gamma_{ij})\\
=&\sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{m=1}^{M}n_{ijm}\log\left(f^*(D=i|\tilde{\mathbf{x}}_m;\boldsymbol{\theta})\right)-\sum_{i=0}^{1}\sum_{j=1}^{J}n_{ij}\log(n_{ij}-\gamma_{ij})\\
&+\sum_{i=0}^{1}\sum_{j=1}^{J}N_{ij}\log(N_{ij}-\gamma_{ij}).
\end{aligned}$$

□

# Appendix B

# Paper 1: Phenprocoumon and risk of intracerebral haemorrhage

**Contribution to the manuscript** I herewith certify that I contributed to the design of the study, performed all statistical analyses, interpreted the results, and drafted the manuscript.

PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2010; **19**: 722–730
Published online 13 May 2010 in Wiley InterScience (www.interscience.wiley.com) **DOI**: 10.1002/pds.1973

ORIGINAL REPORT

# Risk of intracerebral hemorrhage associated with phenprocoumon exposure: a nested case–control study in a large population-based German database†

Sigrid Behr[1]*, Frank Andersohn[2] and Edeltraut Garbe[1]

[1]*Bremen Institute for Prevention Research and Social Medicine (BIPS), Bremen, Germany*
[2]*Institute for Social Medicine, Epidemiology and Health Economics, Charité—Universitätsmedizin, Berlin, Germany*

## SUMMARY

**Purpose** Intracerebral hemorrhage (ICH) is the most serious complication of oral anticoagulation. This study investigated the risk of ICH for phenprocoumon which is the most widely used oral anticoagulant in Germany.

**Methods** We conducted a nested case–control study in a cohort of 13.4 million insurants of 4 German statutory health insurances (SHIs) who were continuously enrolled for 6 months prior to cohort entry. Cases were patients hospitalized for ICH. Ten controls were matched to each case by SHI, birth year, and sex using incidence density sampling. Rate ratios (RR) of ICH for current phenprocoumon use as compared to non-use were estimated from odds ratios calculated by conditional logistic regression analyses considering multiple risk factors.

**Results** Analysis of the full cohort revealed a strong increase in incidence of ICH with increasing age. In the nested case–control study including 8138 cases of ICH and 81 373 matched controls, we observed an increased risk of ICH for current phenprocoumon exposure that varied with age. The phenprocoumon-associated risk of ICH was lower in older age groups with RRs from 4.20 (95% confidence interval (CI) 2.44–7.21) for phenprocoumon users less than 55 years of age to 2.43 (95%CI, 1.81–3.27) for those older than 85 years. Our study confirmed known risk factors of ICH.

**Discussion** Phenprocoumon exposure was associated with an increased risk of ICH. The interaction of risk for phenprocoumon with age was unexpected and needs further study. Copyright © 2010 John Wiley & Sons, Ltd.

KEY WORDS — intracerebral bleeding; oral anticoagulant; pharmacoepidemiological research database; nested case–control study

*Received 22 January 2010; Revised 17 March 2010; Accepted 29 March 2010*

## INTRODUCTION

Oral anticoagulation plays an important role in the primary and secondary prevention and treatment of thromboembolic diseases. The most widely used oral anticoagulants are coumarines such as warfarin or phenprocoumon. Coumarines have a narrow therapeutic index and are quite frequently a cause of bleeding. Intracerebral hemorrhage (ICH) is a very severe complication of coumarine therapy which may lead to irreparable impairment or death. The risk of ICH associated with coumarine therapy has been studied for warfarin[1–3] and also for the combined group of warfarin and phenprocoumon.[4,5] However, there are no published results for phenprocoumon, which is the most widely used substance for oral anticoagulation in Germany. Phenprocoumon accounts for more than 99% of coumarine use in Germany.[6]

Each coumarine has a single chiral center that gives rise to two different enantiomeric forms of which the S-form is more potent than the R-form.[7] There are substantial differences in pharmacokinetic properties between phenprocoumon and warfarin or acenocoumarol. Overall, cytochrom P450 (CYP) 2C9 appears to be most important for the clearance of warfarin and acenocoumarol, whereas it is less important for phenprocoumon due to the involvement of CYP3A4 in its metabolism and significant excretion of unchanged drug in bile and urine.[7] The elimination half lives also differ substantially: acenocoumarol has the shortest half life ranging between 1.8 and 6.6 hours

* Correspondence to: S. Behr, Bremen Institute for Prevention Research and Social Medicine (BIPS), Linzer Str. 10, 28359 Bremen, Germany.
E-mail: behr@bips.uni-bremen.de
† Authors declare no conflict of interest.

for both enantiomeric forms, followed by warfarin (24–58 hours) and phenprocoumon (110–130 hours).[7] The longer half life of phenprocoumon could be associated with a higher risk of bleeding, since it leads to impaired control of drug treatment. Lesser metabolism of phenprocoumon by CYP2C9 may on the other hand result in a lower impact of genetic polymorphisms of CYP2C9 which could reduce the risk of bleeding in case of genetic polymorphisms of CYP2C9.[7]

To study the risk of ICH associated with phenprocoumon, we conducted a nested case–control study in a large insurance population in Germany.

## METHODS

### *Data source*

We used information from the German Pharmacoepidemiological Research Database (GePaRD) consisting of claims data from four German statutory health insurances (SHI). This database includes more than 14 million insurants covering all regions in Germany. The study was conducted with data from the years 2004 to 2006. The database contains demographic information for each insurant as well as information on hospital admissions, ambulatory physician visits, and ambulatory prescriptions. The hospital data comprises information about the admission and discharge dates, the reasons for admission and discharge, and diagnostic and therapeutic procedures with their respective dates. Claims of ambulatory physician visits include ambulatory treatments, procedures, and diagnoses. Since ambulatory physician visits are reimbursed on a quarterly basis, ambulatory diagnoses can only be allocated to a quarter of the year and not to an exact day. All diagnoses, ambulatory as well as inpatient diagnoses, are coded according to the German modification of the International Classification of Diseases (ICD-10 GM). Prescription data are available for all ambulatory prescriptions which are reimbursed by the SHIs. It includes the date of prescription, the date when the prescription was redeemed at the pharmacy, the amount of substance prescribed, and information on the prescribing physician. Prescription data are linked *via* the pharmaceutical reference number to a pharmaceutical reference database which contains information on the anatomical-therapeutic-chemical (ATC) code, the Defined Daily Dose (DDD), packaging size, strength, formulation, generic and trade name. Preliminary analyses regarding age and sex distribution, the number of hospital admissions and drug use have shown the database to be representative for Germany.[8,9]

In Germany, the utilization of health insurance data for scientific research is regulated by the Code of Social Law (SGB X). This study was conducted with permission from the Federal Ministry of Health. Informed consent was not required by law, since the study was based on pseudonymous data.

### *Study design*

We conducted a case–control study nested in a cohort of insurants who were required to be continuously insured for 6 months before cohort entry. Cohort entry was defined as the first day after 6 months of continuous insurance. Cohort exit was the first of the following dates: end of the insurance period, hospitalization for ICH, death or end of study period (30 November 2006). The study period ended in November 2006 to avoid incomplete data for hospitalizations spanning the turn of the year. Cases were defined as insurants who were hospitalized for ICH (main discharge diagnosis with ICD-10 GM code I61 which codes for intracerebral bleeding). The admission day is referred to as the index day.

Ten controls were matched to each case with respect to sex, year of birth, and SHI using incidence density sampling. We assigned an index day to each control that resulted in the same time of follow-up as for the corresponding case. Cohort members who were hospitalized at the index day of the case were excluded from the set of potential controls because they were not at risk of being admitted to hospital due to ICH at that point in time. Cases were eligible to be selected as controls until their hospitalization for ICH.

### *Exposure assessment*

We considered current phenprocoumon exposure (ATC code B01AA04) at the index day which was defined as a prescription overlapping the index day. Because we had no information on the prescribed daily dose, we estimated the average daily dose for each patient by dividing the cumulative dose until the last phenprocoumon prescription before the index day by the number of days corresponding to this time interval. The average daily dose was then used to estimate the duration of exposure for the last phenprocoumon prescription preceding the index day. The cumulative dose was obtained from the number of prescribed tablets and their strength which is 3 mg for most phenprocoumon drugs on the German market. If there was only one phenprocoumon prescription before the index day, the DDD was used instead.

Among patients with current exposure, we distinguished between recent initiators of phenprocoumon



DOI: 10.1002/pds

therapy and those with no recent initiation of therapy. We defined patients as recent initiators of phenprocoumon therapy if they had their first prescription of phenprocoumon recorded within the 90-day period before the index day.

### *Risk factor and confounder assessment*

Potential confounders were assessed in the 6-month period before cohort entry. We considered the following potential confounders in our analyses: diabetes mellitus, systemic hypertension, ischemic heart disease, ischemic cerebral infarction, cerebral amyloid angiopathy, cerebral aneurysm, brain tumor, epilepsy, liver diseases, renal failure, alcohol dependence, epistaxis, previous hospitalization for ICH or for other bleeding events. Concomitant medications were assessed in the 90-day period preceding the index day. We considered current use of the following substances: platelet aggregation inhibitors, heparins, non-steroidal anti-inflammatory drugs (NSAIDs) including acetylsalicylic acid (ASA), selective serotonin reuptake inhibitors (SSRIs), diuretics, corticosteroids, and statins. The exact definition in terms of ICD-10 GM codes and ATC codes is provided in the Appendix.

### *Statistical analysis*

We calculated incidence rates of ICH stratified by sex for different age groups making use of the full cohort data. Confidence intervals for the incidence rates were estimated by the substitution method assuming a Poisson distribution for the number of bleedings.[10]

Based on the case–control data, crude odds ratios (OR) were calculated by using the Mantel–Haenszel estimator to account for the matching. We conducted conditional logistic regression analyses to estimate adjusted ORs and two-sided 95% confidence intervals (CI) for ICH and current use of phenprocoumon. Since control selection was done by incidence density sampling, ORs correspond to incidence rate ratios (RR). The primary model for the multivariate analysis was based on prior knowledge on risk factors and included all variables. We kept two-way interaction terms with phenprocoumon exposure in this model, if the RR for phenprocoumon exposure changed by more than 10% after adding the respective interaction term. Following this approach, only phenprocoumon interaction with age (centered at 68 years) remained in the model. In addition, we used a backward selection procedure to select relevant covariates. While phenprocoumon exposure was forced to stay in the model, covariates were removed from the model step by step in case the Wald test was not significant ($p > 0.05$). We also performed a multivariate analysis in several age strata: <55 years, 55–<65 years, 65–<75 years, 75–<85 years, >85 years.

In a further analysis, we estimated adjusted RRs for ICH associated with current use of phenprocoumon distinguishing between recent and non-recent initiators.

We conducted several sensitivity analyses regarding our definition of current use at the index day. In these sensitivity analyses, a person was defined as exposed at the index day, if phenprocoumon was prescribed within 90 days, 180 days, or 270 days preceding the index day. For all these analyses, the reference category was the absence of current exposure to phenprocoumon.

All statistical analyses were done using SAS 8.2 (SAS Institute Inc., Cary, NC).

## RESULTS

In total, 13.4 million insurants were included in the cohort with a mean follow-up time of 798 days (standard deviation (STD) 208 days). The average age in the cohort was 39.9 (STD: 22.2) years and 45% of cohort members were male. The incidence of ICH was higher in males than in females and rose with increasing age from 6.82 bleedings per 100 000 person years in males younger than 55 to 236.29 bleedings per 100 000 person years in males older than 85 and from 4.67 bleedings per 100 000 person years in females younger than 55 to 189.69 bleedings per 100 000 person years in females older than 85 years (Figure 1).

Within this cohort we identified 8138 cases of ICH to whom we matched 81 373 controls. Only three controls could be matched to one case who was 102 years old at cohort entry. In the case–control sample, the mean age was 68.1 (STD: 14.3) years and proportions of males (51%) and females (49%) were similar. Cases and controls differed with respect to most risk factors (Table 1). The greatest differences were observed for history of ICH, cerebral amyloid angiopathy, cerebral aneurysm, epilepsy, ischemic cerebral infarction, alcohol dependence, and heparin use. Table 2 displays crude and adjusted RRs for phenprocoumon and other risk factors for development of ICH. Among the other risk factors, high risks were observed for cerebral aneurysm, epilepsy, alcohol dependence, cerebral amyloid angiopathy and history of ICH, however, with broad confidence intervals for the latter two factors.

Current phenprocoumon use was associated with a 3.4 fold risk of ICH referring to a 68-year old person (Table 2). A significant interaction between phenprocoumon exposure and age was observed resulting in lower phenprocoumon risk for ICH in persons older than 68 years and in higher risk for those younger than



DOI: 10.1002/pds

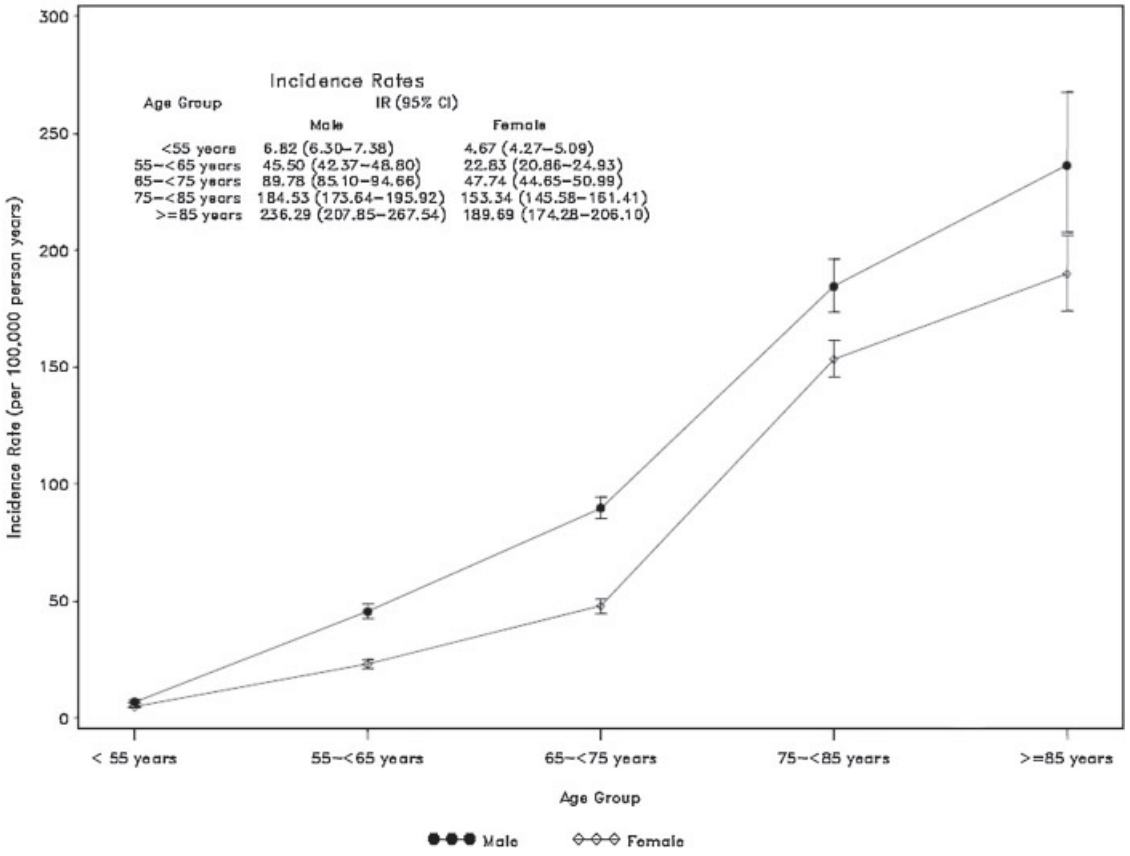



Figure 1. Incidence rates of intracerebral hemorrhage by age group stratified by sex

68 years. The stratified analysis revealed adjusted RRs of 4.20 (95%CI 2.44–7.21), 3.16 (95%CI 2.41–4.15), 3.51 (95%CI 3.00–4.10), 2.52 (95%CI 2.20–2.88), and 2.43 (95%CI 1.81–3.27) for phenprocoumon users in the age groups <55 years, 55–<65 years, 65–<75 years, 75–<85 years, and >85 years, respectively.

The backward selection procedure eliminated brain tumor and ischemic heart disease since both variables were not significant at the 5% level. However, the adjusted RRs in the reduced model did not change for phenprocoumon exposure or the other risk factors.

The distinction between recent and non-recent initiators of therapy among current phenprocoumon users revealed a higher RR for those patients who started their therapy recently (Table 3).

Table 4 shows the results of the sensitivity analyses concerning assessment of current phenprocoumon exposure. Similar risks of ICH for phenprocoumon were found for the different exposure measures. The adjusted RRs varied from 3.26 for phenprocoumon exposure identified by prescriptions within 90 days before the index day to 3.68 for prescriptions issued within 270 days before the index day. The corresponding 95%CIs overlapped widely.

## DISCUSSION

We found a more than three-fold risk of ICH in phenprocoumon users as compared to non-users, which was in the same range as the ORs reported for a combined group of warfarin or phenprocoumon by Johnsen *et al*. and Grønbaek *et al*.[4,5] Johnsen *et al*. conducted a case–control study in North Jutland (Denmark) in a population of *ca*. 490 000 inhabitants and observed an adjusted OR for ICH of 2.15 (95%CI 1.38–3.35) for users of oral anticoagulants compared to non-users. Grønbaek *et al*. reported an adjusted OR for ICH of 2.9 (95%CI 2.5–3.5) for patients treated with oral anticoagulants from a case–control study including about 1.4 million inhabitants of Denmark. Most other epidemiological studies investigating the risk of ICH for warfarin were not designed to compare use of warfarin with non-use of warfarin. These studies usually only included warfarin users and investigated risk factors for ICH in this subgroup of patients.[1,2,11,12]

The role of age as an independent risk factor for ICH is well known. Several previous studies have shown that the risk of ICH increases with age.[1,4,13] Sturgeon *et al.* pooled two cohort studies (ARIC and CHS) and



DOI: 10.1002/pds

Table 1. Characteristics of cases and matched controls

| | Cases $N = 8138$ $n$ (%) | | Controls $N = 81\,373$ $n$ (%) |
|---|---|---|---|
| Age* | | Mean (Std) 68.08 (14.33) | |
| Male sex* | 4178 (51.34%) | | 41 773 (51.34%) |
| Comorbid conditions$ | | | |
| Diabetes mellitus | 1717 (21.10%) | | 13 949 (17.14%) |
| Hypertension | 4340 (53.33%) | | 37 747 (46.39%) |
| Ischemic heart disease | 1806 (22.19%) | | 16 633 (20.44%) |
| Ischemic cerebral infarction | 811 (9.97%) | | 3049 (3.75%) |
| Cerebral amyloid angiopathy | 55 (0.68%) | | 2 (<0.01%) |
| Cerebral aneurysm | 20 (0.25%) | | 37 (0.05%) |
| Brain tumor | 115 (1.41%) | | 859 (1.06%) |
| Epilepsy | 241 (2.96%) | | 726 (0.89%) |
| Liver diseases | 938 (11.53%) | | 7755 (9.53%) |
| Renal failure | 460 (5.65%) | | 3517 (4.32%) |
| Alcohol dependence | 234 (2.88%) | | 804 (0.99%) |
| Epistaxis | 103 (1.27%) | | 554 (0.68%) |
| Previous ICH | 127 (1.56%) | | 20 (0.02%) |
| Other hemorrhage | 207 (2.54%) | | 419 (0.51%) |
| Concomitant medication§ | | | |
| Current use of ... | | | |
| Platelet aggregation inhibitors | 701 (8.61%) | | 5049 (6.20%) |
| Heparin | 223 (2.74%) | | 871 (1.07%) |
| NSAIDs | 1192 (14.65%) | | 10 268 (12.62%) |
| ASA | 80 (0.98%) | | 471 (0.58%) |
| SSRIs | 198 (2.43%) | | 1084 (1.33%) |
| Diuretics | 1832 (22.51%) | | 18 493 (22.73%) |
| Corticosteroids | 310 (3.81%) | | 2611 (3.21%) |
| Statins | 774 (9.51%) | | 7561 (9.29%) |

