Application of Modern Statistical Methods in Worldwide Health Insurance

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

eingereicht von
Andreas Bayerstadler

München, 2017

1. Gutachter: Prof. Dr. Christian Heumann
2. Gutachter: Prof. Dr. Matthias Schmid
Tag der Einreichung: 31. August 2017
Tag der mündlichen Prüfung: 13. Dezember 2017

Acknowledgements

I would like to cordially thank all persons who have supported me in the authoring of this dissertation, especially

- Prof. Dr. Christian Heumann for the highly dedicated and patient supervision of my dissertation throughout all the years, the countless consultation meetings and his advice in statistical research questions.
- Dr. Fabian Winter and Prof. Dr. Franz Benstetter for their continuous professional and personal support of my research activities, their valuable input in insurance and economic questions and their ongoing motivation to finalize the dissertation in parallel to a challenging job.
- Prof. Dr. Thomas Augustin, Prof. Dr. Matthias Schmid and Prof. Dr. Helmut Küchenhoff for supporting my research projects and taking over a role in the doctoral committee.
- Linda van Dijk and Dr. Stefan Pilz for their input to the chapter on fraud and abuse detection and proof-reading of my manuscripts.
- Dr. Ingrid Eberlein and Dr. Suzan Kozak for their input to the chapter on candidate selection for disease management programs and proof-reading of my manuscripts.
- Dr. Stefan Kottmair, Dr. Özer Bebek and Dr. Nihat Canizci for their input on innumerable medical questions.
- all insurance companies and related contact persons who have provided their data and knowledge as indispensable foundation of the statistical analyses in this dissertation.
- Lena and my parents for their unlimited understanding and inestimable moral support throughout many tough years.

Statutory Declaration

(according to §8, Section 2, Item 5 of the doctorate regulation from 12 July 2011)

Here I explain on oath instead of that I made the available work independently and without illicit use of others.

Bayerstadler, Andreas

Munich, January 17, 2018

Published Parts and Contribution of Co-Authors

Parts of this dissertation have already been published:

- Section 2 refers to the article "A predictive modeling approach to increasing the economic effectiveness of disease management programs" published in Health Care Management Science in 2013 (see Bayerstadler et al. [2014]).
- Section 4 refers to the article "Bayesian multinomial latent variable modeling for fraud and abuse detection in health insurance" published in Insurance: Mathematics and Economics in 2016 (see Bayerstadler et al. [2016]).

The contents of these sections essentially concur with the corresponding publications. In both cases, the co-authors have contributed to the articles by proof-reading the manuscripts and supporting in non-statistical questions, especially placing the statistical results in an economic context.

All academic work necessary to answer the underlying research questions, including setup of analysis plans, literature research, development of (new) statistical methods and algorithms, analysis of data and interpretation of statistical results, has been carried out by myself.

Kurzfassung

Mit der wachsenden Verfügbarkeit von internen und externen Daten in der (Kranken-)Versicherungsindustrie steigt die Nachfrage nach neuen Erkenntnissen gewonnen aus analytischen Verfahren. In dieser Dissertation werden vier Anwendungsbeispiele für komplexe regressionsbasierte Vorhersagetechniken im Schaden- und Netzwerkmanagement von Krankenversicherungen präsentiert: Patientensegmentierung für und ökonomische Auswertung von Gesundheitsprogrammen, Betrugs- und Missbrauchserkennung und Messung medizinischer Behandlungsqualität. Basierend auf verschiedenen Krankenversicherungsdatensätzen wird gezeigt, dass maßgeschneiderte Modelle und neu entwickelte Algorithmen, wie bayesianische latente Variablenmodelle, die Geschäftsteuerung von Krankenversicherern optimieren können. Durch das Einbringen und Strukturieren von medizinischem und versicherungstechnischem Wissen können diese maßgeschneiderten Regressionsansätze mit Methoden aus dem maschinellen Lernen und der künstlichen Intelligenz zumindest mithalten. Gleichzeitig bieten diese Ansätze dem Businessanwender ein höheres Maß an Transparenz und Interpretierbarkeit. In allen vier Beispielen werden Methodik und Ergebnisse der angewandten Verfahren ausführlich aus einer akademischen Perspektive diskutiert. Verschiedene Vergleiche mit analytischen und marktüblichen Best-Practice-Methoden erlauben es, den Mehrwert der angewendeten Ansätze auch aus einer ökonomischen Perspektive zu bewerten.

Abstract

With the increasing availability of internal and external data in the (health) insurance industry, the demand for new data insights from analytical methods is growing. This dissertation presents four examples of the application of advanced regression-based prediction techniques for claims and network management in health insurance: patient segmentation for and economic evaluation of disease management programs, fraud and abuse detection and medical quality assessment. Based on different health insurance datasets, it is shown that tailored models and newly developed algorithms, like Bayesian latent variable models, can optimize the business steering of health insurance companies. By incorporating and structuring medical and insurance knowledge these tailored regression approaches can at least compete with machine learning and artificial intelligence methods while being more transparent and interpretable for the business users. In all four examples, methodology and outcomes of the applied approaches are discussed extensively from an academic perspective. Various comparisons to analytical and market best practice methods allow to also judge the added value of the applied approaches from an economic perspective.

Contents

1 The Growing Importance of Statistics in Health Insurance

Which Factors cause the Growing Importance of Statistical Methods?

Basic statistical theory, like linear regression models [Legendre, 1805; Gauss, 1809] or the Bayes theorem [Bayes, 1763], goes back until the late 18th century. Even more advanced techniques, like generalized linear models [McCullagh and Nelder, 1989] have already been introduced in the 1970ies. However, the insurance industry, especially health insurers, only started to use these methods in the last 5 to 15 years. Many health insurers, especially in less developed markets, do not even nowadays base their business steering on data and statistical evaluations. Reasons for the hesitant approach of this field are – among others – regulatory restrictions, a lack of captured data, a lack of statistical capabilities and a lack of belief in the economic impact of analytical findings.

So, what has changed in the last decades that more and more insurance companies address the topic, establish dedicated "Analytics" teams and rely on data-based decisions? Figure 1 summarizes some of the most important promoting factors and developments which lead to this change of directions.

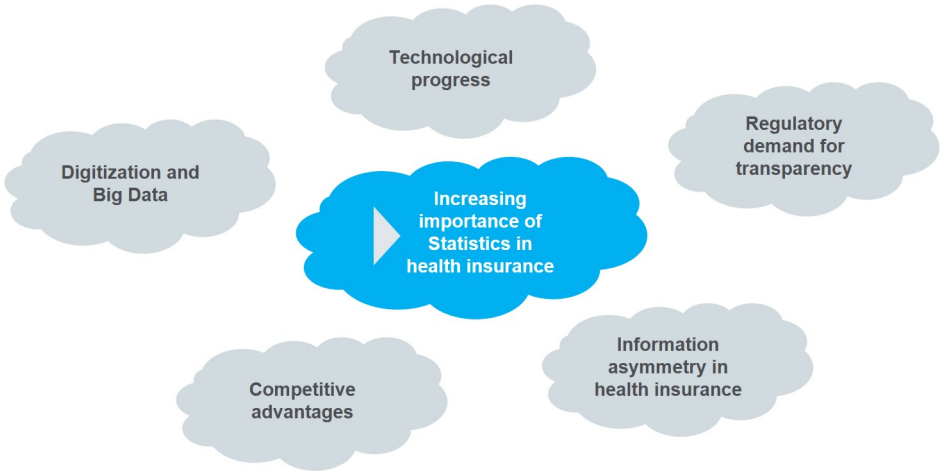

Figure 1: Factors fostering the increasing importance of statistical methods in health insurance.

A very important factor promoting the creation of statistical insights is the exploding *technological progress* over the last two decades. This progress has caused a decrease of prices for data storage, and thereby, an increase in data capturing and data quality. Also, the immense increase in available computing power accelerates the generation of statistical outcomes. While the calculation of simple regression models needed hours and days before the year 2000, nowadays billions of lines can be processed in a few seconds which allows real-time decision making and data-based process steering.

Of course, also the growing *digitization* and the hype around *"Big Data"* strengthen the need for statistical methods. These developments lead to the fact that health insurance datasets grow in two directions. First, insurance datasets get longer as data are captured in a more granular way. Earlier, data were often only captured on aggregated level, e.g. the number of claims in a certain age-band. Today, every claim/invoice and even every invoice item is stored, ideally together with detailed medical information. Due to this more granular data capturing health insurance data can have hundreds millions of lines. With the growing availability of external data, datasets also get wider, i.e. more (co)variates are available for statistical analyses. For example, data from social media, health apps and trackers ("wearables") lead to an enormous amount of additional information. Statistics has the important function to differentiate between relevant and irrelevant information in this vast amount of data. Also, the combination of internal and external data requires statistical know-how. For instance, if a one-to-one matching between internal and external data is not possible from regulatory reasons statistical matching techniques are needed.

Also, regulators in many different countries encourage the application of statistical methods to increase the *transparency* of the health (insurance) system. Important applications in this regard are the data-based detection of fraud, waste and abuse as well as the monitoring of medical outcome quality enabling pay-for-performance systems. Also, the decisions of insurers in risk assessment, for example in medical underwriting, need to be transparent and comprehensible. In this context, the combination of medical expert systems and data-based medical underwriting is important to increase the insurability of diseases and avoid discrimination of applicants.

The opportunity of *competitive advantages* is another important argument to introduce statistical methods in health insurance. In developed markets, there often exists a high cost pressure which causes the need to leverage saving potentials. As outlined in this thesis, statistical methods can be used to identify both medical and operational saving potentials. In emerging markets, the focus of health insurers is on fast and sustainable growth. Also here, statistical applications, for example, intelligent cross- and up-selling models, allow insurance companies to create competitive advantages.

Finally, the *information asymmetry* present in every health insurance system is an immanent argument for health insurers to create more data-based insights. Considering the information triangle displayed in Figure 2, medical providers are in the strongest position as they usually know most about the health status of their patients. Therefore, they can create so called supplier-induced demand [LaBelle et al., 1994]), i.e. create revenues for unnecessary medical treatments. Insured persons may also exploit the system in the knowledge that health insurers cover their treatment (so-called moral hazard [Dembe and Boden, 2000]). Statistical insights can help the insurer to make up the information edge of the other players in the healthcare market and avoid increasing costs.

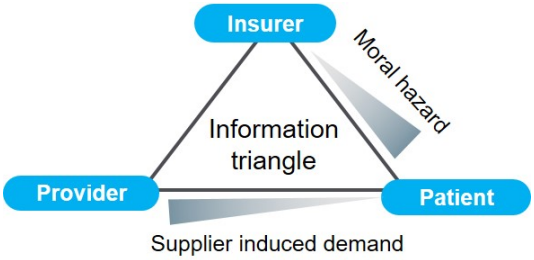

Figure 2: Information asymmetry in health insurance.

What are the Major Changes compared to Traditional Insurance Reporting?

The main differences in analytical business steering based on modern statistical approaches compared to classical (health) insurance reporting are the following:

- "From past to future": Modern statistical approaches, also called "Predictive Models", are not only explaining the past, but also identify systematic patterns which allow a prediction of future developments (e.g. prediction of burning costs).
- "From descriptive to inductive/causal": The goal of applying statistical techniques, especially modern regression techniques, is to use all available and relevant information contained in the data to not only receive an outcome, but also to be able to explain the occurrence of the outcome (e.g. occurrence of a premium loading for a specific disease in medical underwriting).
- "From portfolio to individual": The incorporation of more (granular) data information allows to not only analyze portfolio trends or trends in certain subgroups of the portfolio, but also to make statements on individuals (e.g. up-selling probability, likelihood of hospitalization).

The creation of these additional insights allows a pro-active, instead of a re-active, steering of health insurance.

What are the Prerequisites to create Statistical Insights in Health Insurance?

Of course, the creation of the described insights is also related to efforts and investments health insurers need to take. Figure 3 summarizes the most important preconditions for efficiently creating quality assured statistical outcomes in health insurance.

An indispensable asset for health insurers are their data. Focus of health insurers should be to first exploit the full potential of their internal data before including additional external data which can increase the efficiency of the analysis. Even for small portfolios reliable statistical analyses are possible, as long as data quality is

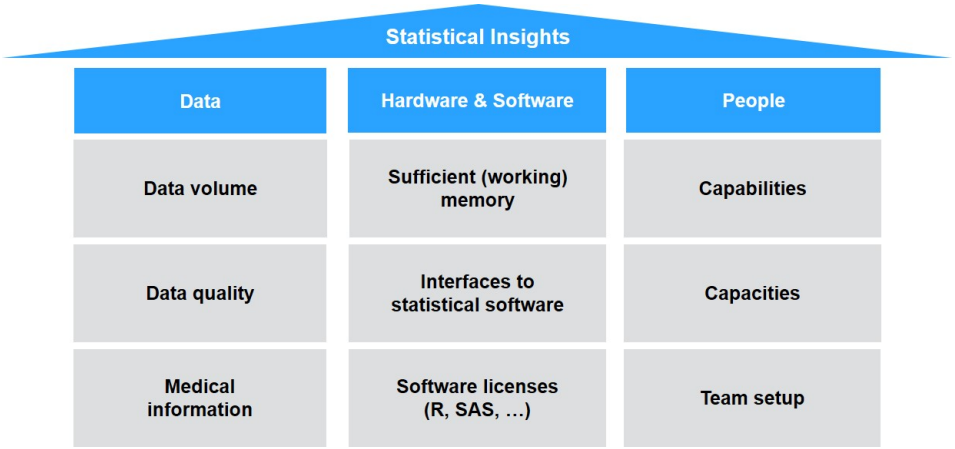

Figure 3: Prerequisites for creating statistical insights in health insurance.

assured and policy and claims data are captured on granular level. Another important prerequisite for valuable statistical insights beside data completeness and consistency is the capturing of detailed medical information. Ideally both diagnoses and medical procedures are coded in procedure coding systems to establish comparability of medical treatments. The usage of international standard coding systems, like ICD coding for diagnoses, further increases the transferability of statistical solutions between markets.

As many modern statistical approaches rely on computationally intensive algorithms, powerful hardware is necessary to process also large data in adequate time. This especially holds for applications in which time-critical decisions are based on statistical outcomes, e.g. real-time scoring of claims. In this context, in-memory technology plays an increasing role which allows to load huge amounts of data in the working memory. Of course, statistical software, like R or SAS, is also needed to implement standard approaches and new algorithms. Important in this regard is a structured data storage in a data warehouse, which allows quick and easy access of the data and acts as "single source of truth".

Last but not least, a core requirement to create value out of Statistics are capable analysts – nowadays called "data scientists" – who have enough time to work on advanced statistical problems. Best results can be achieved if the statistical experts work in multidisciplinary teams with subject-matter experts, like medical doctors, actuaries or underwriters. Especially, for smaller (health) insurance companies the recruiting of such highly sought-after resources can be a challenge, because they compete with nearly all other industries. From an organizational perspective, it is meaningful to build central Analytics units who serve all departments of the insurance company instead of having separate resources for each department.

Where can Statistical Methods be applied in Health Insurance?

As illustrated by Figure 4, statistical methods can be applied along the whole value chain of health insurance.

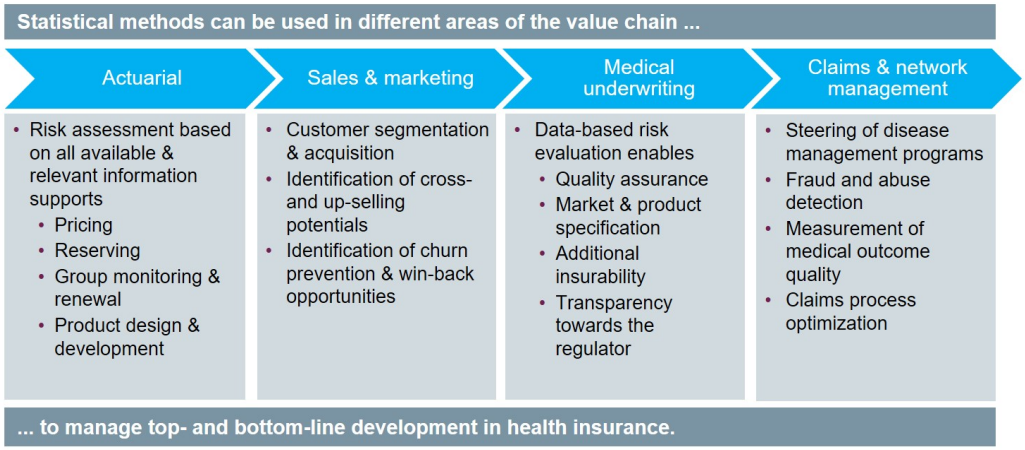

Figure 4: Fields of the health insurance value chain in which statistical methods provide value.

In the *actuarial* sector, statistical methods can be used to increase the risk understanding and quantify the influence of different parameters on claims patterns. For example, insights from GLMs or Random Forests support actuarial decisions in pricing and reserving as well as group monitoring and underwriting. Besides, statistical cost driver analyses help to identify gaps in the existing product landscape and design tailored new products.

In *sales and marketing*, internal and external data are combined to identify new sales potentials and increase the responsiveness of campaigns by addressing the "right" customers. Segmentation and selection models support the customer relationship management from acquisition over up- and cross-selling offerings until churn prevention and win-back contacts. Applied to a comprehensive database, even comparably simple models can already increase the efficiency of sales and marketing activities.

Individual *medical underwriting* is strongly dominated by medical expert systems translated into health questionnaires based on which a risk assessment of new applicants is performed. Statistical models can complement these systems by increasing the risk understanding of prevalent medical conditions and their interactions. In this way, they contribute to a market- and product-specification of medical underwriting decisions.

In the *claims* sector, statistical models based on medical input help to buffer adverse cost trends arising from different reasons. A huge problem in many markets is fraudulent and abusive behavior which leads to a strong increase in healthcare costs. At the same time, poor medical quality and insufficient treatment cause immense costs due to the occurrence of complications and related follow-up costs. Data-based

provider profiling models help to pinpoint such deficits and increase the transparency in health care. Based on statistical insights, health insurance companies can define targeted countermeasures, e.g. a performance-oriented network management. Also, the increasing (mainly lifestyle-driven) incidence of chronic diseases is a severe problem, especially in developed markets. Here, targeted disease management programs are an important preventive measure to avoid uncontrolled courses of a disease and costs. Statistical methods enable an efficient steering of these programs by targeted candidate selection and monitoring of the economic efficiency.

What is the Goal of this Thesis?

The goal of this thesis is to demonstrate the value of bespoke statistical methods and new algorithms using four practical examples from different health insurance operations. The academic research presented in this thesis mainly focuses on the question if statistical techniques which are tailored to make use of insurance-related content knowledge provide an additional value compared to standard approaches. The given examples are focusing on risk-related claims applications, because here the interplay of data availability, medical and statistical knowledge contains a high economic potential which is still relatively untapped in the health insurance industry (see Figure 5).

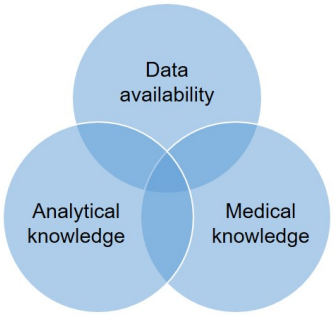

Figure 5: The three core elements of applying statistical methods in claims management.

Section 2 compares various statistical techniques for individual loss prediction in health insurance as foundation for candidate selection in disease management. Also, regression techniques are benchmarked with different machine learning approaches. Based on the outcomes of this analysis, Section 3 presents a comprehensive matched-pair approach for a reliable measurement of the economic impact of disease management programs. Statistical challenges here are to form appropriate control groups based on all cost-relevant information and to adequately account for the uncertainty of the measurement. Sections 4 and 5 introduce new Bayesian algorithms for fraud and abuse detection and medical quality assessment. In both cases, latent variables are used to structure medical and insurance knowledge and ensure – in combination with shrinkage techniques – that all relevant information is used to optimally predict the corresponding target variables.

2 Candidate Selection for Disease Management Programs

(refers to the publication Bayerstadler et al. [2014])

Section 2 illustrates a predictive modeling (PM) approach that enables the economic potential of disease management programs (DMPs) to be fully exploited by optimized candidate selection. The approach is based on a generalized linear model (GLM) that is easy to apply for health insurance companies. By means of a small portfolio from an emerging country, it is shown that the presented GLM approach is stable compared to more sophisticated regression techniques in spite of the difficult data environment. Additionally, it is demonstrated for this example of a setting that the model can compete with the expensive solutions offered by professional PM vendors and outperforms non-predictive standard approaches for DMP selection commonly used in the market.

2.1 Introduction

Chronic diseases are a major driver of rapidly rising health care costs both in developed and in emerging countries. For instance, in the U.S. the care of chronic illness consumes approximately 75% of total healthcare expenditures which makes over 1 trillion U.S. dollars per year [Freeman et al., 2011]. Equally, chronic diseases were responsible for 50% of the disease burden in 23 developing countries in 2005 and will cost those countries 84 billion U.S. dollars by 2015 if nothing is done to slow their growth [Nugent, 2008]. Disease management programs (DMPs) are usually offered to chronic patients by public health initiatives or by specific vendors on behalf of insurance companies. They are meant to improve the medical situation and the quality of life of program participants as well as to reduce expenditure on benefits in the long run. Common measures are regular phone calls to increase compliance with medical plans and disease-specific consultancy ("tele-coaching"), and online monitoring of disease-specific parameters, such as blood pressure, weight or insulin level ("tele-monitoring"). The economic effect of disease management programs for chronic patients is the subject of controversial debate in scientific literature. This chapter describes a data-driven approach that enhances the economic benefit of disease management programs through optimal selection of participants assuming that the related program measures lead to a decrease of disease-related costs. The question of how to measure and quantify this economic effect will be addressed in Section 3.

In most cases it is neither possible nor economically viable to include all patients with certain characteristics in a DMP. In order to exploit the full economic potential of a DMP, it is necessary to choose the patients with the highest possible future savings. It is assumed that the individual saving potential is mainly triggered by two characteristics:

a) the risk of the chronic disease moving on to an uncontrolled state in the near future which usually results in expensive emergency treatment, hospitalizations and the development of co-morbidities,

b) the patient's compliance with the program measures and their ability to manage the chronic disease by themselves.

Assumption a) is based on the finding of different researchers that disease management programs are most cost-effective if the focus is on severely ill patients with continued high utilization and co-morbidities [Freeman et al., 2011; Meyer and Smith, 2008]. Assumption b) is supported by different publications that prove the cost-effectiveness of measures that enhance the self-management abilities of chronic patients [Bodenheimer et al., 2002a; Lorig et al., 2001].

For an insurer, factor b) is hardly assessable in advance. Consequently, the focus is on trigger point a) and it is further assumed that there is a high correlation between the individual saving potential and the future costs of the patient. This implies that optimal up-front identification of high-risk/high-cost patients (according to a)) and inclusion of these patients in the DMP will maximize the medical savings for the insurer. This argumentation is in line with the results of several researchers, such as Billings and Mijanovich [2007] or Meyer and Smith [2008]. Meyer and Smith [2008] have analyzed a large number of peer-reviewed studies on clinical and economic outcomes of DMPs in the U.S. and have identified key factors for the cost-effectiveness of a DMP on this basis. As one of their major results they found that cost-effective DMPs target the intervention to sicker patients who are likely to generate high costs in the future. Billings and Mijanovich [2007] clearly stress the importance of an ex-ante identification of those patients.

Hence, a method is needed that reliably predicts the future medical costs of chronic patients. In the context of insurer-driven DMPs this prediction is usually based on the clients' claims history and policy data. Modern statistical prediction methods, frequently summarized by the term "predictive modeling" (PM), promise to outperform the standard approaches used in the health insurance market. In order to verify this claim and to arrive at an optimal solution for the DMP selection problem,

a) scientific literature has been studied extensively with regard to different PM approaches for the prediction of claims costs at an individual level (see Section 2.2.1).

b) an appropriate regression approach to predict annual claimed amounts at an individual level has been developed (see Section 2.3).

c) the approach has been compared to other (more sophisticated) regression approaches mentioned in literature (see Section 2.3.4).

d) three professional vendors of PM solutions with different methodological backgrounds were asked to participate in a forecasting competition and their solutions were compared to the described regression approach and to two standard selection methods frequently used in the market (see Section 2.4.1).

PM methods for cost prediction are acknowledged as state-of-the-art technology in

the insurance industry for large portfolios and high data quality. By contrast, the presented analyses are based upon a small dataset from a rapidly growing portfolio in order to assess whether PM techniques also yield meaningful prediction results in a difficult data environment (see Section 2.2.2).

Goal of this chapter is to present a model for DMP candidate selection which is

a) easy to implement for health insurers using standard statistics software,

b) stable in spite of the difficult data environment,

c) not prone to overfitting,

d) superior to standard (non-predictive) selection methods commonly used in the market,

e) equivalent to expensive solutions of professional vendors.

The comparison of different cost prediction techniques is based on various measures of predictive quality (see Appendix A) in order to highlight different aspects of the DMP selection problem.

In order to show that the model is applicable for different chronic indications as well as for general health programs, the analyses are based on a full health insurance portfolio without pre-selection of specific chronic patients. However, some subgroup analyses of chronic patients with different indications are presented that illustrate the benefit of the approach with regard to indication-specific DMPs. In Section 2.4.2 the extent to which results can be generalized is discussed and the key factors for an optimal prediction in the context of insurance data are illustrated. In addition, the outcome of another vendor test which has been carried out based on a pre-selected group of chronic patients is summarized.

2.2 Background

This chapter outlines the growing importance of PM methods for the (health) insurance industry and gives an overview of related literature with a special focus on the underlying cost prediction problem (see Section 2.2.1). Furthermore, the test dataset used for the comparison of methods is described and some of its characteristic properties are illustrated (see Section 2.2.2).

2.2.1 Literature on Relevant Predictive Modeling Approaches

Statistical regression models are a well-known and frequently used tool in order to analyze the influence of multiple covariates on an outcome variable and to predict future realizations of this variable. An important advantage of these models is the incorporation of the variability in the data which permits the testing of hypotheses as well as the construction of confidence and prediction intervals. Such models are also increasingly applied in the insurance industry [Haberman and Renshaw, 1998]. Especially in the U.S. market, predictive modeling techniques are used for

risk evaluation/pricing, the management of sales activities and medical management, usually based on large high-quality datasets with long member history. There exists hardly any scientific literature on the application of PM techniques to small health insurance portfolios under difficult conditions. This chapter demonstrates the benefit of classical PM techniques applied in such a situation. Goal of the model is the prediction of annual claimed amounts per insured person for an optimal DMP candidate selection. Of course, the same model can also be used for other purposes like actuarial risk evaluation.

A challenge in predicting annual claimed amounts arises from the characteristic distributional form of this attribute. Simple linear regression models assume that the target variable (asymptotically) follows a normal distribution given all covariates included in the model. The distribution of annual medical costs, however, has several properties that conflict with this normality assumption (see the distribution of individual claimed amounts in the test dataset, left plot of Figure 6):

- The range of possible values of annual claimed amounts is restricted to non-negative numbers.
- A large proportion of annual claimed amounts is equal to zero.
- The distribution of annual claimed amounts is highly right-skewed and non-symmetric.
- The right tail of the distribution is very long.

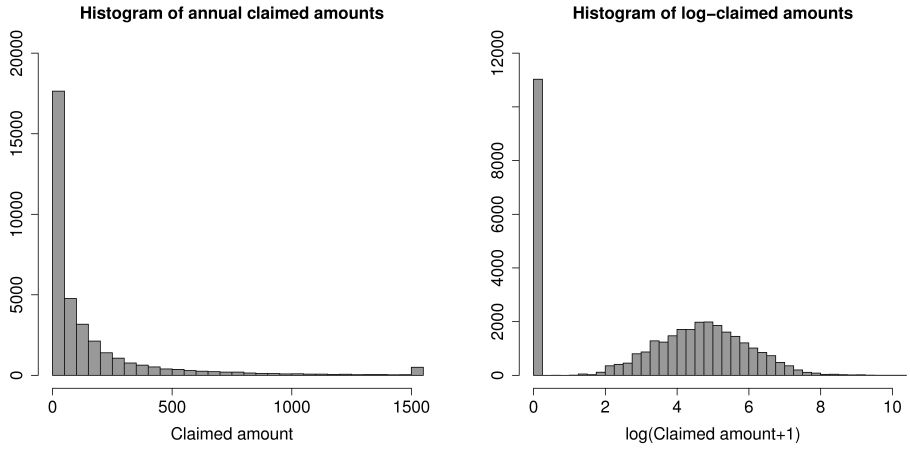

Figure 6: Histograms showing the distribution of individual annual claimed amounts (left, truncated at 1,500 units of the national currency) and log-transformed annual claimed amounts (right) in the test dataset (observation period 1995–2008).