NSAID: Non-steroidal anti-inflammatory drug; ASA: Acetylsalicylic acid; SSRI: Selective serotonin reuptake inhibitor; ICH: Intracerebral hemorrhage.
*Birth year and sex are matching variables.
$Assessed in the 6 months baseline period preceding cohort entry.
§Ambulatory prescriptions assessed in the 90 days prior to the index day.

observed a doubling of risk for ICH with every 10 year increase in age.[14] Ariesen *et al.* [15] conducted a systematic review of cohort and case–control studies investigating risk factors for ICH. For each risk factor, they summarized the crude relative risk or odds ratio from those studies in which the respective information was available. For each 10-year increase in age, they reported a crude relative risk of 1.97 based on 5 cohort studies. Since our case–control study was matched by age, we could not estimate the effect of age as independent risk factor. We, therefore, estimated the incidence of ICH in the underlying cohort and observed a significant increase in ICH incidence with rising age. These results were comparable to the age-related incidence rates published by Bamford *et al*., which were based on a population of approximately 105 000 persons underlying the Oxfordshire Community Stroke Project.[13]

Interestingly, in our study we also observed a significant interaction of phenprocoumon with age that resulted in a lower phenprocoumon-associated risk of ICH at older ages. This issue has not been addressed in other studies investigating the risk of ICH associated with oral anticoagulants. However, a similar age effect can be seen for intracranial hemorrhage in a study by Fang *et al.* which investigated warfarin-associated hemorrhage in a cohort of more than 13 000 patients with atrial fibrillation (ATRIA study).[16] Although the phenprocoumon-associated risk of ICH is lower for older patients we have shown in our full cohort analysis that the incidence of ICH is much higher in older age groups as it is expected based on the results of other studies.[17]

Regarding the time of phenprocoumon initiation we observed a higher risk of ICH associated with recent initiation of phenprocoumon therapy. To our knowledge the effect of the time of phenprocoumon initiation has only been studied for major bleedings but not specifically for ICH. However, several studies on major bleedings also showed an increased risk for recent initiation of therapy.[17–20] Because we did not exclude patients using phenprocoumon before cohort entry, the risk we observed for no recent initiation of therapy is likely to be underestimated as a consequence of depletion of susceptible bias.[21]

Our study confirmed several other well-known risk factors for ICH. For alcohol dependence, an increased



DOI: 10.1002/pds

Table 2. Crude and adjusted incidence rate ratios and 95% confidence intervals for intracerebral bleeding

| Multivariate model | Crude incidence rate ratio | Adjusted incidence rate ratio* | 95% confidence interval* |
|---|---|---|---|
| Phenprocoumon exposure† | 3.00 | 3.42‡ | 3.08–3.79 |
| Comorbid conditions | | | |
| Diabetes mellitus | 1.31 | 1.16 | 1.09–1.23 |
| Hypertension | 1.36 | 1.28 | 1.21–1.35 |
| Ischemic heart disease | 1.12 | 0.95 | 0.89–1.01 |
| Ischemic cerebral infarction | 2.88 | 2.18 | 1.99–2.38 |
| Cerebral amyloid angiopathy | 275 | 279 | 67.4–1156 |
| Cerebral aneurysm | 5.41 | 3.65 | 2.03–6.55 |
| Brain tumor | 1.35 | 1.11 | 0.90–1.37 |
| Epilepsy | 3.38 | 2.43 | 2.07–2.85 |
| Liver diseases | 1.24 | 1.09 | 1.01–1.17 |
| Renal failure | 1.33 | 1.12 | 1.00–1.24 |
| Alcohol dependence | 3.00 | 2.57 | 2.20–3.01 |
| Epistaxis | 1.87 | 1.44 | 1.15–1.80 |
| Previous ICH | 66.8 | 50.8 | 30.7–84.0 |
| Other hemorrhage | 2.04 | 1.57 | 1.22–2.02 |
| Current use of ... | | | |
| Platelet aggregation inhibitors | 1.43 | 1.33 | 1.21–1.45 |
| Heparin | 2.60 | 1.77 | 1.51–2.08 |
| NSAIDs | 1.19 | 1.16 | 1.08–1.24 |
| ASA§ | 1.71 | 1.53 | 1.20–1.97 |
| SSRIs | 1.85 | 1.53 | 1.30–1.80 |
| Diuretics | 0.99 | 0.77 | 0.72–0.82 |
| Corticosteroids | 1.20 | 1.16 | 1.02–1.31 |
| Statins | 1.03 | 0.82 | 0.75–0.89 |

NSAID: Non-steroidal anti-inflammatory drug; ASA: Acetylsalicylic acid; SSRI: Selective serotonin reuptake inhibitor; ICH: Intracerebral hemorrhage.
*Adjusted for age, interaction between phenprocoumon exposure and age, sex and all other covariates included in the table.
†Phenprocoumon exposure is estimated by using the average daily dose and the number of prescribed tablets.
‡Adjusted incidence rate ratio for phenprocoumon exposure refers to a 68 years old patient. The two-way interaction incidence rate ratio is 0.97 (95% confidence interval 0.97–0.98).
§The effect of ASA is adjusted for all other covariates included in the table except for NSAIDs.

risk of similar magnitude was reported by Ariesen *et al.* [15] who calculated a crude OR of 3.36 for high alcohol intake combining the results of eight case–control studies. The increased risks we observed for previous ischemic stroke, epistaxis, or epilepsy are in line with results from a Finish case–control study conducted by Saloheimo *et al.*,[22] although the observed risks were somewhat higher in their study.

Cerebral amyloid angiopathy has been discussed as an important risk factor for ICH by Pezzini *et al.* and Schutz *et al.* [23,24] and was specifically studied for warfarin-associated ICH by Rosand *et al.* [25] who conducted a genetic and pathologic study in 107 patients taking warfarin. In this study, cerebral amyloid angiopathy was diagnosed in 7 out of 11 patients with available tissue samples. Since cerebral amyloid angiopathy is a rare disease most other studies could not investigate the risk of ICH for this risk factor.

### *Strengths and limitations*

To our knowledge this is the largest study analyzing the risk of ICH for phenprocoumon exposure and other risk factors. Previous studies have not specifically analyzed

Table 3. Time of initiation of phenprocoumon therapy and risk of ICH

| Initiation of phenprocoumon exposure (before index day) | Cases $N = 8138$ $n$ (%) | Controls $N = 81\,373$ $n$ (%) | Adjusted incidence rate ratio* | 95% confidence interval |
|---|---|---|---|---|
| No current exposure† | 7323 (89.99%) | 78 392 (96.34%) | 1.00 | — |
| Recent initiation (≤90 days) | 116 (1.43%) | 329 (0.40%) | 4.28 | 3.27–5.59 |
| No recent initiation (>90 days) | 699 (8.59%) | 2652 (3.26%) | 3.30 | 2.95–3.69 |

*Adjusted incidence rate ratios and corresponding confidence intervals are adjusted for age, sex and all covariates included in Table 2. The adjusted incidence rate ratios for phenprocoumon exposure refer to a 68 years old patient. The two-way interaction incidence rate ratios for recent initiation and non-recent initiation are 0.97 (95%CI 0.94–0.99) and 0.98 (95%CI 0.97–0.99), respectively.
†Reference category.



DOI: 10.1002/pds

Table 4. Sensitivity analyses for current phenprocoumon exposure using different definitions for current phenprocoumon exposure

| Phenprocoumon exposure | Main exposure assessment* | Prescription within 90 days | Prescription within 180 days | Prescription within 270 days |
|---|---|---|---|---|
| Exposed cases | 815 (10.01%) | 599 (7.36%) | 944 (11.60%) | 1074 (13.20%) |
| Exposed controls | 2981 (3.66%) | 2143 (2.63%) | 3366 (4.14%) | 3872 (4.76%) |
| Crude incidence rate ratio | 3.00 | 2.96 | 3.11 | 3.12 |
| Adjusted incidence rate ratio† | 3.42 | 3.26 | 3.57 | 3.68 |
| 95% confidence interval | (3.08–3.79) | (2.90–3.66) | (3.24–3.94) | (3.36–4.04) |

*With last phenprocoumon prescription overlapping the index date.
†Adjusted incidence rate ratios and corresponding confidence intervals are adjusted for age, sex and all covariates included in Table 2. The adjusted incidence rate ratios for phenprocoumon exposure refer to a 68 years old patient.

the risk for phenprocoumon, but could only investigate the risk for the combined group of phenprocoumon or warfarin due to limited numbers of patients.[4,5,20] Due to the size of our study we could also investigate the risk of ICH for rare diseases such as cerebral amyloid angiopathy or cerebral aneurysm. Selection bias in the choice of controls is unlikely because this study was designed as a nested case–control study in a defined cohort providing both cases and controls. All information was recorded prospectively so that recall bias was avoided.

Cases were defined by a hospitalization for intracerebral bleeding based on the ICD-coded discharge diagnosis of intracerebral bleeding. Due to the great number of cases and restrictions by German data protection laws, case validation based on medical charts was not feasible. However, imaging procedures as e.g. computed tomography (CT) or magnetic resonance imaging (MRI) of the brain were identified in close temporal relation in 88.1% of cases with phenprocoumon use and 88.4% of cases without phenprocoumon use. CTs do not necessarily need to be coded when the patient is in a stroke unit. We cannot rule out detection bias, i.e. that patients with phenprocoumon use undergo imaging procedures more frequently, even if they have only light symptoms and are therefore diagnosed with ICH more frequently. However, since ICH usually has a very severe course, we think that the potential for detection bias is low.

We did not have information on the prescribed daily dose in our database. We calculated the average daily dose (ADD) of phenprocoumon instead and used it for estimating the duration of phenprocoumon use in our study. Sensitivity analyses using different methods of exposure assessment showed that our main study results were robust.

Our study included information on a great number of risk factors such as diabetes mellitus, hypertension, ischemic heart disease, brain tumor, epilepsy, liver diseases, renal failure, alcohol dependence, or NSAID use. However, we did not have information on other risk factors such as smoking, intensity of anticoagulation, or over-the-counter (OTC) use of ASA. Whereas smoking can be considered as a weak risk factor,[14,15] intensity of anticoagulation was an important risk factor in several studies.[1,2,26] Increases in the bleeding risk were reported for International Normalized Ratio (INR) values of more than 3.0 with a dramatic increase in bleeding risk for INR values of greater than 4.0.[2,27] However, since the target INR is between 2.0 and 3.0 for all indications except for mechanical heart valve replacement where the target INR is between 2.0 and 3.5, the risk estimated for phenprocoumon exposure in our study may not suffer from major bias.[28] In Germany, ASA is available as prescription drug in low doses for the secondary prevention of cardiovascular disease, whereas it is used in higher doses as OTC drug. We could, therefore, only consider the risk for ASA in low dose and it is possible that we did not identify all patients with low dose ASA, since some of them may have bought low dose ASA OTC instead of getting a prescription from their physician.

Concomitant heparin use with phenprocoumon may be related to the initiation phase of phenprocoumon when overlapping heparin is required to bridge the time until phenprocoumon develops its full anticoagulant effect. Heparin use could only be identified from outpatient prescriptions because our database does not include information on inpatient medication. Since phenprocoumon with overlapping heparin therapy is often initiated in the hospital and heparin then is a short-term therapy we most likely underestimated concomitant heparin use.

In summary, we identified in our study a risk of ICH associated with phenprocoumon that was similar to risks reported for the combined group of warfarin and phenprocoumon in Denmark where warfarin use is much more prevalent than phenprocoumon use.[4,5,20,29] Our results indicate that the risk of ICH associated with phenprocoumon or warfarin may be in a similar range despite their different pharmacokinetic and pharmacogenetic properties. Further studies are needed which provide risk estimates for the sole use of warfarin



DOI: 10.1002/pds

KEY POINTS

- Intracerebral hemorrhage (ICH) is a severe complication of oral anticoagulation.
- Phenprocoumon, the most widely used oral anticoagulant in Germany, is associated with an increased risk of ICH.
- The risk of ICH for phenprocoumon use is higher (4-fold) for younger patients (<55 years) as compared to older patients (2-fold for >85 years).
- Despite different pharmacokinetic properties the risk of ICH is similar for phenprocoumon and warfarin.

against non-use for further comparison or which permit to compare both substances head-to-head.

## ACKNOWLEDGEMENTS

The authors are grateful to all SHIs that provided data for this study. Three of these are the Allgemeine Ortskrankenkasse (AOK) Bremen/Bremerhaven, the Deutsche Angestellten-Krankenkasse (DAK), and the Handelskrankenkasse (HKK).

## REFERENCES

1. Hylek EM, Singer DE. Risk factors for intracranial hemorrhage in outpatients taking warfarin. *Ann Intern Med* 1994; **120**(11): 897–902.
2. Fang MC, Chang Y, Hylek EM, *et al.* Advanced age, anticoagulation intensity, and risk for intracranial hemorrhage among patients taking warfarin for atrial fibrillation. *Ann Intern Med* 2004; **141**(10): 745–752.
3. Rosand J, Eckman MH, Knudsen KA, Singer DE, Greenberg SM. The effect of warfarin and intensity of anticoagulation on outcome of intracerebral hemorrhage. *Arch Intern Med* 2004; **164**(8): 880–884.
4. Johnsen SP, Pedersen L, Friis S, *et al.* Nonaspirin nonsteroidal anti-inflammatory drugs and risk of hospitalization for intracerebral hemorrhage: a population-based case-control study. *Stroke* 2003; **34**(2): 387–391. DOI: 10.1161/01.STR.0000054057.11892.5B.
5. Gronbaek H, Johnsen SP, Jepsen P, *et al.* Liver cirrhosis, other liver diseases, and risk of hospitalisation for intracerebral haemorrhage: a Danish population-based case-control study. *BMC Gastroenterol* 2008; **8**: 16. DOI: 10.1186/1471-230X-8-16.
6. Hein L, Schwabe U. Antikoagulantien und Thrombozytenaggregationshemmer. In *Arzneiverordnungsreport* Schwabe U, Paffrath D (eds). Springer Medizin Verlag: Heidelberg, 2007; 425–438.
7. Ufer M. Comparative pharmacokinetics of vitamin K antagonists: warfarin, phenprocoumon and acenocoumarol. *Clin Pharmacokinet* 2005; **44**(12): 1227–1246.
8. Pigeot I, Ahrens W. Establishment of a pharmacoepidemiological database in Germany: methodological potential, scientific value and practical limitations. *Pharmacoepidemiol Drug Saf* 2008; **17**(3): 215–223. DOI: 10.1002/pds.1545.
9. Schink T, Behr S, Garbe E. Externe Validierung von Verschreibungsdaten nichtsteroidaler Antirheumatika anhand des Arzneiverordnungs-Reports. Schink, T, Behr, S, and Garbe, E. 54. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (gmds). (9 September 2009).
10. Daly LE. Confidence limits made easy: interval estimation using a substitution method. *Am J Epidemiol* 1998; **147**(8): 783–790.
11. Landefeld CS, Goldman L. Major bleeding in outpatients treated with warfarin: incidence and prediction by factors known at the start of outpatient therapy. *Am J Med* 1989; **87**(2): 144–152.
12. Berwaerts J, Webster J. Analysis of risk factors involved in oral-anticoagulant-related intracranial haemorrhages. *Quart J Med* 2000; **93**(8): 513–521.
13. Bamford J, Sandercock P, Dennis M, Burn J, Warlow C. A prospective study of acute cerebrovascular disease in the community: the Oxfordshire Community Stroke Project–1981-86. 2. Incidence, case fatality rates and overall outcome at one year of cerebral infarction, primary intracerebral and subarachnoid haemorrhage. *J Neurol Neurosurg Psychiatr* 1990; **53**(1): 16–22.
14. Sturgeon JD, Folsom AR, Longstreth WT Jr, *et al.* Risk factors for intracerebral hemorrhage in a pooled prospective study. *Stroke* 2007; **38**(10): 2718–2725. DOI: 10.1161/STROKEAHA.107.487090.
15. Ariesen MJ, Claus SP, Rinkel GJ, Algra A. Risk factors for intracerebral hemorrhage in the general population: a systematic review. *Stroke* 2003; **34**(8): 2060–2065. DOI: 10.1161/01.STR.0000080678.09344.8D.
16. Fang MC, Go AS, Hylek EM, *et al.* Age and the risk of warfarin-associated hemorrhage: the anticoagulation and risk factors in atrial fibrillation study. *J Am Geriatr Soc* 2006; **54**(8): 1231–1236. DOI: 10.1111/j. 1532-5415. 2006.00828.x.
17. Palareti G, Hirsh J, Legnani C, *et al.* Oral anticoagulation treatment in the elderly: a nested, prospective, case-control study. *Arch Intern Med* 2000; **160**(4): 470–478.
18. Hylek EM, Evans-Molina C, Shea C, Henault LE, Regan S. Major hemorrhage and tolerability of warfarin in the first year of therapy among elderly patients with atrial fibrillation. *Circulation* 2007; **115**(21): 2689–2696. DOI: 10.1161/CIRCULATIONAHA.106.653048.
19. Palareti G, Leali N, Coccheri S, *et al.* Bleeding complications of oral anticoagulant treatment: an inception-cohort, prospective collaborative study (ISCOAT). Italian study on complications of oral anticoagulant therapy. *Lancet* 1996; **348**(9025): 423–428. DOI: 10.1016/S0140-6736(96)01109-9.
20. Steffensen FH, Kristensen K, Ejlersen E, Dahlerup JF, Sorensen HT. Major haemorrhagic complications during oral anticoagulant therapy in a Danish population-based cohort. *J Intern Med* 1997; **242**(6): 497–503. DOI: 10.1111/j. 1365-2796. 1997. tb00023.x.
21. Moride Y, Abenhaim L. Evidence of the depletion of susceptibles effect in non-experimental pharmacoepidemiologic research. *J Clin Epidemiol* 1994; **47**(7): 731–737.
22. Saloheimo P, Juvela S, Hillbom M. Use of aspirin, epistaxis, and untreated hypertension as risk factors for primary intracerebral hemorrhage in middle-aged and elderly people. *Stroke* 2001; **32**(2): 399–404.
23. Pezzini A, Padovani A. Cerebral amyloid angiopathy-related hemorrhages. *Neurol Sci* 2008; **29** (Suppl 2): 260–263. DOI: 10.1007/s10072-008-0957-7.
24. Schutz H, Bodeker RH, Damian M, Krack P, Dorndorf W. Age-related spontaneous intracerebral hematoma in a German community. *Stroke* 1990; **21**(10): 1412–1418.
25. Rosand J, Hylek EM, O'Donnell HC, Greenberg SM. Warfarin-associated hemorrhage and cerebral amyloid angiopathy: a genetic and pathologic study. *Neurology* 2000; **55**(7): 947–951.
26. Neau JP, Couderq C, Ingrand P, Blanchon P, Gil R. Intracranial hemorrhage and oral anticoagulant treatment. *Cerebrovasc Dis* 2001; **11**(3): 195–200. DOI: 10.1159/000047638.
27. Levine MN, Raskob G, Beyth RJ, Kearon C, Schulman S. Hemorrhagic complications of anticoagulant treatment: the Seventh ACCP Conference on Antithrombotic and Thrombolytic Therapy. *Chest* 2004; **126** (3 Suppl): 287–310. DOI: 10.1378/chest.126.3_suppl.287S.
28. Meda Pharma. Fachinformation Marcumar. http://www.pharmnet-bund.de/dynamic/de/index.html [accessed 18 December 2009].
29. Danish Medicines Agency. Statistics on medicinal products. http://dkma.medstat.dk/MedStatDataViewer.php [accessed 5 December 2009].

## APPENDIX

### DEFINITION OF POTENTIAL CONFOUNDING FACTORS

Potential confounding factors were identified *via* ICD-10 codes (German Modification) or ATC codes: diabetes mellitus (ICD-10 codes E10–E14 and prescriptions of antidiabetic treatment: ATC codes A10A and A10B), alcohol dependence (ICD-10 code F10 and prescriptions of disulfiram or acamprosate: ATC code N07BB), hypertension (ICD-10 codes I10–I15), ischemic heart disease (ICD-10 codes I20–I25), liver diseases (ICD-10 codes K70–K77, B15–B19), renal failure (ICD-10 codes N17–N19, P96.0), brain tumor (ICD-10 code C71), epilepsy (ICD-10 code G40), ischemic cerebral infarction (ICD-10 codes I63, I64), cerebral amyloid angiopathy (ICD-10 code I68.0*), cerebral aneurysm (ICD-10 code I67.1), epistaxis (ICD-10 code R04.0), intracerebral hemorrhage (ICD code I61), other hemorrhage (ICD-10 codes K92.0, K92.2, K25.0, K25.2, K25.4, K25.6, K26.0, K26.2, K26.4, K26.6, K27.0, K27.2, K27.4, K27.6, K28.0, K28.2, K28.4, K28.6, K29.0, I85.0, K22.6, K31.82, K55.22, K57.01, K57.03, K57.11, K57.13, K57.21, K57.23, K57.31, K57.33, K57.41, K57.43, K57.51, K57.53, K57.81, K57.83, K57.91, K57.93, K62.5, I60, I62, R04, H92.2, N02.-, N42.1, N83.6, N85.7, N89.7, N93.-, N95.0, R31, H11.3, H21.0, H31.3, H35.6, H43.1, I31.2, J94.2, K66.1, D68.3, D69.8, D69.9, M25.0, R23.3, R58), platelet aggregation inhibitors (ATC code B01AC), heparin (ATC code B01AB), non-steroidal anti-inflammatory drugs (NSAIDs) including acetylsalicylic acid (ASA) (ATC codes M01A, M01BA, N02BA01, N02BA51, N02BA71), selective serotonin reuptake inhibitors (SSRIs) (ATC code N06AB), diuretics (ATC code C03), corticosteroids (ATC code H02), and statins (ATC codes C10AA, C10BA).



DOI: 10.1002/pds

# Appendix C

# Paper 2: Risk of subarachnoid hemorrhage associated with antithrombotic drug use

**Contribution to the manuscript** I herewith certify that I contributed to the design of the study, performed all statistical analyses, interpreted the results, and made critical revision of the manuscript for important intellectual content.

Stroke
Journal of the American Heart Association

American Heart Association | American Stroke Association

# Risk of Subarachnoid Hemorrhage and Early Case Fatality Associated With Outpatient Antithrombotic Drug Use

Edeltraut Garbe, Stefan H. Kreisel and Sigrid Behr

*Stroke*. published online June 27, 2013;
*Stroke* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 0039-2499. Online ISSN: 1524-4628

# Original Contribution

# Risk of Subarachnoid Hemorrhage and Early Case Fatality Associated With Outpatient Antithrombotic Drug Use

Edeltraut Garbe, MD, PhD; Stefan H. Kreisel, MD, MSc; Sigrid Behr, Dipl-Math

***Background and Purpose***—Subarachnoid hemorrhage (SAH) accounts for <7% of all strokes, but is an enormous individual and societal burden. We investigated the risk of SAH associated with prior use of antithrombotic drugs and their influence on 30-day case fatality.

***Methods***—We conducted a nested case–control study in a cohort of 13.4 million members of the German Pharmacoepidemiological Research Database. Ten controls were matched to each case hospitalized for SAH between July 2004 and November 2006 by health insurance, year of birth, and sex using risk set sampling. Exposure was assessed for the warfarin analog phenprocoumon, heparin, clopidogrel/ticlopidine, and acetylsalicylic acid. Multivariable-adjusted odds ratios (ORs) for SAH were estimated by conditional logistic regression. Risk factors for 30-day case fatality were assessed in patients with SAH by logistic regression.

***Results***—The nested case–control study included 2065 SAH cases and 20 649 matched controls. The risk of SAH was significantly increased for phenprocoumon (OR, 1.7; 95% confidence interval [CI], 1.3–2.3), clopidogrel/ticlopidine (OR, 1.7; 95% CI, 1.1–2.5), and for acetylsalicylic acid use (OR, 1.5; 95% CI, 1.2–2.0), but not for outpatient heparin use (OR, 1.2; 95% CI, 0.5–2.7). The early case fatality of 22.8% was associated with an age >70 years (OR, 2.3; 95% CI, 1.8–3.1) and arterial hypertension (OR, 1.3; 95% CI, 1.0–1.6), but not with any of the antithrombotic drugs.

***Conclusions***—Outpatient antithrombotic drug use was associated with an increased risk of SAH, but no association was observed with early case fatality. **(*Stroke*. 2013;44:00-00.)**

**Key Words:** anticoagulant ■ coumarin ■ phenprocoumon ■ platelet aggregation inhibitors

Subarachnoid hemorrhage (SAH) causes only 5% to 7%[1,2] of all incident strokes; however, treatment-related costs and loss in productivity are high.[3,4] One in 2 affected persons is <55 years and dies rapidly or experiences severe disability.[4] In 2000–2008, the risk of death within 30 days after onset was ≈25%[5,6] and 10% to 15% before reaching the hospital.[5,6] Survivors often remain impaired with persistent SAH-related symptoms and experience a greatly reduced quality of life.[7]

A ruptured intracranial aneurysm accounts for 85% of SAH[4] and is more frequent in patients with autosomal-dominant polycystic kidney disease.[4,8,9] Despite genetic predisposition, intracranial aneurysm is usually not congenital, but develops throughout the course of life.[4,9] Thus, modifiable risk factors, such as hypertension, smoking, and alcohol abuse, remain most important in the prevention of SAH.[1,2,8,10]

The influence of antithrombotic drugs on the risk of SAH has not been systematically studied, although these are widely used in the secondary prevention of thromboembolic diseases. In Germany, the warfarin analog phenprocoumon was the only vitamin K antagonist in use during the study period.[11] It has the longest plasma half-life of the coumarins and is associated with high-dose variability.[12] In previous studies, we showed an increased risk of serious bleeding from all causes in phenprocoumon users[13,14] and also for intracerebral hemorrhage.[15]

The present study was conducted to assess the influence of antithrombotic drugs on the risk of SAH and on early case fatality within 30 days.

## Methods

### Study Design and Setting

We conducted a nested case–control study in a cohort of 13.4 million insurance members included in the German Pharmacoepidemiological Research Database who fulfilled the inclusion criteria defined below. This database contains data from 4 German statutory health insurances (SHIs) covering all regions in Germany and representing ≈20% of the German population. The study was based on data from the years 2004–2006 because more recent data were not available to us at the time of the analysis. The database included insurance members' demographic characteristics, information on all hospitalizations, outpatient physician visits, and all refundable outpatient prescriptions. Death of insurance members can be identified (1) as reason for exit from the SHI or (2) as reason for discharge from the hospital for hospitalized patients. The hospital data contain information about the periods of and reasons for admission and discharge with diagnoses, as well as

Received January 15, 2013; final revision received May 3, 2013; accepted May 9, 2013.

From the Departments of Clinical Epidemiology (E.G.), and Biometry and Data Management (S.B.), Leibniz Institute for Prevention Research and Epidemiology–BIPS, Bremen, Germany; Faculty of Social and Health Sciences, University of Bremen, Bremen, Germany (E.G.); and Department of Psychiatry and Psychotherapy Bethel, Evangelisches Krankenhaus Bielefeld, Bielefeld, Germany (S.H.K.).

Correspondence to Edeltraut Garbe, MD, PhD, Leibniz Institute for Prevention Research and Epidemiology–BIPS, Achterstr 30, 28359 Bremen, Germany. E-mail garbe@bips.uni-bremen.de



***Stroke*** **is available at http://stroke.ahajournals.org** **DOI: 10.1161/STROKEAHA.111.000811**

diagnostic and therapeutic procedures. Claims of outpatient physician visits include outpatient treatments, procedures, and diagnoses. All diagnoses are coded according to the German modification of the *International Classification of Diseases, Tenth Revision* (*ICD-10* GM). Prescription data include the date of prescription and drug dispensation at the pharmacy, the amount of substance prescribed, and information on the prescribing physician. Prescription data are linked via the central pharmaceutical reference number to a pharmaceutical reference database, which contains information on the anatomic-therapeutic-chemical code, the defined daily dose, packaging size, strength, formulation, generic and trade name. Preliminary analyses on age and sex distribution, the number of hospital admissions, and drug use have shown the database to be representative for Germany.[16]

In Germany, the use of health insurance data for scientific research is regulated by the Code of Social Law. All involved SHIs, the Federal Ministry of Health (for federal SHI data), and the provincial health authority (for regional SHI data) approved the use of the data for this study. Informed consent was not required by law because the study was based on pseudonymous data.

## Study Population and Outcome Assessment

All insurants with ≥6 months of continuous insurance and no history of hospitalization for SAH during this period were included in the cohort. Cohort entry was defined as July 1, 2004, or the first day after 6 months of continuous enrollment in the SHI. Thereafter, data were collected from cohort entry onward until hospitalization for SAH, death, end of the insurance, or November 30, 2006. The latter was chosen to avoid incomplete data for hospitalizations spanning the end of the year.

Cases of SAH were identified from the main hospital discharge diagnosis (*ICD*-10 GM code of I60). To ensure that SAH was an acute event, a procedure code indicating imaging by computed tomography, MRI, or arteriography or documentation of other intracranial procedures was also required (codes available on request). The day of hospital admission was defined as the index date for the respective case. In a sensitivity analysis we also excluded all cases with head injuries (*ICD-10* GM codes S00-S09) and those without MRI or computed tomography.

Ten controls were matched to each case by sex, year of birth, SHI, and time in cohort using risk set sampling.[17] Thereby, the index date in each control was chosen with the same time of follow-up as for the corresponding case. Cohort members who were hospitalized on the index date of the case were excluded from the set of potential controls because they were not at risk of being admitted to hospital because of SAH at that time.

Early case fatality was defined as the proportion of all patients with SAH who died in the 30-day period after the index date. Information on death was obtained from the hospital discharge and insurance records. In addition, cases without any medical treatment or physician contacts beyond 30 days after the index date were presumed to have died.

## Exposure Assessment

Exposure to the following anticoagulants was assessed: phenprocoumon, unfractionated or low molecular weight heparins, clopidogrel/ticlopidine, and low-dose acetylsalicylic acid (ASA). Exposure was defined as current if the last prescription overlapped with the 7-day period preceding the index date. The duration of a prescription was estimated by the amount of defined daily doses for all anticoagulants except for phenprocoumon where the defined daily dose could not be applied because of high interindividual dose variability. For phenprocoumon, the average daily dose was estimated for each patient by dividing the cumulative phenprocoumon dose until the last outpatient phenprocoumon prescription before the index date by the number of days corresponding to this period. The average daily dose was then used to estimate the duration of exposure for the last prescription preceding the index date. If there was only 1 prescription before the index date, the defined daily dose was used instead. Sensitivity analyses were conducted using fixed exposure assessment periods of 90, 180, and 270 days before the index date.

## Confounder Assessment

The following comorbid conditions were assessed from hospital and outpatient diagnoses in the time period 6 months before cohort entry: diabetes mellitus, systemic hypertension, ischemic heart disease, ischemic cerebral infarction, cerebral aneurysm, brain tumor, epilepsy, liver and renal failure, polycystic kidney disease, alcohol dependence, history of bleeding events, and connective tissue disorders (Ehlers–Danlos syndrome, Marfan syndrome, neurofibromatosis, and fibromuscular dysplasia). Diabetes mellitus and alcohol dependency were identified from diagnoses and prescriptions of antidiabetic substances and disulfiram or acamprosate, respectively. Selective serotonin reuptake inhibitors were also considered in the analyses.

## Statistical Methods

Incidence rates of SAH were calculated in the full cohort for different age groups stratified by sex. Corresponding 95% confidence intervals (CIs) were estimated by the substitution method assuming a Poisson distribution for the number of bleedings.[18] In addition, incidence rates were standardized by age and sex to the 2006 European population[19] using the direct method.[20]

On the basis of case–control data, crude odds ratios (ORs) were calculated using the Mantel–Haenszel estimator to account for matching. Multivariable conditional logistic regression analyses were conducted to estimate adjusted ORs and 2-sided 95% CI for SAH in subjects currently using phenprocoumon, heparin, or platelet aggregation inhibitors (ie, drugs of interest).

The preliminary multivariable model included known risk factors of SAH without interaction terms. Relevant covariates were selected by backward elimination using the Wald test ($P<0.05$) and forcing of all drugs of interest to stay in the model. Two-way interactions between sex or age, and other risk factors were only added to the model if they were significant at the 5% level. In addition, 2-way interactions between all considered antithrombotics were explored.

Predictors for 30-day case fatality were analyzed in subjects with SAH by logistic regression analysis including all considered risk factors. Backward selection was performed at a significance level of $P<0.1$ because of the small sample size.

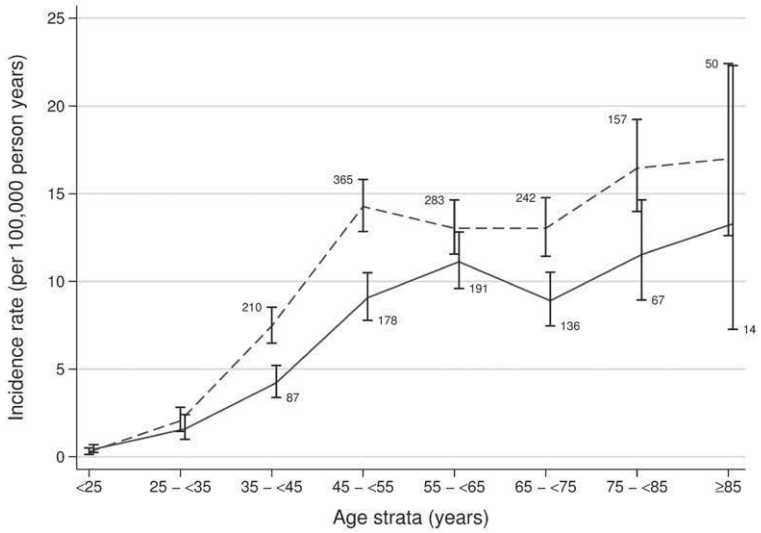
Statistical analyses were performed using SAS/STAT software, version 9.2 of the SAS system for Windows (SAS Institute, Inc, Cary, NC).



## Results

The cohort included 13.4 million insurants with a median follow-up time of 883 days. The average age was 39.9 years with a SD of 22.2 years, and 55% of cohort members were women. The overall crude incidence of SAH in this cohort was 7.1 (95% CI, 6.8–7.4) hemorrhages per 100 000 person-years. It was higher in women (8.39; 95% CI, 7.95–8.85 SAH per 100 000 person-years) than in men (5.42; 95% CI, 5.02–5.83 per 100 000 person-years) and rose with increasing age (Figure). The overall direct standardized incidence rate to the 2006 European population was 6.38 (95% CI, 6.10–6.65) per 100 000 person-years.

Within this cohort, we identified 2065 cases of SAH and 20 649 matched controls. Characteristics of the case–control sample are presented in Table 1. The final multivariable model included the drugs of interest and the covariables diabetes mellitus, arterial hypertension, cerebral aneurysm, epilepsy, polycystic kidney disease, alcohol dependence, and selective serotonin reuptake inhibitors. No interaction terms were added to the final multivariable model because either the main effect or the interaction term was not significant in the respective analysis.




| Age strata (years) | Male | Female |
|---|---|---|
| <25 | 0.42 (0.24 – 0.69) | 0.27 (0.13 – 0.50) |
| 25 – <35 | 1.58 (0.99 – 2.40) | 2.05 (1.44 – 2.82) |
| 35 – <45 | 4.22 (3.38 – 5.20) | 7.45 (6.48 – 8.53) |
| 45 – <55 | 9.06 (7.78 – 10.50) | 14.27 (12.84 – 15.81) |
| 55 – <65 | 11.12 (9.60 – 12.81) | 13.03 (11.56 – 14.64) |
| 65 – <75 | 8.90 (7.46 – 10.52) | 13.02 (11.44 – 14.77) |
| 75 – <85 | 11.53 (8.94 – 14.65) | 16.46 (13.98 – 19.24) |
| ≥85 | 13.29 (7.26 – 22.29) | 16.99 (12.61 – 22.40) |

**Figure.** Incidence of subarachnoid hemorrhage by age group and stratified by sex. The solid line connects incidence rates per 100 000 person-years per age stratum in men, the dashed line in women. 95% confidence interval (CI) are shown. The numbers next to the CI bars are absolute numbers of cases per given age stratum.

Table 2 shows the crude and adjusted ORs based on the multivariable analysis model. The crude ORs were similar to those obtained from the multivariable model. Use of ASA, clopidogrel/ticlopidine, and phenprocoumon was associated with a small, but significantly increased risk of SAH. The risk was also increased for heparin; however, this increase was not significant. Among the other risk factors, high risks were observed for cerebral aneurysm and polycystic kidney disease. The adjusted ORs resulting from the full model, including all covariables and those of the sensitivity analysis excluding cases with head injuries, were similar to those obtained from the final model (results not shown). Similar results were also observed for the sensitivity analyses using fixed exposure assessment periods of 90, 180, and 270 days before the index date (results not shown).

A total of 470 subjects with SAH (22.8%) died within 30 days after hospitalization. Seventy percent of these were women and the average age was 60.5 (SD 15.8) years. A significantly increased risk was observed for ages >70 years (OR, 2.3; 95% CI, 1.8–3.1) and arterial hypertension (OR, 1.3; 95% CI, 1.0–1.6). Antithrombotic drug use was not found to be associated with increased 30-day case fatality.

## Discussion

In our population-based nested case–control study, use of antithrombotics was associated with an increased risk of SAH, but not with early case fatality.

Our findings on phenprocoumon are in line with the results of a case–control study by Risselada et al,[21] which investigated the risk of SAH for phenprocoumon, acenocoumarol, and platelet aggregation inhibitors with the Institute for Drug Outcomes Research database in the Netherlands. Only 1 study has investigated the risk of SAH in patients receiving warfarin.[22] This was a population-based case–control study in Northern Denmark using data from the Danish National Registry of Patients. In contrary to our results, this study did not report an increased risk of SAH for vitamin K antagonist use (90% of patients used warfarin, only 10% had phenprocoumon prescribed). Overall vitamin K antagonist use in the Danish controls was with 1.3% similar to that in the controls of our study (1.6%); however, only 0.8% of the

**Table 1. Characteristics of Cases and Controls**

| Variable | Cases (n=2065) n (%) | Controls (n=20 649) n (%) |
|---|---|---|
| Patient characteristics | | |
| Age, y, mean (SD) | 56.3 (14.6) | |
| Women | 1354 (65.6) | 13 539 (65.6) |
| Antithrombotic and anticoagulant medications | | |
| Acetylsalicylic acid | 72 (3.5) | 488 (2.4) |
| Clopidogrel/ticlopidine | 31 (1.5) | 177 (0.9) |
| Heparin | 8 (0.4) | 50 (0.2) |
| Phenprocoumon | 55 (2.7) | 326 (1.6) |
| Comorbid conditions before cohort entry | | |
| Diabetes mellitus | 172 (8.3) | 2083 (10.1) |
| Arterial hypertension | 710 (34.4) | 6308 (30.6) |
| Ischemic heart disease | 219 (10.6) | 2059 (10.0) |
| Ischemic cerebral infarction | 47 (2.3) | 371 (1.8) |
| Cerebral aneurysm | 20 (1.0) | 9 (<0.1) |
| Brain tumor | 14 (0.7) | 68 (0.3) |
| Epilepsy | 31 (1.5) | 152 (0.7) |
| Liver failure | 158 (7.7) | 1475 (7.1) |
| Renal failure | 57 (2.8) | 428 (2.1) |
| Polycystic kidney disease | 3 (0.1) | 6 (<0.1) |
| Alcohol dependence | 45 (2.2) | 212 (1.0) |
| History of bleeding | 12 (0.6) | 81 (0.4) |
| Connective tissue disorders* | 2 (0.1) | 6 (<0.1) |
| Concomitant medication | | |
| Selective serotonin reuptake inhibitors | 42 (2.0) | 243 (1.2) |

*Connective tissue disorders: Ehlers–Danlos syndrome, Marfan syndrome, neurofibromatosis, and fibromuscular dysplasia.