A large number of different approaches have been proposed to account for these properties of claims data. Diehr et al. [1999], Buntin and Zaslavsky [2004] and Powers et al. [2005] give an an overview of methods proposed in scientific literature. A common approach to dealing with the distributional form of claims data is to normalize the target variable y by an appropriate transformation and to subsequently apply linear models. Especially the log-transformation $\log(y + 1)$ (log denotes the

natural logarithm to the base e) is frequently used [Duan et al., 1983; Manning, 1998; Manning and Mullahy, 2001]. Veazie et al. [2003] recommend using the square root of y to account for the non-normality of the response variable.

Even though the $\log(y+1)$-transformation widely normalizes the distribution of annual claimed amounts, a large proportion of zeros in the target variable can still disturb the assumption of normality (see the individual log-transformed annual claimed amounts in the test dataset, right plot of Figure 6). In order to deal with this challenge, different authors apply two-stage models based on the law of total expectation [Duan et al., 1983; Mullahy, 1998; Blough et al., 1999; Powers et al., 2005]. These models separately predict

- the probability of having at least one health claim and
- the expected amount of health expenses given the fact that the member has at least one claim.

Duan et al. [1983] extend this approach to a four-stage model differentiating between inpatient and outpatient cases.

Generalized linear models (GLMs) [McCullagh and Nelder, 1989] are a more general regression approach to explain and predict non-normally distributed outcome variables that frequently occur in claims data (see Section 2.3.2 for more details on the structure of GLMs). De Jong and Heller [2008] give a broad overview of possible applications of GLMs to insurance data in general. Frees and Valdez [2008] present a hierarchical three-stage approach using different GLMs to reliably predict automobile claims. Blough et al. [1999] and Buntin and Zaslavsky [2004] apply GLMs in the health insurance context to predict annual medical expenses. They also consider the more flexible quasi-likelihood approach based on Wedderburn [1974] which relaxes the distributional assumptions of GLMs.

An important assumption of GLMs is the independence of observations, which no longer holds if claims data are observed over a longer period of time. More precisely, there may exist a serial correlation between the annual claimed amounts of the same member. To account for this correlation, statistical theory basically offers two approaches. The marginal or GEE approach (generalized estimating equation) is based on the additional specification of a so-called working covariance that reflects the intra-individual correlation structure [Liang and Zeger, 1986]. The idea of the conditional or GLMM approach (generalized linear mixed model) is to incorporate individual-specific random effect parameters in the linear predictor of a GLM [McCulloch and Searle, 2001]. Frees et al. [1999] and Antonio and Beirlant [2007] have successfully applied the GLMM methodology to insurance/claims data. Yau et al. [2002] propose a two-stage approach using, first, a binomial GLMM to estimate the probability of observing an annual claimed amount greater than zero. A gamma GLMM is then applied to estimate the expected annual claimed amount given that it is positive.

Newhouse et al. [1989] focus on the right-hand side of the regression equation and identify health measures (based on diagnoses) and other attributes having a high predictive value for utilization data. Similarly, Lamers [1999] describes indicators for

chronic conditions based on pharmacy claims and derives "pharmacy cost groups" that can be used as covariates in a predictive model for medical expenses.

Beside regression models, further statistical techniques, like Neural Networks [Ripley, 1996] and Random Forests [Breiman, 2001], are increasingly used by insurance companies. These methods can be applied in many different fields of insurance business, such as cost prediction, fraud detection, treatment management and customer relationship management. Some publications have assessed the benefit of these approaches for insurance companies [Viaene et al., 2002; Francis, 2001, 2003]. The main reasons for the increasing attractiveness of these approaches is that they overcome some well-known shortcomings of traditional methods, like the automated detection of interactions between covariates [Kolyshkina et al., 2004]. However, a crucial drawback of these techniques is that, unlike regression models, they remain a black box to the user. Moreover, such methods can tend to overfitting if they are applied to small datasets [Freitag, 2002]. Additionally, regression techniques are easy to implement in all kinds of standard statistics software, also by less experienced users. For these reasons, the focus of the following methodology section (see Section 2.3) is to develop an optimal solution for the DMP selection problem. The chosen regression model is compared to various other, partly not regression-based prediction techniques in the results section (see Section 2.4).

2.2.2 Overview of the Data Environment

The dataset used for model development and comparative analysis was provided by a private insurance company in an emerging country. Due to electronic invoicing, the data quality is excellent. The prediction of medical expenses is exacerbated by a considerably small, but rapidly growing portfolio and by the high and unstable medical inflation rates that are typical for an emerging market (see Figure 7).

The original dataset consists of insured members who were enrolled between 1 January 1995 and 31 December 2009. Out of this group, all members who were enrolled on 1 January 2009 or later and all members who were no longer under risk on 31 December 2009 were excluded. As the provision of the full dataset to any external vendor was not possible, a random sample of the remaining members was drawn. All analyses presented in this chapter are based on this sample. The random sample was generated by allocating uniformly distributed random numbers between 0 and 1 to all members (using SPSS random number generator) and excluding all members with a value greater than 0.4. In this way, 39.76% of all members in the portfolio have been selected into the random sample. In order to check whether random sampling was successful, several tests comparing the distribution of important parameters (e.g. medical costs, age, diagnosis groups, ...) between the random sample and the full dataset were performed. The sample comprises 9,150 members who have submitted 423,162 claims in the observation period (1995–2008). The development of the model and the vendor test are exclusively based on claims and policy data between 1995 and 2008 for the members in the random sample. The claims data for the year 2009 were held back as a test sample for final evaluation of results.

Figure 7 shows the average annual claimed amounts per member in the observation period (1995–2008, solid line) and in the prediction year 2009 (dotted line) for all members within the random sample. The alternating gradient of the graph shows the unstable development of annual medical inflation, which was 10.31% on average between 1995 and 2009. In total, the average medical costs per member almost quadrupled between 1995 and 2009 (up by 394.86%). Figures need to be handled with care until the year 2000 as they are based on less than 1,000 insured members. Table 1 illustrates the strong increase in members under risk within the observation period. Accordingly, there are many patients with a short data history, which exacerbates the prediction. Only 20% of the patients were under risk for more than 5 years and only 40% for more than 2 years.

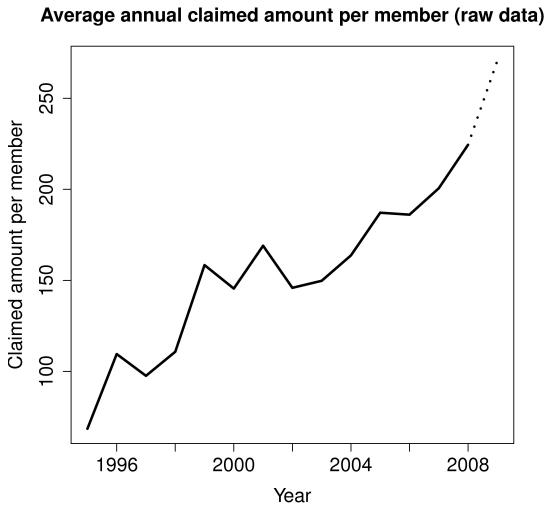

Figure 7: Average annual claimed amounts per member in units of the national currency (1995–2009).

Year	Members under risk
1996	63
1997	255
1998	392
1999	717
2000	882
2001	1,139
2002	1,400
2003	1,797
2004	2,263
2005	2,685
2006	3,652
2007	4,629
2008	6,606
2009	9,150

Table 1: Number of members under risk in the random sample as at 1st January of the corresponding year.

The distribution of age and gender within the random sample is illustrated by the population pyramid in Figure 8. It shows that there is a slight preponderance of male members throughout all age groups. Besides, the average age in the portfolio is quite low compared to the portfolios of most European health insurers.

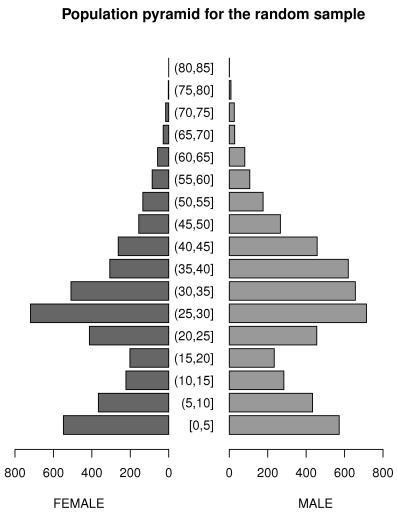

Figure 8: Population pyramid of all members in the random sample.

The available dataset additionally includes the following information:

Information on member level:

- Age and gender
- Country of residence
- Type of employment
- Relation to insured person (e.g. co-insured spouse, child, ...)
- Insurance plan (defining the individual cover for each member)

Information on invoice line/claim level:

- Type of procedure (e.g. CPT or service code)
- Diagnosis (using ICD-9 codes)
- Type of provider (e.g. pharmacy, general practitioner, hospital, ...)

All these factors have a potential impact on the annual claimed amount to be predicted. In order to incorporate them in the predictive model for DMP selection, various data transformations of raw data had to be performed which are described in Section 2.3.3.

2.3 Methodology

In the following, the criteria used for model selection and calibration (Section 2.3.1) in order to identify the best model for DMP participant selection among different regression models are defined. Furthermore, a short introduction to the theory of the selected model (see Section 2.3.2), a specific GLM, is given. In Section 2.3.3,

the construction of model covariates out of the raw data information on potential influence factors is described. In order to avoid overfitting caused by selecting too many covariates, a variable selection approach that identifies relevant influencing factors is illustrated. Finally, the selection of the GLM is constituted by comparing it to several other (more complex) regression approaches with regard to the criteria defined (see Section 2.3.4).

2.3.1 Criteria for Model Selection and Calibration

In order to find the optimal model/model calibration and to analyze the stability of the model at the same time a re-sampling approach (see Figure 11 in Section 2.4) according to the principal idea of predictive modeling has been used: detect systematic patterns in past data that explain the behavior of the target variable and apply these patterns in order to predict future values of the target variable.

Following this principle, the original sample was split into a training sample for model calibration and a test sample for model validation. This procedure is meant to avoid overfitting to the training data, i.e. to filter out systematic patterns in the data. Two re-sampling methods that are frequently used for the optimization of predictive models are cross validation and bootstrapping [Good, 2005]. Cross validation methods split the full sample into k equal parts in order to iteratively train different models using $k-1$ parts as a training sample and the missing k-th part as a test sample. Similarly, bootstrapping methods iteratively construct training samples by drawing with replacement out of the full sample and using the non-selected elements (out-of-sample elements) as a validation sample.

The data information in 2009 has only been used for a comparison to the vendors' solutions and the standard methods, but not for model selection or calibration. Consequently, the training models are constructed based on the data information from 1995 to 2007 of all members in the random sample who were enrolled before 1 January 2008 (n = 6,606). The real claimed amounts in 2008 were held back for validation purposes. The models were optimized based on different predictive quality measures (see Appendix A) comparing the actually observed claimed amounts in 2008 with the predicted values. In order to reduce the effect of single observations with high leverage and to receive a model that can be transferred from the random sample to the whole portfolio, a bootstrapping approach was applied: 100 bootstrap samples at member-level were drawn from the remaining 6,606 members and used to predict the claimed amounts in 2008 (out-of-bag-sample). As with the construction of the original random sample, SPSS's random selection algorithm was used for re-sampling. Beside the control of overfitting, the re-sampling approach permits analysis of the stability of prediction models by considering the variance of the predictive measures between the re-samples.

2.3.2 Basic Structure of the Selected Model

The starting point for the analysis was a classical linear model assuming a normal distribution for the target variable y given the vector of covariates $\boldsymbol{x}$. This model, however, showed an insufficient adaptation to the training data (see Section 2.3.4) due to the specific distributional form of annual claimed amounts (see Section 2.2.1). Consequently, more complex structures including generalized linear models (GLMs), different two-stage approaches, Generalized estimating equation models (GEEs) and generalized linear mixed models (GLMMs) were assessed.

GLMs extend the concept of linear models by relaxing the assumption of normality ($y|\boldsymbol{x}$ may follow any distribution from the exponential family) and by introducing a link function g. The model equation is then defined by

$$g(y) = \boldsymbol{x}'\boldsymbol{\beta} \tag{1}$$

where y is the target variable, $\boldsymbol{x} = (1, x_1, \ldots, x_p)'$ a set of covariates or predictors and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ the vector of unknown regression parameters that is estimated by an iterative numeric algorithm.

A prediction y^* can be obtained by plugging in the estimation $\hat{\boldsymbol{\beta}}$ into the regression equation where $\boldsymbol{x}$ is replaced by the corresponding covariate information $\boldsymbol{x}^*$:

$$y^* = g^{-1}(\boldsymbol{x}^{*\prime}\hat{\boldsymbol{\beta}}). \tag{2}$$

To determine the optimal link function for the prediction model, different distributional assumptions for $y|\boldsymbol{x}$ (normal and gamma distribution) and different link functions g (identity, log and inverse) were tested using the re-sampling approach described. In the process, it turned out that a GLM that assumes a normal distribution for $y|\boldsymbol{x}$ and uses the logarithmic function $g(y) = \log(y)$ as a link function yields the most accurate predictions for 2008 (see Section 2.3.4). In order to account for annual claimed amounts equal to zero in the observation period the link function was slightly modified to $g(y) = \log(y + 1)$, which avoids the exclusion of these observations.

The aim of the model is to predict the claimed amount y_{it} for the individual i ($i = 1, \ldots, n$) in a future period t for which no covariate information $\boldsymbol{x}_{it}$ is available. For this purpose, all preceding claims information was aggregated at member-year level assuming that the claimed amount in t depends on the average predictor values from all previous years since the enrolment year of the member $\boldsymbol{x}_{i,t_{\text{enrolled}}}, \ldots, \boldsymbol{x}_{i,t-1}$ (so-called mean-lag approach).

In this setting the independence of the observations $y_{i,t_{\text{enrolled}},\ldots,y_{i,t-1}}$ for the same member i is questionable and an approach which reflects the intra-individual correlation (see Section 2.2.2) is preferable. Accordingly, different GEEs with various assumptions on the structure of the intra-individual correlation (AR(1), exchangeable and unstructured) as well as different GLMMs were tested. The GEEs yield similar prediction results like the GLMs. The (small) differences arise from the fact

that smaller standard errors lead to another linear predictor after the variable selection process described in Section 2.3.3). Due to the fact that GLMs are easier to apply (for example, there is no automated variable selection approach for GEEs implemented in standard statistics software in comparison to GLMs) the simpler model structure was chosen. In contrast to GEEs, GLMMs yielded a substantially worse prediction result (see Section 2.3.4).

Though the GLM approach does not directly reflect the intra-individual correlation included in the annual claimed amounts, this structure was indirectly accounted for by including "mean-lags" for annual inpatient and outpatient costs in the linear predictor. This means assuming that the claimed amount for an individual i in year t y_{it} directly depends on the previous claimed amounts $y_{i,t_{\text{enrolled}}}, \ldots, y_{i,t-1}$. This approach also helps to control the strong medical inflation described in Section 2.2.2 and clearly improves the prediction results.

2.3.3 Definition of Covariates and Selection of Relevant Predictors

This section explains how the data information on potential influence factors described in Section 2.2.2 is made available for the prediction model. Additionally, it deals with the selection of relevant predictors using an appropriate variable selection approach.

Basic member information was incorporated into the vector $\boldsymbol{x}$ of covariates by constructing categorical variables in the following way:

- Gender
- Age (16 categories representing age bands of 5 years)
- Country of residence (in 3 geographical categories)
- Type of employment (executive or manager, office staff in 5 categories, undefined or other)
- Relation of co-insured to insured person (applicant, child, personnel, spouse, undefined or other)
- Insurance plan for outpatient treatments (outpatient plan group 1, ..., outpatient plan group 4, other plans, undefined)

For all categorical variables dummy coding [Tutz, 2000] with the first category as reference category was used. In order to reduce the number of categories, outpatient plans have been clustered into 4 outpatient plan groups using a k-means clustering algorithm [MacKay, 2003]. The algorithm allocates plans with similar cover and co-payment conditions to the same group. No differentiation was made for inpatient plans due to their similar cover structure. Time-dependent covariates x_t were used in order to account for changes in a member's properties throughout the observation period, such as a change in the outpatient plan. For member information, it was assumed that the category observed in year t affects the claimed amount in the same year. If the value of a covariate x_{it} was unknown, like in the validation year 2008,

the assumption was made that the category has not changed since the last year and x_{it} was set to $x_{i,t-1}$.

To enrich the vector of predictors $\boldsymbol{x}_i$ by incorporating historic claims information (procedures, diagnoses, specialist consultations and claimed amounts), the following strategy was applied: for the procedures a patient underwent within a year, count variables that reflect the frequency of observed inpatient and outpatient procedures within several procedure groups (surgery, anesthesia, consultation, ...) were built. In order to take account of the large variation in outpatient costs, outpatient count variables were subdivided into four cost groups. Similarly, count variables for inpatient and outpatient diagnoses were separately constructed for each of the 19 ICD-9 chapters. Further count variables indicating how often a patient visited a certain provider type (dermatologist, cardiologist, pediatrist, ...) were defined. Following the assumption that the claimed amount y_{it} in year t depends on preceding claims information $\boldsymbol{x}_{i,t_{\text{enrolled}}}, \ldots, \boldsymbol{x}_{i,t-1}$, the mean-lag approach described in Section 2.3.2 was applied to the annual count variables for diagnoses, procedures and visits of specific provider types. Accordingly, mean-lags were calculated for preceding inpatient and outpatient costs.

For prediction purposes, however, it is not meaningful to use the large amount of all available predictors because covariates that do not have an influence on the target variable can disturb the prediction. By including more and more covariates into the model, the adaptation to the training data can steadily be increased. The predictive quality of the model, by contrast, starts to decrease after a certain break-even point. The intention of variable selection approaches is to find this break-even point in order to optimize the prediction. Additionally, they help to identify important influence factors on the target variable. Another advantage of the GLM approach is that automated variable selection approaches are already implemented for this class of models in all kinds of standard statistics software. For example, in the statistics software package R [R Development Core Team, 2009], the function "stepAIC" (library "MASS", Venables and Ripley [2002]) can be used for covariate selection. For other regression models, like GEEs, variable selection techniques had to be implemented manually.

To configure the GLM, a stepwise variable selection approach [Miller, 1990] was applied starting from the intercept model. As an inclusion, exclusion and stop criterion, the Akaike information criterion (AIC [Akaike, 1974]) was used. The application of the Schwarz Bayesian criterion (SBC [Schwarz, 1978]), often also denoted as BIC, was also evaluated by means of the described re-sampling approach. Considering several measures of predictive quality, the AIC models showed better prediction results than the BIC models that include fewer covariates and, correspondingly, seem to underfit the training data. This is in line with previous results of other authors who argue that the AIC is more appropriate for prediction purposes because it uses prediction of future data as the key criterion for the quality of a model [Lamers, 2004].

In the AIC selection for all 100 re-samples, mean-lags of procedure and provider counts have been selected less frequently than mean-lags of diagnosis counts. This

means that procedure and provider information hardly improves the prediction result where diagnosis information is available. This is not surprising, since procedure, diagnosis and provider specialty variables contain a lot of overlapping information. Consequently, procedure and provider type covariates were completely excluded to demonstrate that a good prediction result can also be reached without this information. This result might be useful for many insurance companies, especially in emerging countries and in countries where no electronic billing is used, as insurance companies in those countries often do not capture provider type and procedures in an appropriate way. Considering data storage in health insurance companies, all other necessary variables should be available in most insurance companies, even in emerging countries. In combination with the simple handling of the data aggregation and modeling concept for which only basic statistics skills and standard software is required, the implementation effort for health insurance companies is considerably low.

Table 2 shows a list of those covariates that have most often been selected by AIC selection in the re-sampling process. Consequently, this set of covariates was used for the final model, which is compared to the vendors' solutions and to two non-predictive standard techniques in Section 2.4.1. According to the number of selections, the member-specific covariates, the mean lag of outpatient costs and the mean lags of the frequencies of inpatient diagnoses belonging to ICD-9 chapter 7 (diseases of the circulatory system) and to ICD-9 chapter 6 (diseases of the nervous system and the sense organs) have the highest predictive value. It is not recommended using exactly this set of covariates for application to other portfolios. Nevertheless, a log-normal GLM in combination with an AIC selection can be used to determine the best set of covariates for predicting individual annual claimed amounts.

2.3.4 Comparison to Other Regression Techniques

In the following, the model selection process based on the re-sampling approach described in Section 2.3.1 is illustrated. Furthermore, the advantages of applying the log-normal GLM presented in Section 2.3.2 for DMP selection compared to many other (more complex) regression approaches are demonstrated. As already mentioned in Section 2.2.2, the focus is on regression approaches due to their availability in all kinds of statistics software, the high level of interpretability, the various possibilities for controlling overfitting and the high stability for small datasets. All models, especially the log-normal GLM approach with AIC variable selection, can be implemented with the open source statistics software R [R Development Core Team, 2009], which requires basic programming skills. If the user prefers to work with menu-based interfaces, commercial software packages, like SAS, Statistica or SPSS, can equally be used. Only for the Bayesian GLMM approach described below BayesX [Belitz et al., 2009], a specific software for Bayesian regression models, was used due to computational advantages compared to the frequentist approaches implemented in standard software. For those models for which no automated variable selection algorithms were available, a forward AIC selection was manually implemented in R.

Covariate	# selected
Age category	100
Country of residence	100
Relation to insured person	100
Type of employment	100
Insurance plan for outpatient treatments	100
Mean lag of outpatient costs	100
Mean lag of # inpatient diagnoses within ICD-9 chapter 7 (diseases of the circulatory system)	98
Mean lag of # inpatient diagnoses within ICD-9 chapter 6 (diseases of the nervous system and the sense organs)	97
Gender	91
Mean lag of # outpatient diagnoses within ICD-9 chapter 7 (diseases of the circulatory system) / first cost quartile	91
Mean lag of # outpatient diagnoses within ICD-9 chapter 3 (endocrine, nutritional and metabolic diseases, and immunity disorders) / third cost quartile	87
Mean lag of # outpatient diagnoses within ICD-9 chapter 10 (diseases of the genitourinary system) / fourth cost quartile	85
Mean lag of # inpatient diagnoses within ICD-9 chapter 8 (diseases of the respiratory system)	83
Mean lag of # inpatient diagnoses within ICD-9 chapter 10 (diseases of the genitourinary system)	81

Table 2: Covariates most often chosen by AIC selection in the re-sampling process.

Four kinds of models in the model selection process were tested: GLMs, GEEs, two-stage GLMs and GLMMs. For each model type, different distributional assumptions and different link functions were compared. Table 3 summarizes all regression approaches assessed by re-sampling. Before the results are outlined, the structure of the applied two-stage models is briefly described. Further methodological details on GLMs, GEEs and (Bayesian) GLMMs can be found in Tutz and Fahrmeir [2001], Liang and Zeger [1986] and Fahrmeir and Kneib [2010].

Abbr.	Model
M1	Linear model $\hat{=}$ GLM with identity link
M2	Normal GLM with log-link
M3	Normal GLM with inverse link
M4	Gamma GLM with inverse link
M5	Gamma GLM with log-link
M6	Normal GEE with log-link (correlation unstructured)
M7	Gamma GEE with inverse link (correlation unstructured)
M8	Two-stage model (Binomial GLM with logit-link & Normal GLM with log-link)
M9	Two-stage model (Binomial GLM with logit-link & Gamma GLM with inverse link)
M10	Normal GLMM with log-link (random intercept)
M11	Gamma GLMM with inverse link (random intercept)
M12	Normal GLMM with log-link (random intercept & slope)
M13	Gamma GLMM with inverse link (random intercept & slope)

Table 3: Overview of all regression approaches considered.

As mentioned above, many claimed amounts were zero, which is a challenge for

parametric models that assume a specific distribution for the response variable y given the vector of covariates $\boldsymbol{x}$. Zero-inflation models [Cameron and Trivedi, 1998] are a class of statistical models that is specifically designed for count data with many values equal to zero. These models assume a discrete mixture distribution for $y|\boldsymbol{x}$. The two-step models described in Section 2.2.2 transfer this idea to a continuous response variable y (like, in the existing situation, a monetary amount). The underlying principle is to decompose the expectation $\mathbb{E}(y|\boldsymbol{x})$ into

- $\pi := \mathbb{E}(I(y > 0)|\boldsymbol{x})$ (where I is an indicator function that is equal to 1 if $y > 0$ and 0 else) and
- $\mu := \mathbb{E}(y|y > 0, \boldsymbol{x})$.

Then, both components can be estimated separately. By way of example, two approaches (model M8 and M9) that assume a binomial distribution for estimating π and a normal and a gamma distribution, respectively, for estimating μ are presented here. Changing the link function of the binomial GLM hardly had an influence on the results. The predictions $\boldsymbol{y}^*$ were calculated by multiplying $\boldsymbol{\mu}^*$ and $\boldsymbol{\pi}^*$ as described in Blough et al. [1999]. AIC selection was applied separately for both models.

Figure 9 illustrates the predictive quality of all regression approaches by means of the predictive R-squared R^{2^*} (see Appendix A) for all 100 bootstrap samples. As already mentioned above, the log-linear GLM yields the highest predictive quality. In addition, the small variance of the results based on different re-samples shows the high stability of the GLM approach. The log-linear GEE is almost on the same level, but does not improve the prediction results by accounting for the intra-individual correlation. Consequently, it is preferable to use the less complex GLM that is easier to apply for health insurers. All other models show significantly lower predictive quality and, except for the linear model, a higher variation within the re-samples. This indicates that they are more sensitive to outliers in the training data and prone to overfitting. Throughout all model types, assuming a normal distribution for the individual annual claimed amounts instead of a gamma distribution seems to be the better choice for the underlying portfolio. Considering all of the aspects mentioned, it turned out that a log-linear GLM with AIC variable selection is the most appropriate regression approach for selecting DMP candidates in the existing setting. The application of other predictive measures (see Appendix A), also of those that directly measure the sorting capacity of the models, does not change the overall picture and leads to the same conclusion.

If the adaptation of the model to the training data is also considered (for example by looking at the model R-squared or related measures for generalized models), the overfitting problem in the identification of an optimal prediction model becomes apparent. Figure 10 illustrates the conclusion from the assessment of different models based on the test dataset: With increasing model complexity, the adaptation to the training data can be improved without limit. The predictive accuracy, however, ascends up to a certain optimal level, in this case the selected GLM, and descends after this point due to overfitting to the training data.

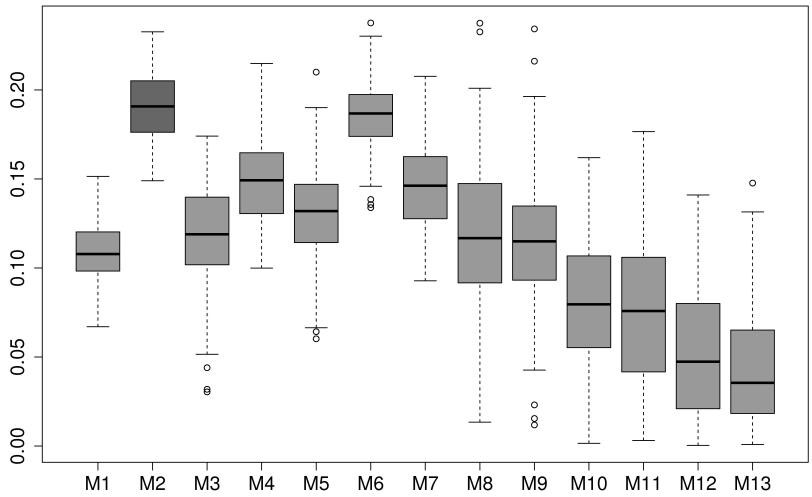

Figure 9: Distribution of predictive R-squared values based on re-sampling for all regression approaches considered.

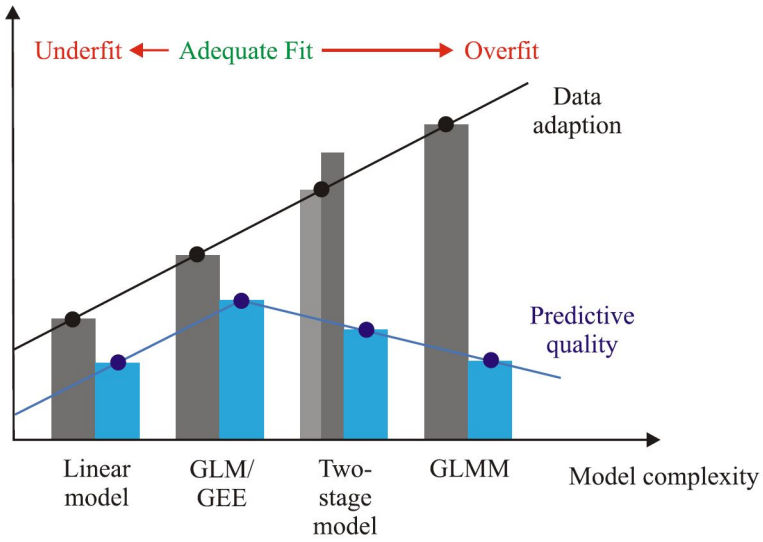

Figure 10: Schematic illustration of the different development of data adaption and predictive quality with increasing model complexity.

2.4 Results

In order to be able to provide the best possible solution for a cost effective selection of DMP participants, three professional providers of PM solutions were asked to participate in a forecasting competition. The vendors who are active in the German and international markets applied different, partly not regression-based PM techniques without revealing methodological details. All vendors (referred to below as A, B or C) received the random sample described above including data from 1995

to 2008 and were asked to predict the individual claimed amount in 2009. In the following, the results of the providers' approaches and the GLM approach described in Section 2.3.2 (referred to as O below) are compared to each other. For this comparison, the selected model was applied to all members in the random sample, also incorporating the claims information from 2008. As a benchmark for the PM solutions, also two non-predictive standard methods for selecting DMP participants were included in the comparison. Neither method explicitly predicts future costs, but both aim to put members in order according to their individual risk potential based on historic data. The first method (referred to as S1 below) simply uses the (order of) individual claimed amounts in 2008 to identify the DMP candidates with the highest future saving potential. The second method (referred to as S2 below) lists members in the order of a chronic score order that is based on the average number of hospitalizations and outpatient visits due to chronic conditions in the observation period. Figure 11 summarizes the sampling approach applied to the underlying portfolio and all prediction methods described.

In Section 2.4.1, various comparisons based on all individuals within the random sample are presented. All methods are compared on the basis of different predictive measures with regard to their capability of selecting the best DMP participants (see Appendix A for a detailed explanation of measures and their interpretation and table 4 for a summary). For the same purpose, different subgroups in the random sample are analyzed: high-cost cases and members with preceding chronic diagnoses. Subsequently, the key factors that determine the predictive quality of a forecasting approach (see Section 2.4.2) are summarized and the extent to which the presented results are generally applicable is discussed. This summary also considers the results of a second vendor performed on the basis of a group of chronic asthma patients.

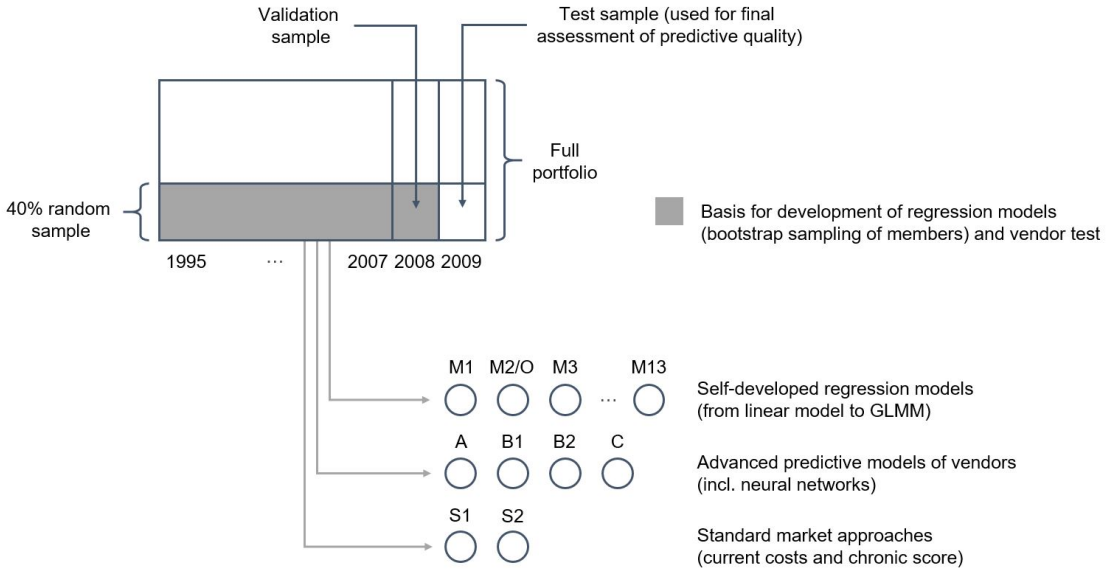

Figure 11: Overview of the described sampling approach and all prediction methods applied.

Measures for accuracy of prediction			
Abbr.	Pred. measure	Interpretation	Range
MPSE	Mean predictive squared error	Deviation of $\boldsymbol{y}^*$ from $\boldsymbol{y}_o$ avoiding large errors	$[\mathbf{0}; +\infty]$
MPAE	Mean predictive absolute error	Deviation of $\boldsymbol{y}^*$ from $\boldsymbol{y}_o$ favoring small average error	$[\mathbf{0}; +\infty]$
R^{2^*}	Predictive R-squared	(Squared) linear correlation between $\boldsymbol{y}^*$ and $\boldsymbol{y}_o$	$[0; \mathbf{1}]$

Measures for sorting capacity			
Abbr.	Pred. measure	Interpretation	Range
R_{Sp}	Spearman correlation coefficient	Monotonic correlation between $\boldsymbol{y}^*$ and $\boldsymbol{y}_o$	$[-1; \mathbf{1}]$
AUC_m AUC_m	Area under the matching curve	Sorting capacity throughout the whole portfolio	$[0; \mathbf{1}]$
$m(i)$	identification or hit ratio	% of identified cases among i most expensive members	$[0; \mathbf{1}]$

Table 4: Overview of predictive measures used (optimal value within possible range marked by boldface).

2.4.1 Comparison to Solutions of Vendors and Standard Approaches

Table 5 shows different predictive measures introduced in Appendix A for the prediction results of all vendors, the GLM and the two standard methods for DMP risk scoring. For all PM solutions, the measures compare the predicted and actually observed claimed amounts in 2009 for all members in the random sample (n = 9,150). The standard methods S1 and S2 are based on risk scores rather than an explicit cost prediction. Therefore, MPSE and MPAE, which measure the deviation between predicted and actual costs are not considered here. However, measures analyzing the correlation of the risk scores with actual costs as well as direct sorting capacity measures are applied. The prediction of vendor C included 21 missing values so that the measures are only based on n_{C} = 9,129 observations for vendor C. Vendor B submitted two predictions B1 and B2, based on estimating respectively the mean and the median of the distribution of annual claimed amounts.

Method	MPSE	MPAE	R^{2^*}	R_{Sp}	AUC_m
A	660,108	268.87	0.0671	0.3025	0.6202
B1	533,484	212.10	0.1709	0.6007	0.7065
B2	593,104	**194.13**	0.1329	**0.6695**	**0.7285**
C	618,994	226.62	0.0733	0.4133	0.6386
O	**516,363**	246.20	**0.1932**	0.3084	0.6166
S1	–	–	0.0598	0.2531	0.5706
S2	–	–	0.0432	0.2021	0.5525

Table 5: Predictive measures for the predictions of all vendors, the GLM and the standard DMP selection methods based on the random sample from the portfolio (best prediction according to the corresponding measure marked by boldface).

In general, the predictive quality of the vendor solutions is quite heterogeneous and none of the professional vendors was able to clearly outperform the GLM approach. As regards the MPSE and the predictive R-squared, the GLM shows the highest predictive quality followed by the mean prediction of vendor B, i.e. it minimizes the probability of large discrepancies between predicted and real data. This property is crucial for DMP selection because not identifying members that are at high risk

of developing an uncontrolled chronic disease means clearly decreasing the overall saving potential related to the DMP. The predictions of vendor B, especially the median prediction, yield the best results in terms of minimization of the MPAE and in terms of sorting capacity (R_{Sp} and AUC_m). In this context, the developed model is on a similar level to the PM approaches of vendors A and C. Considering all measures presented, the non-predictive standard methods cannot compete with any of the PM solutions in terms of future risk assessment. This clearly shows that modern statistical prediction techniques are preferable for an optimal selection of DMP participants in small portfolios and a difficult data environment. It is assumed that the benefit is greater if more training data (more members with longer history) are available as long as overfitting is controlled. The experience with larger portfolios in a more stable environment shows that predictive R-squared values up to 0.35 are possible.

To illustrate the sorting capacity of the methods throughout the random sample, figure 12 shows the matching curves (defined in Appendix A) for all approaches. Here, the PM approach of vendor B clearly delivers the best results in every cost region leading to an AUC_m of 0.7285, which is remarkable bearing in mind the difficult data environment.

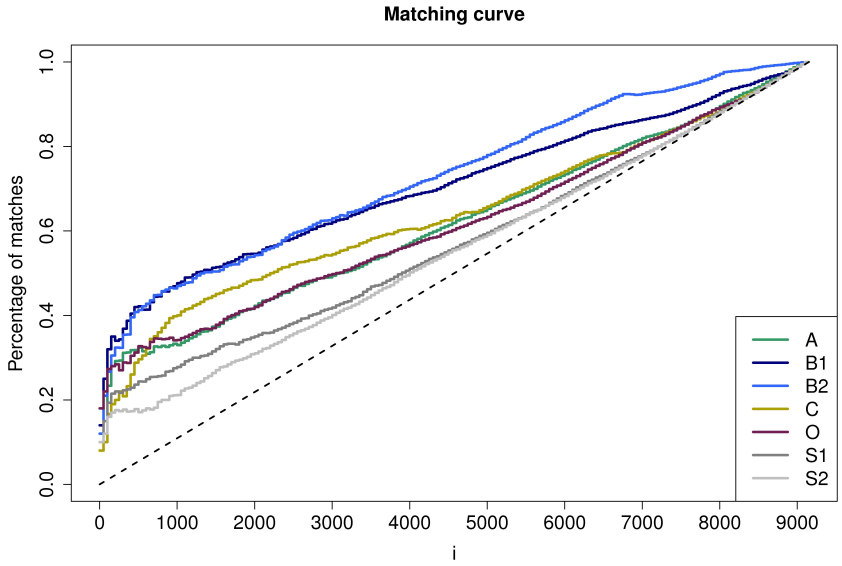

Figure 12: Matching curves of all predictions for the annual claimed amounts in 2009 based on the random sample from the portfolio.

In addition, a closer look is taken at the high-cost region and different values of the matching curve $m(i)$ are compared. Also, the percentage of members that could correctly be identified in advance as high-risk members in 2009 is determined. Based on these figures and some experience-driven assumptions on program costs and average individual saving potential (usually offered by DMP vendors) the economic benefit of a DMP is directly calculable.

The absolute number of members that are selected for a DMP basically depends on two factors:

a) the investment budget of the health insurer and

b) the economic break even point until which the individual saving potential exceeds the program costs.

Depending on the scope of the program, typically, up to 20% of all insured members are included in a DMP. Indication-specific programs usually have fewer participants than general health programs. In order to illustrate different scenarios, table 6 compares the identification ratios $m(i)$ of all methods for those members who actually had the highest claimed amounts in 2009 (top 1%, 5% and 10% of all members in the random sample, respectively). For the top 1% of members by expense, the models of vendor B and the GLM deliver the best identification results. For example, method B1 identifies more than twice the number of actual high-cost members (25%) as the standard approach S2 (12%) based on the chronic score. For instance, assuming a 20% higher saving potential in this cost group than in the next-highest cost group, the overall saving potential of a DMP increases by at least $(25\% - 12\%) \cdot 120\% \approx 17\%$ using method B1 instead of S2. The PM solutions of vendor A and C are on a similar level to the standard approaches. For the top 5% and top 10% of members by expense, the efficiency of the PM solutions of vendor B becomes more apparent. All other PM solutions, including the GLM, are still at a good level compared to the standard approaches, which perform significantly worse. This confirms the conclusion that appropriate PM solutions achieve a better selection of DMP candidates than non-predictive standard methods.

Method	$m(92)$ (1%)	$m(458)$ (5%)	$m(915)$ (10%)
A	15%	32%	33%
B1	**25%**	**42%**	**47%**
B2	21%	41%	47%
C	10%	29%	39%
O	22%	31%	35%
S1	15%	24%	27%
S2	12%	18%	21%

Table 6: Identification ratios for those members in the random sample with the highest real costs in 2009 (best prediction among all methods according to the identification ratio marked by boldface).

In addition, subgroups of the sample population are separately investigated in order to detect the strengths and weaknesses of the different prediction tools in respect of DMP selection. First, the predictive quality of the methods is compared with a focus on the high-cost sector. Table 7 shows MPSEs, MPAEs and predictive R-squared values for all members above the 99%, 95% and 90% quantile of annual claimed amounts in 2009, respectively. Looking at the top 1% of members with the highest actual costs in 2009, the GLM clearly yields the best results in terms of all measures. Especially with regard to the predictive R-squared, the GLM is the only approach with significant positive linear correlation between observed and predicted values. The professional vendors are on a similar level to the standard approaches here. For the members above the 95% and the 90% cost quantile, the overall picture

is similar. In terms of MPSE, the developed model clearly permits the most precise prediction of claimed amounts. The fact that large losses are avoided for high-cost cases confirms that the GLM provides a reliable risk evaluation for DMP selection. Vendor B delivers the most accurate prediction results "on average" indicated by the smallest MPAEs for the top 5% and top 10% of members by expense in the random sample.

Top 1% ($n = 92$) of members with the highest actual costs in 2009

Method	MPSE	MPAE	R^{2^*}
A	47,067,518	4,885.47	0.0128
B1	45,014,918	4,573.25	0.0220
B2	50,106,040	5,031.31	0.0033
C	51,055,836	5,143.98	0.0157
O	**38,614,019**	**4,531.84**	**0.2060**
S1	–	–	0.0103
S2	–	–	0.0076

Top 5% ($n = 458$) of members with the highest actual costs in 2009

Method	MPSE	MPAE	R^{2^*}
A	10,779,042	1,787.58	0.0240
B1	9,894,721	**1,636.38**	0.0195
B2	11,158,029	1,865.91	0.0015
C	11,609,443	2,000.01	0.0116
O	**8,907,106**	1,777.05	**0.1625**
S1	–	–	0.0155
S2	–	–	0.0112

Top 10% ($n = 915$) of members with the highest actual costs in 2009

Method	MPSE	MPAE	R^{2^*}
A	5,607,409	1,153.29	0.0372
B1	5,084,827	**1,044.67**	0.0428
B2	5,767,319	1,212.10	0.0108
C	5,996,250	1,300.58	0.0182
O	**4,632,960**	1,161.31	**0.1825**
S1	–	–	0.0208
S2	–	–	0.0166

Table 7: Measures of predictive accuracy for those members in the random sample with the highest real costs in 2009 (best prediction among all methods according to the respective measure marked by boldface).

Finally, the ability of all approaches to accurately predict the claims costs of members suffering from five of the most common chronic diseases (chronic heart failure, diabetes, chronic respiratory disease, hypertension and chronic back pain) is examined. This group is especially relevant for health insurers because an already high and still growing percentage of total health expenditures results from the treatment of chronic diseases. For instance, healthcare expenditures on diabetes alone accounted for 11.6% of the total healthcare expenditures in the world in 2010 (based on figures of the World Health Organization (WHO) made available by the IDF [2011]). In North Africa and the Middle East health expenditures for diabetes will have approximately doubled by 2030. One reaction to this development is the establishment of indication-specific DMPs in order to attenuate the cost impact of chronic diseases.

In order to evaluate whether PM solutions can optimize the selection for indication-specific programs, the predictive accuracy of all methods in the disease subgroups mentioned were compared. Table 8 summarizes the MPSEs, MPAEs and predictive R-squared values of patients having at least one diagnosis in the observation period with an ICD9-code identifying the corresponding disease. High absolute values of R^{2^*} must be handled with care as some of the chronic diseases rarely occur in the random sample (e.g. $n = 18$ for chronic heart failure and $n = 189$ for diabetes). For the existing dataset, the predictive R-squared of the developed model is particularly high for chronic heart failure (0.7502), diabetes (0.6737), chronic respiratory disease (0.4615) and hypertension (0.5027). In terms of predictive R-squared, the predictive accuracy of the vendor solutions is considerably lower in most cases, but usually higher than the predictive accuracy of the standard approaches. Among these, the chronic score (S2) performs better than the risk evaluation based on costs in the preceding year (S1). In terms of the MPSE, the GLM yields optimal prediction results for all chronic diseases analyzed and is permanently among the best solutions as regards the MPAE. Here, the predictions of vendor B show a similar predictive quality. These findings are in line with the results for the high-cost sector, because chronic patients, especially those with an uncontrolled chronic disease, are likely to generate higher claims costs. In either case, appropriate PM solutions, such as the model of vendor B and the GLM, can increase the economic efficiency of indication-specific DMPs through an optimized selection. In order to validate this finding, a second vendor test with three other international vendors has recently been performed based on a portfolio comprising only chronic asthma patients. The findings are summarized in Section 2.4.2.

Further analyses of subgroups (using simulation techniques) have shown that the predictive quality of the GLM approach is especially high for members with ascending and stagnating (high) claims costs over time compared to members with descending and highly unstable costs. This result confirms that future claimed amounts for patients with both controlled and uncontrolled chronic diseases can be predicted with a high degree of precision.

2.4.2 General Applicability of Results

Apart from the example dataset presented, a lot of other health insurance datasets of various size and quality have been used in order to predict different target variables, such as different cost types, utilization and likelihood of hospitalization. Based on this experience, it can be concluded that the prediction of medical costs (like of all other outcomes) is mainly driven by three factors: sample size, completeness of data and length of individual claims history. Figure 13 schematically illustrates the interactive effect of these components on predictive quality and classifies the test dataset applied in this context. By comparing their data situation to the test setting used, health insurers can obtain a first rough estimation of the reliability of cost prediction models and PM-based DMP selection if they are to be applied to their data. Especially the small sample size and the short data history of many members in the random sample lead to a comparably low absolute predictive quality. Nevertheless,

Chronic heart failure ($n = 18$)			
Method	MPSE	MPAE	R^{2*}
A	68,766	3.56	**0.8368**
B1	86,632	3.90	0.4464
B2	105,337	4.17	0.2250
C	106,849	4.56	0.3380
O	**48749**	**3.52**	0.7502
S1	–	–	0.1716
S2	–	–	0.2733

Diabetes ($n = 189$)			
Method	MPSE	MPAE	R^{2*}
A	146,226	16.15	0.1756
B1	132,910	**14.38**	0.2285
B2	162,690	16.24	0.1188
C	173,834	19.77	0.1173
O	**68,786**	14.59	**0.6737**
S1	–	–	0.1506
S2	–	–	0.2188

Chronic respiratory disease ($n = 697$)			
Method	MPSE	MPAE	R^{2*}
A	193,524	30.89	0.1121
B1	164,494	**25.00**	0.2293
B2	192,087	26.37	0.1280
C	199,037	29.55	0.1396
O	**127,777**	26.33	**0.4615**
S1	–	–	0.0837
S2	–	–	0.1888

Hypertension ($n = 426$)			
Method	MPSE	MPAE	R^{2*}
A	194,463	30.30	0.1518
B1	170,644	**26.80**	0.2126
B2	204,897	28.31	0.1257
C	225,003	35.56	0.0954
O	**112,487**	29.03	**0.5027**
S1	–	–	0.1076
S2	–	–	0.2388

Chronic back pain ($n =$ **1,180**)			
Method	MPSE	MPAE	R^{2*}
A	190,901	50.55	0.0350
B1	**131,506**	40.38	**0.1642**
B2	145,848	**40.28**	0.1490
C	153,546	45.67	0.0925
O	139,764	44.70	0.1213
S1	–	–	0.0226
S2	–	–	0.0933

Table 8: Measures of predictive accuracy for members with 5 typical chronic diseases (best prediction among all methods according to the respective measure marked by boldface).

Section 2.4.1 demonstrates that predictive models are still more effective in detecting high-cost patients in such a data environment than non-predictive approaches common in the market.

As other studies have already shown, the (theoretically possible) predictive quality of analytical approaches to forecasting health insurance claims is limited, even in more stable markets [Mehmud and Winkelman, 2007]. This is mainly due to the

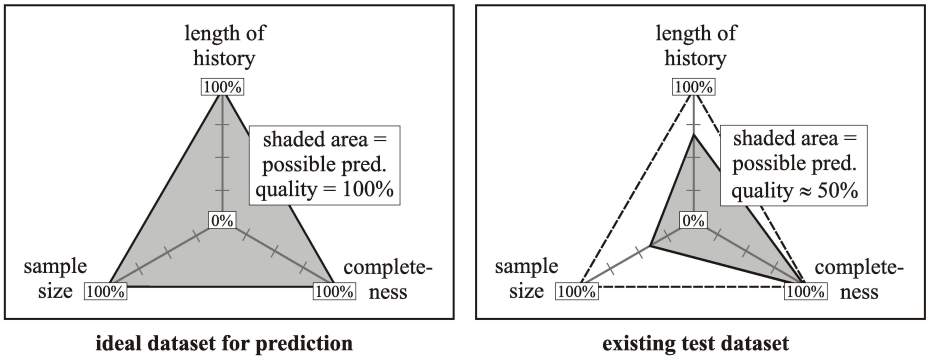

Figure 13: Schematic illustration of the three data dimensions relevant for good prediction results.

considerable influence of unpredictable/random events and the complexity of human health. In order to obtain the best possible prediction result, it is necessary to fully exploit the medical information included in claims data, especially if the aim of the prediction is to identify participants for indication-specific DMPs. For example, it is vital to account for the correlations of medical diagnoses. A further improvement in predictive quality can be attained by incorporating additional member information such as smoking status or clinical records if this information is available to the health insurer.

In order to demonstrate that PM solutions can enhance the saving potential of effective DMPs even under difficult data conditions, a small portfolio of insured members with short claims history from an emerging market was chosen as a test setting. Looking at the results from Section 2.4.1 and considering the stability of the GLM approach illustrated in Section 2.3.4, appropriate PM solutions make a significant contribution to increasing the economic benefit of general and indication-specific DMPs for similar settings. As mentioned above, the absolute predictive quality increases rapidly with the size of the portfolio (if overfitting is avoided). Based on the experience of applying the described methodology to different larger datasets, it is assumed that the advantage of PM solutions over standard selection methods grows with the size of the portfolio. More advanced PM solutions are assumed to perform better for larger portfolios with longer claims history.

In 2012 a second test of professional PM vendors was performed to confirm the results of the first comparison. In order to analyze the economic benefit of PM techniques, especially for indication-specific DMPs, a portfolio of approximately 10,000 asthma patients was chosen. The homogeneity of this group lead to a significant increase in absolute predictive quality ($R^{2^*} \approx 0.30$), even though the data environment was similar (emerging market, strong medical inflation and short data history). Certainly, the smaller percentage of annual claimed amounts equal to zero for chronic patients, facilitated the prediction, too. It was possible to achieve an improvement in predictive quality by incorporating more medical knowledge into the GLM approach than for the generic approach of covariate construction presented in this chapter. Detailed analyses have shown that the combination of medical knowl-

edge (e.g. interactions of disease-related diagnoses or severity grouping based on pharmaceuticals) and empirical methods (e.g. algorithms for procedure clustering) can further improve the identification of future high-cost patients. Like in the first test, none of the professional vendor solutions outperformed the developed model. Furthermore, the advantage of DMP selection based on PM techniques over non-predictive standard methods became even more apparent in this indication-specific setting.

2.5 Summary and Outlook

The analyses on DMP selection show that, for small portfolios and a difficult data environment, appropriate predictive modeling techniques clearly identify more future high-cost patients in advance than non-predictive standard methods. For an example of a setting, it is demonstrated that such solutions can increase the saving potential of general or disease-specific DMPs. In addition, a generic regression-based approach with low requirements regarding data quality and volume is illustrated. These requirements are met by most health insurance companies, even in emerging countries. Hence, the GLM approach can easily be implemented following the data aggregation and modeling strategy described above for which only basic statistics skills and standard software is required. In spite of the difficult data situation, this approach delivers more stable prediction results than more complex regression techniques which tend to overfitting. Furthermore, the approach can, at least in this setting, compete with expensive professional solutions, which usually remain a black box to the user. A further comparison of data-driven selection techniques based on a portfolio of asthma patients confirms that the GLM approach presented (enriched with medical knowledge) optimizes the selection of participants for indication-specific DMPs.

Considering the results of the analyses, regression models that estimate the mean of the response distribution cannot further improve the predictive quality of medical cost forecasts. Instead the application of non- and semi-parametric regression techniques, such as density regression or quantile regression, need to be analyzed in the context of DMP selection. The quantile regression approach, which estimates a certain quantile of the response distribution (for instance, the 95% quantile of the distribution of annual claimed amounts) is especially interesting for the identification of high-cost cases.

One important assumption made in order to prove the economic benefit of PM solutions is the ability of the DMP to reduce the costs for patients with high saving potential. However, as already mentioned in Section 2.1, the economic effect of DMPs is subject to controversial discussion in scientific literature. For some indications, like chronic heart failure, the economic effect of DMPs has been confirmed by medical studies [Inglis et al., 2010]. For other indications, such as diabetes, an economic effect is questionable, at least in the short term. In practice, measuring the economic effect through randomized controlled trials, which is the standard procedure for medical treatment evaluation, is difficult. Health insurers, who usually

install such programs, compete for clients. Consequently, they do not want to offer a program to clients and then deprive them of participation because they have randomly been allocated to the control group. In order to still perform a valid cost comparison, various (statistical) methods have been proposed in related literature. However, the DMP effect depend seems to depend heavily on the method applied and some techniques obviously produce biased results. In order to solve this unsatisfactory situation, a fair and stable approach to measure the economic effect of DMPs is needed. A solution proposed in statistical literature is the "matched-pair method", which allocates a "twin" from an external control group to every DMP participant based on "similarity" in respect of all cost-relevant covariates. These covariates can, for example, be determined and ranked by means of a cost prediction model. For this purpose, the overall predictive quality of a model is relevant. Due to the fact that the GLM approach yields good results in this regard, several matched-pair approaches based on this model are introduced in the following chapter. Furthermore, the stability of the results, especially the sensitivity to the predictive quality of the model are evaluated.

3 Economic Evaluation of Disease Management Programs

Throughout the world, disease management programs (DMPs) are a popular way of improving healthcare provision to chronic patients. The question whether such programs are viable for reducing medical costs in the long run is the subject of controversial debate in the healthcare literature. In Europe in particular, DMPs are usually operated by or on behalf of health insurance companies. For operational reasons, the economic effect of these insurer-driven DMPs cannot normally be measured using classic randomized controlled trials. Many different measurement methods producing extremely heterogeneous and partly biased results have therefore been proposed in the healthcare literature. Section 3 illustrates a stable and consistent matched-pair approach based on modern statistical methods (predictive regression analysis and dimension reduction) to quantify the financial impact of DMPs, based on health insurance data. Using three different DMPs as examples, it is demonstrated that the proposed matched-pair method is able to increase the precision of the measurement compared with several benchmark methods. Also, the relationship between the method's uncertainty and the estimated risk of financial loss is analyzed as a prerequisite for a reinsurance solution for future DMPs.