**Table 2. Crude and Adjusted Odds Ratios of Subarachnoid Hemorrhage for Antithrombotics and Risk Factors**

| Variable | Crude OR | Adjusted* OR | 95% CI** | *P* Value** |
|---|---|---|---|---|
| Antithrombotic medications | | | | |
| Acetylsalicylic acid | 1.5 | 1.5 | 1.2–2.0 | 0.001 |
| Clopidogrel/ticlopidine | 1.8 | 1.6 | 1.1–2.4 | 0.016 |
| Heparin | 1.6 | 1.2 | 0.6–2.8 | 0.604 |
| Phenprocoumon | 1.7 | 1.7 | 1.3–2.3 | <0.001 |
| Comorbid conditions | | | | |
| Diabetes mellitus | 0.8 | 0.7 | 0.6–0.9 | <0.001 |
| Arterial hypertension | 1.2 | 1.2 | 1.1–1.4 | <0.001 |
| Cerebral aneurysm | 22.2 | 19.5 | 8.8–43.3 | <0.001 |
| Epilepsy | 2.1 | 1.7 | 1.1–2.5 | 0.014 |
| Polycystic kidney disease | 5.0 | 4.8 | 1.2–19.4 | 0.026 |
| Alcohol dependence | 2.2 | 2.0 | 1.4–2.8 | <0.001 |
| Concomitant medication | | | | |
| Selective serotonin reuptake inhibitors | 1.8 | 1.7 | 1.2–2.3 | 0.003 |

CI indicates confidence interval; and OR, odds ratio
*Adjusted for all other covariables.
**95% confidence intervals and *P* values refer to the adjusted odds ratios.

Danish cases had used vitamin K antagonists compared with 2.7% of the cases in our study. The age and sex distribution was nearly the same in both studies. Comorbidity of the study populations could not be well compared because the Danish study lacked outpatient diagnoses.

The increased risk observed for low-dose ASA is in line with the results of another Danish population–based nested case–control study, which reported a 2.5-fold risk of SAH associated with new use of low-dose ASA, whereas this study did not find an increased risk for long-term ASA use.[23] Because of low numbers of clopidogrel users, this Danish study was inconclusive with respect to the risk of clopidogrel. In the study by Risselada et al,[21] the use of platelet aggregation inhibitors was not associated with an increased risk of SAH. This study did, however, not differentiate between new or long-term use of platelet aggregation inhibitors or between ASA and clopidogrel.

In our study, heparin use resulted in a slight, but nonsignificant increase in the risk of SAH. Because only outpatient drug use was available to us for analysis, and this was rather low, our study had limited power to detect any increased risk. There are no other studies which have reported on the risk of heparin use.

We did not observe an increased risk for 30-day case fatality with any of the antithrombotic drugs used; however, power was limited for some of the antithrombotic exposures because of the small sample size for this analysis. Mortality was, however, significantly increased in patients >70 years and with arterial hypertension. Age has previously been shown to be a strong predictor for 60-day case fatality in a prognostic model presented by Risselada et al.[24] This prognostic model was based on data from the randomized International Subarachnoid Aneurysm Trial which provided considerably more clinical detail for risk prediction than the health insurance data we had available for our study. Antithrombotic drug use and arterial hypertension were not considered as prognostic factors in the model reported by Risselada et al.[24]

Early case fatality (ie, within 30 days of the event) was 22.8% and is at the lower end of the estimates reported in a recent systematic review for high income countries (covering the period between 2000 and 2008, excluding Germany).[5] Because of the advancement of diagnostic and treatment strategies for SAH, a constant decline of case fatality has been observed over time,[5,6] which is in line with the rather low observed 30-day case fatality in our study.

Our results were consistent with previous research in terms of the crude incidence of SAH[1] and the median age of patients with SAH.[23] The standardized incidence rate was slightly lower than that provided in another study.[25] Our study also confirmed several well-known risk factors for SAH, such as arterial hypertension,[2,8] alcohol dependence,[2,8,10] and autosomal-dominant polycystic kidney disease.[10] Our findings also confirmed the reduced risk for diabetes mellitus reported for case–control studies.[2] The reason for this association is not well understood, but it was suggested that patients with diabetes mellitus might have a higher risk of mortality because of other causes, reducing the chances of developing SAH compared with controls.[2]

## Strengths and Limitations

The study was conducted in a large database representative for Germany[26,27] including >17 million subjects. It provides data on the practice of antithrombotic prescribing and the occurrence of SAH in a real-life setting on a population level. The large size of German Pharmacoepidemiological Research Database enabled us to also investigate single drugs and not only, for example, combined drug classes and rare diseases as risk factors for SAH. Because prescription data are available with the exact date of dispensal, there is low potential for misclassification of drug exposure when compared with field studies on the basis of interview data. Selection bias in the choice of controls is unlikely because this study was designed as a nested case–control study in a defined cohort providing both cases and controls. In addition, all information was recorded prospectively, thereby avoiding recall bias.

Cases of SAH were identified by the main hospital discharge diagnosis which provides the reason for the hospitalization. We did not consider secondary discharge diagnoses to avoid misclassifying prevalent as incident cases. Cases also had to have specific imaging and surgical procedures for SAH or one of the two to ensure an acute event. Because of the large amount of cases, but foremost because of restrictions of German data protection laws, we could not validate the cases on the basis of medical charts. However, with the case definition chosen, the incidence of SAH we observed was comparable with that of other studies.[1,4,23] A sensitivity analysis excluding SAH cases with head injuries and cases without MRI or computed tomography corroborated the findings.

Because the database does not contain information on the prescribed dose and duration of use, the duration of exposure for phenprocoumon was estimated on the basis of average daily dose. Sensitivity analyses using fixed exposure assessment periods of 90, 180, and 270 days before the index date showed that our results were robust. Although we included information on many potential risk factors of SAH, we could not control for other potential confounders, such as smoking or over the-counter use of high-dose ASA because this information is

not available in the database. However, whether a risk factor is truly a confounder depends on whether it is also associated with the exposure under study. Besides, the potential for confounding also depends on the magnitude of the risk of the potential risk factor. Smoking has been shown to be an important risk factor for SAH; however, the magnitude of the risk has varied between 1.2[27] and 2.2 in a systematic review of longitudinal studies.[2] The association of smoking with oral anticoagulant exposure is probably weak, given that oral anticoagulation is prescribed for many conditions which seem rather unrelated to smoking. We, therefore, do not expect major confounding by lack of adjustment for smoking in our analyses. This is in line with the results of one of our previous studies on phenprocoumon use and serious bleeding where additional information on smoking obtained for a subsample of patients in a 2-phase analysis did not result in a relevant change of the risk estimate for phenprocoumon.[13] Information on anticoagulation intensity is also lacking in the database. However, as it is a prerequisite that a confounder does not lie on the causal pathway between exposure and outcome, we are not concerned by this lack of information because we believe that high international normalized ratio values are on the causal pathway between phenprocoumon use and bleeding. Therefore, adjustment for anticoagulation intensity as a confounder in the statistical analysis is not appropriate. We could not provide information on the new generation of anticoagulants, such as rivaroxaban or dabigatran, as they had not yet been marketed during the study period.

## Conclusions

We found that outpatient use of antithrombotic drugs increased the risk of SAH. We did not observe an increase in 30-day case fatality for the antithrombotic drugs under study; however, the power for this analysis was limited because of the small sample size for this analysis. Early case fatality was associated with an age >70 years and arterial hypertension.



## Acknowledgments

We thank J. Böse who helped us with the preparation of the article.

## Disclosures

Dr Garbe is running a department that occasionally performs studies for pharmaceutical industries with the full freedom to publish. Companies include Mundipharma, Bayer, Stada, SanofiAventis, SanofiPasteur, Novartis, Takeda, Celgene, and GlaxoSmithKline. Dr Garbe has been consultant to Bayer, Nycomed, Teva, GlaxoSmithKline, and Novartis unrelated to this work. The other authors report no conflict.

## References


1. de Rooij NK, Linn FH, van der Plas JA, Algra A, Rinkel GJ. Incidence of subarachnoid haemorrhage: a systematic review with emphasis on region, age, gender and time trends. *J Neurol Neurosurg Psychiatry*. 2007;78:1365–1372.
2. Feigin VL, Rinkel GJ, Lawes CM, Algra A, Bennett DA, van Gijn J, et al. Risk factors for subarachnoid hemorrhage: an updated systematic review of epidemiological studies. *Stroke*. 2005;36:2773–2780.
3. Johnston SC, Selvin S, Gress DR. The burden, trends, and demographics of mortality from subarachnoid hemorrhage. *Neurology*. 1998;50:1413–1418.
4. van Gijn J, Kerr RS, Rinkel GJ. Subarachnoid haemorrhage. *Lancet*. 2007;369:306–318.
5. Feigin VL, Lawes CM, Bennett DA, Barker-Collo SL, Parag V. Worldwide stroke incidence and early case fatality reported in 56 population-based studies: a systematic review. *Lancet Neurol*. 2009;8:355–369.
6. Nieuwkamp DJ, Setz LE, Algra A, Linn FH, de Rooij NK, Rinkel GJ. Changes in case fatality of aneurysmal subarachnoid haemorrhage over time, according to age, sex, and region: a meta-analysis. *Lancet Neurol*. 2009;8:635–642.
7. Greebe P, Rinkel GJ, Hop JW, Visser-Meily JM, Algra A. Functional outcome and quality of life 5 and 12.5 years after aneurysmal subarachnoid haemorrhage. *J Neurol*. 2010;257:2059–2064.
8. Kissela BM, Sauerbeck L, Woo D, Khoury J, Carrozzella J, Pancioli A, et al. Subarachnoid hemorrhage: a preventable disease with a heritable component. *Stroke*. 2002;33:1321–1326.
9. Rinkel GJ, Djibuti M, Algra A, van Gijn J. Prevalence and risk of rupture of intracranial aneurysms: a systematic review. *Stroke*. 1998;29:251–256.
10. Ruigrok YM, Buskens E, Rinkel GJ. Attributable risk of common and rare determinants of subarachnoid hemorrhage. *Stroke*. 2001;32:1173–1175.
11. Hein L, Schwabe U. Antikoagulantien und Thrombozytenaggregationshemmer. In: Schwabe U, Paffrath D, eds. *Arzneiverordnungsreport*. Heidelberg: Springer Medizin Verlag; 2007:425–438.
12. Ufer M. Comparative pharmacokinetics of vitamin K antagonists: warfarin, phenprocoumon and acenocoumarol. *Clin Pharmacokinet*. 2005;44:1227–1246.
13. Behr S, Schill W, Pigeot I. Does additional confounder information alter the estimated risk of bleeding associated with phenprocoumon use–results of a two-phase study. *Pharmacoepidemiol Drug Saf*. 2012;21:535–545.
14. Jobski K, Behr S, Garbe E. Drug interactions with phenprocoumon and the risk of serious haemorrhage: a nested case-control study in a large population-based German database. *Eur J Clin Pharmacol*. 2011;67:941–951.
15. Behr S, Andersohn F, Garbe E. Risk of intracerebral hemorrhage associated with phenprocoumon exposure: a nested case-control study in a large population-based German database. *Pharmacoepidemiol Drug Saf*. 2010;19:722–730.
16. Schink T, Garbe E. Representativity of dispensations of non-steroidal antiinflammatory drugs (NSAIDs) in the German Pharmacoepidemiological Research Database. *Pharmacoepidemiol Drug Saf*. 2010;19:S294. Abstract.
17. Rothman KJ, Greenland S, Lash TL. Case-Control Studies. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3 rd. Philadelphia, PA: Lippincott Williams & Wilkens; 2008:111–127.
18. Daly LE. Confidence limits made easy: interval estimation using a substitution method. *Am J Epidemiol*. 1998;147:783–790.
19. EUROSTAT-Europe in Figures. *Eurostat yearbook 2008*. Available at: URL: http://epp.eurostat.ec.europa.eu. Accessed: April 3, 2013.
20. International Agency for Research on Cancer. Statistical Methods in Cancer Research - Volume II - The Design and Analysis of Cohort Studies. IARC Scientific Publications No. 82. Available at: URL:http://www.iarc.fr/en/publications/pdfs-online/stat/sp82/SP82.pdf.1987. Accessed: April 3, 2013.
21. Risselada R, Straatman H, van Kooten F, Dippel DW, van der Lugt A, Niessen WJ, et al. Platelet aggregation inhibitors, vitamin K antagonists and risk of subarachnoid hemorrhage. *J Thromb Haemost*. 2011;9:517–523.
22. Olsen M, Johansen MB, Christensen S, Sørensen HT. Use of vitamin K antagonists and risk of subarachnoid haemorrhage: a population-based case-control study. *Eur J Intern Med*. 2010;21:297–300.
23. Schmidt M, Johansen MB, Lash TL, Christiansen CF, Christensen S, Sørensen HT. Antiplatelet drugs and risk of subarachnoid hemorrhage: a population-based case-control study. *J Thromb Haemost*. 2010;8:1468–1474.
24. Risselada R, Lingsma HF, Bauer-Mehren A, Friedrich CM, Molyneux AJ, Kerr RS, et al. Prediction of 60 day case-fatality after aneurysmal subarachnoid haemorrhage: results from the International Subarachnoid Aneurysm Trial (ISAT). *Eur J Epidemiol*. 2010;25:261–266.
25. Pigeot I, Ahrens W. Establishment of a pharmacoepidemiological database in Germany: methodological potential, scientific value and practical limitations. *Pharmacoepidemiol Drug Saf*. 2008;17:215–223.
26. Schink T, Garbe E. Assessment of the representativity of in-patient hospital diagnoses in the German Pharmacoepidemiological Research Database. *Pharmacoepidemiol Drug Saf*. 2010;19:S178–S179. Abstract.
27. de Rooij NK, Rinkel GJ, Dankbaar JW, Frijns CJ. Delayed cerebral ischemia after subarachnoid hemorrhage: a systematic review of clinical, laboratory, and radiological predictors. *Stroke*. 2013;44:43–54.

# Appendix D

# Paper 3: Two-phase study on bleeding risk under phenprocoumon use

**Contribution to the manuscript** I herewith certify that I conceived and designed the study, performed all statistical analyses, interpreted the results, and drafted the manuscript.

PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2012; **21**: 535–545
Published online 2 February 2012 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/pds.3193

ORIGINAL REPORT

# Does additional confounder information alter the estimated risk of bleeding associated with phenprocoumon use—results of a two-phase study

Sigrid Behr*, Walter Schill and Iris Pigeot

*Bremen Institute for Prevention Research and Social Medicine, Bremen University, Bremen, Germany*

## ABSTRACT

**Purpose** Claims databases are an important source for pharmacoepidemiological studies although they often lack information on some confounders. Two-phase methodology was used to estimate the bleeding risk in patients treated with phenprocoumon from claims data combined with additional information on body mass index (BMI) and smoking.

**Methods** We conducted a nested case–control study using claims data from 2004 to 2007 (phase 1). Additional information was obtained from interviews in a subset of 505 insurants (phase 2). Adjusted bleeding OR were calculated using logistic regression using data from the complete case–control dataset. Furthermore, a two-phase analysis was conducted, taking into consideration phase 2 data on BMI and smoking.

**Results** The phase 1 sample included 1248 cases and 24 960 controls. In phase 1, we observed an adjusted bleeding ORs of 3.93 (95%CI: 2.75–5.61) for male subjects aged 55 years taking phenprocoumon. The bleeding risk associated with phenprocoumon use decreased with increasing age. The two-phase analysis revealed smoking and a high BMI as risk factors for bleeding. The OR for phenprocoumon obtained from the two-phase analysis was of similar size as the phase 1 estimate.

**Discussion** Phase 2 data added valuable information on smoking and BMI. However, phase 1 results did not change dramatically after accounting for phase 2 information, which is reassuring for the validity of database studies. Copyright © 2012 John Wiley & Sons, Ltd.

KEY WORDS—two-phase design; validation sample; missing confounder information; database study

*Received 22 July 2011; Revised 14 October 2011; Accepted 21 November 2011*

## BACKGROUND

Claims databases are an important source for pharmacoepidemiological studies, although they often lack information on potentially relevant confounders. The German Pharmacoepidemiological Research Database (GePaRD) consists of claims data from several German statutory health insurances. A large number of risk factors can be identified from diagnosis codes and medications included in this database. However, the database does not contain information on body mass index (BMI) and smoking status, which are important risk factors for many diseases. In previous research, we conducted a study investigating the risk of intracerebral hemorrhage associated with phenprocoumon exposure using the GePaRD.[1] A limitation of this study was that BMI, smoking status, intensity of anticoagulation and over-the-counter (OTC) use of acetylsalicylic acid (ASA) could not be considered in the analysis because this information was not available in the claims database. One idea to overcome this limitation is to collect additional data from patient interviews in a subset of patients and to include this information into the database study by using two-phase methodology. Two-phase designs were introduced into epidemiology in the 1980s by Walker and White to reduce costs of covariate collection in field studies.[2,3] The basic idea of a two-phase case–control study is that disease status and at least crude information on exposure and covariates are available for all cases and controls in phase 1, whereas precise or additional information on exposure and covariates is collected for a subsample of cases and controls in phase 2. The statistical analysis is then conducted on the complete phase 2 data taking into

*Correspondence to: S. Behr, Achterstr. 30, 28359 Bremen, Germany. E-mail: behr@bips.uni-bremen.de

account the phase 1 information, which is available for the full case–control sample. Although the two-phase design has been recognized as efficient design for large database studies,[4] until now, there are only very few studies adopting this methodology.[5] The present study uses a two-phase design to investigate whether additional confounder information on BMI and smoking status obtained from patient interviews alters the results of a database study on the risk of serious bleedings associated with phenprocoumon use. The focus of the paper is to describe the methods and results of this two-phase study to provide an example of how two-phase methodology can be exploited to better cope with missing confounder information.

## METHODS

### *Study design*

We conducted a two-phase case–control study in a population of inhabitants of the State of Bremen who were insured at the regional healthcare provider AOK Bremen/Bremerhaven between 2004 and 2007. Phase 1 data were obtained from insurance claims comprising information on demographic characteristics, hospital admissions, ambulatory physician visits, and ambulatory prescriptions. Hospital diagnoses as well as ambulatory diagnoses were coded according to the German modification of the International Classification of Diseases (ICD-10 GM). Ambulatory prescribed drugs were characterized by the central pharmaceutical number, which was linked to the corresponding anatomical–therapeutic–chemical (ATC) code and the Defined Daily Dose (DDD). In phase 1, cases and controls were sampled from a cohort of insurants who were required to be continuously insured for 6 months before cohort entry. Cohort entry was defined as the first day after 6 months of continuous insurance. Cohort exit was defined as the first of the following dates: end of the insurance period, hospitalization for serious bleedings as defined below, death or end of study period (31 December 2007). To ensure that all cohort members were able to participate in the telephone interviews, which were conducted for phase 2, we excluded insurants with a diagnosis of dementia during the study period and insurants younger than 18 years or older than 75 years at the time of the interview. Cases were defined as cohort members hospitalized for serious bleeding including gastrointestinal bleedings, intracerebral bleedings, urogenital bleedings and other bleedings. The ICD-10 GM codes used for identification of serious bleedings are specified in the appendix. The respective hospital admission date is referred to as the index date for cases. Cohort members who did not become a case during the study constitute the set of potential controls. Controls were randomly selected from this set with a case : control ratio of 1:20. A random index date was chosen from the period between cohort entry and cohort exit for each control. This control selection strategy was chosen instead of the usual risk-set sampling because two-phase methodology cannot be applied for matched case–control pairs. Because the set of potential controls, which were under risk at each time point during the study, was very large, both selection strategies would lead to similar control groups.