3.1 Introduction

Chronic diseases are a major driver of rapidly rising health care costs in both developed and in emerging countries. In the US, for example, the care of chronic illness consumes approximately 75% of total healthcare expenditure and amounts to over 1 trillion US dollars per year [Freeman et al., 2011]. Chronic diseases were also responsible for 50% of the disease burden in 23 developing countries in 2005 and will cost those countries 84 billion US dollars by 2015 if nothing is done to slow their growth [Nugent, 2008]. According to many studies [Okunade and Murthy, 2002; Smith et al., 2009; Manning, 1991; Leigh and Fries, 1992], the main reasons for this development are an unhealthy lifestyle and a high medical inflation rate. As insurance companies have little influence on the latter, the steering of the behavior of insured persons is an important trigger. An example of such steering measure is disease management. Disease management programs (DMPs) are usually offered to chronic patients by public health initiatives or by specific vendors on behalf of insurance companies. They are meant to improve the medical situation and the quality of life of program participants, as well as reduce expenditure on benefits in the long term. Common measures are, for example, regular phone calls to increase compliance with medical plans and disease-specific consultancy ("telecoaching") and the online monitoring of disease-specific parameters such as blood pressure, weight or insulin level ("telemonitoring"). Other programs include specific fitness classes or regular consulting appointments with specialized doctors.

Section 2 demonstrates how modern statistical methods (predictive models) can be applied to optimize the saving potential of DMPs by ex-ante identification of

potential future high-cost patients. As a necessary prerequisite for this optimization, the program needs to be viable for reducing medical costs in the long run. This cost-reducing effect of DMPs is, however, subject to controversial discussion in the scientific literature. For some indications, like chronic heart failure, the economic effect of programs has been confirmed by various medical studies [Inglis et al., 2010; Goetzel et al., 2005]. For other indications, such as diabetes, an economic effect is questionable, at least in the short term [Goetzel et al., 2005]. Even for programs with similar target groups and program measures scientific studies arrive at very different results as regards the economic saving potential. For example, Sidorov et al. [2002] report a significant short-term effect of a diabetes program, whereas other authors like Leatherman et al. [2003] and Bodenheimer et al. [2002b] state that medical savings for similar programs can only be realized after 10 program years.

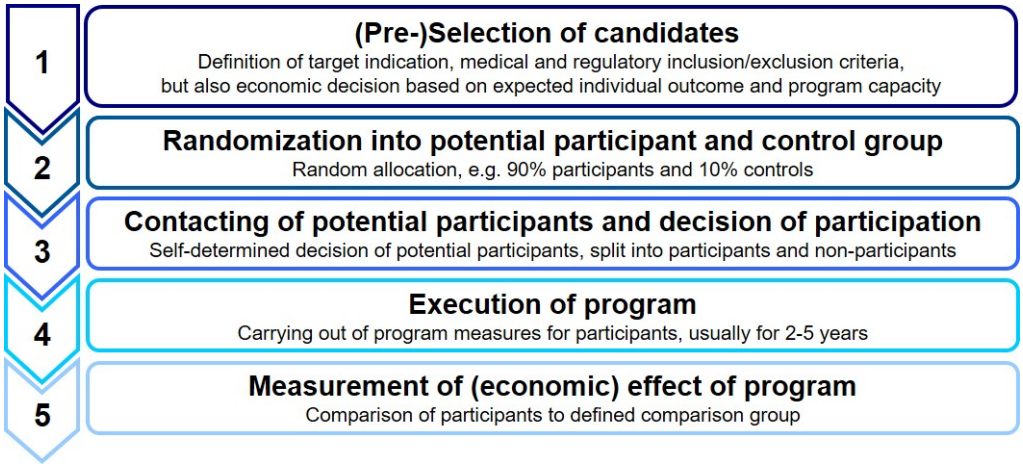

Figure 14: Usual operational process of a disease management program.

One of the main reasons for this strong variation in results lies in the usual operational process of DMPs offered by insurance companies, which is illustrated in Figure 14. In a situation in which the effect of a medical treatment procedure is to be determined, randomized controlled trials are the acknowledged standard procedure for reducing the measurement bias to a minimum. Looking at the results of different metastudies, like Goetzel et al. [2005], even the results of randomized controlled trials as the standard proceeding can vary considerably. Randomized controlled trials require everyone meeting the study's inclusion criteria to first be asked whether they want to take part. Those wishing to do so are randomized into two study arms – the treatment arm and the control arm. As health insurers are competing for clients, they do not want to offer a program and then deprive clients of the chance to take part because they have randomly been allocated to the control group. This is why the randomization is usually carried out first in insurer-driven DMPs, followed by the insured's decision to take part. This procedure though leads to a serious measurement problem: in the control group there is no self-selection process like the one observed in the treatment group. This means that any direct comparison between actual participants and controls may be biased (so-called "self-selection bias"). This bias cannot be neglected because many studies argue that a

patient's motivation to actively improve his or her health status is a decisive factor for the success of a DMP [Kralik et al., 2004].

This measurement problem leads to various methods being used to determine the (economic) effect of DMPs. Many of these methods have clear disadvantages and produce biased results (see Section 3.2.1). The lack of a standard measurement method for DMPs without adequate randomization also means that it is difficult to reliably quantify the financial risk of loss for an insurer offering a DMP. In order to estimate this risk of loss, it is necessary to have a precise estimate of the individual saving potential per participant. Moreover, the risk of loss depends on the uncertainty of the measurement method. In order to quantify this risk for future programs, the relationship between the uncertainty of the measurement method and the estimated financial risk of loss needs to be analyzed. The ex-ante quantification of a program's risk of loss can at the same time form the basis for a reinsurance solution protecting the primary insurer from the potential financial loss caused by the DMP.

In view of this situation, the main goals of this chapter are

a) to define a consistent and bias-minimizing standard measurement approach for (insurer-driven) DMPs where classic randomized controlled trials do not apply and

b) to quantify the relationship between the uncertainty of the defined measurement approach and the estimated risk of financial loss as a basis for evaluating the insurability of similar programs in the future.

For various reasons outlined in Section 3.2.1, the "matched-pair approach" (Section 3.2.2) is considered as the theoretical basis of an adequate measurement approach in the sense of target a). The matched-pair approach has its origins in observational medical and epidemiological studies [Rubin, 2006] where it reduces the most severe drawbacks of alternative measurement methods. In particular, it controls the previously mentioned self-selection bias by allocating so-called "twins" or "matched pairs" from a control group to each DMP participant. The costs of participants and their allocated twins can then be compared. Sections 3.3.1 and 3.3.2 present an intuitive way of combining this approach with a cost-prediction model that can also be used for the selection of DMP candidates (see Section 2). The allocation of twins based on model covariates with a significant impact on future costs ensures that only patients with a similar medical situation and similar long-term cost prognosis will be compared. In Section 3.3.3 the theoretical foundation for estimating the risk of financial loss related to a DMP is established by deriving the variance and the distribution of individual savings per participant.

In Section 3.4, three different DMPs are evaluated, in order to

a) demonstrate the stability and precision of the matched-pair approach presented compared with different benchmarks (see Section 3.4.1),

b) analyze its robustness to methodological variations and variations in the data input (see Section 3.4.2) and

c) assess the relationship between the uncertainty of the measurement method and the estimated risk of financial loss (see Section 3.4.3).

3.2 Background

In this section, an extensive overview of the statistical methods that are applied, especially in Central European markets and the US market, to measure the economic effect of DMPs is given (see Section 3.2.1). Some of these methods are used as a benchmark for the matched-pair approach introduced in Section 3.3. Also, some theoretical background on the matched-pair method and some examples from the literature on the use of matched-pair approaches in the evaluation of medical treatment is provided (see Section 3.2.2).

3.2.1 Existing Approaches for DMP Measurement

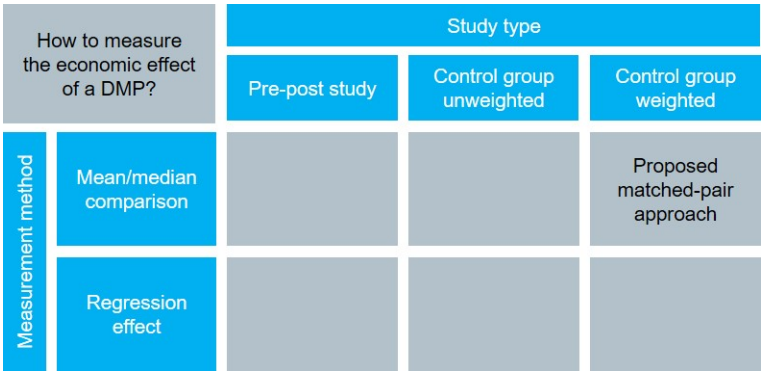

Figure 15: Overview of frequently applied measurement schemes for the (economic) effect of DMPs.

Extensive metastudies on the economic measurement of DMPs, like Goetzel et al. [2005] or Dove and Duncan [2004], show that various measurement schemes are used, especially in observational studies. Most of these schemes can be classified into the clusters illustrated in Figure 15. The measurement of the economic effect of a DMP is generally based either on the repeated measurement of costs for the same population before and during intervention, or on a simultaneous cost comparison between a group with intervention and a separate control group [Goetzel et al., 2005; Dove and Duncan, 2004]. If a control group is used, either all individuals of this group can be used for a cost comparison or only those individuals which are sufficiently similar to the participants. The latter case means that different weights are assigned to the controls, according to their relevance for the cost comparison. In this regard, some authors, like Abadie et al. [2004], distinguish between unweighted and weighted control group design. Moreover, statistical theory offers different methods for performing the cost comparison. One of the most common methods is to estimate the average individual saving from the difference in mean or median costs between both groups. Alternatively, the effect of a DMP can be assessed by including a

treatment effect in a regression model and using the estimated regression parameter as an estimate for the average individual saving. In the following, an overview of different concrete approaches will be given without any claim to be exhaustive.

In quasi-experimental studies where no external control group is available some authors (like Lorig et al. [2001]) use a pre-post comparison of participants' costs in order to measure the effect of the DMP. However, depending on the chronic disease being evaluated, a patient's risk and cost structure can change significantly with increasing age, even if the DMP is successful in slowing down the progress of the disease. Moreover, the healthcare situation (new drugs/therapies, better/worse access to healthcare etc.) can change considerably over time. Different adjustments to account for these distortions are therefore necessary [Villagra and Tamim, 2004]. As statistical adjustment methods can usually not fully offset the confounding effect of changed conditions, most authors prefer other study designs [Dove and Duncan, 2004].

Controlled studies which compare different patients at the same point in time are generally considered to be more reliable [Goetzel et al., 2005]. Some authors (e.g. vanVonno et al. [2005] and Riegel et al. [2000]) use a controlled pre-post setting, i.e. a mixture of a pre-post study and an unweighted control group study. They compare the cost development in the group of participants within a certain timeframe with the cost development in a control group. This method may also produce biased results if there is no risk stratification to ensure comparability between participants and controls with respect to the expected future cost development [Goetzel et al., 2005].

As outlined in Section 3.1, controlled studies are hampered by the fact that the randomization into participants and controls is usually carried out before the candidates are contacted and asked if they wish to take part. This leads to various definitions of comparison groups (see Figure 16).

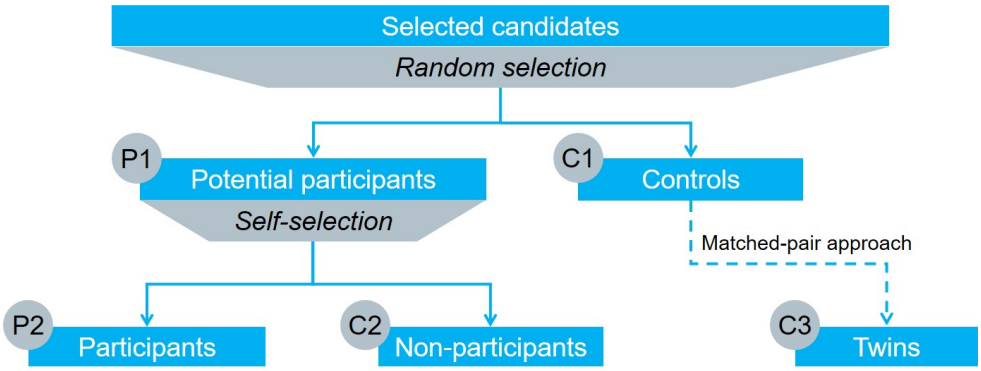

Figure 16: Different potential comparison groups for the (economic) evaluation of a DMP.

A straightforward approach for determining the effect of the DMP intervention is to compare the costs of potential participants (P1) – regardless of whether they actually took part – with the costs of all controls (C1). This approach follows the intention-to-treat principle which postulates that all patients who have been

allocated to the treatment group need to be included in the analysis, even if the treatment was not carried out in accordance with the study protocol. If the DMP has a cost-saving effect on participants, this effect will be underestimated because more "expensive" non-participants are included in the calculation (often the percentage of non-participants is greater than 50%). Conversely, if no or even a negative treatment effect exists, the intention-to-treat approach may spuriously suggest the existence of a saving effect because more "cheap" non-participants are included in the calculation [Porta et al., 2007]. If the evaluation is performed according to the complementary per-protocol or as-treated principle, only the actual participants (P2) are compared to the controls (C1). As no self selection takes place in the control group, patients in the intervention group may – on average – be more motivated to improve their health status. Because patient motivation plays a crucial role in the development of future costs and the success of the DMP [Kralik et al., 2004], this proceeding can also lead to a biased assessment of the treatment effect (self-selection bias). For the same reason, the direct comparison of participants (P2) and non-participants (C2) is even more inadequate. The intention-to-treat and the per-protocol approaches are both unweighted control group approaches, as no members of the control group are excluded or weighted higher than others (see Figure 15).

The so-called matched-pair approach aims to compensate for the self-selection bias in the P2-C1 comparison by restricting the set of controls considered. To determine the economic effect of the DMP, the costs of actual participants (P2) are compared with the costs of so-called "twins" or matched pairs (C3). Twins are controls who are sufficiently similar to the participants in terms of all (available) factors that are relevant to the patients' future cost development. In this way, the matched-pair approach ensures a balanced distribution of cost-relevant parameters between the intervention and control groups. An unbalanced distribution can occur if no stratification is carried out in the randomization process, or if cost-relevant parameters are not considered in the stratification. In the matched-pair approach, not all members of the control group will be used as twins and other controls will be used more than once. Consequently, the average costs of the allocated twins are a weighted average of the individual costs of all members of the control group. The weights are equal to zero for non-allocated persons and grow proportionally which each allocation. This is why the matched-pair approach belongs to the weighted control group approaches (see Figure 15).

A weighted control group approach closely related to the introduced matched-pair approach is the so-called "propensity scoring" method [Guo and Fraser, 2010]. Propensity scoring has been applied successfully for many years in clinical and observational studies [Hirano and Imbens, 2001]. The propensity score is the probability that an individual is allocated to the treatment group. This probability is modeled, normally by a logistic regression model, depending on all potential confounding factors that can be observed. In case the estimated propensity scores are unbalanced between participant and control group, a weighting of controls can be applied to establish a balance in both groups with regard to the confounding factors. For example, propensity score can be used as a criterion for matched-pair allocation [Imbens and Abadie, 2002]. If the decision of participation is taken by the pa-

tient, like in the DMP context, major confounding factors are the patients' health awareness and motivation. As important behavioral parameters such as smoking status or physical activity are usually not contained in insurance data, it is unclear whether propensity scoring or the matching approach based on cost prediction introduced in this chapter can control a potential self-selection bias. However, as such behavioral parameters are usually correlated with observed socio-demographic and medical covariates, the self-selection bias is indirectly controlled. In the existing practical situation, the matched-pair approach based on cost prediction has been preferred to classical propensity matching, because the available test datasets contain a large amount of categorical covariates (especially indicators for diagnosis occurrence) which destabilized the binary propensity models.

Another weighted control group approach which is applied in the health insurance market is illustrated in Figure 17. The approach seeks to control the self-selection effect by virtually reproducing this effect in the control group. The economic effect of the DMP is determined by comparing the costs c_P of actual participants (P2) to the costs c_{VP} of virtual participants (C4). In order to calculate the costs of virtual participants c_{VP}, two assumptions are made: first that the ratio between participants and non-participants r_P (selection ratio) is the same and, second, that the costs of non-participants c_N and virtual non-participants c_{VN} are the same. Thus, the costs of virtual participants c_{VP} can be calculated based on the observed costs of controls c_C and non-participants c_N using the following formula:

$$c_{VP} = \frac{c_C - c_N \cdot (1 - r_P)}{r_P}. \tag{3}$$

The formula is based on the assumption that the costs of controls c_C are a weighted average of the costs of virtual participants c_{VP} and virtual non-participants c_{VN}. The drawback of this approach is that actual participants have been asked to take part, unlike virtual participants. Even if no program measures have been carried out for non-participants, the act of simply contacting them can create awareness of their current health status and cause a change in behavior [Fishbein and Yzer, 2003]. The assumption that costs of actual and virtual non-participants are equal is therefore questionable. Assuming that the costs of the actual participants decrease compared with the costs of virtual participants as a result of the psychological effect of being contacted, the method underestimates the costs of virtual non-participants. At the same time, the costs of virtual participants are overestimated under this assumption, and so is the program effect. This theoretical drawback of the virtual participant method has been confirmed in a practical test based on health insurance data (compare the benchmarking analysis in Section 3.4.1).

Another (weighted or unweighted) control group approach is the so-called "postponed-treatment" method, which avoids patients being deprived from taking part. Here, one group of patients immediately receives program treatment, whereas another group is contacted but receives treatment only after a certain period of time. During this period, the patients with postponed treatment serve as control group. As with the virtual self-selection approach, the act of contacting the patients with post-

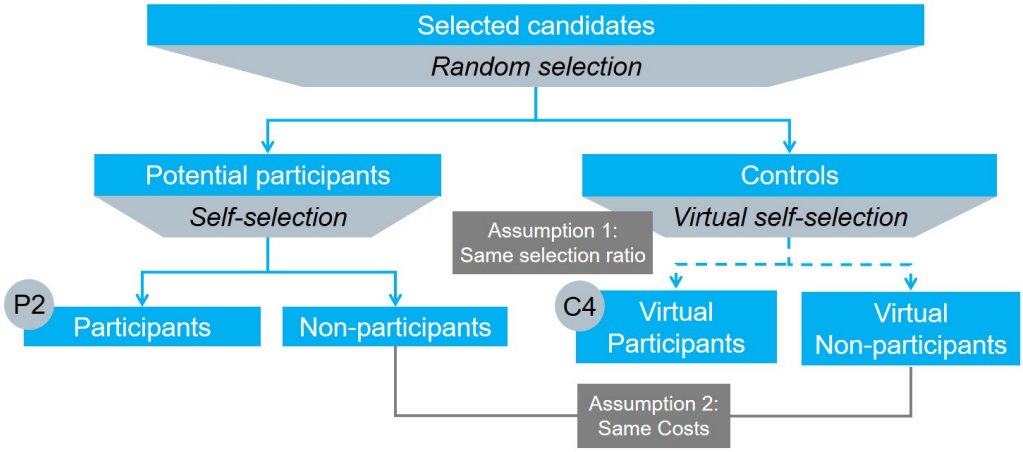

Figure 17: Another possible method of economic evaluation is to compare participants with virtual participants from the control group.

poned treatment may already cause a change in behavior which is accompanied by a cost reduction. In the case of the postponed-treatment approach, this leads to the program effect being underestimated. Moreover, the treatment cannot be postponed for a longer period of time, which means that only a short-term effect can be measured. Most chronic programs, however, are long-term measures that are designed to sustainably reduce costs. For those programs, an appropriate measurement technique needs to be able to determine the economic effect over a longer period of time.

Beside the various possibilities for defining comparison groups, different methods for calculating the economic effect are also conceivable (see Figure 15). In the literature, two major approaches are proposed for this purpose. The first method uses the difference in raw means or medians of observed costs (or of cost development) between both groups as an estimate of the average individual saving, from which the economic effect can be derived (see, for example, Berg and Wadhwa [2009], Esposito et al. [2008] or Henderson et al. [2013]). A comparison based on raw means is very sensitive to cost outliers in both groups, so the median or some kind of truncated mean is usually applied to control these outliers. Statistical tests can also be used to assess the significance of the cost effect. If raw means are used, a t-test for independent samples (and unknown sample variances) is the most appropriate way to determine the significance of the effect. If a truncated mean or median is used, the non-parametric Wilcoxon rank-sum test – also known as Mann-Whitney U test – is preferable. The matched-pair approach proposed in Section 3.3 uses the mean cost difference of participants and their twins to estimate the cost effect. As outliers in the cost difference are excluded, the resulting estimate can be seen as truncated mean.

The second method is based on a regression model that includes all cost-relevant parameters, plus a treatment parameter which indicates if an observation stems from the intervention or control group (see, for example, Conti [2011], Villagra and Tamim [2004] or Ninot et al. [2011]). A cumulative cost difference over several years

can be estimated by introducing a time-varying treatment effect and adding up the corresponding regression parameters over several years. The regression approach has the advantage that cost-relevant parameters are implicitly controlled which supports a risk-adjusted comparison. Moreover, an assessment of the significance of the effect is straightforward, because the standard tests (Wald test, score test, likelihood ratio test) on the significance of regression parameters can be used. For the approach proposed in Section 3.3, the mean comparison method is used to calculate the economic effect, because the regression method lead to an unstable economic effect over time (compare Section 3.4.1). However, the idea to control all cost-relevant parameters using a regression model in the matching process is applied. In this way, the advantages of both approaches are combined.

3.2.2 Theory and Application of the Matched-Pair Method

Rubin [2006] defines the matching approach in the following way: "Matched sampling is a method of data collection and organization designed to reduce bias and increase precision in observational studies, i.e. in those studies in which the random assignment of treatments to units (subjects) is absent. Rubin [2006] gives an extensive overview of scientific literature on the topic and includes many important articles on matching theory. Like many other statistical approaches, the matching approach has a long history, but its importance only grew with the increase in computing power in the last decade of the 20th century. At the start of the 1950s, Cochran [1953] was the first researcher to systematically study the matching approach, although some isolated articles on the topic did already exist in the 1930s, like the contribution from Wilks [9]. Nowadays, matched-pair sampling is a standard measurement approach for medical treatments in observational studies. In the DMP context, matching is used for measuring clinical outcomes like survival [Miksch et al., 2010], or economic outcomes like service utilization or costs [Berg et al., 2004].

As described by Rubin [2006], the aim of matching is to reduce the bias that arises from non-random allocation to treatment and control groups, like the described self-selection bias in the economic evaluation of insurer-driven DMPs. The basic assumption of all matching approaches is that this bias is caused by one or more influencing factors $\boldsymbol{x} = (x_1, \ldots, x_p)$ on the response y which are differently distributed in the treatment and control groups. The idea of all matching approaches is to balance the distribution of the influencing factors or matching variables $x_1, \ldots, x_p$ in both groups to reduce the bias, i.e. to carry out a covariate adjustment. This is done by allocating comparable/similar controls with regard to the matching variables to the persons in the treatment group. This reduces the bias in the comparison of y values between both groups. In a setting with only one matching variable, Rubin [2006] defines the bias B in the following way:

$$B = \frac{\mathbb{E}(x_C) - \mathbb{E}(x_P)}{\sqrt{\frac{\mathrm{Var}(x_C)+\mathrm{Var}(x_P)}{2}}}. \tag{4}$$

This means that the bias grows if the expectation of x differs significantly between participant (P) and control (C) groups, or if the variances in the groups are small.

Two types of matching approaches are generally differentiated [Rubin, 2006] – mean matching or balancing and individual or pair matching. The first type aims at minimizing the mean difference in matching variables between the treatment and control group, i.e. to reduce $|\bar{\boldsymbol{x}}_C - \bar{\boldsymbol{x}}_P|$. Individual or pair matching searches for a "twin" or matched pair for each participant that is as similar as possible with regard to the matching variables, i.e. it minimizes $\sum_{i=1}^{n_P} |\boldsymbol{x}_{C,i} - \boldsymbol{x}_{P,i}|$ for all persons $(i = 1, \ldots, n_P)$ in the treatment group. Mean matching is clearly simpler, but works only if the treatment effect is constant over the complete range of $\boldsymbol{x}$-values. However, this assumption does not hold in most practical situations, especially if more than one matching variable is considered. If the treatment effect varies with (one of) the matching variables, Rubin [2006] recommends using the individual matching approach in order to get a less biased estimate of the average treatment effect over the whole treatment group.

Based on these considerations, many more or less complex matching algorithms have been developed, like, for example, "nearest available neighbor matching", "caliper matching" or "radius matching" [Coca-Perraillon, 2007]. All matching algorithms can be divided into matching with replacement and matching without replacement. In the first case, each member of the control group can be assigned to different members of the treatment group, whereas in the latter case the controls can only be assigned to one member of the treatment group. Matching without replacement usually only works in situations where the control group clearly has more members than the treatment group. If the control group is too small, there will be no similar controls left for some elements of the treatment group and the bias increases because the numerator in Equation (4) grows. The drawback of matching with replacement is that the variance in the control group $\mathrm{Var}(x_C)$ decreases if the same control is allocated to many members of the treatment group. Here too, the bias grows because the denominator in Equation (4) decreases. Especially if the control group is smaller than the treatment group (as the insurer does not want to deprive too many clients from participation) matching without replacement is not meaningful.

Another differentiation can be made between 1:1 matching algorithms and 1:l matching algorithms. Allocating more than one control to each member of the treatment group (and averaging the controls' response) increases the bias because also less similar controls are allocated to members of the treatment group, especially if l is large. The allocation of several controls, however, has a stabilizing effect on matching and controls outliers in the response. In the literature on matching, the choice of l is heterogeneous and depends on the size of the control group (usual choices are $l = 4, 5, 8, 10, 16$, see Hollenbeck [2005], D'Agostino et al. [2001] and Blackman et al. [2010]). Most authors, like D'Agostino et al. [2001], Grant et al. [2012] and Fairfax et al. [1976], use $l = 4$ and argue that this matching ratio offers enough stability and hardly increases the bias, especially for small control groups.

The quality of matching algorithms can be measured and compared using the percentage of bias reduction, which Rubin [2006] defines as

$$100 \cdot \left(1 - \frac{\text{expected bias for matched sampling}}{\text{expected bias for random sampling}}\right). \tag{5}$$

The aim of matching algorithms is to maximize this quality criterion. For multiple matching variables (multivariate matching) Rubin [2006] postulates that optimal matching needs to be equal percent bias reducing. This means that the percentage of bias reduction is the same for each matching variable. In multivariate matching problems with continuous matching variables $\boldsymbol{x}$, the similarity between members of the treatment group and members of the control group can be quantified using a multivariate distance measure, like the Euclidean distance. For example, Rubin [1979] proposes a multivariate matching algorithm based on the Mahalanobis distance [Mahalanobis, 1936] which is equal percent bias reducing.

The regression-based matching approach for economic DMP evaluation introduced in Section 3.3 also considers the relevance of the matching variables $\boldsymbol{x}$ which can implicitly be derived from the underlying cost prediction model. The idea is to focus on balancing the distribution of those covariates with highest relevance for the outcome instead of equal percent bias-reduction. Consequently, the percentage of explained variance in future costs (measured by the predictive R-squared of the cost-prediction model, see Section 2) is used as a quality measure for the matching. This quality measure also considers that there may be unknown influence factors on y that are not considered and lead to a certain percentage of unexplained variance or hidden bias. Linden et al. [2006] discuss the problem of such hidden bias in the context of DMPs and propose a strategy to reduce it.

An extensive theoretical introduction to the very common propensity score matching method (see Section 3.2.1), which is also regression-based, can be found in Guo and Fraser [2010], Rubin [2006] and Dehejia and Wahba [2002]. Propensity score matching is applied by different authors in the DMP context, like Linden et al. [2005] and Stock et al. [2010].

A clear advantage of regression-based methods is that they can deal with covariates on different scales. Methods which require a distance calculation in the space of matching variables $\boldsymbol{x}$, like Rubin's Mahalanobis distance matching [Rubin, 1979], are usually only defined for continuous covariates. Statistical theory offers some approaches for calculating distances with regard to nominal and ordinal scaled covariates [Boriah et al., 2008]. However, it is difficult to aggregate distances of variables on different scales [Fahrmeir et al., 1996].

A recent development in matching theory is "coarsened exact matching" (CEM) [Iacus et al., 2012], the idea of which is to define the lower and upper bounds of allowed deviation for each matching variable. A control is assigned to a member of the treatment group if all x-values fall within the defined bounds. Members for which no controls can be found who fulfill this condition are excluded from the calculation of the treatment effect. The bounds can be estimated from the data but can also be determined from medical knowledge or best practice. This can be seen as a strength of the approach but can also be a weakness in situations where medical experience is missing, because the data-driven definition of bounds is clearly

less reliable. For example, Wells et al. [2013] have applied the CEM approach to measure the treatment effect of DMPs.

3.3 Methodology

In the following, the methodology of the matched-pair approach suggested for measuring the economic effect of disease management programs is presented. The general idea of the approach is to reduce the self-selection effect that distorts any direct cost comparison between participants $(p_1, \ldots, p_n)$ and controls $(c_1, \ldots, c_m)$, by allocating matched pairs or twins $(t_1, \ldots, t_{m'}$ with $m' \leq m)$ to every participant. The basis for the allocation of twins is the definition of a measure of similarity or distance between individuals. Sections 3.3.1 and 3.3.2 describe several measures of similarity in cost-relevant parameters based on cost-prediction models. Also, concrete allocation functions to match controls and participants based on their similarity are defined. In Section 3.3.3, it is described how the economic effect for programs that have already been carried out can be calculated based on the selected twins. Then, the (cumulative) average saving per participant based on the difference in average costs between participants and allocated twins is defined. Section 4.4 outlines how the financial risk of loss for programs that have already been carried out can be estimated in a reliable way. For this end, the variance of the economic effect needs to be estimated and an assumption about its distributional form is needed. Finally, conditions under which the financial risk of loss can be determined for future programs too, are defined as the basis for a reinsurance solution for DMPs.

As in most programs not every participant is included at the same point in time, the program time τ is defined. τ starts for every participant p_i with the time of the first contact or, if this date is not available, with the start of program measures (t_{start,p_i}), and ends with the time of dropout from the program (t_{end,p_i}). The program period usually ends after a fixed number of years $t_{\max}$. If this is not the case, for the purpose of stable measurement the end of the program period $t_{\max}$ is defined as the program year in which there are at least 100 participants under observation.

To assess the economic effect of a disease management program, the average annual saving per participant added up over all program years is quantified. This cumulative average saving up to program year τ $(\tau = 1, \ldots, t_{\max})$ is defined by

$$\bar{D}_\tau = \sum_{\theta=1}^{\tau} \bar{d}_\theta \tag{6}$$

where $\bar{d}_\tau$ is the average saving per participant in program year τ. $\bar{d}_\tau$ is defined by

$$\bar{d}_\tau = \frac{1}{n} \sum_{i=1}^{n} d_{p_i,\tau} \tag{7}$$

where $d_{p_i,\tau}$ is the individual saving for participant p_i.

The final economic effect $\bar{D}_\tau^c$ is obtained from subtracting the annual program costs c (which is assumed to be constant over the program years) from the cumulative average savings:

$$\bar{D}_\tau^c = \bar{D}_\tau - \tau \cdot c. \tag{8}$$

To evaluate the risk of financial loss related to the DMP, some assumptions on the distribution of the economic effect (see Section 4.4) are needed and an appropriate probability measure $\mathcal{P}$ needs to be defined. The financial risk of loss can then be written as

$$\mathcal{P}(\bar{D}_\tau^c < 0). \tag{9}$$

3.3.1 Cost Prediction Models as Basis of Distance Measures

The allocation of matched pairs is based on maximizing the similarity or minimizing the distance between participants and controls. In Section 3.3.2, two different ways of calculating distances are defined (see Figure 18). Both are based on a cost-prediction model for annual claimed amounts (see Section 2). For matching purposes, both a linear and a log-linear regression model are considered. Both models describe the relationship between future costs y_t and p covariates $x_{1,1}, \ldots, x_{t-1,1}$, $\ldots$, $x_{1,p}, \ldots, x_{t-1,p}$ captured from the start of the observation period up to $t-1$. More precisely, the linear model equation

$$\mathbb{E}(y_{a,t} | \underbrace{x_{a,1,1}, \ldots, x_{a,t-1,1}}_{:=\boldsymbol{x}_{a,t,1}}, \ldots, \underbrace{x_{a,1,p}, \ldots, x_{a,t-1,p}}_{:=\boldsymbol{x}_{a,t,p}}) = \underbrace{(\boldsymbol{x}_{a,t,1}, \ldots, \boldsymbol{x}_{a,t,p})}_{:=\boldsymbol{x}_{a,t}} \boldsymbol{\beta}_{\text{lin}} \tag{10}$$

is established where $a = p_1, \ldots, p_n, c_1, \ldots, c_m$ and

$$t = \begin{cases} t_{\text{firstobs},a}, \ldots, t_{\text{start},a} - 1 & \forall a \in \{p_1, \ldots, p_n\} \\ t_{\text{firstobs},a}, \ldots, t_{\text{lastobs},a} & \forall a \in \{c_1, \ldots, c_m\} \end{cases}. \tag{11}$$

$t_{\text{firstobs},a}$ and $t_{\text{lastobs},a}$ are the first and last year of the observation period for individual a and $t_{\text{start},a}$ is the year of the program start for participant a. This means that the models are based on all participants before the start of the intervention and all controls.

In parallel, the log-linear model equation

$$\log(\mathbb{E}(y_{a,t} | \boldsymbol{x}_{a,t}) + 1) = \boldsymbol{x}_{a,t} \boldsymbol{\beta}_{\log}. \tag{12}$$

is established. The model with log costs as response usually delivers the better predictive R-squared, whereas the linear model is more stable as regards outliers, which is a desirable property for matching.

In order to avoid overfitting and to identify only the cost-relevant covariates for matching, a stepwise variable selection approach [Miller, 1990] is applied. This leads to a reduction in the covariate vector from $(\boldsymbol{x}_{a,t,1}, \ldots, \boldsymbol{x}_{a,t,p})$ to $(\tilde{\boldsymbol{x}}_{a,t,1}, \ldots, \tilde{\boldsymbol{x}}_{a,t,p'}) := \tilde{\boldsymbol{x}}_{a,t}$ with $p' < p$. The covariates selected for the linear and the log-linear model may differ, of course, so that $\tilde{\boldsymbol{x}}_{\text{lin},a,t} \neq \tilde{\boldsymbol{x}}_{\text{log},a,t}$.

3.3.2 Allocation of Matched Pairs to Participants

The matched-pair approach allocates l_i twins to participant i $(i = 1, \ldots, n)$ based on similarity or distance. Mathematically, this allocation is a transformation s of the following form:

$$\begin{aligned} s \quad : \quad & \{p_1, \ldots, p_n\} \rightarrow \{c_1, \ldots, c_m\}^{l_i} \\ & p_i \mapsto s(p_i) = \{t_{i,1}, \ldots, t_{i,l_i} \mid \\ & t_{i,1}, \ldots, t_{i,l_i} \in \{c_1, \ldots, c_m\}; i = 1, \ldots, n\}. \end{aligned} \tag{13}$$

The strategy of allocating multiple twins to each participant increases the stability of the approach and reduces the probability of a "wrong" allocation which distorts the resulting cost comparison [Blackman et al., 2010].

The transformation s is based on a measure of distance $\delta(p_i, c_u)$ between a participant p_i and a control c_u $(i = 1, \ldots, n,\ u = 1, \ldots, m)$. As the matched-pair approach is based on a cost-prediction model for annual claimed amounts, it is straightforward to calculate annual distance measures $\delta_t(p_i, c_u)$ with $t = 1, \ldots, t_{i,u}$ and $t_{i,u}$ the number of years in which participant p_i and control c_u have been under common observation. Consequently, the transformation s must include some kind of aggregation of $\delta_t(p_i, c_u)$ over time t.

The concrete transformation function s is defined as

$$\begin{aligned} s_1(p_i) \quad := \quad & \{c_u \mid \text{rank}(\Delta(p_i, c_u)) \leq l = 4; \\ & i = 1, \ldots, n; u = 1, \ldots, m\} \end{aligned} \tag{14}$$

with

$$\Delta(p_i, c_u) = \begin{cases} \frac{1}{t_{i,u}} \sum_{t=1}^{t_{i,u}} \delta_t(p_i, c_u) & \forall t_{i,u} \geq 2 \\ \infty & \forall t_{i,u} < 2 \end{cases} \tag{15}$$

This means that the $l = 4$ controls that have the smallest average deviations over time (based on at least two years of common observation) are allocated to a participant p_i.

Alternatively,

$$s_2(p_i) := \left\{ c_u \,\middle|\, \sum_{t=1}^{t_{i,u}} \mathbb{1}(\text{rank}(\delta_t(p_i, c_u)) \leq k) \geq 2; \quad i = 1, \ldots, n; u = 1, \ldots, m \right\} \tag{16}$$

is defined where $\mathbb{1}(\cdot)$ is an indicator function that is equal to one if the condition in brackets is fulfilled and 0 if not. This means that a valid twin must be among the k controls having the smallest annual deviations from the participant in at least two years. k is a hyperparameter that directly steers the number of allocated controls l_i for each participant. k depends on the number of controls. Based on tests of the robustness of the approach (see Section 3.4.2), k should be chosen so that approximately 80% of all participants have at least one twin allocated.

For the first transformation s_1, $l_i \equiv l = 4$ is used (1:4-matching, see D'Agostino et al. [2001], Grant et al. [2012] and Fairfax et al. [1976]), and for the second transformation s_2, $l_i \in \{0; \ldots; m\}$ is used which is more flexible, especially if participants have very few controls that are sufficiently similar. In an extreme case, no twins are allocated to a participant, in which case the participant is excluded from the cost comparison. The strategy of excluding participants from the cost comparison for which no adequate matched pair can be found increases the stability of the approach. It is also in line with recent findings from authors like Iacus et al. [2012] who deal with the matching approach.

Finally, two concrete methods $\delta_t(p_i, c_u)$ of annual distance calculation between p_i and c_u are defined based on the regression models (12) and (10) introduced in Section 3.3.1. Both methods use a transformation of the covariates that were identified as relevant for future costs through the applied variable selection approach. Figure 18 gives a schematic overview of distance calculation based on transformed covariates and on the ideas behind each transformation step.

The first method directly uses the distance in predicted costs between participant p_i and control c_u in year t ($t = 1, \ldots, t_{i,u}$). These predicted costs $y^*_{a,t}$ are a linear or log-linear function of all cost-relevant covariates, i.e.

$$\begin{aligned} \delta_t^{1,\text{lin}}(p_i, c_u) &= |y^{*,\text{lin}}_{p_i,t} - y^{*,\text{lin}}_{c_u,t}| = \\ &\quad |\tilde{\boldsymbol{x}}^T_{\text{lin},p_i,t}\boldsymbol{\beta}_{\text{lin}} - \tilde{\boldsymbol{x}}^T_{\text{lin},c_u,t}\boldsymbol{\beta}_{\text{lin}}| \end{aligned} \tag{17}$$

$$\begin{aligned} \delta_t^{1,\log}(p_i, c_u) &= |y^{*,\log}_{p_i,t} - y^{*,\log}_{c_u,t}| = \\ &\quad |\exp(\tilde{\boldsymbol{x}}^T_{\log,p_i,t}\boldsymbol{\beta}_{\log}) - \exp(\tilde{\boldsymbol{x}}^T_{\log,c_u,t}\boldsymbol{\beta}_{\log})| \end{aligned} \tag{18}$$

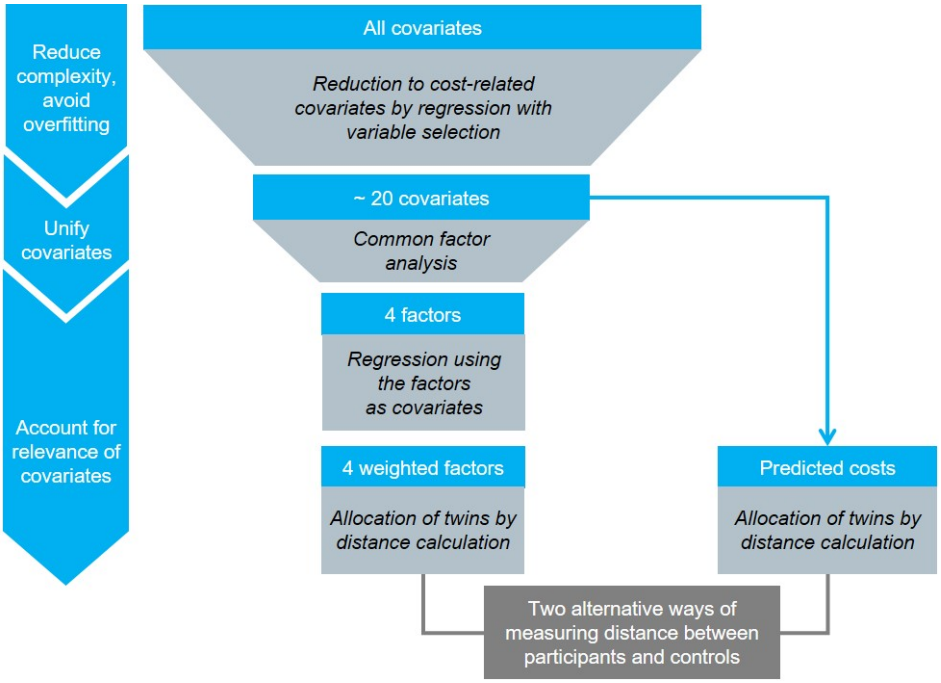

Figure 18: Schematic overview of both covariate-based distance calculation methods.

The idea of the second method is to directly calculate the distance between values of the cost-relevant covariates. However, those covariates are not on the same scale level. Especially for categorical covariates, it is difficult to measure the distance between categories. This is why a common factor analysis is applied to reduce the dimension of the covariate space and obtain a number q of continuous factors to which classical distance measures, like the Euclidean distance, can be applied. The reduction in dimension is, of course, always related to a loss of information. The principle of factor analysis [Thompson, 2004] is to minimize this loss of information by preserving as much as possible of the covariance included in the original covariates. Considering a scree plot, and looking at the variance explained by the factors, a reduction of the covariate space to $q = 4$ continuous factors is considered a reasonable choice for the underlying datasets. In order to obtain orthogonal factors varimax factor rotation [Thompson, 2004] is applied. In this way, four continuous factor values $(\tilde{\boldsymbol{f}}_{a,t,1}, \ldots, \tilde{\boldsymbol{f}}_{a,t,4}) := \tilde{\boldsymbol{f}}_{a,t}$ are obtained for each individual and year.

As the factor analysis reduces the dimension without considering the covariates' impact on annual costs $y_{a,t}$, these four factors do not explain the same percentage of the variance of future costs. Consequently, weighting the factors according to their predictive power is suggested based on new regression models using the four factors as covariates:

$$\mathbb{E}(y_{a,t}|\tilde{\boldsymbol{f}}_{a,t}) = \tilde{\boldsymbol{f}}_{\text{lin},a,t}\boldsymbol{\gamma}_{\text{lin}} \tag{19}$$

$$\log(\mathbb{E}(y_{a,t}|\tilde{\boldsymbol{f}}_{a,t}) + 1) = \tilde{\boldsymbol{f}}_{\log,a,t}\boldsymbol{\gamma}_{\log}. \tag{20}$$

Where cost-relevant covariates have been determined by means of a linear regression model, also a linear regression model (19) is used to weight the resulting factors $\tilde{\boldsymbol{f}}_{\text{lin},a,t}$. Accordingly, if a log-linear model has been used to determine the cost-relevant covariates, a log-linear model (20) is applied for weighting the resulting factors $\tilde{\boldsymbol{f}}_{\text{log},a,t}$.

Finally, the standardized regression coefficients $\gamma_b^{st} = |\gamma_b / \operatorname{se}(\gamma_b)|$ ($b = 1, \ldots, 4$, $\operatorname{se}(\gamma_b)$ denotes the standard error of regression coefficient γ_b) from the regression models (19) and (20) are used as weights for (Euclidean) distance calculation:

$$\delta_t^{2,\text{lin}}(p_i, c_u) = \sqrt{\sum_{b=1}^{4} \gamma_{\text{lin},b}^{st} (\tilde{f}_{p_i,t,b} - \tilde{f}_{c_u,t,b})^2} \quad \text{and} \tag{21}$$

$$\delta_t^{2,\text{log}}(p_i, c_u) = \sqrt{\sum_{b=1}^{4} \gamma_{\text{log},b}^{st} (\tilde{f}_{p_i,t,b} - \tilde{f}_{c_u,t,b})^2}. \tag{22}$$

The two measures of distance calculation δ_t^1 and δ_t^2 are only based on covariates $\boldsymbol{x}_t := x_1, \ldots, x_{t-1}$ that have an impact on future costs y_t (see the definition of the prediction model (10)). This means that no matching by outcome, but a matching by drivers of the future outcome takes place. This is a basic principle of the matching theory [Rubin, 2006] which ensures that only individuals with similar expected future cost development are matched.

This is especially important, as not indication-related claimed amounts are modeled in the described setting but total claimed amounts. Consider, for example, two insureds suffering from diabetes: insured A with a well-controlled status of disease but a severe car-accident in year t, and insured B with an uncontrolled status of disease and several co-morbidities in year t. Both insureds have similarly high claims costs in year t. Insured A's claims costs will very likely decrease in the next few years, however, whereas patient B is expected to have at least constantly high costs over the next few years. If current costs were used as a criterion for matching, patient B would be allocated to patient A and the estimate of the program effect would be distorted. The integration of a time component, i.e. predicting annual costs and measuring similarity over time, ensures that participants are only compared to controls with a similar progress of the chronic disease.

In practical tests of the methodology it turned out that the application of transformations s_1 and s_2 with defined distance measures that do not include actual costs y_t has some drawbacks. In particular, outliers cannot be controlled sufficiently without considering actual costs before the start of intervention. This is why one exception from the rule of covariate-based allocation is made. To ensure at least a similar starting point, a criterion for outlier control is defined based on the cost difference between participants and controls prior to the start of the intervention.

First, the set of all assignable controls $\mathcal{C}_i$ for participant p_i is defined:

$$\mathcal{C}_i := \Big\{ c_u \Big| \max_{t=1,\ldots,t_{i,u}} |y_{p_i,t} - y_{c_u,t}| < b; \quad i = 1,\ldots,n; u = 1,\ldots,m \Big\}. \tag{23}$$

This means that controls can only be allocated to a participant if the raw cost difference prior to the start of the intervention has not exceeded the limit b. In this way outliers can be controlled more efficiently and do not have to be excluded completely. Different values of b have been tested. $b = \bar{y}_t$ with $\bar{y}_t$ the average annual costs of all participants before the start of the intervention and all controls seems to be a reasonable boundary (see Section 3.4.2).

The transformations $\tilde{s}_1$ and $\tilde{s}_2$ denote the modifications of s_1 and s_2 so that only assignable controls from $\mathcal{C}_i$ are considered in the allocation process:

$$\tilde{s}_1(p_i) := \big\{ c_u \in \mathcal{C}_i \,|\, \text{rank}(\Delta(p_i, c_u)) \leq l = 4; \quad i = 1,\ldots,n \big\} \text{ and} \tag{24}$$

$$\tilde{s}_2(p_i) := \left\{ c_u \in \mathcal{C}_i \middle| \sum_{t=1}^{t_{i,u}} \mathbb{1}(\text{rank}(\delta_t(p_i, c_u)) \leq k) \geq 2; \quad i = 1,\ldots,n \right\}. \tag{25}$$

3.3.3 Cost Comparison – Participants vs. Matched Pairs

In order to obtain an estimate of the average cumulative savings, we estimate the individual saving of participant p_i in program year τ ($\tau = 1,\ldots,t_{\max}$)

$$\hat{d}_{p_i,\tau} = \underbrace{\frac{1}{l_i} \sum_{c=1}^{l_i} y_{s_c(p_i),\tau}}_{\bar{y}_{s_c(p_i),\tau}} - y_{p_i,\tau} \tag{26}$$

where $s_c(p_i)$ denotes the c-th element of the set of allocated controls for participant p_i based on the allocation s.

It is now straightforward to insert this estimate of the individual saving $d_{p_i,\tau}$ into the definition of the average saving $\bar{d}_\tau$ (see Equation (7)). However, it has to be considered that for the transformation $\tilde{s}_2$ defined in (25) not every participant will have allocated twins. Moreover, not every participant will stay in the program for the same period of time, but could potentially drop out before program year τ. This is why the index set

$$\mathcal{R}_\tau := \{i | s(p_i) \neq \emptyset \wedge \tau <= t_{\text{end},p_i}; i = 1,\ldots,n\} \tag{27}$$

is defined where t_{end,p_i} is the year in which participant p_i leaves the program. $|\mathcal{R}_\tau|$ is the cardinality of the set $\mathcal{R}_\tau$ or the number of participants in program year τ with at least one allocated twin. Then, the estimate for $\bar{d}_\tau$ is defined in the following way:

$$\begin{aligned}\hat{\bar{d}}_\tau &= \frac{1}{|\mathcal{R}_\tau|}\sum_{r\in\mathcal{R}_\tau}\hat{d}_{p_r,\tau} \\ &= \frac{1}{|\mathcal{R}_\tau|}\sum_{r\in\mathcal{R}_\tau}(\bar{y}_{s(p_r),\tau} - y_{p_r,\tau}) \\ &= \underbrace{\frac{1}{|\mathcal{R}_\tau|}\sum_{r\in\mathcal{R}_\tau}\bar{y}_{s(p_r),\tau}}_{:=S_\tau} - \underbrace{\frac{1}{|\mathcal{R}_\tau|}\sum_{r\in\mathcal{R}_\tau}y_{p_r,\tau}}_{:=P_\tau}.\end{aligned} \tag{28}$$

This means that $\hat{\bar{d}}_\tau$ can be written as the difference between the averages P_τ and S_τ, where P_τ are the average costs of the participants in program year τ (with at least one allocated twin) and S_τ are the average costs of the "virtual twins" that result from averaging the costs of all allocated twins per participant.

At the same time, S_τ can be seen as a weighted average of the costs in the control group. The weighting is meant to make the control group more comparable to the participant group by imitating the participants' risk structure. This works by assigning a higher weight to the controls that are similar to many of the participants. However, the weighting factor depends on the way the average costs of the assigned controls are calculated.

In the following, two alternative ways of weighting the assigned controls are defined which leads to two alternative estimates for $\bar{d}_\tau$.

The first estimate $\hat{\bar{d}}^+_\tau$ leads to a higher weighting of controls who are similar to many participants. It is calculated by replacing S_τ in Equation (28) with S^+_τ where

$$S^+_\tau := \frac{1}{\sum_{r\in\mathcal{R}_\tau} l_r}\sum_{r\in\mathcal{R}_\tau}\sum_{c=1}^{l_r} y_{s_c(p_r),\tau}. \tag{29}$$

The second estimate $\hat{\bar{d}}^-_\tau$ leads to a lower weighting of controls who are similar to many participants. $\hat{\bar{d}}^-_\tau$ is obtained by replacing S_τ in Equation (28) with S^-_τ where

$$S^-_\tau := \frac{1}{|\mathcal{S}|}\sum_{v\in\mathcal{S}} y_{c_v,\tau}. \tag{30}$$

The index set $\mathcal{S}$ is defined as

$$\mathcal{S} := \left\{ v \mid c_v \in \bigcup_{r \in \mathcal{R}} s(p_r); v = 1, \ldots, m \right\} \tag{31}$$

and $|\mathcal{S}|$ is the cardinality of $\mathcal{S}$.

Both estimates $\hat{\bar{d}}_\tau^+$ and $\hat{\bar{d}}_\tau^-$ yield similar results to $\hat{\bar{d}}_\tau$ for all programs for which the estimates were calculated (see Section 3.4.2). Considering these results and the intuitive derivation given in Equation (28), it is recommended using $\hat{\bar{d}}_\tau$ as an estimate for the average saving per participant, which implies a "medium" weighting for controls who are similar to many participants.

Finally, an estimate for the cumulative average saving per participant $\hat{\bar{D}}_\tau$ is calculated by summing up the estimates over all program years:

$$\hat{\bar{D}}_\tau = \sum_{\theta=1}^{\tau} \hat{\bar{d}}_\theta. \tag{32}$$

If the transformations $\tilde{s}_1$ and $\tilde{s}_2$ are combined with the four different distance measures $\delta_\tau^{1,\text{lin}}$, $\delta_\tau^{1,\log}$, $\delta_\tau^{2,\text{lin}}$ and $\delta_\tau^{2,\log}$, eight different estimates $\hat{\bar{d}}_\tau^1, \ldots, \hat{\bar{d}}_\tau^8$ (see Figure 19 for an overview) are obtained which lead to broadly consistent saving curves (see Section 3.4.1). Because there is no way of checking which of these eight estimates is closest to the real underlying savings, it is suggested using the median of all eight methods $\hat{\bar{d}}_\tau^{\text{med}}$ as the final estimate. This procedure clearly increases the stability of the approach and reduces the probability of a substantially wrong estimate.

3.3.4 Distribution of Savings and Financial Risk of Loss

The results shown in Section 3.4 indicate that the defined estimate $\hat{\bar{D}}_\tau^c$ is a reliable estimate for the economic effect of a DMP. For this reason, it is suggested measuring this effect using the matched-pair approach described, at least in situations where no randomized controlled trial is feasible. Accepting the proposed measurement approach as a standard procedure implies that the defined estimate is equated with the actual economic effect ($\bar{D}_\tau^c \equiv \hat{\bar{D}}_\tau^c$) and that the only uncertainty included in this effect is the uncertainty of the measurement method. This also means that the question of the program's profitability is answered by the question whether the estimate $\hat{\bar{D}}_\tau^c$ is greater than zero. In this situation, the uncertainty of the approach needs to be analyzed for two reasons:

1. For programs which have already been carried out, the uncertainty of the measurement approach directly determines the financial risk of loss related to the program. In this case, the financial risk of loss corresponds to the statistical type I error, which is the probability of a DMP being rated as profitable though it actually is not.

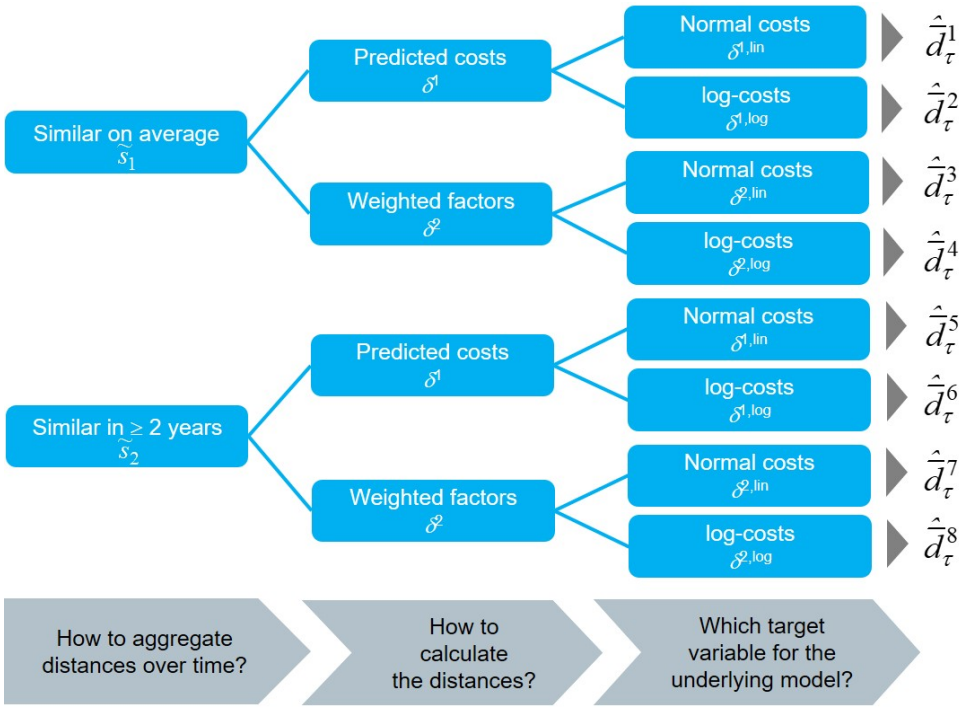

Figure 19: Overview of all defined matched-pair methods.