We identified current phenprocoumon exposure on the index date by ambulatory prescriptions with ATC code B01AA04 overlapping the index date. The duration of exposure associated with the last prescription before the index date was calculated using the estimated average daily dose as described in Behr *et al.*[1]

The following risk factors for serious bleedings and potential confounders were extracted from the claims data in phase 1 of the study: sex, age at cohort entry, comorbid conditions (assessed in the 6 months before cohort entry), and concomitant medications (assessed by prescriptions overlapping the index date). Comorbid conditions included diabetes mellitus, systemic hypertension, ischemic heart disease, ischemic cerebral infarction, epilepsy, cancer, liver disease, renal failure, chronic obstructive pulmonary disease (COPD), alcohol dependency, diverticular disease, gastrointestinal disease (e.g., ulcers), and history of serious bleedings. We considered current use of the following substances as concomitant medications: platelet aggregation inhibitors (i.e., clopidogrel, ticlopidine), heparins, non-steroidal anti-inflammatory drugs (NSAIDs), prescribed ASA, selective serotonin reuptake inhibitors (SSRIs), diuretics, corticosteroids, statins, and gastroprotective drugs (i.e., H2-receptor antagonists, proton pump inhibitors, sucralfat, misoprostol, gastrozepin). Details on the specification of diseases and substances are given in the appendix.

### *Phase 2 data*

Additional information for phase 2 was gained from computer-assisted telephone interviews, which were performed in a subsample of persons included in phase 1. The selection process for the subsample aimed at achieving a balanced design regarding case–control status and phenprocoumon exposure on the index date. According to Breslow and Cain, the balanced design is almost always the most efficient design for sampling phase 2.[6,7] For this purpose, the phase 1 sample was stratified according to case–control and exposure status, and two batches of 2000 persons were selected by



DOI: 10.1002/pds

oversampling cases and exposed persons. For adopting a two-phase approach, overall representativeness of the phase 2 sample is not required. The representativeness in each stratum was achieved by selecting persons randomly from each stratum. We collected information on BMI, current and past smoking behavior, general health status, medication use for pain treatment (including OTC drugs), hospital admissions, and on gastrointestinal disorders for 505 insurants who completed the full interview.

The study was conducted with permission from the ethical committee. Furthermore, we obtained individual informed consent from each participant of the phase 2 sample.

### *Statistical analysis*

The statistical analysis was accomplished in two steps. The first step included a full analysis of the phase 1 case–control sample considering all information available in phase 1. In the second step, a two-phase analysis was conducted that made use of the combined data from phase 1 and phase 2.

For the analysis of the phase 1 case–control sample, the adjusted OR of serious bleedings comparing phenprocoumon users with non-users and the corresponding 95%CI were calculated by means of multivariable logistic regression. The regression model included all risk factors and potential confounders specified above, age (centered at 55 years), sex and interactions between age and phenprocoumon use as well as between sex and phenprocoumon use. Further interaction terms for phenprocoumon use and comorbid conditions or concomitant medications were added to the model if they were significant (Wald test, $p$-value $< 0.05$). A sensitivity analysis was conducted to assess the relevance of the covariate selection by adopting a backward selection process that excluded non-significant ($p \geq 0.05$) covariates stepwise. Age, sex, phenprocoumon use, and the interactions with phenprocoumon use were forced to stay in the model. All phase 1 analyses were done using SAS 8.2 and SAS 9.2 (SAS Institute Inc., Cary, NC).

The two-phase analysis was conducted as an exploratory analysis investigating the effect of BMI and smoking behavior on the OR of serious bleedings associated with phenprocoumon use. Several two-phase logistic regression models were fitted including current phenprocoumon use, age (centered at 55 years), sex, smoking status at cohort entry, BMI ($< 30$ versus $\geq 30$ kg/m²), and two-way interaction terms with phenprocoumon use as explanatory variables.

One simple approach to analyze a two-phase study is as follows: first, the full case–control sample is stratified according to information available for each person in phase 1 of the study. Second, sampling probabilities for participation in phase 2 are calculated for each stratum. Then, a weighted logistic regression analysis is conducted based on cases and controls with complete information in the phase 2 sample using the inverse sampling probabilities as weights. More details on the weighted logistic regression are given in Flanders and Greenland.[8] Alternatives to the weighted likelihood approach are pseudo-likelihood and maximum likelihood approaches.[9] Because all three approaches involve complex formulas for the variance estimation, standard logistic regression procedures cannot be used for estimation. However, the three approaches have now been implemented in the R package osDesign.[10] In this study, the maximum likelihood estimator as described in Breslow and Holubkov was used to estimate ORs and corresponding SEs.[11] Ninety-five percent CIs were calculated based on the asymptotic normality of the maximum likelihood estimates.

Because the two-phase methodology requires non-empty strata among cases and controls in the phase 2 sample, stratifications had to be constructed that avoided empty cells. In a first step, phase 1 data were stratified in 12 strata by cross-classifying age at cohort entry ($<$50 years, 50– $<$ 65 years, and $\geq$65 years), sex, and phenprocoumon exposure on the index day. Inclusion of further covariate information in additional stratifications was only considered for common diseases as arterial hypertension and diabetes mellitus. ORs and corresponding SEs were estimated by adopting the R package osDesign.[10] Results were validated with the SAS-based program package by Schill *et al.*[12]

## RESULTS

The study cohort consisted of 186 438 insurants of the AOK Bremen/Bremerhaven, which had a mean follow-up time of 2.83 years (SD 1.10 years). Within this cohort, 1248 cases of serious bleedings were identified, and 24 960 controls were sampled for inclusion in the phase 1 case–control sample. Of the 4000 persons who were initially selected to participate in the phase 2 survey, 3280 were contacted by mail. The remaining persons could not be contacted because they were no longer insured with the AOK Bremen/Bremerhaven, died, or lived in nursing homes at the time of the survey. A total of 505 interviews could be completed, resulting in a response proportion of 15.35%. Two patients had to be excluded from phase 2 because they were not entitled to receive benefits during the study period. In addition, one patient was excluded because of lack of data



DOI: 10.1002/pds

quality.[1] Thus, the phase 2 sample consisted of 156 cases and 346 controls. Because information on BMI was missing for one control and information on smoking was not observed for one case and two further controls, regression analyses including BMI and smoking status were based on 498 patients.

Characteristics of the phase 1 and phase 2 case–control sample are summarized in Table 1. Compared with phase 1, patients in the phase 2 sample were older and had more claims for comorbidities and comedications. The differences were more pronounced among controls than among cases. A logistic regression analysis modeling the probability of participation in phase 2 revealed age as an important predictor with a higher probability of participation for older age. No association with probability of participation was observed for most comorbidities and most comedications; weak associations were found for sex, diverticular disease, and use of statins. The two-phase analysis accounted for the observed differences between the phase 1 and the phase 2 sample by including information on age and sex in all stratifications.

Analysis of the additional information from the patient interviews disclosed that most of the participants in phase 2 were overweight (70% with BMI $\geq$25 kg/m²), and 66% were past or current smokers (Table 2). Cases and controls in the phase 2 sample differed from each other with respect to their smoking behavior with more smokers and also more packyears of smoking among cases.

The multivariable analysis of the phase 1 data revealed that current phenprocoumon use was associated with a 3.9-fold risk of serious bleeding with 55-year-old male subjects as a reference group. The most important risk factors were increasing age, a history of serious bleedings, alcohol dependency, and use of antithrombotic medications (heparin, platelet aggregation inhibitors). A significant interaction between phenprocoumon use and age was observed, indicating that increased age resulted in a decreased risk of serious bleedings associated with phenprocoumon use. No further interactions with phenprocoumon met the criteria for inclusion in the model. The results of the final model are summarized in Table 3. Similar risk estimates were obtained from the backward selection model from which the covariates ischemic heart disease, COPD, statin use, ischemic stroke, diverticular disease, hypertension, and cancer were eliminated. Neither the main effect for sex nor the interaction between sex and phenprocoumon use were significant in the fully adjusted model. However, when adjusting only for arterial hypertension, diabetes mellitus, and age (model used for comparison with the two-phase analysis), male sex turned out as a significant risk factor for bleeding, and the interaction between sex and phenprocoumon use also was significant with a higher risk for female phenprocoumon users (Table 5).

Three different stratifications as depicted in Table 4 were used for the two-phase analysis. Stratification A comprised phase 1 information on age, sex, and phenprocoumon exposure cross-classified into 12 strata for cases and controls. These strata were further classified by underlying hypertensive disease in stratification B and by diabetes mellitus in stratification C. To avoid empty cells in stratifications B and C, some strata needed to be combined. All two-phase logistic regression models included the phase 1 covariates age, sex, phenprocoumon exposure, hypertension, and diabetes mellitus. However, the effects of hypertension and diabetes mellitus on serious bleedings could only be estimated with sufficient precision in the two-phase analysis if the stratification considered phase 1 information on hypertension (stratification B) or diabetes mellitus (stratification C), respectively. Table 6 also shows estimates for the effects of hypertension and diabetes mellitus using the three stratifications. Results of two different two-phase models based on stratification A (one model with interaction between phenprocoumon and smoking and the other model without this interaction) and results of the respective phase 1 analysis are compared in Table 5. In both two-phase models, current smoking was a significant risk factor for serious bleedings. The interaction between phenprocoumon use and smoking was not significant. However, the risk of bleeding associated with phenprocoumon use regardless of smoking status (i.e., estimated from the two-phase model without the interaction between phenprocoumon and smoking) was 4.96 (95%CI: 2.91–8.45), which was different from the risk estimated for phenprocoumon use among non-smokers in a model with interaction term (OR: 7.40, 95%CI: 3.31–16.54). Considering the estimate for the interaction term (OR: 0.37, 95% CI: 0.10–1.36), the corresponding risk of bleeding among phenprocoumon users who smoke was 2.75. Using stratification B, the OR for the interaction between current smoking and phenprocoumon use was 0.33 (95%CI: 0.09–1.17). The two-phase models omitting the smoking interaction revealed similar risk estimates for phenprocoumon use compared with the phase 1 analysis as can be seen from Figure 1.

An increased BMI ($\geq$30 kg/m²) was associated with a 1.6-fold increased risk of bleeding. No interaction between BMI and phenprocoumon use was observed in the two-phase analysis, indicating that the risk of

[1]Most survey items for this patient were missing. In particular, the patient could not remember any dates.



DOI: 10.1002/pds

Table 1. Characteristics of cases and controls in phase 1 and phase 2 sample

| | Phase 1 sample | | Phase 2 sample* | |
|---|---|---|---|---|
| | Cases | Controls | Cases | Controls |
| | $N = 1248$ | $N = 24960$ | $N = 156$ | $N = 346$ |
| | *n* (%) | *n* (%) | *n* (%) | *n* (%) |
| Age: mean (SD) | 54.6 (13.4) | 42.8 (16.2) | 59.2 (9.9) | 56.1 (12.9) |
| Age categories | | | | |
| <30 years | 88 (7.1%) | 6301 (25.2%) | 2 (1.3%) | 20 (5.8%) |
| 30–<50 years | 279 (22.4%) | 9423 (37.8%) | 24 (15.4%) | 61 (17.6%) |
| 50–<65 years | 525 (42.1%) | 6364 (25.5%) | 69 (44.2%) | 160 (46.2%) |
| ≥ 65 years | 356 (28.5%) | 2872 (11.5%) | 61 (39.1%) | 105 (30.3%) |
| Male sex | 686 (55.0%) | 12792 (51.3%) | 71 (45.5%) | 169 (48.8%) |
| Phenprocoumon exposure | 117 (9.4%) | 313 (1.3%) | 26 (16.7%) | 74 (21.4%) |
| Bleeding type | | | | |
| Gastrointestinal bleeding | 639 (51.2%) | — | 79 (50.6%) | — |
| Urogenital bleeding | 216 (17.3%) | — | 33 (21.2%) | — |
| Cerebral bleeding | 189 (15.1%) | — | 21 (13.5%) | — |
| Other bleeding | 204 (16.3%) | — | 23 (14.7%) | — |
| Comorbid conditions[$] | | | | |
| Diabetes mellitus | 224 (17.9%) | 1526 (6.1%) | 35 (22.4%) | 54 (15.6%) |
| Hypertension | 397 (31.8%) | 3298 (13.2%) | 64 (41.0%) | 108 (31.2%) |
| Ischemic heart disease | 137 (11.0%) | 839 (3.4%) | 23 (14.7%) | 42 (12.1%) |
| Ischemic cerebral infarction | 25 (2.0%) | 94 (0.4%) | 4 (2.6%) | 8 (2.3%) |
| Epilepsy | 31 (2.5%) | 218 (0.9%) | 2 (1.3%) | 0 (0.0%) |
| Cancer | 76 (6.1%) | 542 (2.2%) | 12 (7.7%) | 19 (5.5%) |
| Liver diseases | 164 (13.1%) | 1136 (4.6%) | 18 (11.5%) | 27 (7.8%) |
| Renal failure | 43 (3.4%) | 140 (0.6%) | 3 (1.9%) | 7 (2.0%) |
| COPD | 80 (6.4%) | 678 (2.7%) | 12 (7.7%) | 20 (5.8%) |
| Alcohol dependence | 118 (9.5%) | 492 (2.0%) | 9 (5.8%) | 9 (2.6%) |
| Diverticular disease | 31 (2.5%) | 151 (0.6%) | 7 (4.5%) | 8 ( 2.3%) |
| Gastrointestinal diseases | 107 (8.6%) | 930 (3.7%) | 14 (9.0%) | 21 (6.1%) |
| History of serious bleeding | 22 (1.8%) | 22 (0.1%) | 2 (1.3%) | 0 (0.0%) |
| Concomitant medication[§] | | | | |
| Current use of the following: | | | | |
| Platelet aggregation inhibitors | 23 (1.8%) | 57 (0.2%) | 4 (2.6%) | 1 (0.3%) |
| Heparin | 27 (2.2%) | 35 (0.1%) | 4 ( 2.6%) | 2 (0.6%) |
| NSAIDs | 126 (10.1%) | 1014 (4.1%) | 17 (10.9%) | 23 (6.6%) |
| ASA | 90 (7.2%) | 422 (1.7%) | 10 (6.4%) | 17 (4.9%) |
| SSRIs | 19 (1.5%) | 124 (0.5%) | 2 (1.3%) | 0 (0.0%) |
| Diuretics | 286 (22.9%) | 1970 (7.9%) | 37 (23.7%) | 89 (25.7%) |
| Corticosteroids | 54 (4.3%) | 306 (1.2%) | 5 (3.2%) | 6 (1.7%) |
| Statins | 118 (9.5%) | 825 (3.3%) | 21 (13.5%) | 45 (13.0%) |
| Gastroprotective drugs[#] | 188 (15.1%) | 1005 (4.0%) | 24 (15.4%) | 33 (9.5%) |

*Information shown for phase 2 sample was obtained from the claims data.
[$]Assessed in the 6-month baseline period preceding cohort entry.
[§]Ambulatory prescriptions overlapping the index day.
[#]Gastroprotective drugs include H2-receptor antagonists, proton pump inhibitors, sucralfat, misoprostol, and gastrozepin.
COPD, chronic obstructive pulmonary disease; NSAID, non-steroidal anti inflammatory drug; ASA, acetylsalicylic acid; SSRI, selective serotonin reuptake inhibitor; ICH, intracerebral hemorrhage.

bleeding associated with phenprocoumon use was the same for high and low BMI (Tables 5 and 6). Models including BMI information as continuous covariate or as dichotomous variable with more than two categories provided consistent results for the effect of BMI on bleeding (results not shown).

The different stratifications had little effect on the OR estimates for phenprocoumon use (Figure 1). However, the precision with which ORs were estimated varied between the stratifications, leading to significant results for BMI and hypertension when employing stratification B and significant results for diabetes mellitus when employing stratification C (Table 6).

The effect of smoked packyears on the bleeding risk was investigated in a sensitivity analysis. A higher risk was observed for patients with more than 10 packyears of smoking, but no interaction between the number of packyears and phenprocoumon use was detected (results not shown).



DOI: 10.1002/pds

Table 2. Additional information in phase 2

| | Phase 2 sample* | |
|---|---|---|
| | Cases | Controls |
| | $N = 156$ | $N = 346$ |
| BMI[#] | | |
| Mean (SD) | 28.6 (5.9) | 28.1 (6.8) |
| Minimum–maximum | 16–48 | 13–69 |
| Q1–Median–Q3 | 25–28–32 | 24–28–30 |
| BMI categories[#] | | |
| Severe underweight (<16 kg/m²) | 0 (0.0%) | 1 (0.3%) |
| Underweight (16– < 18.5 kg/m²) | 2 (1.3%) | 11 (3.2%) |
| Normal weight (18.5– < 25 kg/m²) | 42 (26.9%) | 96 (27.7%) |
| Overweight (25– < 30 kg/m²) | 56 (35.9%) | 141 (40.8%) |
| Severe overweight (≥ 30 kg/m²) | 56 (35.9%) | 96 (27.7%) |
| Smoking status[$] | | |
| Never smoked | 45 (28.8%) | 121 (35.0%) |
| Past smoker | 49 (31.4%) | 128 (37.0%) |
| Current smoker | 61 (39.1%) | 95 (27.5%) |
| Packyears of smoking[§] | | |
| Never smoked | 45 (28.8%) | 121 (35.0%) |
| > 0–10 packyears | 21 (13.5%) | 61 (17.6%) |
| > 10–20 packyears | 25 (16.0%) | 40 (11.6%) |
| > 20 packyears | 62 (39.7%) | 122 (35.3%) |

*Information shown for phase 2 sample was obtained from the survey data.
[#]BMI was not observed for one control.
[$]Assessed at cohort entry. For one case and two controls, no smoking status was recorded.
[§]Information on packyears was missing for three cases and three controls.
BMI, body mass index.

Table 3. Adjusted odds ratios and 95% confidence intervals for serious bleedings based on phase 1 data

| Multivariable model | Adjusted odds ratio[#] | 95%CI[#] | Wald test |
|---|---|---|---|
| ($N = 26208$) | | | $p$-value |
| Phenprocoumon exposure | 3.93[§] | 2.75–5.61 | < 0.001 |
| Age (centered at 55 years) | 1.04 | 1.03–1.04 | < 0.001 |
| Interaction: age × phenprocoumon exposure | 0.97 | 0.94–0.99 | 0.015 |
| Female sex | 0.89 | 0.79–1.01 | 0.072 |
| Interaction: sex × phenprocoumon exposure | 1.33 | 0.83–2.15 | 0.235 |
| Comorbid conditions | | | |
| Diabetes mellitus | 1.27 | 1.07–1.52 | 0.008 |
| Hypertension | 1.13 | 0.97–1.32 | 0.118 |
| Ischemic heart disease | 1.05 | 0.84–1.32 | 0.670 |
| Ischemic cerebral infarction | 1.28 | 0.78–2.11 | 0.333 |
| Epilepsy | 1.67 | 1.09–2.56 | 0.018 |
| Cancer | 1.26 | 0.96–1.64 | 0.096 |
| Liver diseases | 1.31 | 1.07–1.61 | 0.009 |
| Renal failure | 1.83 | 1.23–2.71 | 0.003 |
| COPD | 0.96 | 0.74–1.25 | 0.775 |
| Alcohol dependence | 3.42 | 2.69–4.35 | < 0.001 |
| Diverticular disease | 1.37 | 0.89–2.11 | 0.149 |
| Gastrointestinal diseases | 1.37 | 1.09–1.73 | 0.008 |
| History of serious bleeding | 9.45 | 4.91–18.19 | < 0.001 |
| Current use of the following: | | | |
| Platelet aggregation inhibitors | 2.74 | 1.60–4.71 | < 0.001 |
| Heparin | 5.88 | 3.27–10.60 | < 0.001 |
| NSAIDs | 1.44 | 1.16–1.78 | < 0.001 |
| ASA | 1.75 | 1.33–2.29 | < 0.001 |
| SSRIs | 1.78 | 1.03–3.09 | 0.039 |
| Diuretics | 1.23 | 1.04–1.45 | 0.016 |
| Corticosteroids | 1.77 | 1.27–2.45 | < 0.001 |
| Statins | 0.90 | 0.71–1.14 | 0.389 |
| Gastroprotective drugs* | 1.77 | 1.46–2.14 | < 0.001 |

[§]Adjusted odds ratio for phenprocoumon exposure refers to a 55-year-old male patient.
[#]Adjusted for all covariates included in the table and for two-way interactions for phenprocoumon use with age and sex.
*Gastroprotective drugs include H2-receptor antagonists, proton pump inhibitors, sucralfat, misoprostol, and gastrozepin.
NSAID, non-steroidal anti inflammatory drug; ASA, acetylsalicylic acid; SSRI, selective serotonin reuptake inhibitor.

## DISCUSSION

Phase 2 information on BMI and smoking behavior added valuable information to the database study investigating phenprocoumon use and the risk of serious bleedings. The phase 1 analysis revealed an increased risk of serious bleedings associated with current phenprocoumon use that decreased with increasing age. The risk of bleeding rose with increasing age among non-users, whereas the effect of age nearly vanished among phenprocoumon users. Because the study only included patients below 75, the observed age effect should not be extrapolated to an elderly population above 75 years.

The effect of sex estimated in the full phase 1 model was not significant, although the results implied a higher risk of bleeding for male subjects in general and a higher risk of bleeding associated with phenprocoumon use for female subjects. Few studies examined the effect of sex on bleeding risk and particularly on anticoagulant-related bleeding because most studies were adjusted for sex by matching or stratification. Results of other studies, in which a significant effect of sex was detected, were consistent with our results.[13–15] If not adjusted for BMI, an observed higher risk for anticoagulant-related bleeding in female subjects also could be related to their lower BMI. However, the two-phase analysis, which was adjusted for BMI also indicated male sex as an independent risk factor for bleeding, whereas female sex was a risk factor for bleeding for phenprocoumon users. Furthermore, we did not observe an interaction between phenprocoumon use and BMI in the two-phase analysis. These results imply that consideration of information on BMI did not alter the risk for bleeding associated with phenprocoumon use estimated from phase 1 data without adjusting for BMI.

Additional knowledge on BMI and smoking was gained from the two-phase analysis. The two-phase analysis revealed an increased bleeding risk for patients with a high BMI. This finding is consistent with the literature. A prospective cohort study conducted by Strate *et al*. in 47 228 male health professionals demonstrated that



DOI: 10.1002/pds

Table 4. Stratification of phase 1 and phase 2 data. a) Stratification A by age, sex and phenprocoumon exposure. b) Stratification B by age, sex, phenprocoumon exposure and hypertension (Hyp). c) Stratification C by age, sex, phenprocoumon exposure, and diabetes mellitus (DM)

| Phase 1 (Phase 2) | Cases | | | | Controls | | | |
|---|---|---|---|---|---|---|---|---|
| | Male | | Female | | Male | | Female | |
| **Stratification A** | | | | | | | | |
| **Exposed**[$] | | | | | | | | |
| **< 50 years** | 9 (2) | | 4 (2) | | 21 (4) | | 20 (5) | |
| **50 - < 65 years** | 28 (7) | | 16 (4) | | 80 (13) | | 45 (12) | |
| **≥ 65 years** | 28 (5) | | 32 (6) | | 94 (22) | | 53 (18) | |
| **Not exposed**[$] | | | | | | | | |
| **< 50 years** | 214 (14) | | 140 (8) | | 8181 (29) | | 7502 (43) | |
| **50 - < 65 years** | 259 (24) | | 222 (34) | | 3143 (69) | | 3096 (66) | |
| **≥ 65 years** | 148 (19) | | 148 (31) | | 1273 (32) | | 1452 (33) | |
| **Stratification B** | **Hyp-** | **Hyp+** | **Hyp+** | **Hyp-** | **Hyp-** | **Hyp+** | **Hyp+** | **Hyp-** |
| **Exposed**[$] | | | | | | | | |
| **< 50 years** | 9 (2) | | 4 (2) | | 21 (4) | | 20 (5) | |
| **50 - < 65 years** | 20 (5) | 8 (2) | 10 (3) | 6 (1) | 47 (6) | 33 (7) | 22 (7) | 23 (5) |
| **≥ 65 years** | 12 (2) | 16 (3) | 19 (3) | 13 (3) | 44 (8) | 50 (14) | 32 (11) | 21 (7) |
| **Not exposed**[$] | | | | | | | | |
| **< 50 years** | 191 (12) | 40 (2) | | 123 (8) | 7894 (27) | 592 (3) | | 7197 (42) |
| **50 - < 65 years** | 182 (15) | 77 (9) | 92 (16) | 130 (18) | 2445 (48) | 698 (21) | 863 (16) | 2233 (50) |
| **≥ 65 years** | 84 (13) | 64 (6) | 68 (20) | 80 (11) | 859 (20) | 414 (12) | 584 (15) | 868 (18) |
| **Stratification C** | **DM-** | **DM+** | **DM+** | **DM-** | **DM-** | **DM+** | **DM+** | **DM-** |
| **Exposed**[$] | | | | | | | | |
| **< 50 years** | 9 (2) | | 4 (2) | | 21 (4) | | 20 (5) | |
| **50 - < 65 years** | 24 (5) | 4 (2) | 3 (1) | 13 (3) | 60 (7) | 20 (6) | 11 (3) | 34 (9) |
| **≥ 65 years** | 28 (5) | 14 (3) | | 18 (3) | 94 (22) | 9 (3) | | 44 (15) |
| **Not exposed**[$] | | | | | | | | |
| **< 50 years** | 206 (14) | 18 (1) | | 130 (7) | 8028 (29) | 289 (4) | | 7366 (39) |
| **50 - < 65 years** | 199 (17) | 60 (7) | 41 (5) | 181 (29) | 2768 (56) | 375 (13) | 336 (3) | 2760 (63) |
| **≥ 65 years** | 110 (12) | 38 (7) | 35 (9) | 113 (22) | 1053 (24) | 220 (8) | 236 (8) | 1216 (25) |

$ Exposed means exposed to phenprocoumon on the index day.
Hyp-: no hypertension in baseline period, Hyp+: hypertension in baseline period
DM-: no diabetes mellitus in baseline period, DM+: diabetes mellitus in baseline period


DOI: 10.1002/pds

Table 5. Results of the two-phase analyses

| Multivariable model[§] | Phase 1 analysis ($N = 26\,208$) | | Two-phase analysis ($n = 498$[#]) Stratification A[$] Model without smoking interaction | | Two-phase analysis ($n = 498$[#]) Stratification A[$] Model with smoking interaction | |
|---|---|---|---|---|---|---|
| | ln OR[§] | OR[§] (95%CI) | ln OR[§] | OR[§] (95%CI) | ln OR[§] | OR[§] (95%CI) |
| Phenprocoumon exposure | 1.42 | 4.14 (2.95–5.81) | 1.60 | 4.96 (2.91–8.45) | 2.00 | 7.40 (3.31–16.54) |
| Age (centered at 55 years) | 0.043 | 1.044 (1.039–1.049) | 0.062 | 1.064 (1.051–1.077) | 0.063 | 1.065 (1.051–1.080) |
| Interaction between phenprocoumon and age | −0.04 | 0.96 (0.94–0.99) | −0.03 | 0.97 (0.94–1.00) | −0.05 | 0.95 (0.91–0.99) |
| Female sex | −0.21 | 0.81 (0.72–0.92) | −0.14 | 0.87 (0.74–1.03) | −0.13 | 0.88 (0.74–1.04) |
| Interaction between phenprocoumon and sex | 0.46 | 1.59 (1.01–2.49) | 0.56 | 1.75 (1.07–2.86) | 0.44 | 1.55 (0.93–2.56) |
| BMI $\geq 30$ kg/m² | — | — | 0.45 | 1.57 (1.00–2.47) | 0.46 | 1.58 (1.00–2.49) |
| Interaction between phenprocoumon and BMI | — | — | −0.39 | 0.68 (0.25–1.82) | −0.43 | 0.65 (0.24–1.75) |
| Current smoker | — | — | 0.83 | 2.30 (1.50–3.53) | 0.96 | 2.61 (1.63–4.17) |
| Interaction between phenprocoumon and smoking | — | — | — | — | −0.99 | 0.37 (0.10–1.36) |

[#]One control without information on BMI and one case and two controls without information on smoking were excluded from the analysis.
[§]ORs also are adjusted for the comorbid conditions hypertension and diabetes mellitus.
[$]Stratification A includes information on age (three categories), sex, and phenprocoumon exposure.
ln, natural logarithm.

obesity is a risk factor for diverticulitis and diverticular bleeding.[16] Furthermore, BMI $\geq 30$ kg/m² was associated with an increased risk of hemorrhagic stroke in a study by Kurth *et al*. who analyzed a cohort of 21 414 US male physicians.[17] In addition, current smoking was associated with an increased bleeding risk in the two-phase analysis. The effect of smoking on the bleeding risk is discussed controversially in the literature. On the one side, smoking has been identified as an independent risk factor for bleeding in several studies. Kaplan *et al*. observed a twofold increased risk for upper gastrointestinal bleeding in a case–control study including 1020 members of an American health maintenance organization who were treated with specific antihypertensive medications.[18] A similar increase in risk was detected in two studies by Kurth *et al*. exploring the association between smoking and hemorrhagic stroke in men and in women.[19,20] Kurth *et al*. suggested

Table 6. Comparison of different stratifications in the two-phase analyses

| Multivariable model | Two-phase analyses ($n = 498$*) | | |
|---|---|---|---|
| | Stratification A[$] | Stratification B[#] | Stratification C[§] |
| | OR (95%CI) | OR (95%CI) | OR (95%CI) |
| Phenprocoumon exposure | 4.96 (2.91–8.45) | 4.92 (3.02–8.04) | 5.67 (3.44–9.35) |
| Age (centered at 55 years) | 1.06 (1.05–1.08) | 1.07 (1.05–1.08) | 1.06 (1.05–1.07) |
| Interaction between phenprocoumon and age | 0.97 (0.94–1.00) | 0.97 (0.94–1.00) | 0.96 (0.93–0.99) |
| Female sex | 0.87 (0.73–1.03) | 0.85 (0.72–1.01) | 0.86 (0.73–1.02) |
| Interaction between phenprocoumon and sex | 1.75 (1.07–2.85) | 1.77 (1.09–2.87) | 1.69 (1.02–2.78) |
| BMI $\geq 30$ kg/m² | 1.57 (1.00–2.47) | 1.64 (1.07–2.51) | 1.56 (1.00–2.42) |
| Interaction between phenprocoumon and BMI | 0.68 (0.25–1.82) | 0.67 (0.28–1.61) | 0.47 (0.18–1.25) |
| Current smoker | 2.30 (1.50–3.53) | 2.47 (1.64–3.72) | 2.24 (1.47–3.42) |
| Comorbid conditions | | | |
| Hypertension | 1.22 (0.80–1.88) | 1.30 (1.06–1.59) | 1.24 (0.82–1.87) |
| Diabetes mellitus | 1.16 (0.72–1.89) | 1.10 (0.68–1.77) | 1.53 (1.22–1.91) |

[$]Stratification A includes information on age (three categories), sex, and phenprocoumon exposure.
[#]Stratification B includes information on age (three categories), sex, phenprocoumon exposure, and hypertension.
[§]Stratification C includes information on age (three categories), sex, phenprocoumon exposure, and diabetes mellitus.
*One control without information on BMI and one case and two controls without information on smoking were excluded from the analysis.



DOI: 10.1002/pds

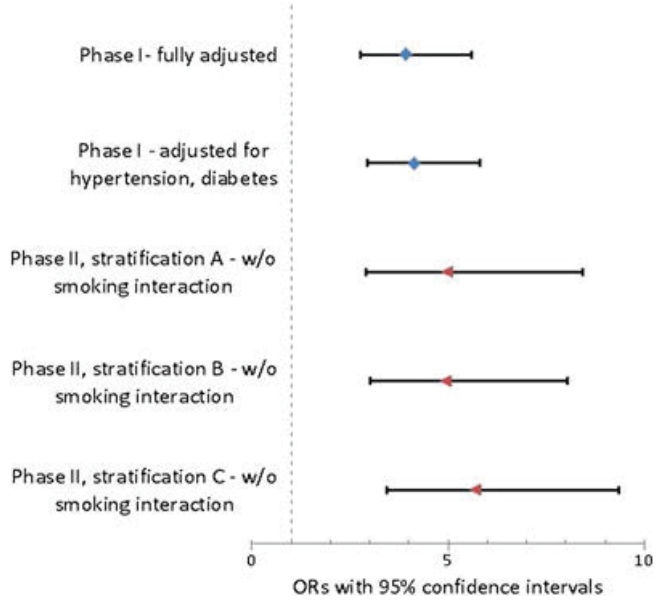



Figure 1. Risk of bleeding associated with phenprocoumon use estimated from phase 1 and phase 2 models using different stratifications

that the structural damage of the arterial wall caused by smoking could be the mechanism by which smoking increases the bleeding risk. On the other side, it is suspected that smoking has a short-term effect of reducing blood flow, which may result in a negative interaction with anticoagulant use. The results of a case–control study investigating the effect of lifestyle and diet on overanticoagulation supported this hypothesis.[21] In our study, we also observed a negative but non-significant interaction between smoking and phenprocoumon use.

From a methodological point of view, the study results demonstrated that the precision of the OR estimates in two-phase analyses depends on the chosen stratification. Stratification on phase 1 variables incorporates the information of these variables into the two-phase analysis. If, for example, age is not used in the stratification, the age distribution in phase 1 would be ignored in the two-phase analysis leading to imprecise estimates for age. This point is illustrated in Table 6; although the effects of hypertension and diabetes mellitus were not significant when using stratification A, both effects became significant when using the alternative stratifications B and C, which incorporate the respective phase 1 distributions.

Database studies have several strengths including the large sample size and the availability of a multitude of covariates and potential confounders. In our study, the lack of important confounder information, for example, BMI, was rectified using additional data from a health survey in a two-phase approach. Because BMI and smoking behavior were self-reported in the survey, there is a potential for misclassification error. However, because only rough categorizations of BMI and smoking were used in the analyses, we do not expect that results were biased by the misclassification error. We did not include survey data on OTC use of pain killers in our analysis because a comparison of survey data with claims data with respect to prescribed pain killers revealed a large amount of reporting error. We also could not account for the intensity of anticoagulation. However, we do not expect major bias because the target international normalized ratio (INR) is between 2.0 and 3.0 for most indications of phenprocoumon and between 2.0 and 3.5 only for mechanical heart valve replacement.[22] Increases in bleeding risk were reported for INR values above 3.0, with the greatest increase for INR values of more than 4.0.[14,23] The prescribed dose of phenprocoumon also was not available in our study, which might lead to misclassification of the exposure status at the index day. Sensitivity analyses involving different definitions of phenprocoumon exposure revealed similar risk estimates for all exposure measures.[1] A further limitation of our two-phase analysis was certainly the small sample size of phase 2, precluding more informative stratifications and inclusion of further covariates in the two-phase analysis. Because we could not include the full set of covariates, the two-phase analyses can only be deemed as sensitivity analyses. Further research will address how to include a maximum of phase 1 information on multiple risk factors in the two-phase analysis.

In conclusion, the phase 2 sample, although very small, added valuable information on the role of smoking and BMI. Hence, the two-phase approach can



DOI: 10.1002/pds

be useful even with small phase 2 samples to cope with missing confounder information. Conversely, it is reassuring that phase 1 results did not change dramatically when phase 2 information on BMI was added because it often is not possible to collect supplemental information for database studies.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

KEY POINTS

- Two-phase methodology can be used to include missing confounder information in database studies.
- The two-phase analysis revealed that high BMI and smoking are risk factors for serious bleedings; information on both factors was lacking in the claims data but obtained via patient interviews.
- The pure phase 1 analysis and the two-phase analysis yielded similar risk estimates for phenprocoumon use and bleeding.

## ACKNOWLEDGEMENT

This work was financially supported by the German Research Foundation (DFG) within the priority program survey methodology (DFG grant: PI345/5-1).

We are grateful to the AOK Bremen/Bremerhaven for providing insurance claims data and collaborating on the survey. Furthermore, we would like to thank Johannes Eggs and the field work unit of the Bremen Institute for Prevention Research and Social Medicine for organizing the survey and Dirk Enders for support in data management and analysis.

## REFERENCES

1. Behr S, Andersohn F, Garbe E. Risk of intracerebral hemorrhage associated with phenprocoumon exposure: a nested case–control study in a large population-based German database. *Pharmacoepidemiol Drug Saf* 2010; **7**(19): 722–730. DOI: 10.1002/pds.1973
2. Walker AM. Anamorphic analysis - sampling and estimation for covariate effects when both exposure and disease are known. *Biometrics* 1982; **4**(38): 1025–1032.
3. White JE. A two stage design for the study of the relationship between a rare exposure and a rare disease. *Am J Epidemiol* 1982; **1**(115): 119–128.
4. Collet JP, Schaubel D, Hanley J, Sharpe C, Boivin JF. Controlling confounding when studying large pharmacoepidemiologic databases: A case study of the two-stage sampling design. *Epidemiology* 1998; **3**(9): 309–315.
5. Sharpe CR, Collet JP, McNutt M, Belzile E, Boivin JF, Hanley JA. Nested case–control study of the effects of non-steroidal anti-inflammatory drugs on breast cancer risk and stage. *Br J Cancer* 2000; **1**(83): 112–120. DOI: 10.1054/bjoc.2000.1119
6. Breslow NE, Cain KC. Logistic-regression for 2-stage case–control data. *Biometrika* 1988; **1**(75): 11–20.
7. Cain KC, Breslow NE. Logistic-regression analysis and efficient design for 2-stage studies. *Am J Epidemiol* 1988; **6**(128): 1198–1206.
8. Flanders WD, Greenland S. Analytic methods for two-stage case–control studies and other stratified designs. *Stat Med* 1991; **10**: 739–747. DOI: 10.1002/sim.4780100509
9. Schill W, Drescher K. Logistic analysis of studies with two-stage sampling: a comparison of four approaches. *Stat Med* 1997; **16**: 117–132. DOI: 10.1002/(SICI)1097-0258(19970130)16:2 < 117::AID-SIM475 > 3.0.CO;2–5
10. Haneuse S, Saegusa T, Lumley T. osDesign: An R package for the analysis, evaluation, and design of two-phase and case–control studies. *J Stat Software* 2011; **43**(11): 1–29. URL: http://www.jstatsoft.org/v43/i11/
11. Breslow NE, Holubkov R. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *J Roy Stat Soc B Meth* 1997; **2**(59): 447–461. DOI: 10.1111/1467-968.00078
12. Schill W, Wild P, Pigeot I. A planning tool for two-phase case–control studies. *Comput Methods Programs Biomed* 2007; **2**(88): 175–181. DOI: 10.1016/j.cmpb.2007.08.002
13. Ariesen MJ, Claus SP, Rinkel GJ, Algra A. Risk factors for intracerebral hemorrhage in the general population: a systematic review. *Stroke* 2003; **8**(34): 2060–2065. DOI: 10.1161/01.STR.0000080678.09344.8D
14. Levine MN, Raskob G, Beyth RJ, Kearon C, Schulman S. Hemorrhagic complications of anticoagulant treatment: the Seventh ACCP Conference on Antithrombotic and Thrombolytic Therapy. *Chest* 2004; **3**(126): 287S-310S. DOI: 10.1378/chest.126.3_suppl.287S
15. Shireman TI, Mahnken JD, Howard PA, Kresowik TF, Hou Q, Ellerbeck EF. Development of a contemporary bleeding risk model for elderly warfarin recipients. *Chest* 2006; **5**(130): 1390–1396. DOI: 10.1378/chest.130.5.1390
16. Strate LL, Liu YL, Aldoori WH, Syngal S, Giovannucci EL. Obesity increases the risks of diverticulitis and diverticular bleeding. *Gastroenterology* 2009; **1**(136): 115–122. DOI: 10.1053/j.gastro.2008.09.025
17. Kurth T, Gaziano JM, Berger K, *et al*. Body mass index and the risk of stroke in men. *Arch Intern Med* 2002; **22**(162): 2557–2562.
18. Kaplan RC, Heckbert SR, Psaty BM. Risk factors for hospitalized upper or lower gastrointestinal tract bleeding in treated hypertensives. *Prev Med* 2002; **4**(34): 455–462. DOI: 10.1006/pmed.2002.1008
19. Kurth T, Kase CS, Berger K, Gaziano JM, Cook NR, Buring JE. Smoking and risk of hemorrhagic stroke in women. *Stroke* 2003; **12**(34): 2792–2795. DOI: 10.1161/01.STR.0000065200.93070.32
20. Kurth T, Kase CS, Berger K, Schaeffner ES, Buring JE, Gaziano JM. Smoking and the risk of hemorrhagic stroke in men. *Stroke* 2003; **5**(34): 1151–1155. DOI: 10.1161/01.STR.0000100165.36466.95
21. Penning-van Beest FJ, Geleijnse JM, van ME, Vermeer C, Rosendaal FR, Stricker BH. Lifestyle and diet as risk factors for overanticoagulation. *J Clin Epidemiol* 2002; **4**(55): 411–417. DOI: 10.1016/S0895-4356(01)00485-1
22. Meda Pharma. Fachinformation Marcumar. http://www pharmnet-bund de/dynamic/de/index html 2009; (accessed 18 December 2009)
23. Fang MC, Chang Y, Hylek EM, *et al*. Advanced age, anticoagulation intensity, and risk for intracranial hemorrhage among patients taking warfarin for atrial fibrillation. *Ann Intern Med* 2004; **10**(141): 745–752.