2. For future programs that are similar to ones that have already been carried out, the risk of financial loss can be predicted based on the relationship between the uncertainty of the approach and the financial risk of loss. The form of this relationship needs to be derived from programs that have already been carried out.

In order to estimate the financial risk of loss of a DMP $\mathcal{P}(\hat{\bar{D}}_\tau^c < 0)$, an estimate for the variance of the (estimate for) the economic effect $\hat{\bar{D}}_\tau^c$ needs to be defined and an assumption about its distributional form needs to be made.

$\hat{\bar{D}}_\tau$ is an average of n independent random variables with existing first and second moments. Therefore it is assumed, based on the central limit theorem [Grimmett and Stirzaker, 1982], that $\hat{\bar{D}}_\tau$ is asymptotically normal distributed. In other words, if the number of participants is large enough, the assumption of a normal distribution holds. This is why the normal distribution $\mathcal{N}$ can be used as an approximation of the unknown probability measure $\mathcal{P}$. In order to estimate the loss risk based on the probability function of the normal distribution, the parameters of the normal distribution μ_τ and σ_τ^2 need to be estimated which represent the expectation $\mathbb{E}(\hat{\bar{D}}_\tau)$ and the variance $\text{Var}(\hat{\bar{D}}_\tau)$ of the distribution, respectively. Because the annual program costs c and the program time τ are deterministic, the economic effect $\hat{\bar{D}}_\tau^c$ is also (asymptotically) normal distributed with parameters $\mu_\tau - \tau \cdot c$ and σ_τ^2.

As the estimate for μ_τ, the median estimate described in Section 3.3.3 is used, i.e.

$$\hat{\mu}_\tau = \hat{\bar{D}}_\tau^{\text{med}} = \sum_{\theta=1}^{\tau} \hat{\bar{d}}_\theta^{\text{med}}. \tag{33}$$

Estimating the variance σ_τ^2 is more complex. Starting point is again an estimate for the variance of the average annual cost difference per participant $\hat{\bar{d}}_\tau$:

$$\begin{aligned} \widehat{\text{Var}}(\hat{\bar{d}}_\tau) &= \widehat{\text{Var}}(S_\tau - P_\tau) \\ &= \widehat{\text{Var}}(S_\tau) + \widehat{\text{Var}}(P_\tau) - 2 \cdot \widehat{\text{Cov}}(S_\tau, P_\tau). \end{aligned} \tag{34}$$

The estimates for the single components of $\widehat{\text{Var}}(\hat{\bar{d}}_\tau)$ are defined in the following way:

$$\begin{aligned} \widehat{\text{Var}}(S_\tau) &= \frac{1}{|\mathcal{R}_\tau|}\left(\frac{1}{|\mathcal{R}_\tau|-1} \right. \\ &\qquad \left. \sum_{r\in\mathcal{R}} (\bar{y}_{s(p_r),\tau} - \bar{\bar{y}}_{s(p_r),\tau})^2 \right), \end{aligned} \tag{35}$$

$$\widehat{\text{Var}}(P_\tau) = \frac{1}{|\mathcal{R}_\tau|}\left(\frac{1}{|\mathcal{R}_\tau|-1} \sum_{r\in\mathcal{R}} (y_{p_r,\tau} - \bar{y}_{p_r,\tau})^2 \right), \tag{36}$$

$$\begin{aligned} \widehat{\text{Cov}}(S_\tau, P_\tau) &= \frac{1}{|\mathcal{R}_\tau|}\left(\frac{1}{|\mathcal{R}_\tau|-1} \right. \\ &\qquad \left. \sum_{r\in\mathcal{R}} (\bar{y}_{s(p_r),\tau} - \bar{\bar{y}}_{s(p_r),\tau})(y_{p_r,\tau} - \bar{y}_{p_r,\tau}) \right). \end{aligned} \tag{37}$$

where $\bar{y}_{p_r,\tau}$ are the average costs of all participants and $\bar{\bar{y}}_{s(p_r),\tau}$ are the average costs of all "virtual twins" in year τ. In order to derive the estimates (35) and (36), it is assumed that the costs of single participants and single "virtual twins" are independent and have the same variance. In parallel, it is assumed for estimate (37) that the costs of a participant and of a "virtual twin" allocated to another participant are independent and that the covariance between participant and allocated "virtual twin" is the same for all pairs. It has to be borne in mind that the assumptions of independence may be questionable for the "virtual twins", as the same control can be part of different "virtual twins". However, in the practical settings presented in Section 3.4, only a few controls were allocated to multiple participants. For this reason, the dependencies between "virtual twins" are not expected to significantly influence the variance of the estimated average savings per participant $\widehat{\text{Var}}(\hat{\bar{d}}_\tau)$ and are ignored.

Based on the estimated variance of the annual average savings per participant, an estimate of the cumulative average savings per participant is obtained by plug-in estimation:

$$\begin{aligned}\hat{\sigma}_\tau^2 &= \widehat{\mathrm{Var}}(\hat{\bar{D}}_\tau) = \widehat{\mathrm{Var}}\left(\sum_{\theta=1}^{\tau}\hat{\bar{d}}_\theta\right) \\ &= \sum_{\theta=1}^{\tau}\widehat{\mathrm{Var}}(\hat{\bar{d}}_\theta) + 2\cdot\sum_{\substack{\theta,\theta'=1\\ \theta<\theta'}}^{\tau}\widehat{\mathrm{Cov}}(\hat{\bar{d}}_\theta,\hat{\bar{d}}_{\theta'}).\end{aligned} \tag{38}$$

As the annual savings in the individual program years are obviously correlated, the covariances $\mathrm{Cov}(\hat{\bar{d}}_\theta, \hat{\bar{d}}_{\theta'})$ are certainly not negligible. An estimate of these covariances is defined in Appendix B.

Based on the eight different allocation functions of twins defined in Section 3.3.3, eight different estimates for the variance are obtained. The stabilizing property of the median leads to the fact that the variance of the median potential savings will be smaller than the median of the single variances. Nevertheless it is proposed using the median of the single variances as the estimate for σ_τ^2, because, from an insurer's perspective, an underestimation of the variance – and thus of the risk of financial loss – is clearly more disadvantageous than an overestimation.

With the estimates for μ_τ and σ_τ^2, the financial risk of loss $\mathcal{P}(\hat{\bar{D}}_\tau^c < 0)$ for programs that have already been carried out can be calculated:

$$\begin{aligned}\hat{\mathcal{P}}(\hat{\bar{D}}_\tau^c < 0) &= \mathcal{N}(\hat{\bar{D}}_\tau^c < 0) \\ &= \mathcal{N}(\hat{\bar{D}}_\tau - \tau\cdot c < 0) \\ &= \mathcal{N}\left(\frac{\hat{\bar{D}}_\tau - \hat{\mu}_\tau}{\hat{\sigma}_\tau} < \frac{\tau\cdot c - \hat{\mu}_\tau}{\hat{\sigma}_\tau}\right) \\ &= \Phi\left(\frac{\tau\cdot c - \hat{\mu}_\tau}{\hat{\sigma}_\tau}\right)\end{aligned} \tag{39}$$

where Φ is the probability function of the standard normal distribution and $\hat{\sigma}_\tau$ is the square root of the defined estimate $\hat{\sigma}_\tau^2$. As mentioned above, $\hat{\mathcal{P}}(\hat{\bar{D}}_\tau^c < 0)$ can be interpreted as the probability of the type I error, i.e. the probability of assuming the program to have a positive economic effect, even though it does not. The distributional assumptions made for $\hat{\bar{D}}_\tau^c$ can also be used to derive the probability of the type II error, i.e. wrongly assuming that the program results in a negative financial outcome. In this context, however, the focus is on the type I error, because a program with an expected negative economic effect $\hat{\bar{D}}_\tau^c$ would not be considered for a reinsurance solution.

μ_τ and σ_τ^2 cannot directly be estimated for programs that have not yet been carried out. For the prediction of the financial risk of loss for such new programs, the economic effect and the relationship between the uncertainty of measurement and the financial risk of loss need to be transferred from a known program. Such transfer

is only possible if the future program is sufficiently comparable in terms of program design and underlying portfolio to the one already carried out and measured. In addition, the uncertainty of the measurement that depends on several factors – like data quality, length of claims history and portfolio size – needs to be quantifiable for the new program.

All in all, the following three conditions need to be fulfilled to reliably predict the financial risk of loss of a new program and determine its reinsurability:

i) The future program has the same design (same target indication, same inclusion and exclusion criteria for participants and controls, and same program measures) as one which has already been carried out and measured using the described matched-pair approach.

ii) The group of potential participants (before self-selection) in the new program is similar to that in the one already carried out and measured with regard to all cost-relevant parameters (determined by the regression model).

iii) Historical member and claims data on potential participants are available over a period of at least three years, which means that the predictive quality of a cost-prediction model and thus the uncertainty of the measurement method can be evaluated.

Condition i) and ii) justify the assumption that a similar outcome $\hat{\bar{D}}_\tau$ to that obtained in the program already carried out will be realized in the future program, which means that $\hat{\bar{D}}_\tau$ from the one already carried out can be used as an estimate for μ_{D_τ}.

Condition iii) makes it possible to quantify the uncertainty of the measurement for the future program, which is, in the described setup, directly related to the predictive quality of the underlying cost-prediction model. As a measure of predictive quality, we use the predictive R-squared R^{2^*}, which is the squared correlation between the values predicted and those actually observed in a back-testing scenario (see Section 2 for further details on measures of predictive quality). The predictive R-squared ranges between 0 and 1 and describes the percentage variance of future costs that can be explained by the model.

For future programs, $\mathrm{Var}(\hat{\bar{d}}_\tau)$ cannot directly be estimated. Therefore, the relationship between the uncertainty of the measurement represented by R^{2^*} and the financial risk of loss driven by the variance $\mathrm{Var}(\hat{\bar{d}}_\tau)$ needs to be quantified based on programs that have already been carried out. Then, only the predictive quality R^{2^*} needs to be determined based on the new portfolio and the derived relationship can be applied to predict the risk of loss for the new program. The connection between the predictive R-squared and the variance of the estimated average savings per participant becomes obvious looking at the derivation of $\mathrm{Var}(\hat{\bar{d}}_\tau)$ in (38):

$$\mathrm{Var}(\hat{\bar{d}}_\tau) = \mathrm{Var}(S_\tau) + \mathrm{Var}(P_\tau) - 2 \cdot \mathrm{Cov}(S_\tau, P_\tau). \tag{40}$$

The higher the percentage of explained variance in future costs (i.e. the predictive R-squared) or as better the matching, the higher is the covariance between the average costs of participants P_τ and the average costs of "virtual twins" S_τ. The variances of S_τ and P_τ are not directly influenced by R^{2^*}, however. This shows that both $\text{Var}(\bar{d}_\tau)$ and σ_τ^2 decrease with the increase in the predictive R-squared R^{2^*}. Thus, σ_τ^2 can be written as a (monotonically decreasing) function f of predictive R-squared:

$$\sigma_\tau^2 = f(R^{2^*}). \tag{41}$$

As the functional form f of the relationship between σ_τ^2 and R^{2^*} is quite complex, it cannot be derived analytically. Therefore, an estimate of the functional form $\hat{f}$ is needed for estimating the variance of the cumulative savings of future programs σ_τ^2. In Section 3.4.3 an estimate $\hat{f}$ is simulated based on the results of three programs which have already been carried out.

If all three preconditions i)-iii) are fulfilled, the approach described can be used to reliably estimate the financial risk of loss $\mathcal{P}(\bar{D}_\tau^c < 0)$ for future programs, and their reinsurability can be determined based on $\hat{\mathcal{P}}(\hat{\bar{D}}_\tau^c < 0)$. A program can, of course, only be reinsured if $\hat{\mathcal{P}}(\hat{\bar{D}}_\tau^c < 0)$ is sufficiently small.

In situations where the above conditions i) and ii) are only partially fulfilled – for example where the inclusion criteria or the distribution of cost-relevant parameters differ between old and new programs – the risk of financial loss can grow by these additional uncertainties. It is then advisable to additionally increase the estimated variance by a multiplicative factor v larger than 1 in order to take account of the additional uncertainty:

$$\hat{\sigma}_{\tau,v}^2 = v \cdot \hat{f}(R^{2^*}). \tag{42}$$

A correction for additional uncertainty is only reasonable up to a certain point. For example, it is not possible to apply conclusions from a known program to a new program if the programs have different indications (compare the different results for chronic heart failure and diabetes in Section 3.4). Also, the correction of the financial risk of loss is only meaningful if the base risk is sufficiently small (above a base risk of 50%, the correction has the opposite effect).

3.4 Results for Different DMPs

To test the measurement approach described in Section 3.3, datasets from two Central European insurance companies are used. Those two companies provided claims data on their insureds over a timeframe longer than five years and program data on three different programs, two for diabetes and one for chronic heart failure (CHF). In the following, several details on the programs are provided, especially inclusion and exclusion criteria, as well as program measures.

Number of participants			
τ	n (Diab. I)	n (Diab. II)	n (CHF)
0	1,358	8,738	600
1	1,100	8,620	548
2	944	7,953	479
3	770	3,231	386
4	665	–	305
5	574	–	273
6	368	–	219

Number of controls		
$m_{\min}$ (Diab. I)	$m_{\min}$ (Diab. II)	$m_{\min}$ (CHF)
52,757	1,801	35,218

Table 9: Number of participants n for the analyzed programs per program year τ and minimum number of available controls $m_{\min}$ within the program phase (without deceased persons).

For the first insurance company (company A), a diabetes and a chronic heart failure program was evaluated. Both programs are based on disease-specific telecoaching, i.e. regular phone calls to check health status, and the telemonitoring of disease-related parameters (like blood pressure or blood sugar). The inclusion criteria for the programs were either a confirmed diagnosis of diabetes type II (ICD-10 codes: E11 to E14) or of chronic heart failure (ICD-10 codes: I50 or I11.0). In addition, only insured persons in certain tariffs were included. The most important exclusion criteria for both programs were:

- age younger than 40 or older than 75 years (for diabetes) and age older than 85 years (for CHF),
- existence of other diseases (e.g. acute myocardial infarction within last 6 months, need for dialysis, AIDS, cancer, alcohol or drug dependency),
- inability to cooperate or communicate due to other diseases, disabilities or care needs,
- simultaneous participation in another DMP,
- cancellation of contract during observation period and
- delay in payment.

For the second insurance company (company B), the economic effect of a diabetes program was analyzed. The program aims at improving the insured persons' health status by optimizing the medical support and infrastructure of those persons. The basic inclusion criteria were inpatient and/or outpatient diagnoses of diabetes and the prescription of specific diabetes medication. Basic exclusion criteria were participation in other healthcare programs, age under 19 or over 89 years, need for care and need for dialysis. Out of all the insureds matching these basic criteria, those with the highest likelihood of hospitalization (which is very cost-intensive) were selected in order to determine all candidates for the DMP.

Table 9 gives an overview of the number of actual participants n per program year τ and the total number of controls m who match all inclusion criteria. Company A's

diabetes program (denoted as Diabetes I) starts with $n = 1{,}358$ participants in the year of the first program call (defined as $\tau = 0$) and has more than 500 participants after 5 program years. The chronic heart failure program (denoted as CHF) includes $n = 600$ participants in the year of the first program call (defined as $\tau = 0$), of which more than 50% are still under observation in program year $\tau = 4$. The last year that can be reliably evaluated for both programs is program year $\tau = 6$. For both programs a large number of controls is available (at least $m_{\min} = 35{,}218$ for diabetes and at least $m_{\min} = 29{,}342$ for CHF).

The selection process for company B's diabetes program (denoted as Diabetes II) leads to a participant group of $n = 8{,}738$ in the year of the first DMP-related contact with the insured person (defined as $\tau = 0$). The last year of evaluation is program year $\tau = 3$, in which there are still more than 3,000 participants under observation. The minimum number of available controls within the program phase is $m_{\min} = 1{,}801$.

The following information was used as an input for the cost-prediction models on which the matched-pair allocation is based (see Section 2 for more information on the construction of cost-prediction models):

- information on the insured person (like age and gender),
- contract information (like years since start of contract, deductibles and scope of cover) and
- claims history (like previous diagnoses, procedures and claims costs).

For evaluating all three programs, people for whom substantial information, like gender or age, was not available and for which the matching algorithm could therefore not be applied were excluded (only a negligible number of individuals were affected). Also, the programs were evaluated twice. First, people who died during the observation period, were included and then excluded to avoid distorting the economic effect caused by strongly increased claims costs in the months before death.

It is important to note that the cost-prediction models used for allocating twins are based on the claimed amount per individual instead of on the amount actually paid by the insurer. The reason for this is that paid amounts are triggered in a complex way by the feautures of the insured's specific tariff and are therefore more difficult to predict. In contrast, claimed amounts can usually be reliably predicted based on the medical information available. However, the question as to whether the cost difference between participants and allocated controls will be based on claimed or paid amounts depends on the underlying perspective. From a health economics perspective, the saving in terms of claimed amount, i.e. full medical costs, is certainly more interesting. From an insurer's perspective, however, the actual paid amount is the deciding quantity, especially with regard to the risk of financial loss related to the program. The average savings presented in the following are nevertheless based on claimed amounts, because the paid amount was not available in the data from insurance company B. For insurance company A, the final cost difference has been calculated using both the paid and the claimed amount. As a result, it turned out that the savings based on the paid amount are very similar to

the savings based on the claimed amount. The risk of financial loss also barely differs between the two approaches. This outcome can be explained by the fact that insured persons in the portfolio analyzed have comprehensive health cover and comparable tariffs. Therefore, paid and claimed amount do not differ strongly and are highly correlated which is certainly not the case for portfolios with mainly supplementary cover. Furthermore, tariff information is considered as a matching criterion in the allocation model.

The three programs introduced are quite heterogeneous in terms of program measures, indication, number of participants and controls, as well as length of the program phase. Also, available information and data quality differ between insurance companies A and B. Accordingly, the different programs provide a comprehensive test environment for various measurement techniques. Sections 3.4.1 and 3.4.2 compare the described measurement approach to alternative methods and show its stability and consistency for all three scenarios. Finally, the programs described are used to characterize the relationship between the uncertainty of the measurement approach and the financial risk of loss for the insurer (see Section 3.4.3) based on the assumptions outlined in Section 3.3.4.

3.4.1 Stability of Matched-Pair Approach

First, the average savings per participant resulting from the different matched-pair options are illustrated (see Figure 19 for an overview of the distance calculations, aggregation functions and target variables used). Figure 20 shows the average savings per participant $\hat{\bar{d}}_\tau^1, \ldots, \hat{\bar{d}}_\tau^8$ in euros as well as the median $\hat{\bar{d}}_\tau^{\text{med}}$ of the saving curves for all evaluated programs.

The parallelism of the curves indicates the consistency of the measurement approach. Nevertheless, there is a considerable variation between the eight different curves in the potential savings for the single points in program time. This variation is caused by the uncertainty of the measurement approach, i.e. the part of the variation in future costs that cannot be explained by the underlying cost-prediction models. This inexplicable part of the variation arises from unavailable information, like disease-related clinical parameters, or immeasurable information, like the patient's degree of motivation, and can hardly be reduced by applying statistical techniques. Consequently, a single measurement method will not be able to produce a reliable estimate for the economic effect of a DMP. Therefore, it is recommended using the median estimate $\hat{\bar{d}}_\tau^{\text{med}}$, which robustly represents different aspects of the measurement problem, and also considering a possible range of potential savings. Using the variance estimation given in Section 3.3.4 it is also possible to estimate pointwise confidence intervals for the saving curves (not shown here).

Comparing the saving curves in Figure 20, there is no clear trend that single methods tend to systematically produce higher or lower savings. For the two diabetes programs, the median of all methods does not show a significant saving potential, also bearing in mind that the curves do not include program costs. One possible explanation for this is that the observation period of 3 or 6 program years, respec-

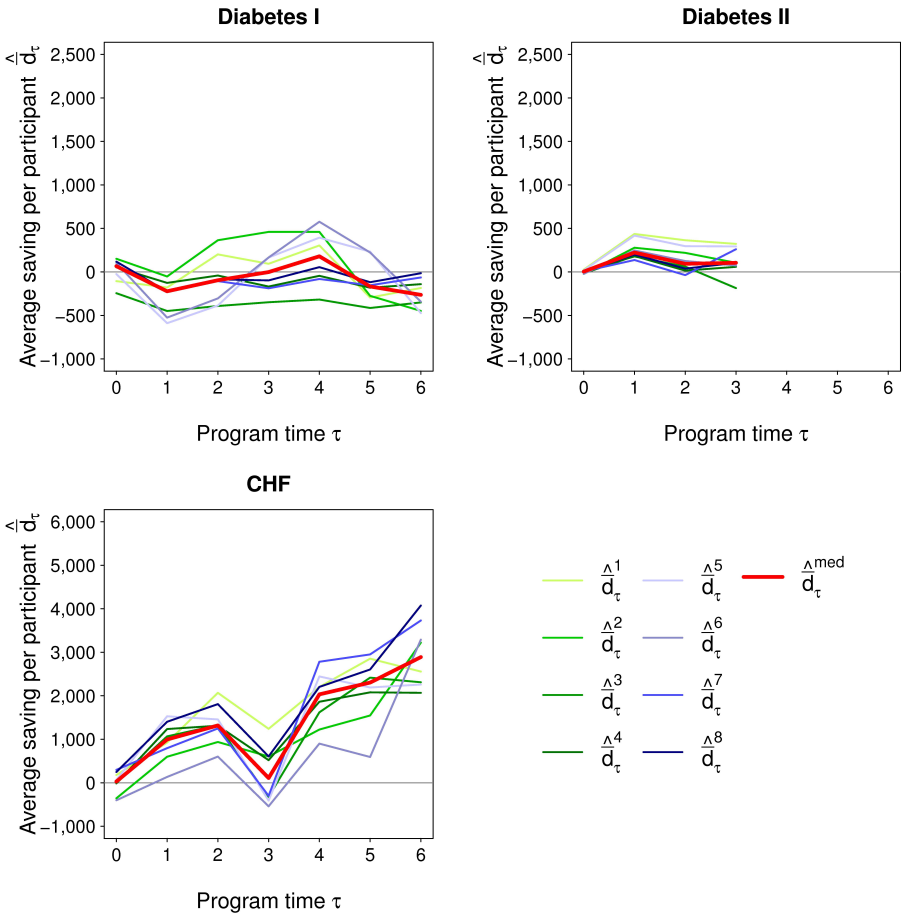

Figure 20: Matched-pair estimates (green/blue lines) including the median (bold red line) of all methods for the average saving per participant (in euros) of all DMPs analyzed.

tively, may be too short to measure any financial effect of a diabetes program. Even though there are only a few studies examining the long-term effect of diabetes DMPs, some authors, like Dove and Duncan [2004], argue that positive effects might not be visible within the first ten years because they require a long-term change in lifestyle.

For chronic heart failure, on the contrary, a clear saving potential was measured that almost linearly increases with program time, apart from a small break in the third program year. This break is visible in the curves of all calculated methods and is related to reduced costs in the population of controls which cannot be explained from the available data information. As the break is visible in all methods, including ones which are not matched-pair-based (compare the benchmarking analysis below), it is assumed that this inconsistency is mainly data-driven. Especially after $\tau = 4$ years, considerable savings are observable for the CHF program (up to 2,900 euros in the sixth program year). However, it has to be remembered that the number of participants that have been in the program for more than 4 years is quite small (see Table 9), so the uncertainty of the corresponding estimates is clearly larger than for the first program years. The existence of a verifiable saving potential from CHF

programs is generally in line with other authors, like Inglis et al. [2010].

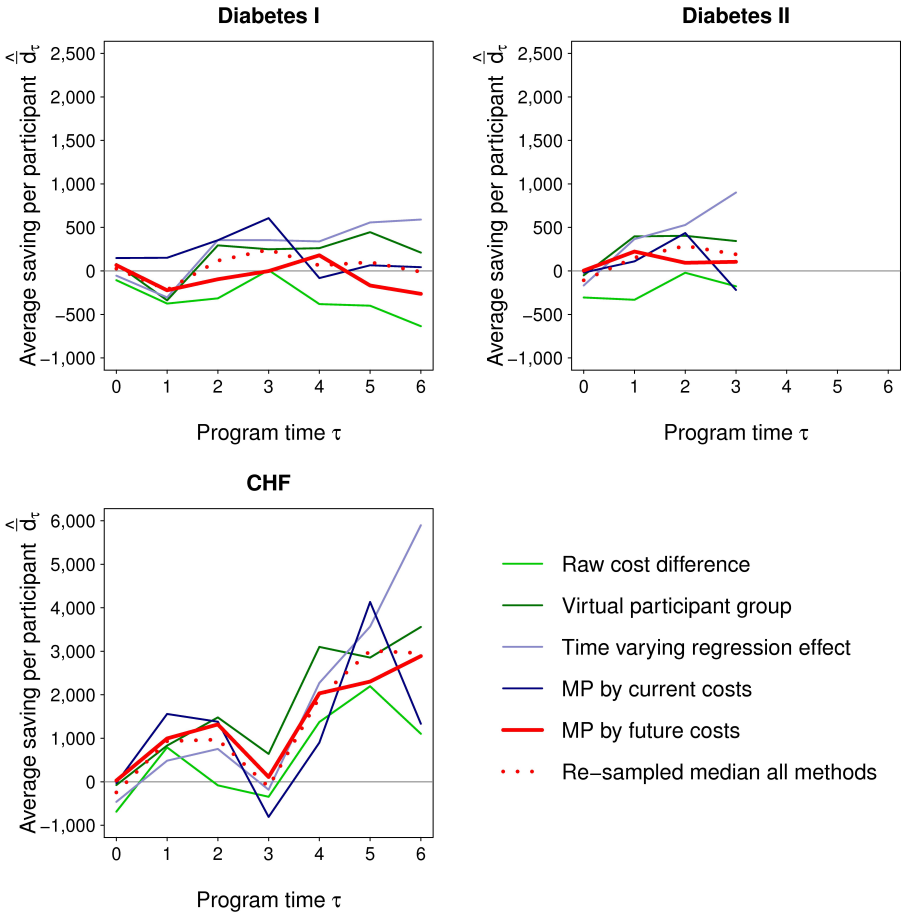

Figure 21: Matched-pair estimate (bold red line) for the average saving per participant (in euros) compared to several other measurement methods (green/blue lines).

In order to assess the stability of the approach, the final estimate $\hat{\bar{d}}_\tau^{\text{med}}$ is compared with different benchmark methods described in Section 3.2.1. The benchmarks represent different measurement schemes (compare Figure 15) and, as described in Section 3.2.1, some of them theoretically tend to under- or overestimate the treatment effect. The various benchmarks thus give an idea of the range of results possible. Figure 21 shows the matched-pair estimate for the average saving per participant over program time τ compared with the benchmarks tested.

The estimates related to different measurement methods are evaluated based on two quality criteria. The first criterion is the deviation from the real underlying savings, i.e. the extent to which a method can reduce the measurement bias. Of course, the real underlying savings are unknown and may depend on single observations, even though outliers are controlled. They cannot therefore be determined exactly by any statistical measurement approach. In this situation, it is assumed that a "mixture" of all the methods tested yields the best estimate for the real underlying savings. In order to reduce the influence of single individuals, a re-sampling approach

based on 100 random re-samples from the original portfolio is applied (by drawing n participants and m controls with replacement). As the final estimate for the real underlying saving, the median $\hat{\bar{d}}_{\tau}^{\text{med,all}}$ of the estimates resulting from all tested methods based on all re-samples is used (bold dotted red lines in Figure 21). Then, the sum of absolute deviations from the estimated real savings is considered as criterion Q_1 for the precision of the method. The second criterion assesses the stability of the method. This is a crucial property from an insurer's perspective, especially if the risk of loss and the reinsurability of the program are to be assessed based on this estimate. One (measurable) aspect of stability is the smoothness of the saving curve. In the theory of discrete time series, the "curvature" of the time series at a certain point in time is measured by the second-order difference $\nabla_2(\tau) = (d_\tau - d_{\tau-1}) - (d_{\tau-1} - d_{\tau-2})$ [Box et al., 2013]. Therefore, the sum of absolute second-order differences over the whole observation period is used as criterion Q_2 for the smoothness and stability of the saving curves. Table 10 summarizes the defined quality criteria for bias reduction and stability for all the methods tested. The optimal (smallest) values among the methods tested are marked in bold face.

Diabetes I

Method	Q_1	Q_2
Raw cost difference	2,527	1,895
Virtual participant group	**1,120**	2,374
Time varying regression effect	1,849	1,999
MP by current costs	1,322	2,197
MP by future costs	1,139	**1,316**

Diabetes II

Method	Q_1	Q_2
Raw cost difference	1,358	804
Virtual participant group	570	508
Time varying regression effect	1,223	581
MP by current costs	688	1,176
MP by future costs	**462**	**490**

CHF

Method	Q_1	Q_2
Raw cost difference	5,095	7,754
Virtual participant group	3,459	8,697
Time varying regression effect	4,780	7,471
MP by current costs	5,726	15,231
MP by future costs	**1,782**	**7,274**

Table 10: Quality criteria for bias reduction and stability for all methods tested and all programs analyzed (best method in terms of corresponding criterion marked in bold face).

1. The first benchmark is the raw cost difference between participants and all controls (light green lines in Figure 21) following the per-protocol measurement approach described in Section 3.2.1 (unweighted control group approach). For the controls, a weighted cost average of the calendar years corresponding to the participants' program years is used. As mentioned in Section 3.2.1, this approach (P2-C1) does not consider any potential self-selection effect in the treatment group. It tends to underestimate the treatment effect because the participants may have a higher level of motivation to improve their health status. This theoretical drawback is confirmed by the analysis performed. Compared with all the other measurement approaches shown in Figure 21,

the raw cost difference yields clearly lower average savings and the values of the bias reduction criterion Q_1 are worse than for most other methods. In particular, the large negative difference in $\tau = 0$ already shows that such unweighted approaches may lead to incomparable groups being compared. In the three programs analyzed, the participant group is clearly more expensive at the start of the intervention, which might indicate a more serious status of the chronic disease. The values of Q_2 for the programs analyzed also indicate a lack of stability.

2. The second benchmark is the weighted control group approach simulating a virtual participant group within the control group (compare Section 3.2.1). The method's estimated economic effect of the programs analyzed (dark green lines in Figure 21) is mostly higher than the estimated real effect. This is in line with the hypothesis that the method tends to overestimate the treatment effect because of the questionable assumption that the costs of actual and virtual non-participants are equal (compare Section 3.2.1). However, the absolute deviation from the real effect (criterion Q_1) is considerably lower than for all other benchmark methods. Considering the smoothness of the saving curves based on quality criterion Q_2, the approach seems to be less stable than the proposed matched-pair method.

3. For the third benchmark, no mean or median comparison is used, but the treatment effect is directly estimated from a regression model. The cost-prediction model (linear model) also used for matched-pair allocation is extended with a time-varying coefficient of the treatment effect. To allow maximum flexibility reference coding [Tutz, 2000] of the program time τ is used and an interaction with the treatment effect is included in the model. In this way, another benchmarking curve is obtained (light blue lines in Figure 21) which is similar to the other estimates in the first program phase, but clearly rises more sharply than all other estimates for a longer program duration. If a linear or quadratic time trend is used, and especially if a log-linear model is applied, the increasing trend is even stronger. Q_1 reflects the strong deviation from the estimated real treatment effect at the end of the observation period, which may cause an overestimation of the overall treatment effect. Even though the strongly increasing trend is quite linear, with Q_2-values being lower than for other benchmarks, the stability of this measurement approach is questionable.

4. The fourth benchmark is also a matched-pair estimate. Compared with the approach proposed, it is not the relevant covariates for future costs that are used as matching variables here, but actual costs. More precisely, the final estimate is based on the same distance aggregation methods over time ($\tilde{s}_1$ and $\tilde{s}_2$) as the proposed estimate, but distances are calculated with actual costs instead of predicted costs or weighted factors. The resulting curves (dark blue lines in Figure 21) are rather volatile for all three programs and, in part, do not follow the trend of the other methods tested (compare the high Q_1- and Q_2-values in Table 10). This demonstrates that it is very important to use the drivers of future costs instead of actual costs for matched-pair allocation. In doing so, it is ensured that only individuals with a similar progress of the

chronic disease and a similar expected cost development are compared.

5. The proposed matched-pair estimate produces a quite smooth saving curve over time (bold red lines in Figure 21) for all the programs analyzed, compared with the benchmarks tested. This observation is supported by the lowest values of the stability criterion Q_2 in Table 10. These results indicate the relative stability of the approach compared with the alternative methods, at least for the programs considered. Furthermore, the method's saving curves lie close to the estimated real saving curves, which is an indicator that the proposed matched-pair approach avoids both under- and overfitting of the program effect. The method's Q_1-values are accordingly almost always lower than the Q_1-values of the benchmarks. The extent to which the methods actually reduce the self-selection bias arising from the study design cannot ultimately be assessed. However, assuming that the programs' real savings can be reasonably approximated using the re-sampling approach described, the matched-pair method yields the most precise estimation and the smallest bias of all the benchmark methods calculated.

3.4.2 Robustness of Matched-Pair Approach

In the following, it is evaluated how the proposed estimate responds to changes in the data input, especially its robustness to extreme observations. Also, the effect of modifying the defined proceeding is analyzed and reasons for the choice of hyperparameters are given.

An important question is how the estimate for the average savings per participant changes if persons who have died during the observation phase are included or excluded, because medical costs usually increase sharply in the last month of life [Emanuel and Emanuel, 1994]. It is important to note that those individuals are not automatically excluded by the outlier criterion defined in Section 3.3.2 because outlying observations are only controlled prior to the start of the intervention. It must also be taken into consideration that the question of excluding deceased persons is closely related to the question of whether the DMP has an impact on mortality in the intervention group (which is not a focus of this work). For the estimates shown in Section 3.4.1 the deceased persons were excluded in order to avoid the economic comparison being skewed by an unequal distribution of deceased persons in the participant and control groups. In order to assess the sensitivity of the method to the inclusion of deceased people, the estimates $\hat{\bar{d}}_{\tau}^{\text{med}}$ with deceased persons (light green lines in Figure 22) and without deceased persons (bold red lines in Figure 22) are compared.

For the DMPs of company A (Diabetes I and CHF) the general trends do not change, although there are variations of up to 500 euros for the single program years. These variations can broadly be explained by individuals who cause very high medical costs for a short period before their death. The mortality rates are quite stable between 2% and 5% for both the intervention and control group. For company B's diabetes program (Diabetes II) the curves differ considerably, especially in the first program

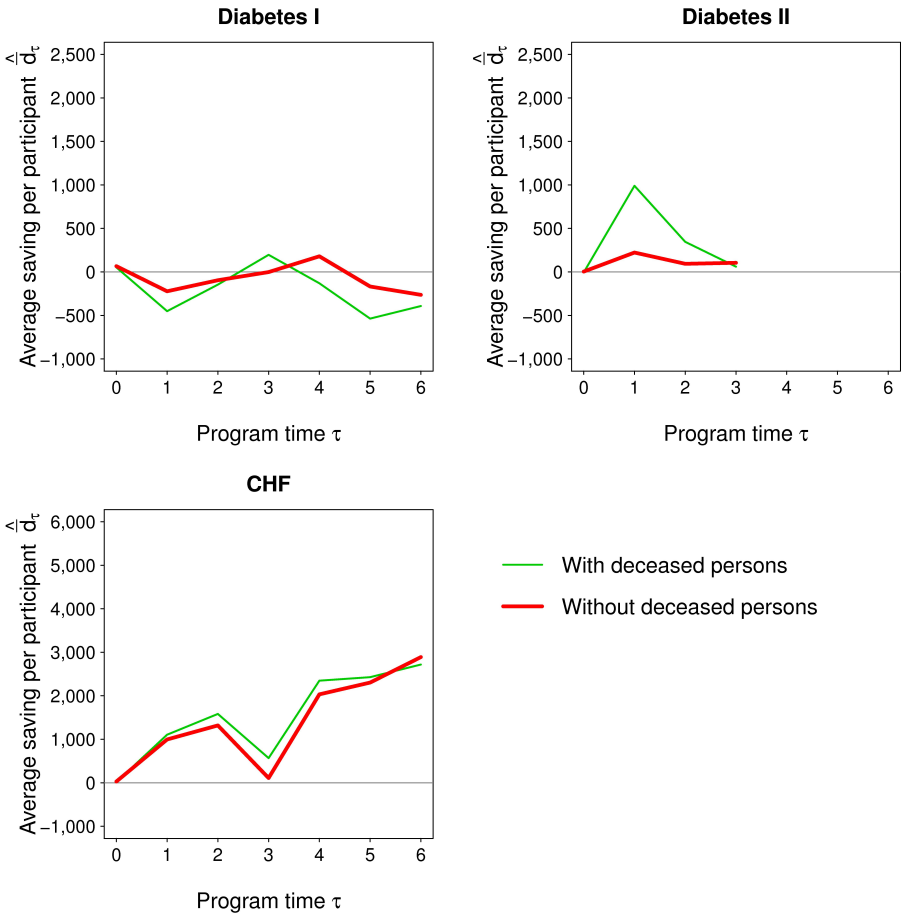

Figure 22: Matched-pair estimate for the average saving per participant (in euros) with exclusion (bold red line) and inclusion (green line) of deceased persons.

year. Here, the average saving per participant is approximately 750 euros higher if deceased people are included. Further analyses showed that this observation can broadly be explained by an increased mortality rate in the control group that leads to increase in costs. In addition, the average costs per deceased person are lower in the intervention group. Overall, the values of the quality criterion Q_2 (values not shown) strongly increase and the stability of the saving curves decreases if deceased people are included.

Irrespective of the question of improved survival due to the DMP, it is recommended excluding deceased persons from the measurement in order to avoid any distortion caused by non-observable factors influencing mortality. Another option for dealing with the question of including deceased people in a retrospective analysis is to use a person's survival as a matching variable in the allocation process.

For the defined criterion to control outliers in the allocation process, the maximum cost difference b between participants and controls prior to the start of the intervention needs to be defined. For greater transferability to other measurement problems, b is determined depending on the average annual claimed amounts by participants and controls prior to the start of the intervention $\hat{y}_t$. Figure 23 shows the proposed

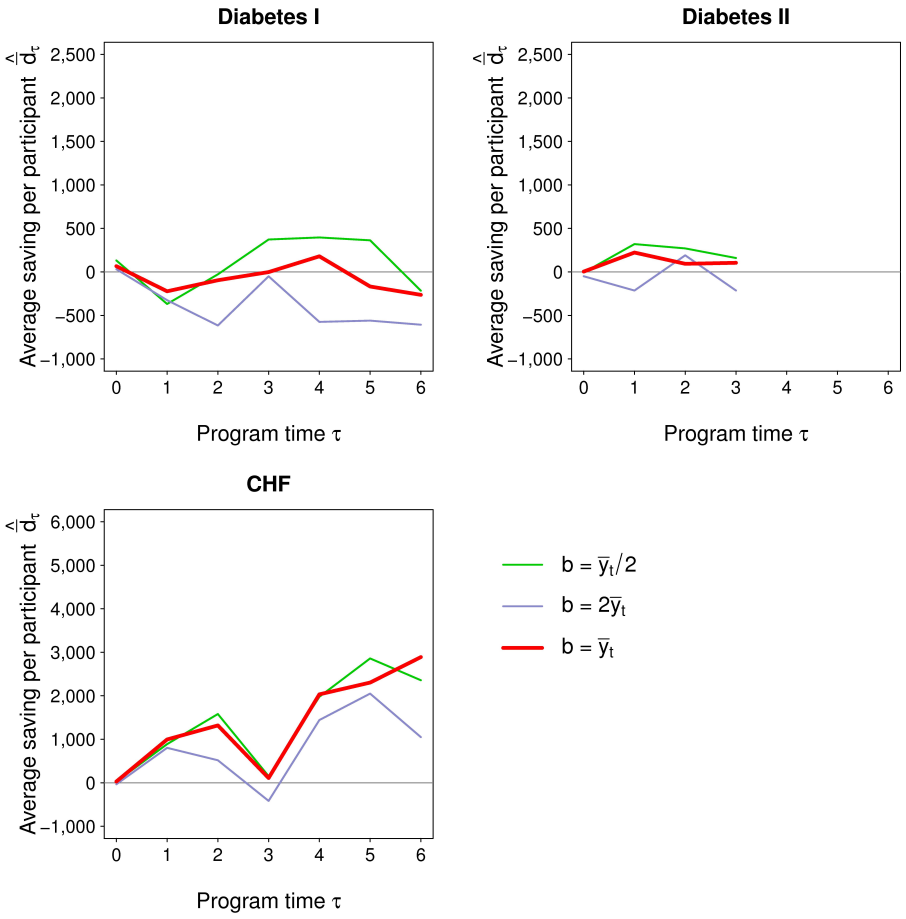

Figure 23: Matched-pair estimate for the average saving per participant (in euros) with lower (green line), higher (blue line) and medium cost boundary b (bold red line) for outlier control.

estimate $\hat{\bar{d}}_\tau^{\text{med}}$ for all three DMPs analyzed with different cost boundaries b. For the analyses shown in Section 3.4.1, $b = \hat{y}_t$ is used (bold red lines in Figure 23). For all three programs analyzed, larger values of b, like $b = 2\hat{y}_t$ (blue lines in Figure 23) lead to higher average savings, whereas smaller values of b, like $b = \hat{y}_t/2$ (green lines in Figure 23) reduce the average individual savings. In this regard it must me remembered that if the value of b is chosen smaller, the number of participants for which an adequate twin can be found decreases. More participants consequently need to be excluded, which means that the uncertainty of the measurement grows. Conversely, if b is chosen too large, outliers may not be controlled sufficiently. Based on the analyses performed, it is suggested using $b = \hat{y}_t$.

As outlined in Section 3.2.1, the matched-pair technique can be seen as a weighted control group method. The weights correspond to how often a control is used as a matched-pair for the participants. In Section 3.3.3, three alternative ways of weighting are defined using different ways of calculating the average costs of the allocated controls (S_τ^+, S_τ^- and S_τ, defined in Equations (29), (30) and (28), respectively). Figure 24 shows the resulting estimates $\hat{\bar{d}}_\tau^{\text{med},+}$ (green lines), $\hat{\bar{d}}_\tau^{\text{med},-}$ (blue lines) and

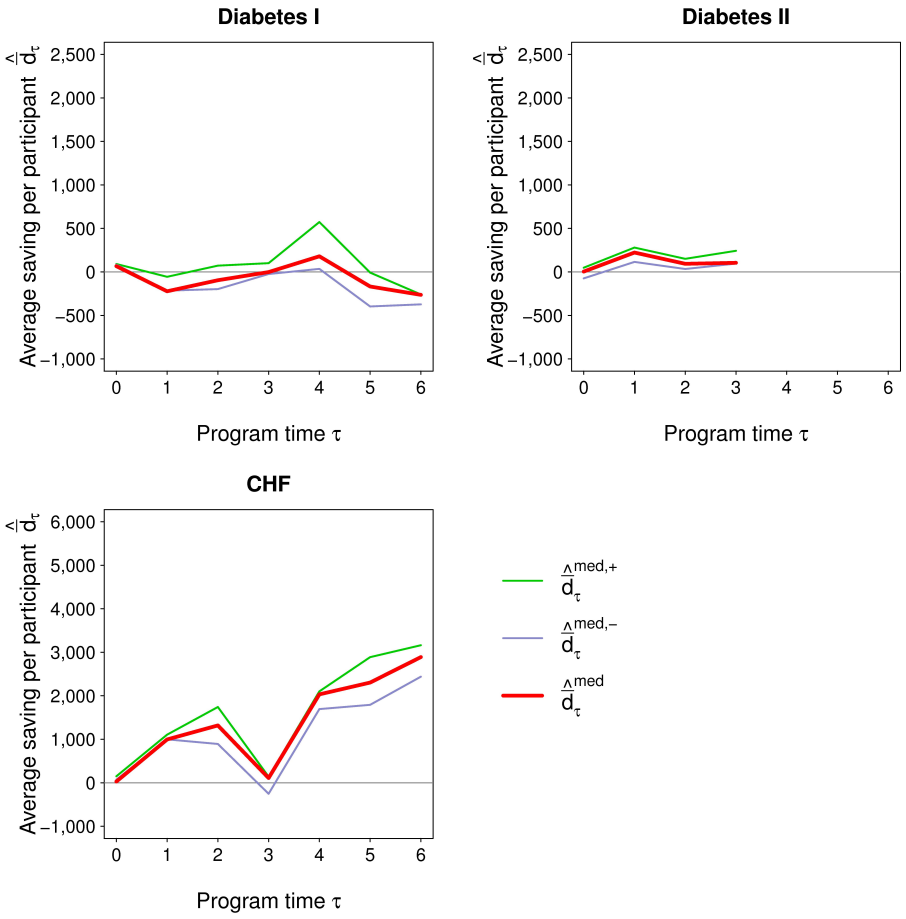

Figure 24: Matched-pair estimate for the average saving per participant (in euros) with stronger (green line), lower (blue line) and medium weighting (bold red line) of assigned controls.

$\hat{\bar{d}}_\tau^{med}$ (bold red lines) which imply a high, low or medium weighting, respectively, of controls who are similar to many participants. For all three programs, the higher weighting leads to slightly increased average savings per participant, whereas the lower weighting reduces them to a similar extent. For most points in program time, however, the deviations are rather small and the overall trend remains the same. Overall, the final estimate is not heavily dependent on the weighting applied, and it is suggested using the medium approach.

Finally, the robustness of the estimates towards hyperparameters l and k of the matching approach is examined, which control the number of allocated matches to every participant and justify the choice made. In allocation function $\tilde{s}_1$ (see Equation (24)) the hyperparameter l determines how many of the controls with the smallest distance to the participant are allocated. Similarly, hyperparameter k in allocation function $\tilde{s}_2$ (see Equation (25)) defines the maximum rank that a control must not exceed in at least two years in order to be an adequate twin for a participant. Figure 25 shows the cumulative average saving per participant after $\tau = 6$ program years depending on the choice of hyperparameters k and l using

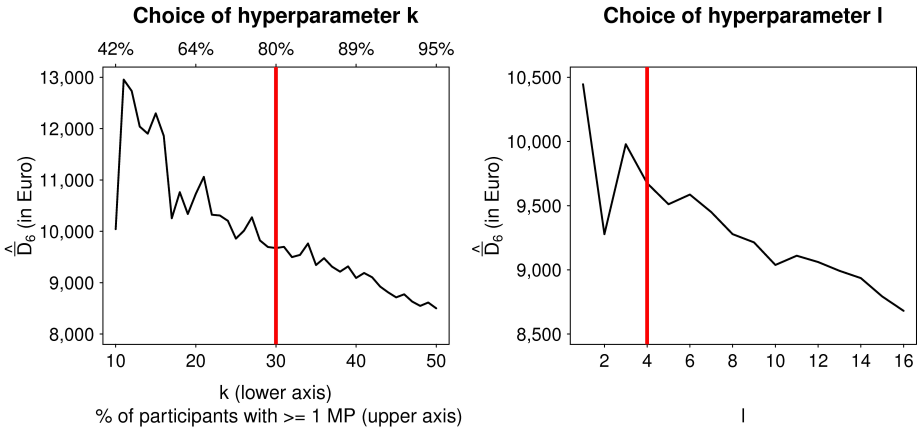

Figure 25: Choice of hyperparameters k and l (bold vertical line) steering the number of allocated matches in transformations $\tilde{s}_2$ and $\tilde{s}_1$, respectively, using the example of the CHF program.

the example of the CHF program. As not all participants necessarily receive a twin with allocation function $\tilde{s}_2$, the percentage of participants with at least one twin depending on k is also displayed (upper x-axis of the left plot). The cumulative savings generally decrease with increasing k. The volatility of the cumulative average savings also strongly decreases with growing k. $k = 30$ (bold red line) is chosen in order to obtain a stable estimate that is still based on a reasonable number of allocated controls per participant. Also, the number of excluded participants without allocable twins, at 20% for $k = 30$, is still manageable. For allocation function $\tilde{s}_1$, the cumulative average savings per participant and their volatility similarly decrease as l grows. Based on the same criteria as for the choice of k, $l = 4$ (bold red line) seems to be a reasonable choice here. Though the saving curves are robust to slight modifications in k (between 25 and 40) and l (between 3 and 8), the analyses show that it is crucial to control these parameters for a stable measurement.

Overall, the analyses carried out suggest that the proposed matched-pair approach is widely robust towards outlying observations. Nevertheless, it is recommended excluding deceased persons in order to ensure the stability of the approach. It is also vital to choose the right hyperparameters for the matched-pair approach proposed, even though slight changes to the procedure suggested do not change the overall trend.

3.4.3 Uncertainty of the Measurement and Risk of Loss

Next, the relationship between the uncertainty of the measurement approach presented and the estimated risk of loss of a DMP is evaluated. Based on the three programs carried out, this relationship is quantified in order to be able to predict the risk of loss of similar programs that will be applied to new portfolios with different data volumes and quality in the future. If the new program and the portfolio to which it is applied meet the pre-conditions defined in Section 3.3.4, the methodology

presented makes it possible to reliably predict the insurer's risk of loss and decide on the program's reinsurability.

As outlined in Section 3.3.4, the uncertainty of the described measurement approach is directly related to the predictive quality of the regression model on which matched-pair allocation is based. This predictive quality is measured using the predictive R-squared R^{2^*} (see Section 2). Section 3.3.4 also shows that the predictive R-squared drives the variance of the matched-pair estimate and therefore the estimated risk of loss of the program. The functional form f (see Equation (41)) of the relationship of R^{2^*} and the variance of the cumulative average savings per participant σ_τ^2 is estimated based on the three programs carried out.

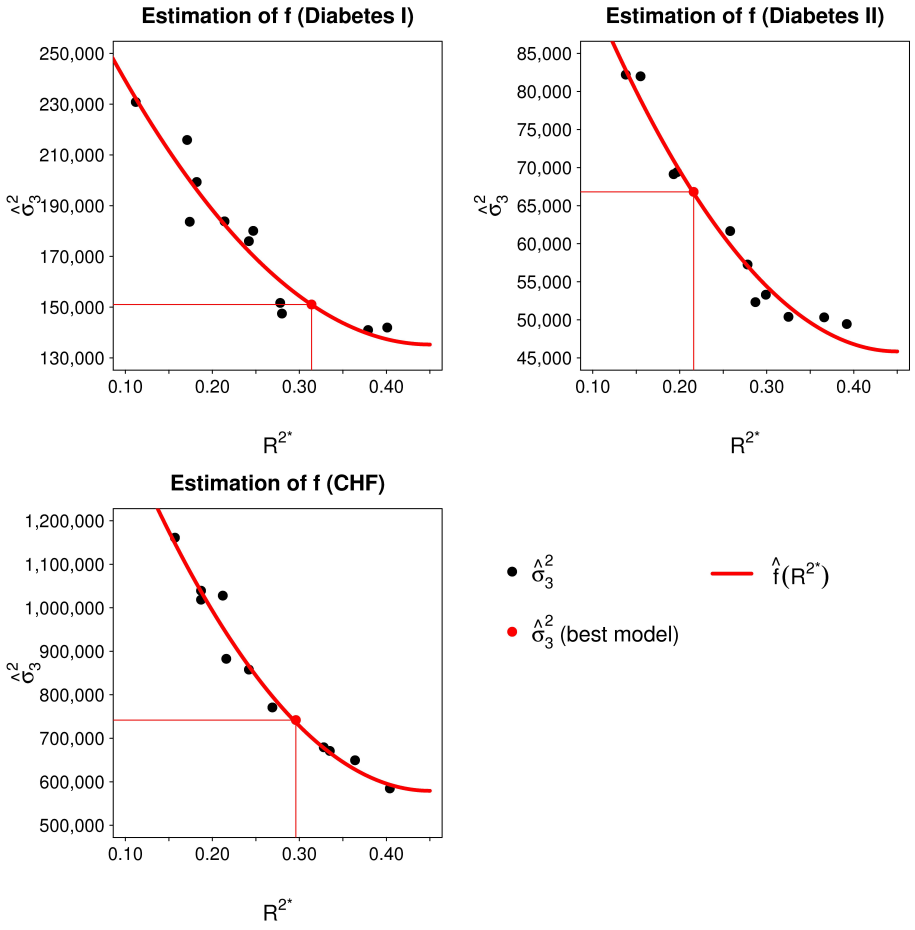

Figure 26: Estimated variance of cumulative average savings after 3 program years $\hat{\sigma}_3^2$ (black dots) and estimated quadratic function $f(R^{2^*})$ (bold red line) for all programs analyzed.

For analyzing the financial risk of loss, the cumulative average saving after $\tau = 3$ program years is considered, because a reliable number of observations for this timeframe is available for all three DMPs. Analyses for different timeframes, however, lead to similar results. As f needs to be transferred to new portfolios of different size and data quality, allocation models with different data input were calculated. To simulate portfolios with worse data quality or a smaller volume of data, several

variables with high predictive value were removed, which leads to a decrease in the predictive R-squared. Conversely, different simulated variables with a high correlation to the target variable were added in order to artificially increase the predictive quality of the allocation models. Figure 26 shows the R-squared values R^{2^*} of the calculated allocation models and the resulting variances $\hat{\sigma}_3^2$ of the corresponding matched-pair estimates. The covariate configuration of the models is chosen in such a way that the R^{2^*} values range between 10% and 45%. According to previous studies (see Section 2) this is a realistic range for the predictive quality of individual cost-prediction models applied to groups of chronic patients in health insurance portfolios. The red dots mark the models with the maximum possible predictive quality that could be achieved, based on the respective population. The black dots mark the models with covariates removed or added as described above.

As expected and justified in Section 3.3.4, the variance of the cumulative average savings per participant clearly decreases with increasing predictive quality. The plots also indicate that the variance reduction decreases with growing R^{2^*}. For the two diabetes programs especially, the variance $\hat{\sigma}_3^2$ even seems to converge towards a certain minimum variance level. This means that it is necessary to construct a cost-prediction model with a high degree of predictive quality in order to control the variance of the measurement. From a cost-benefit perspective, however, it is not necessary to squeeze out the last percent of improvement in terms of R^{2^*} which, in practice, usually requires the most effort.

In order to formalize the relationship f between the predictive quality of the allocation model and the variance considered, the following candidate functions for f are defined which follow the properties suggested by the observed data from the three DMPs described (see Figure 26):

$$
\begin{aligned}
f_{\text{lin}}(R^{2^*}) &= \omega + (R^{2^*} - 0.45) \cdot \xi \\
f_{\text{squ}}(R^{2^*}) &= \omega + (R^{2^*} - 0.45)^2 \cdot \xi \\
f_{\text{cub}}(R^{2^*}) &= \omega + (R^{2^*} - 0.45)^3 \cdot \xi \\
f_{\log}(R^{2^*}) &= \omega + \log\left(\frac{1}{R^{2^*} - 0.45}\right) \cdot \xi
\end{aligned}
\tag{43}
$$

The candidate functions f_{lin}, f_{squ}, f_{cub} and $f_{\log}$ are all monotonically decreasing and convex on a definition range from 0 to 0.45. The parameter ω corresponds to the minimum variance level that is reached for the maximum possible predictive quality of $R^{2^*} = 0.45$. The second parameter ξ of the candidate functions steers the speed of the variance decrease with growing R^{2^*}.

Next, the candidate function f is determined by analyzing the data observed from the three DMPs considered. Therefore, the residual errors of four linear regression models with target variable $\hat{\sigma}_3^2$ and a transformation of R^{2^*} according to the definition of the candidate functions (see Equations (43)) as only covariate are calculated. The parameters ω and ξ can be interpreted as the intercept and slope of these models, respectively. Table 11 shows the residual errors for the three DMPs observed

indicating that a quadratic function best describes the relationship between R^{2^*} and the variance $\hat{\sigma}_3^2$. The bold red curves in Figure 26 illustrate the quadratic function $\hat{f}$ with minimum residual error. Also, the cubic function seems to be an adequate choice.