## APPENDIX A

## DEFINITION OF SERIOUS BLEEDINGS:

The following ICD-10 GM codes were used to identify serious bleedings from the main discharge diagnosis:

- Gastrointestinal bleedings: K25.0, K25.2, K25.4, K25.6, K25.8, K26.0, K26.2, K26.4, K26.6, K26.8, K27.0, K27.2, K27.4, K27.6, K27.8, K28.0, K28.2, K28.4, K28.6, K28.8, K29.0, I85.0, K22.6, K31.82, K55.22, K57.01, K57.03, K57.11, K57.13, K57.21, K57.23, K57.31, K57.33, K57.41, K57.43, K57.51, K57.53, K57.81, K57.83, K57.91, K57.93, K62.5, K92.0, K92.2
- Cerebral bleedings: I60.-, I61.-, I62.-
- Urogenital bleedings: N02.-, N42.1, N83.6, N85.7, N89.7, N93.-, N95.0, R31
- Other bleedings: D68.3, D69.8, D69.9, H11.3, H21.0, H31.3, H35.6, H43.1, H92.2, I31.2, J94.2, K66.1, M25.0, R04.-, R23.3, R58

## DEFINITION OF RISK FACTORS AND POTENTIAL CONFOUNDERS FROM CLAIMS DATA:

Risk factors and potential confounding factors were identified via ICD-10 GM codes from ambulatory and inpatient diagnoses or ATC codes: diabetes mellitus (ICD-10 codes E10 – E14 and prescriptions of antidiabetic treatment: ATC codes A10A and A10B), alcohol dependence (ICD-10 code F10 and prescriptions of disulfiram or acamprosate: ATC code N07BB), systemic hypertension (ICD-10 codes I10–I15), ischemic heart disease (ICD-10 codes I20–I25), liver diseases (ICD-10 codes K70–K77, B15–B19), renal failure (ICD-10 codes N17 – N19, P96.0), cancer s(ICD-10 codes C00-C97), epilepsy (ICD-10 code G40), ischemic cerebral infarction (ICD-10 codes I63, I64), chronic obstructive pulmonary disease (COPD) (ICD-10 code J44), diverticular disease (ICD-10 codes K57, Q43, K38.2), gastrointestinal disease (ICD-10 codes K20, K21, K22.1, K25.1, K25.3, K25.5, K25.7, K25.9, K26.1, K26.3, K26.5, K26.7, K26.9, K27.1, K27.3, K27.5, K27.7, K27.9, K28.1, K28.3, K28.5, K28.7, K28.9, K29.1, K29.3, K29.5, K29.7, K29.9), platelet aggregation inhibitors (ATC codes B01AC04, B01AC05), heparin (ATC code B01AB), non-steroidal anti inflammatory drugs (NSAIDs) (ATC codes M01A, M01BA), acetylsalicylic acid (ASA) (ATC codes N02BA01, N02BA51, N02BA71, B01AC06) selective serotonin reuptake inhibitors (SSRIs) (ATC code N06AB), diuretics (ATC codes C02L, C03, C07B-D, C08G, C09BA, C09DA ), corticosteroids (ATC code H02), statins (ATC codes C10AA, C10BA), and gastroprotective drugs (ATC codes A02BA, A02BC, A02BX02, A02BB01, A01BX03). Concomitant medications were identified by ambulatory prescriptions overlapping the index date or the week before the index date. The duration of a prescription was estimated under consideration of the amount of prescribed substance and the DDD.



DOI: 10.1002/pds

# Appendix E

# Paper 4: Stratification in two-phase database studies with a rich phase 1 data set

Since this manuscript has not been published and most journals prohibit previous publication of substantial parts of the manuscript, only the abstract and a selection of tables and figures are included here.

**Contribution to the manuscript** I herewith certify that I conceived and designed the research, performed all statistical analyses, interpreted the results, and drafted the manuscript.

# Stratification in two-phase database studies with a rich phase 1 data set

Sigrid Behr, Walter Schill

*Department of Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Achterstr. 30, 28359 Bremen, Germany*

**Abstract**

In logistic two-phase studies a dichotomous outcome and some covariate information are available for a large phase 1 data set whereas the full vector of covariates is only observed for a subsample. Information on phase 1 covariates is utilized in the analysis of the phase 2 sample via stratification. Whereas in traditional two-phase studies only few phase 1 variables are collected in a field study, two-phase database studies encompass a multitude of covariates in phase 1. The traditional stratification strategy, cross-classification of all available phase 1 variables, reaches its limits due to tremendously large numbers of strata. New stratification strategies are needed which account for relevant phase 1 covariates. In this work we propose to stratify on percentiles of a disease score, which summarizes information on multiple phase 1 covariates. The new approach is compared to cross-classification in an empirical example of a two-phase pharmacoepidemiological database study as well as by means of a simulation study based on the empirical example.

*Keywords:* Database study, post stratification, stratified sampling, study design, two-phase methodology

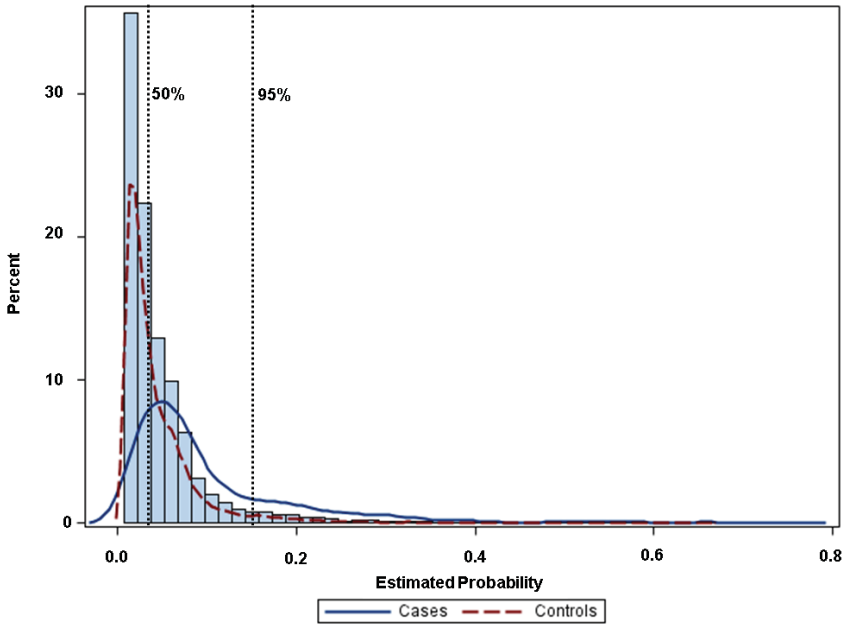



**Figure E.1:** Empirical distribution of the disease score in cases and controls. The histogram in the background shows the marginal distribution. Vertical reference lines correspond to the 50th and 95th percentile of the marginal distribution.

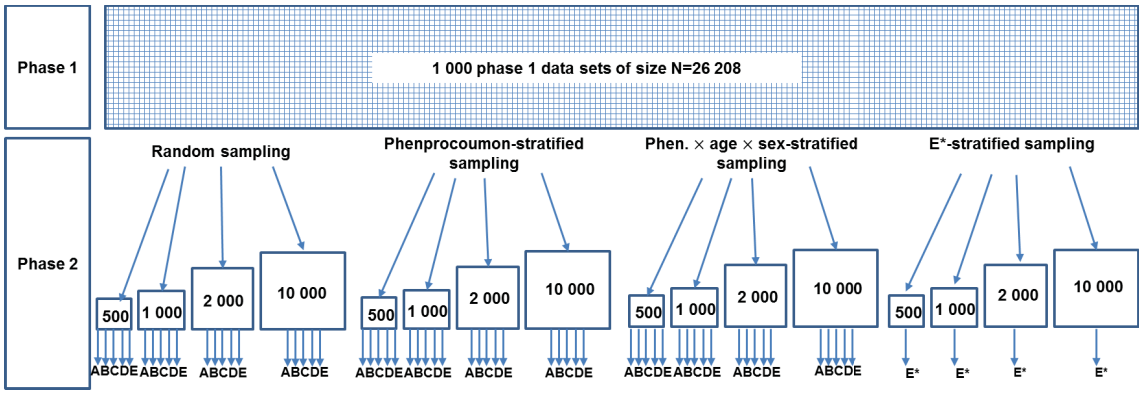
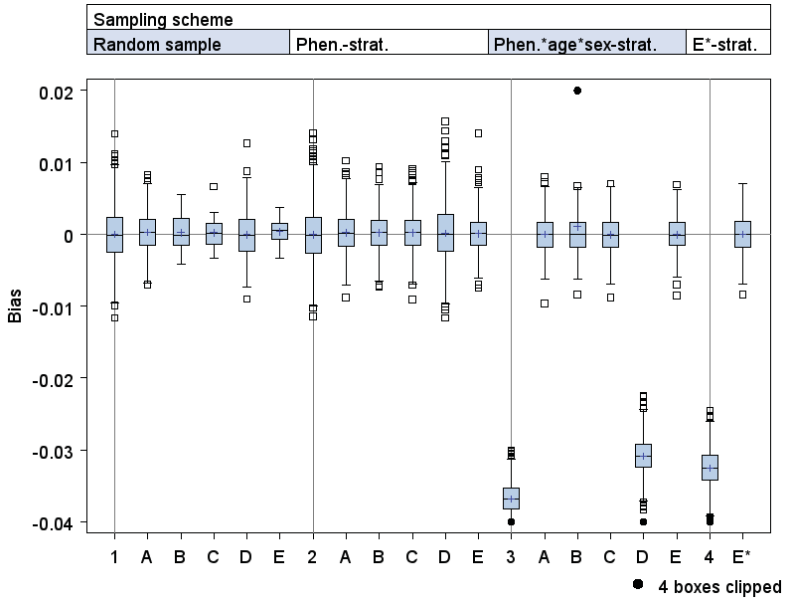
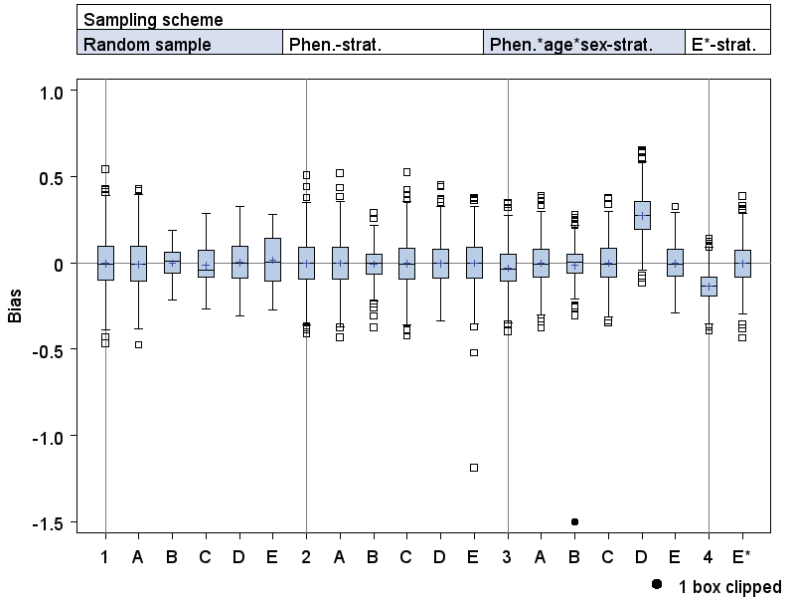



**Figure E.2:** Set-up of the simulation study. Phase 2 samples of sizes 500-10 000 are sampled according to four sampling schemes from each of the 1 000 phase 1 data sets. Phase 2 data sets are analysed with respect to stratifications A-E and E*.




(a) Age




(b) Hypertension

**Figure E.3:** Bias in complete-case and two-phase analyses of phase 2 samples of size 2 000. Vertical reference lines mark boxes showing bias in the complete-case analysis, all other boxes refer to bias from the two-phase analysis. The numbers 1-4 denote the sampling scheme, where 4 refers to sampling according to stratification E*. Boxes are clipped if values exceed 10 times the interquartile range.

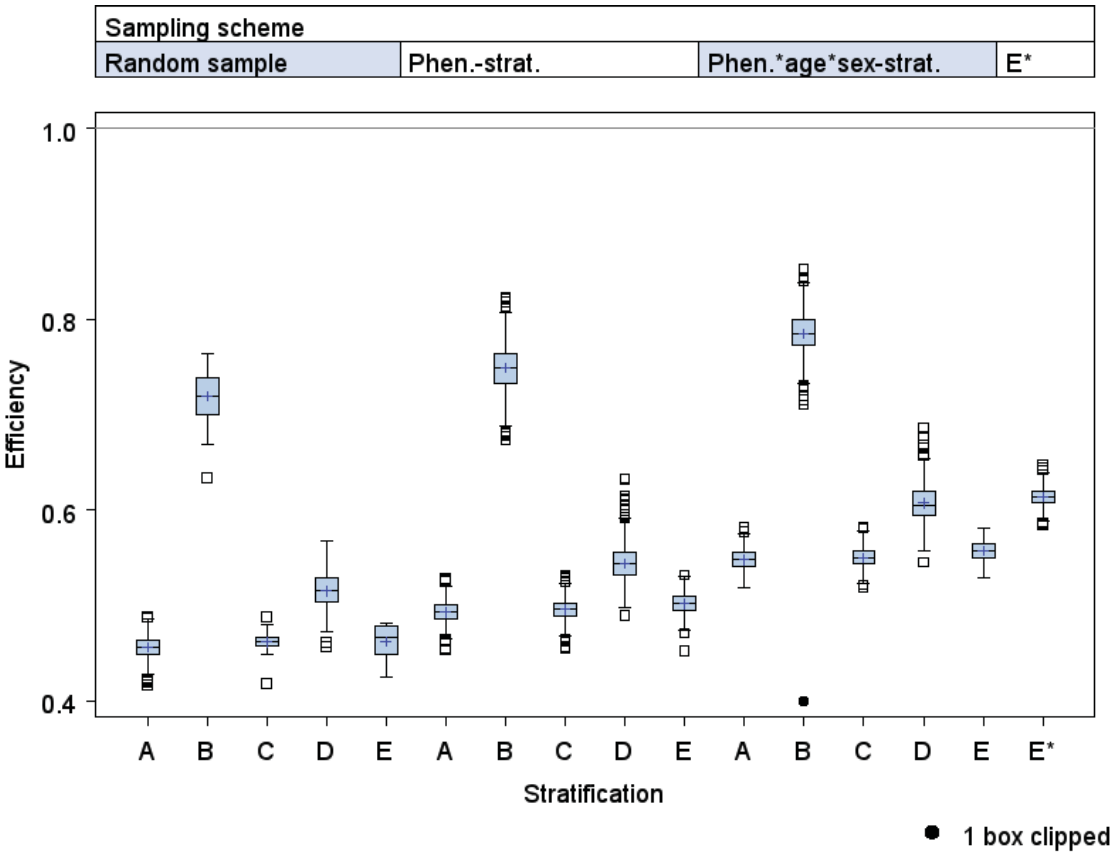



**Figure E.4:** Efficiency of two-phase estimators for hypertension in phase 2 samples of size 2 000. Stratifications are denoted by the capitals A-E. E* denotes the box corresponding to the sample drawn and analysed with respect to stratification E*. Boxes are clipped if values exceed 10 times the interquartile range.

**Table E.1:** Definition of stratifications applied in the empirical example and in the simulation study

| Name | Stratified by... | No of strata in the empirical study | No of strata in the simulation study[a] |
|---|---|---|---|
| A | Phenprocoumon, age (<50, 50-<65, ≥65 years), sex | 12 | 10 |
| B | Phenprocoumon, age (<50, 50-<65, ≥65 years), sex, hypertension | 21 | 20 |
| C | Phenprocoumon, age (<50, 50-<65, ≥65 years), sex, diabetes | 20 | 20 |
| D | Phenprocoumon, percentiles of the disease score (50th, 75th, 90th, 95th, 99th %ile) | 8 | 12 |
| E | Phenprocoumon, age (<50, 50-<65, ≥65 years), sex, 90th percentile of the disease score | – | 20 |
| E* | Phenprocoumon, age (<50, 50-<65, ≥65 years), sex, percentiles of the disease score (90th, 95th %ile) | 20 | 21(20-22) |
| F | Phenprocoumon, age (<65, ≥65 years), sex, percentiles of the disease score (50th, 75th, 90th, 95th, 99th %ile) | 24 | – |

[a] If the number of strata varies, the median number of strata is presented with the quartiles Q1 and Q3 in parentheses.

**Table E.2:** Results of two-phase analyses with different stratifications compared to results from the phase 1 analysis

| Multivariable model | Log OR (SE) | | | | | | |
|---|---|---|---|---|---|---|---|
| | Phase 1 N=26 208 | Two-phase analysis n=498[a] (155 cases, 343 controls) | | | | | |
| | | Strat. A | Strat. B | Strat. C | Strat. D | Strat. E* | Strat. F |
| Phenprocoumon use | 1.37 (0.18) | 1.46 (0.22) | 1.45 (0.22) | 1.51 (0.22) | 1.75 (0.34) | 1.44 (0.22) | 1.32 (0.24) |
| Age[b] | 0.04 (<.01) | 0.07 (0.01) | 0.07 (0.01) | 0.06 (0.01) | 0.06 (0.01) | 0.06 (0.01) | 0.06 (0.01) |
| IA: phen.×age | -0.04 (0.01) | -0.03 (0.02) | -0.03 (0.02) | -0.04 (0.02) | -0.06 (0.03) | -0.04 (0.02) | -0.03 (0.02) |
| Female sex | -0.18 (0.06) | -0.13 (0.10) | -0.16 (0.10) | -0.13 (0.10) | 0.03 (0.22) | -0.14 (0.10) | -0.23 (0.09) |
| IA: phen.×sex | 0.39 (0.24) | 0.58 (0.27) | 0.61 (0.27) | 0.51 (0.27) | 0.06 (0.52) | 0.55 (0.28) | 0.54 (0.30) |
| *Comorbid conditions:* | | | | | | | |
| Diabetes mellitus | 0.29 (0.09) | 0.09 (0.27) | 0.05 (0.26) | 0.40 (0.16) | 0.28 (0.26) | 0.14 (0.25) | 0.31 (0.26) |
| Hypertension | 0.16 (0.08) | 0.31 (0.25) | 0.37 (0.14) | 0.28 (0.25) | 0.45 (0.24) | 0.50 (0.23) | 0.52 (0.23) |
| Ischemic heart disease | 0.10 (0.11) | 0.10 (0.34) | 0.05 (0.33) | 0.23 (0.35) | <.01 (0.33) | 0.21 (0.28) | 0.11 (0.30) |
| Liver disease | 0.56 (0.10) | 0.42 (0.35) | 0.44 (0.33) | 0.46 (0.34) | 0.60 (0.33) | 0.67 (0.29) | 0.55 (0.28) |
| Gastrointestinal disease | 0.44 (0.12) | 0.25 (0.40) | 0.19 (0.39) | 0.39 (0.39) | 0.44 (0.38) | 0.61 (0.35) | 0.61 (0.33) |
| *Current use of:* | | | | | | | |
| NSAIDs | 0.38 (0.11) | 0.37 (0.36) | 0.29 (0.34) | 0.46 (0.37) | 0.54 (0.34) | 0.40 (0.31) | 0.32 (0.29) |
| ASA | 0.60 (0.13) | -0.10 (0.47) | -0.13 (0.46) | -0.03 (0.46) | 0.09 (0.38) | 0.51 (0.40) | 0.34 (0.41) |
| Diuretics | 0.21 (0.08) | -0.57 (0.27) | -0.62 (0.26) | -0.48 (0.27) | -0.38 (0.26) | -0.56 (0.25) | -0.50 (0.24) |
| Statins | -0.05 (0.12) | -0.15 (0.35) | -0.17 (0.34) | -0.23 (0.35) | -0.12 (0.35) | -0.28 (0.34) | -0.43 (0.33) |
| Gastroprotective drugs | 0.73 (0.09) | 0.55 (0.33) | 0.49 (0.30) | 0.47 (0.32) | 0.94 (0.29) | 0.65 (0.31) | 0.97 (0.28) |
| *Phase 2 variables:* | | | | | | | |
| BMI $\geq$30kg/m$^2$ | — | 0.45(0.22) | 0.45 (0.21) | 0.32 (0.22) | 0.35 (0.23) | 0.41 (0.21) | 0.31 (0.21) |
| Current smoker | — | 0.85 (0.22) | 0.87 (0.22) | 0.81 (0.23) | 0.74 (0.23) | 0.72 (0.22) | 0.79 (0.21) |

[a] One control without information on BMI and one case and two controls without information on smoking are excluded from the analysis.

[b] Age is centred around 55 years.

SE: standard error, IA: interaction, NSAID: non-steroidal anti-inflammatory drug, ASA: acetylsalicylic acid, BMI: body mass index

**Table E.3:** ML-estimates of two-phase analyses with different a priori stratifications

| **Analysis:** | | **Phase 1** | **Two-phase** | | | |
|---|---|---|---|---|---|---|
| *Sample size:* | | N=26 208 | n=500 | n=2 000 | | |
| *Sampling scheme:* | | | E*-strat. | phen.×age×sex-strat. | | E*-strat. |
| *Stratification:* | | | E* | B | E | E* |
| *Number of simulations:* | | *1 000* | *1 000* | *967* | *364* | *1 000* |
| **Multivariable model** | True $\beta^a$ | $\overline{\beta}$ (SE)$^b$ | $\overline{\beta}$ (SE)$^b$ | $\overline{\beta}$ (SE)$^b$ | $\overline{\beta}$(SE)$^b$ | $\overline{\beta}$(SE)$^b$ |
| Phenprocoumon use | 1.37 | 1.20 (0.17) | 1.37 (0.27) | 1.33 (0.20) | 1.35 (0.20) | 1.36 (0.20) |
| Age$^c$ | 0.04 | 0.04 (<.01) | 0.04 (<.01) | 0.04 (<.01) | 0.04 (<.01) | 0.04 (<.01) |
| IA: phen.×age | -0.03 | -0.03 (0.01) | -0.03 (0.02) | -0.03 (0.01) | -0.03 (0.01) | -0.03 (0.01) |
| Female sex | -0.12 | -0.12 (0.05) | -0.12 (0.10) | -0.14 (0.07) | -0.12 (0.06) | -0.12 (0.06) |
| IA: phen.×sex | 0.29 | 0.28 (0.23) | 0.29 (0.28) | 0.30 (0.24) | 0.30 (0.23) | 0.29 (0.23) |
| *Comorbid conditions:* | | | | | | |
| Diabetes mellitus | 0.24 | 0.23 (0.08) | 0.23 (0.22) | 0.24 (0.15) | 0.24 (0.14) | 0.23 (0.12) |
| Hypertension | 0.12 | 0.11 (0.07) | 0.12 (0.21) | 0.10 (0.09) | 0.12 (0.12) | 0.12 (0.11) |
| Ischemic heart disease | 0.05 | 0.05 (0.10) | 0.04 (0.28) | 0.03 (0.18) | 0.07 (0.18) | 0.05 (0.15) |
| Liver disease | 0.27 | 0.26 (0.09) | 0.29 (0.27) | 0.27 (0.18) | 0.26 (0.18) | 0.28 (0.14) |
| GI disease | 0.32 | 0.31 (0.11) | 0.33 (0.31) | 0.32 (0.21) | 0.35 (0.21) | 0.32 (0.16) |
| *Current use of:* | | | | | | |
| NSAIDs | 0.37 | 0.36 (0.09) | 0.37 (0.26) | 0.38 (0.19) | 0.37 (0.18) | 0.37 (0.14) |
| ASA | 0.56 | 0.54 (0.12) | 0.56 (0.31) | 0.60 (0.26) | 0.55 (0.24) | 0.56 (0.16) |
| Diuretics | 0.21 | 0.21 (0.07) | 0.22 (0.21) | 0.22 (0.13) | 0.21 (0.13) | 0.21 (0.11) |
| Statins | -0.11 | -0.11 (0.11) | -0.11 (0.30) | -0.13 (0.19) | -0.12 (0.18) | -0.11 (0.15) |
| Gastroprotective drugs | 0.57 | 0.55 (0.09) | 0.56 (0.21) | 0.58 (0.17) | 0.59 (0.16) | 0.57 (0.12) |
| *Phase 2 variables:* | | | | | | |
| BMI ≥30kg/m$^2$ | 0.45 | — | 0.47 (0.22) | 0.46 (0.11) | 0.46 (0.11) | 0.46 (0.11) |
| IA: phen.×BMI | -0.39 | — | -0.39 (0.50) | -0.39 (0.27) | -0.39 (0.27) | -0.39 (0.27) |
| Current smoker | 0.83 | — | 0.87 (0.19) | 0.83 (0.10) | 0.84 (0.10) | 0.84 (0.10) |

[a] Parameter that is used for the simulation of the disease status.

[b] Parameter estimate and standard error (SE) are averaged over all simulations.

[c] Age is centred around 55 years.

# Bibliography


Aitchison, J. and Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics*, 29(3): 813–828.

Allison, P. D. (2005). "Imputation of categorical variables with PROC MI". *SUGI 30 Proceedings*.

Behr, S. and Schill, W. (2013). Stratification in two-phase database studies with a rich phase 1 data set. Unpublished manuscript.

Behr, S., Andersohn, F., and Garbe, E. (2010). Risk of intracerebral haemorrhage associated with phenprocoumon exposure: a nested case-control study in a large population-based German database. *Pharmacoepidemiology and Drug Safety*, 7(19): 722–730. DOI: 10.1002/pds.1973.

Behr, S., Schill, W., and Pigeot, I. (2012). Does additional confounder information alter the estimated risk of bleeding associated with phenprocoumon use - results of a two-phase study. *Pharmacoepidemiology and Drug Safety*, 21(5): 535–545. DOI: 10.1002/pds.3193.

Breslow, N. E. and Cain, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, 75(1): 11–20.

Breslow, N. E. and Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Applied Statistics*, 48(4): 457–468.

Breslow, N. E. and Holubkov, R. (1997a). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society B*, 59(2): 447–461.

Breslow, N. E. and Holubkov, R. (1997b). Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Statistics in Medicine*, 16(1): 103–116.

Breslow, N. E., McNeney, B., and Wellner, J. A. (2003). Large sample theory for semiparametric regression models with two-phase, outcome-dependent sampling. *The Annals of Statistics*, 31(4): 1110–1139.

Breslow, N. E., Amorim, G., Pettinger, M. B., and Rossouw, J. (2013). Using the whole cohort in the analysis of case-control data. *Statistics in Biosciences*. DOI: 10.1007/s12561-013-9080-2.

Charlson, M. E., Pompei, P., Ales, K. L., and MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Diseases*, 40: 373–383.

Collet, J., Schaubel, D., Hanley, J., Sharpe, C., and Boivin, J. (1998). Controlling confounding when studying large pharmacoepidemiologic databases: a case study of the two-stage sampling design. *Epidemiology*, 9(3): 309–315.

Deville, J. C., Särndal, C. E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88: 1013–1020.

Flanders, W. D. and Greenland, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine*, 10(5): 739–747.

Garbe, E., Kreisel, S., and Behr, S. (2013). Risk of subarachnoid hemorrhage and early case fatality associated with outpatient antithrombotic drug use. *Stroke*. DOI: 10.1161/STROKEAHA.111.000811.

Gill, R. D., Vardi, Y., and Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Annals of Statistics*, 16: 1069–1112.

Groenbaek, H., Johnsen, S., Jepsen, P., Gislum, M., Vilstrup, H., Tage-Jensen, U., and Soerensen, H. (2008). Liver cirrhosis, other liver diseases and risk of hospitalisation for intracranial haemorrhage: a Danish population-based case-control study. *BMC Gastroenterology*, 8: 16. DOI: 10.1186/1471-230Y-8-16.

Haneuse, S., Saegusa, T., and Lumley, T. (2011). osDesign: An R package for the analysis, evaluation, and design of two-phase and case-control studies. *Journal of Statistical Software*, 43(11): 1–29.

Hein, L. and Schwabe, U. (2007). Antikoagulantien und Thrombozytenaggregationshemmer. *Arzneiverordnungsreport*. Ed. by U. Schwabe and D. Paffrath. Heidelberg: Springer Medizin Verlag: 425–438.

Hernandez-Diaz, S. and Rodriguez, L. Garcia (2002). Incidence of serious upper gastrointestinal bleedings/perforation in the general population: Review of epidemiologic studies. *Journal of Clinical Epidemiology*, 55: 157–163.

Hirose, Y. and Lee, A. J. (2012). Reparametrization of the least favorable submodel in semi-parametric multisample models. *Bernoulli*, 18(2): 586–605.

Holubkov, R. (1995). "Maximum likelihood estimation in two-stage case-control studies". PhD thesis. University of Washington.

Horvitz, D. and Thompson, D. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of of the American Statistical Association*, 47(11): 663–685.

Hsieh, F. Y., Bloch, D. A., and Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine*, 17: 1623–1634.

Jobski, K., Behr, S., and Garbe, E. (2011). Drug interactions with phenprocoumon and the risk of serious haemorrhage: A nested case-control study in a large population-based German database. *European Journal of Clinical Pharmacology*, 67(9): 941–951. DOI: 10.1007/s00228-011-1031-6.

Johnsen, S., Pedersen, L., Friis, S., Blot, W., McLaughlin, J., Olsen, J., and Soerensen, H. (2003). Nonaspirin nonsteroidal anti-inflammatory drugs and risk of hospitalisation for intracerebral haemorrhage: a population-based case-control study. *Stroke*, 34(2): 387–391. DOI: 10.1161/01.STR.0000054057.11892.5B.

Korff, M. von, Wagner, E. H., and Saunders, K. (1992). A chronic disease score from automated pharmacy data. *Journal of Clinical Epidemiology*, 45: 197–203.

Kreuter, F., Mueller, G., and Trappmann, M. (2010). Nonresponse and measurement error in employment research - Making use of administrative data. *Public Opinion Quarterly*, 74(5): 880–906.

Lavori, P. W., Dawson, R., and Shera, D. (1995). A multiple imputation strategy for clinical trials with truncation of patient data. *Statistics in Medicine*, 14: 1913–1925.

Lee, A. J. (2007). On the semiparametric efficiency of the Scott-Wild estimator under choice-based and two-phase sampling. *Journal of Applied Mathematics and Decision Sciences*. Article ID 86180.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*. Second. New Jersey: John Wiley & Sons.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9: 1–19.

March, S., Rauch, A., Thomas, D., Bender, S., and Swart, E. (2012). Procedures according to data protection laws for coupling primary and secondary data in a cohort study: the lidA study. *Gesundheitswesen*. DOI: 10.1055/s-0031-1301276.

Martel, M., Rey, E., Malo, J., Perreault, S., Beauchesne, M., Forget, A., and Blais, L. (2009). Determinants of the incidence of childhood asthma: a two-stage case-control study. *American Journal of Epidemiology*, 169(2): 195–205.

Marti, H. and Chavance, M. (2011). Multiple imputation analysis of case-cohort studies. *Statistics in Medicine*, 30: 1595–1607.

Myers, J. A. and Louis, T. A. (2007). *Optimal propensity score stratification*. Working papers Working Paper 155. Johns Hopkins University, Dept. of Biostatistics.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33(201): 101–116.

Pigeot, I. and Ahrens, W. (2008). Establishment of a pharmacoepidemiological database in Germany: Methodological potential, scientific value and practical limitations. *Pharmacoepidemiology and Drug Safety*, 3(17): 215–223. DOI: 10.1002/pds.1545.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66(3): 403–411.

Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82(2): 299–314.

Reilly, M. and Pepe, M. (1997). The relationship between hot-deck multiple imputation and weighted likelihood. *Statistics in Medicine*, 16: 5–19.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89: 846–866.

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.

Rubin, D. B. (1978). "Multiple imputations in sample surveys". *Proceedings of the Survey Research Methods Section, American Statistical Association*: 20–34.

– (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley.

– (1996). Multiple imputation 18+ years. *Journal of the American Statistical Association*, 91(434): 473–489.

Rudolph, H. and Trappmann, M. (2007). Design und Stichprobe des Panels "Arbeitsmarkt und Soziale Sicherung" (PASS). *IAB Forschungsbericht*, 12/2007: 60–101.

*SAS/STAT® 9.2 User's Guide* (2008). SAS Institute Inc. Cary, NC.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1): 3–15.

Schaubel, D., Hanley, J., Collet, J., Boivin, J., Sharpe, C., Morrison, H. I., and Mao, Y. (1997). Two-stage sampling for etiologic studies. *American Journal of Epidemiology*, 146(5): 450–458.

Schill, W. and Drescher, K. (1997). Logistic analysis of studies with two-stage sampling: a comparison of four approaches. *Statistics in Medicine*, 16: 117–132.

Schill, W. and Wild, P. (2006). Minmax designs for planning the second phase in a two-phase case-control study. *Statistics in Medicine*, 25: 1646–1659.

Schill, W., Joeckel, K.-H., Drescher, K., and Timm, J. (1993). Logistic analysis in case-control studies under validation sampling. *Biometrika*, 80(2): 339–352.

Schill, W., Enders, D., and Drescher, K. (2013). sas-twophase-package: A SAS-package for logistic two-phase studies. *Journal of Statistical Software*. submitted.

Scott, A. J. and Wild, C. J. (1991). Fitting logistic regression models in stratified case-control studies. *Biometrics*, 47: 497–510.

– (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84(1): 57–71.

– (2001). Maximum likelihood for generalised case-control studies. *Journal of Statistical Planning and Inference*, 96(1): 3–27.

– (2006). Calculating efficient semiparametric estimators for a broad class of missing-data problems. *Festschrift for Tarmo Pukkila on his 60th birthday*. Ed. by E. P. Liski, J. Isotalo, J. Niemelae, S. Putanen, and G. P. H. Styan. University of Tampere: 301–314.

– (2011). Fitting regression models with response-biased samples. *Canadian Journal of Statistics*, 39: 519–536.

Scott, A. J., Lee, A. J., and Wild, C. J. (2007). On the Breslow-Holubkov estimator. *Lifetime Data Analysis*, 13: 545–563.

Sharpe, C., Collet, J., McNutt, M., Boivin, J., and Hanley, J. (2000). Nested case-control study of the effects of non-steroidal anti-inflammatory drugs on breast cancer risk and stage. *British Journal of Cancer*, 83(1): 112–120.

Vaart, A. van der and Wellner, J. A. (2001). Consistency of semiparametric maximum likelihood estimators for two-phase sampling. *The Canadian Journal of Statistics*, 29(2): 269–288.

Walker, A. M. (1982). Anamorphic analysis: Sampling and estimation for covariate effects when both exposure and disease are known. *Biometrics*, 38(4): 1025–1032.

White, J. E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, 115(1): 119–128.

Whittemore, A. S. (1981). Sample size for logistic regression with small response probability. *Journal of the American Statistical Association*, 76(373): 27–32.

Wild, C. J. and Jiang, Y. *Description of the "missreg" library*. Accessed on 2013-01-12. URL: `http://www.stat.auckland.ac.nz/~wild/software.html`.

# CURRICULUM VITAE


## Sigrid Behr

Leibniz Institute for Prevention Research and Epidemiology - BIPS
Department of Biometry and Data Manangement
Achterstr. 30, 28309 Bremen, Germany
Phone: +49 421 218-56951
Email: behr@bips.uni-bremen.de


## EDUCATION

| | |
|---|---|
| 2007-present | Ph.D. candidate in Statistics, Bremen University |
| 1998-2004 | Diploma in mathematics, RWTH Aachen |

## PROFESSIONAL POSITIONS

| | |
|---|---|
| 2007-present | Research associate at BIPS |
| 2004-2007 | Statistician at Bayer Healthcare, Wuppertal, Germany |

## PUBLICATIONS

### Peer-reviewed publications


**2013** E. Garbe, S. Kreisel, and S. Behr. Risk of subarachnoid hemorrhage and early case fatality associated with outpatient antithrombotic drug use. *Stroke*, 2013. doi: 10.1161/STROKEAHA.111.000811

**2012** S. Behr, W. Schill, and I. Pigeot. Does additional confounder information alter the estimated risk of bleeding associated with phenprocoumon use - results of a two-phase study. *Pharmacoepidemiology and Drug Safety*, 21(5):535–545, 2012. doi: 10.1002/pds.3193

| | |
|---|---|
| | K. Jobski, U. Schmid, S. Behr, F. Andersohn, and E. Garbe. 3-year prevalence of alcohol-related disorders in German patients treated with high-potency opioids. *Pharmacoepidemiology and Drug Safety*, 21(10):1125 – 1129, 2012. doi: 10.1002/pds.3268 |
| **2011** | K. Jobski, S. Behr, and E. Garbe. Drug interactions with phenprocoumon and the risk of serious haemorrhage: A nested case-control study in a large population-based german database. *European Journal of Clinical Pharmacology*, 67(9):941–951, 2011. doi: 10.1007/s00228-011-1031-6 |
| **2010** | S. Behr, F. Andersohn, and E. Garbe. Risk of intracerebral haemorrhage associated with phenprocoumon exposure: a nested case-control study in a large population-based german database. *Pharmacoepidemiology and Drug Safety*, 7(19):722–730, 2010. doi: 10.1002/pds.1973 |

## Other publications

| | |
|---|---|
| **2011** | R. Mikolajczyk, S. Behr, and E. Garbe. Re: "Auswirkungen leitlinienkonformer Therapie auf das berleben von Patientinnen mit primrem Mammakarzinom - Ergebnisse einer retrospektiven Kohortenstudie." by R. Wolters, A. Wockel, M. Wischnewsky and R. Kreienberg (Z Evid Fortbild Qual Gesundhwes 2011;105(6):468-475). *Zeitschrift fr Evidenz, Fortbildung und Qualitt im Gesundheitswesen*, 105(6): 468–475, 2011. doi: doi:10.1016/j.zefq.2011.10.010. Letter to the editor |
| **2008** | S. Behr. New drug development: Design, methodology, and analysis by J. R. Turner. *Biometrics*, 64(1):313–314, 2008. doi: 10.1111/j.1541-0420.2008.00962_6.x. Book review |