Estimation of f			
Candidates for f	RE (Diab. I)	RE (Diab. II)	RE (CHF)
f_{lin}	12,178	3,851	51,135
$\boldsymbol{f_{\text{qua}}}$	**9,238**	**2,054**	**32,746**
f_{cub}	10,460	2,103	39,382
f_{log}	18,073	6,442	89,472

Table 11: Residual errors of candidate models to determine the form of the functional relationship between σ_3^2 and R^{2^*} for all analyzed programs after 3 program years (best candidate function f marked in bold face).

Now, the estimated function $\hat{f}$ can be used to predict the variance of the matched-pair estimate for similar programs applied to new portfolios. For this purpose, a prediction model needs to be calculated based on the new portfolio the predictive quality in terms of R^{2^*} needs to be determined. Together with the assumption that the cumulative average savings per participant are reproducible for a sufficiently similar portfolio, the financial risk of loss can then be predicted for the new program by applying the probability function of the suggested normal distribution.

According to the developed measurement approach, the Diabetes I program has a negative cumulative saving after $\tau = 3$ program years even without considering program costs. The type I error probability (see Equation (39)) of this program is therefore greater than 50% due to the symmetry of the assumed normal distribution around the mean. The Diabetes II program yields a cumulative saving of 420 euros after $\tau = 3$ program years. Even though the annual program costs only amount to 100 euros per participant, the type I error probability of the measurement for the best allocation model ($R^{2^*} = 21.6\%$) is still above 30%. Consequently, similar programs like the two diabetes programs analyzed cannot be considered for a reinsurance solution because the predicted risk of financial loss is too high. In contrast, the CHF program yields cumulative average savings per participant of 2,420 euros after $\tau = 3$ program years. As the exact annual program costs are unknown, an annual expenditure of 180 euros per participant is assumed, based on studies of similar programs (like Giordano et al. [2009]). Assuming a normal distribution for the economic effect parametrized by the estimates $\hat{\mu}_3$ and $\hat{\sigma}_3$ (see Equations (33) and (38)), a type I error probability of 1.4% for the best allocation model ($R^{2^*} = 29.6\%$) is obtained. This probability corresponds to the predicted financial risk of loss if the same program was applied to a comparable portfolio in terms of cost-relevant covariates as well as same data quality and volume.

Finally, Figure 27 illustrates how the predicted risk of financial loss for future programs changes under different conditions, provided the prerequisites i)-iii) defined in Section 3.3.4 are fulfilled. First, it is assumed that exactly the same program is applied to a comparable portfolio in terms of cost-relevant covariates, but with lower data quality and/or volume leading to an R^{2^*} of 20%. In this case, the predicted risk of financial loss increases to 2.9% (green area under the curve in Figure 27).

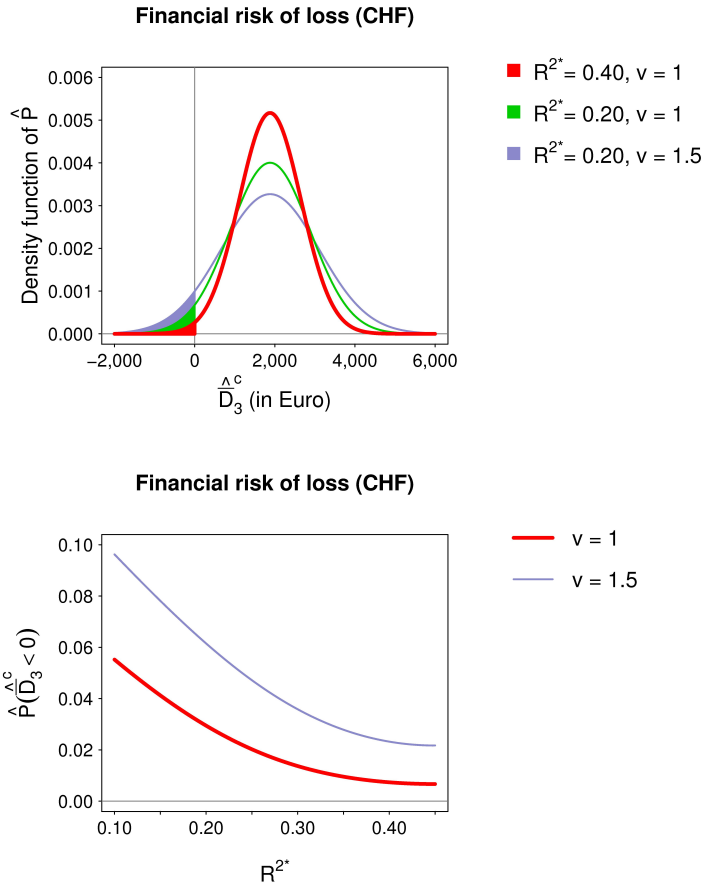

Figure 27: Estimated risk of loss after 3 program years $\hat{\mathcal{P}}(\bar{D}_3^c < 0)$ (upper plot: shaded area under the density function of $\hat{\mathcal{P}}$, lower plot: value in the y-axis) for different values of R^{2*} and v using the example of the CHF program.

Conversely, the predicted risk of loss decreases to 0.7%, if the predictive quality for a new portfolio reaches $R^{2*} = 40\%$ (red area under the curve in Figure 27). If there is an additional source of uncertainty due to a changed program setting (like different inclusion and exclusion criteria), the multiplicative factor v is used to adapt the variance of the measurement (see Equation (42)). Based on a variance correction of $v = 1.5$ and assuming a predictive quality of $R^{2*} = 20\%$, the predicted risk of financial loss grows to 6.2% (blue area under the curve in Figure 27). For this result, it is implicitly assumed that the changed setting does not affect the cumulative average saving itself, but only its variance. Based on these outcomes, the CHF program can be considered for a reinsurance solution. The methodology presented for predicting the financial risk of loss can be used to define the conditions of such solution.

3.5 Summary and Outlook

This chapter presents a matched-pair approach based on predictive regression analysis for a reliable measurement of the economic effect of DMPs. The approach can

be applied in situations where no randomized control trial is feasible, especially for insurer-driven DMPs. Due to the operational setup of these programs, an adequate measurement technique must deal with the self-selection bias disturbing the comparison between treatment and control groups. The results of the analyses for three programs carried out (two diabetes programs and one chronic heart failure program) suggest that the matched-pair approach presented delivers a consistent measurement of the financial impact of insurer-driven DMPs over time. Compared with the benchmarks tested, it increases the stability of the measurement for the programs analyzed. Due to the fact that the real underlying economic effect is unknown, a final evaluation of the method's ability to reduce the existing self-selection bias is not possible. In this situation, a re-sampling approach is applied to approximate the real underlying savings based on all methods tested using different methodological approaches. Under the assumption that this approximation is correct, the matched-pair approach avoids an under- and overestimation of the treatment effect. It also minimizes the existing self-selection bias among all methods tested through matched-pair risk adjustment based on all available information. Even though there may be other approaches which fulfill these requirements, no superior measurement technique for insurer-driven DMPs could be identified.

Also, a method for determining the probability of a wrong measurement is presented, especially the type I error probability which corresponds to the financial risk of loss related to the DMP. Based on an estimate of the variance of the average saving potential and an assumption on its distributional form, first, the risk of financial loss for programs that have already been carried out is determined. Second, the relationship between the uncertainty of the measurement method and the financial risk of loss related to these programs is characterized. These results can then be used to predict the risk of financial loss for similar programs which have not yet been carried out, as long as certain defined conditions on the similarity of the programs are fulfilled. Based on this prediction, it is possible to assess a program's suitability for a reinsurance solution protecting the insurance company against financial loss. The evaluation of three observed DMPs shows that not all programs are appropriate for such a solution because the estimated saving potential is partly not high enough.

Further research effort needs to be spent on testing the measurement technique for additional programs and indications to which the methodology can easily be transferred. Equally, the economic evaluation of case management programs with a sufficient number of participants is an interesting field of application. The methodology presented can also be used to compare other outcomes – like mortality or the frequency of claims – between treatment and control groups. For every new scope, it is vital to spend some effort on appropriately calibrating the underlying models and hyperparameters of the approach, in order to control its stability and efficiency. The analyses performed also show that the underlying measurement problem does not have just one correct solution, but a range of possible values which represents the uncertainty included in the measurement needs to be considered. This uncertainty can be reflected by specifying a confidence interval for the average savings or by varying the methodology as proposed in Section 3.3.

Another vital conclusion from the analysis of the economic saving potential of DMPs

is the growing importance of appropriate data capturing and appropriate analytical techniques in the healthcare management sector. In Section 2, it is shown that high predictive quality depends on the right statistical methods being applied to data of adequate quality and size. In the economic evaluation of DMPs, high predictive quality is particularly important because for all three programs analyzed the accuracy of the measurement strongly decreases with the predictive quality of the underlying models. Higher data quality will therefore allow a more precise measurement in the future.

The optimization of DMPs and their economic evaluation is only one example of the successful application of modern predictive techniques in healthcare management. Another promising field of application is efficient data-driven fraud and abuse detection for cost control in health insurance which will be treated in Section 4.

4 Fraud and Abuse Detection

(refers to the publication Bayerstadler et al. [2016])

Healthcare fraud and abuse are a serious challenge to healthcare payers and to the entire society. Section 4 presents a predictive model for fraud and abuse detection in health insurance based on a training dataset of manually reviewed claims. The goal of the analysis is to predict different fraud and abuse probabilities for new claims. The prediction is based on a wide framework of fraud and abuse reports which examine the behavior of medical providers and insured members by measuring systematic deviation from usual patterns in medical claims data. In this chapter, it is shown that models which directly use the results of the reports as model covariates do not exploit the full potential in terms of predictive quality. Therefore, a multinomial Bayesian latent variable model which summarizes behavioral patterns in latent variables, and calculates different fraud and abuse probabilities is proposed. The estimation of model parameters is based on a Markov Chain Monte Carlo (MCMC) algorithm using Bayesian shrinkage techniques. The improved prediction results of the developed model are illustrated, compared to alternative approaches, and its transferability to other markets by specification of adequate prior distributions is discussed.

4.1 Introduction

Fraud, abuse and waste in healthcare strongly contribute to the increase in total healthcare expenditure. Therefore, they are serious issues for public and private payers of healthcare and, as costs are usually transferred to the collective of insured persons, also to the insured persons themselves and to the entire society. According to a global study including 92 separate loss measurement exercises in 33 organizations from 6 countries (the UK, USA, France, Belgium, the Netherlands and New Zealand), the average loss due to fraud and abuse is 7% (range of 3% to 15%) of total healthcare expenditure [Gee and Button, 2014]. In the US, loss estimations from fraud and abuse range from 9% to 19% of total healthcare expenditure [Berwick and Hackbarth, 2012]. Assuming an average of 14% this means a total loss of 369 billion US dollars in 2011, or more than 1,000 US dollars per US citizen. In Europe similar rates are assumed [European Union Commission, 2013]. It is self-evident that this additional burden leads to increased taxes and higher health insurance premiums for individuals. Possible consequences are an eroding social solidarity and an individualization in health insurance.

It is important to note that the definitions of fraud and abuse are quite heterogeneous in the literature, and depend on market and regulatory environments. In most articles, the terms waste, abuse, and fraud are used, with distinctions being fluid (see Figure 28).

In this chapter, the focus will be on fraud and abuse, i.e. intentional behavior by patients and/or medical providers to create unjustified benefits for themselves or

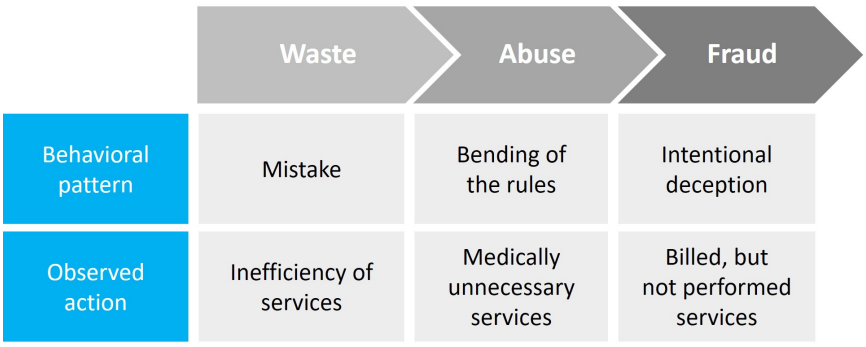

Figure 28: Definition of fraud and abuse in healthcare.

related persons. However, as a clear differentiation between waste and abuse is not possible, also some aspects of unintentional behavior are covered by the analytical modeling approach. Overall, three behavioral categories are used:

- *Unperformed services (fraud):* Medical services which are documented and charged, but not performed (e.g. consultation fees for a doctor's visit that did not take place, insured members' faking prescriptions).
- *Unjustified services (abuse):* Medical services performed without medical necessity/justification, and which deviate from medical best practice, in US literature often denoted as overutilization (e.g. x-ray examinations regardless of symptoms, prescription of antibiotics for mild respiratory diseases).
- *Other billing issues (fraud/abuse):* All other kinds of intentional misbehavior by medical providers and/or insured members (e.g. concealment of pre-existing conditions in medical underwriting, masking of uncovered services, unbundling of procedure codes, etc.).

In healthcare, fraud and abuse are not only committed by medical providers. Insured patients, approvers of services and other healthcare players are also involved in fraudulent and abusive actions [Busch, 2008]. It is also important to note that fraud and abuse are often based on the cooperation or at least complicity between different players in the health market (e.g. doctor and pharmacist, provider and insured, etc.). Nevertheless, many publications on fraud and abuse, and related analytical detection methods, focus on medical providers [Busch, 2008]. Reasons for this are the assumption of a higher recovery potential and a broader data basis than for single insured persons.

As major contributors of healthcare funding, private and public insurers have a strong motivation to prevent losses arising from fraud and abuse. However, there is often no systematic approach in place to deal with this issue. In many insurance companies, fraud and abuse detection is limited to opportunistic checks of single patterns within the standard claims handling process. In addition, the claims handler usually only has an invoice-by-invoice perspective which does not take into account the (common) claims history of members and providers. The main focus is usually on historic provider behavior as well as diagnoses and procedures frequently related

to fraud and abuse cases, whereas member behavior and the interaction of different players play a minor role. In this setup, new and more complex patterns can only be detected by coincidence. Another difficulty is the lack of data experience on fraud and abuse cases, i.e. a small sample size of reviewed cases, especially in smaller insurance companies, due to a lack of resources for the topic. Other challenges are the lack of defined actions if a suspicious case is identified, and regulatory conditions that impede the recovery of money once claims are paid.

Therefore, a systematic analytical approach for the identification of fraudulent and abusive behavior is proposed which is based on

a) a vast data basis of reviewed fraud and abuse cases in the target market to be used as quality assured response,

b) comprehensive knowledge of fraud and abuse patterns in different markets (generating new input for the target market),

c) a "reporting factory" which provides a 360-degree view on invoices, including reports on provider and member behavior, reports on network behavior (i.e. the interaction between different players) and reports on invoice properties (e.g. diagnoses, procedures, number of invoice lines, ...), and translates the knowledge of fraud and abuse patterns into quantitative measures,

d) a predictive scoring model which allocates fraud and abuse probabilities (in the three categories described above) to incoming invoices (if necessary, in real time) and outputs reasons for high probabilities as a starting point for further investigation.

This chapter focuses on the methodology of the predictive scoring model (item d)). More precisely, a Bayesian multinomial latent variable model is presented which was specifically developed to make optimal use of the knowledge described in items a) to c) (see Section 4.3). The idea behind the latent variable approach is to summarize different observations from the reporting factory in behavioral scores for providers, members and networks which stabilize the model. The target of the scoring model is to reach high predictive power and transferability by avoiding overfitting to the training data, which is a challenge due to the underlying data situation (see Section 4.2). For the insurance company, higher predictive power means a more efficient claims adjudication process by a more targeted investigation of invoices. Section 4.4 presents the prediction results of the model based on two different test datasets and compares them to alternative scoring techniques. Moreover, some thoughts are presented on how the scoring model can be transferred to other markets with different regulatory and data conditions, especially if no or little data knowledge is available in the target market.

4.2 Literature and Background

4.2.1 Literature

Analytical fraud detection methods are being successfully applied in many areas of the financial industry. Bolton and Hand [2002] as well as Ngai et al. [2011] provide a comprehensive overview of applications and statistical techniques. Statistical techniques gain importance where mass data ("Big Data") need to be analyzed, such as in credit card fraud detection. Bhattacharyya et al. [2011] summarize and compare different fraud detection techniques in this field. Other important areas of application in the financial industry are money laundering, telecommunications fraud, computer intrusion, medical/healthcare fraud and scientific fraud.

In general, the approaches applied can be divided into supervised techniques, like classification and regression, where a learning dataset of identified fraud cases is available, and unsupervised techniques, like clustering and outlier detection, where the focus is on detecting abnormal patterns [Bolton and Hand, 2002; Phua et al., 2005]. Concrete techniques are, for example, Neural Networks, (logistic) regression, Naive Bayes, Decision Trees, Fuzzy Logic, CART, Genetic Algorithm, k-Nearest Neighbors and Bayesian Belief Networks [Ngai et al., 2011]. Moreover, increasing digitization and "Big Data" generation require the usage of text [Holton, 2009] and image mining [Brown et al., 2005] techniques. Bayesian methods are gaining importance because they offer the opportunity to involve expert knowledge by adequate prior specification [Kirkos et al., 2007], and are computationally stable [Phua et al., 2005].

Within the insurance industry, analytical fraud and abuse detection is most widespread in motor and health insurance, but there are also examples from other lines of business, like crop insurance [Jin et al., 2005] or individual disability income insurance [Peng et al., 2007]. An extensive overview of statistical fraud detection methods in motor insurance, together with a comparative case study, can be found in [Viaene et al., 2002]. Bayesian learning is, for instance, applied by Viaene et al. [2004a,b, 2005] and Bermúdez et al. [2008]. Similar to health insurance, relations between different players are very important for motor insurance. This is why recent research in this area is often focused on the analysis of (social) networks [Šubelj et al., 2011].

In health insurance, various classification techniques to identify fraudulent and abusive behavior are applied. Joudaki et al. [2014] and Li et al. [2008] give a general overview of data mining techniques supporting fraud identification in health insurance, Dua and Bais [2014] focus on supervised classifications methods. For instance, Thornton et al. [2013] propose a multidimensional classification model for different kinds of healthcare fraud and abuse in the US Medicaid system. As medical knowledge and experience is very important in setting up fraud and abuse measures, which are the input to analytical prediction/classification models, most authors base their approaches on some kind of expert system, like Major and Riedinger [2002] or Musal [2010]. In addition, machine learning techniques, like Neural Networks or Feature Selection, are applied to identify new patterns and increase automation [He et al.,

1997; Yang and Hwang, 2006; Aral et al., 2012]. Some approaches focus purely on provider behavior [Shin et al., 2012] whereas others also involve patient behavior and interactions of players [Thornton et al., 2013; Parente et al., 2012]. Also in the field of health insurance, some researchers have started to make use of the advantages of Bayesian techniques. For example, Ekina et al. [2013] use Bayesian co-clustering to identify fraudulent providers and beneficiaries.

Latent variable approaches have already been used in different research areas. An overview can be found in Skrondal and Rabe-Hesketh [2004]. Bayesian latent variable approaches, similar to the one introduced in this chapter, have been applied in AIDS prevention [Adebayo et al., 2011], in the examination of human birth defects [Sammel et al., 1997] and in social science [Fahrmeir and Raach, 2007; Fahrmeir and Steinert, 2006]. So far, Bayesian latent variable models have not yet been applied to detect fraud and abuse in health insurance. Bayesian latent variable models with multinomial response as well as the combination with Bayesian shrinkage techniques are new fields of statistical research.

4.2.2 Background

As already mentioned in the introduction, one key element of an efficient analytical fraud and abuse detection method is the availability of a sufficient number of cases with quality assured response that can be used for supervised learning. As basis of the analyses, a fraud and abuse dataset from an insurance company in the Middle East is used. This dataset includes more than 100,000 manually reviewed cases (collected over a period of four years), which have been assigned to one of the three response categories "unperformed services", "unjustified services", "other billing issues", or the reference category "no irregularities".

Depending on the type of invoice (outpatient, inpatient, pharmaceuticals or dental), fraud and abuse patterns strongly vary and other reports and measures need to be applied. Therefore, separate models for each invoice type with different covariates need to be fitted. However, the model structure described in Section 4.3 can be applied to all kinds of invoices, as long as a sufficient number of reports for members, providers and interactions of players is available. In the following, the model development will be described using the example of outpatient invoices, of which n = 36,796 are in the training dataset.

Even though the analyses performed are based on a comparably large training dataset, there are some obstacles in the development of a prediction model that can be applied to new invoices of the same company and also be transferred to other markets.

The n = 36,796 outpatient invoices have already been pre-selected by claims experts based on a simple rule model (focus on providers who have already been involved in fraud and abuse, as well as on certain diagnoses and procedures). Thereby, the share of invoices in the training dataset with detected irregularities is, at 36.9%, very high and not realistic compared to the assumptions in literature on fraud and abuse rates (about 10% according to Gee and Button [2014], Berwick and Hackbarth [2012] and

European Union Commission [2013]). In addition, the pre-selection leads to a high danger of overfitting to training data if the model is to be applied to non-preselected claims in the same market, or to data from other markets.

Moreover, the pre-selection of invoices leads to the fact that the distribution of response categories in the sample is not representative for the full set of invoices of the same company and for other markets. In the training dataset, 33.3% of invoices belong to the category "unperformed services", 1.1% to the category "unjustified services" and 2.4% to the category "other billing issues". The reason for the high percentage of found cases of unperformed services is that the reviewers focused on this kind of misbehavior because it usually leads to a direct recovery of payments. The other fraud and abuse categories are usually harder to prove and need a different, more extensive investigation process. However, many authors, like Ezekiel and Fuchs [2008], assume that in most markets the major part of the loss is related to abusive behavior or overutilization. Even if the loss is not directly proportional to the saving potential for the insurer, as investigation effort may be higher and direct recovery is difficult for abuse, it is necessary to get a grip on the topic. Therefore, some kind of weighting is needed in the development of a scoring model to give more importance to the categories "unjustified services" and "other billing issues".

From these reasons, a sampling strategy was applied that helps to control overfitting and to increase the transferability of the model to the full set of invoices of the same company and to other companies and markets (see Figure 29).

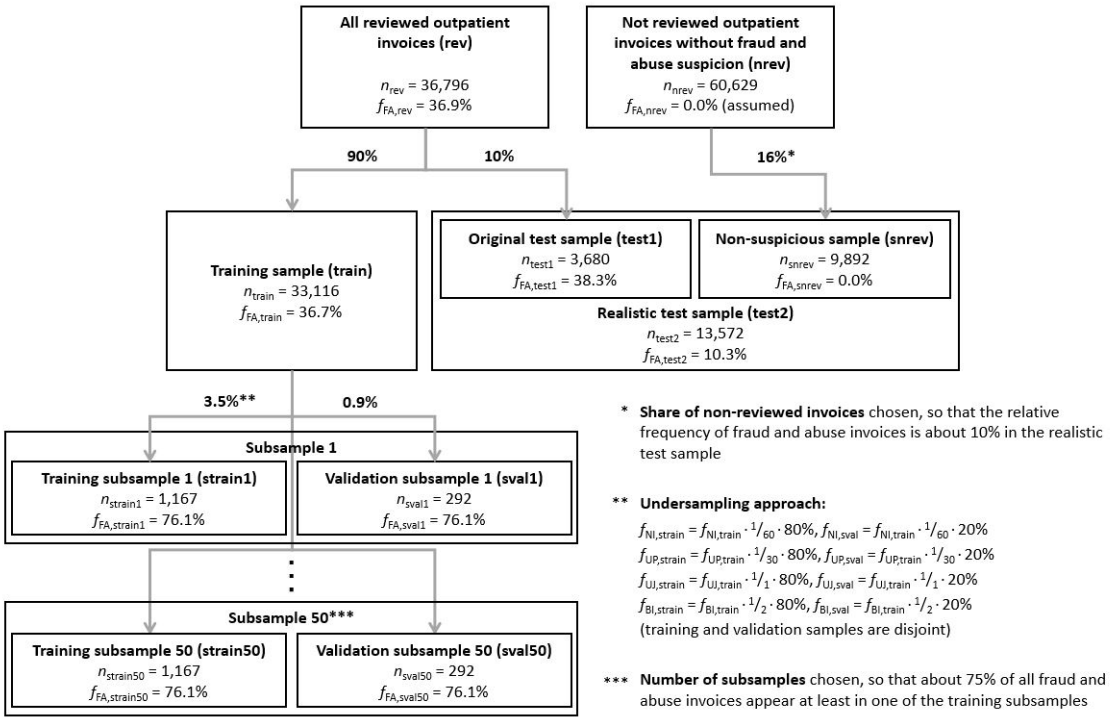

Figure 29: Sampling strategy to control overfitting and to increase model transferability.

As a first step, a 10% test sample (test1) was randomly branched off from the

original dataset of outpatient invoices which is only used to measure the predictive performance of the models. As the test dataset also has an unrealistically high proportion of fraud and abuse cases, also a second test dataset (test2) was used. This second test dataset results from combining the original test dataset with a sample from a pool of approximately 60,000 comparable, but not manually reviewed outpatient invoices. For these invoices, it is assumed that they are not fraudulent or abusive, as there have been no findings in previous reviews of related providers and members. Even if this assumption may be incorrect for some cases, the relative frequency of fraud and abuse cases $f_{\text{FA,test2}}$ is clearly more realistic than for the original test sample. The number of non-reviewed, presumably clean invoices added to test1 is chosen, so that $f_{\text{FA,test2}} \approx 10\%$, which is in line with scientific literature [Gee and Button, 2014; Berwick and Hackbarth, 2012; European Union Commission, 2013].

The remaining 90% of the training sample ($n_{\text{train}} = 33{,}116$) is used for model building. To avoid overfitting and increase the transferability of the model, a subsampling approach (bagging, see Breiman [1996]) is applied based on repeated undersampling of the training sample. More precisely, several stratified subsamples are drawn from the training sample so that the relative frequencies for the response categories "unperformed services" ($f_{\text{UP,strain}}$), "unjustified services" ($f_{\text{UJ,strain}}$), "other billing issues" ($f_{\text{BI,strain}}$) and "no irregularities" ($f_{\text{NI,strain}}$) are approximately equal, by randomly excluding cases from more frequent response categories (exact sampling ratios are given in Figure 29). This balancing approach is proposed by several authors, e.g. Wallace et al. [2011], and has already been applied in (credit card) fraud detection based on highly unbalanced samples [Sahin and Duman, 2011a,b]. In this way, the predictive performance is improved for rarer response categories and especially abusive behavior can better be examined.

The number of subsamples B is chosen so that not too much data information, especially from the more frequent categories "unperformed services" and "no irregularities" is lost. With $B = 50$ more than 75% of all fraud and abuse cases in the training sample occur in at least one of the subsamples. Final parameter estimates and predicted probabilities are calculated by averaging over all 50 subsamples [Breiman, 1996] (see more details in Section 4.3.2). In order to monitor performance, the non-selected invoices from each subsample are used as a validation sample. An additional measure to avoid overfitting is the shrinkage approach integrated in the Bayesian latent variable model introduced in Section 4.3.

In the sampling process, the balancing of response categories is only applied to the training and validation subsamples, but not to the test samples. This proceeding may lead to a model calibration bias and a lower predictive performance on the defined test samples. This decrease in performance is accepted because it is assumed that the balanced subsampling approach reduces overfitting to the training data biased through the pre-selection of invoices and increases the generalizability of the model.

The choice of the described sampling approach is based on different tests of the predictive quality of resulting models and their stability. First, the number of

subsamples ($B = 50$) and the non-category-specific sampling ratio of 80% have been varied which has not lead to a major change of model outcomes. Second, the category-specific balancing weights (1/60, 1/30, 1, 1/2) have been increased (anti-proportionally) to decrease the degree of undersampling. In the extreme case of no undersampling (i.e. weights of 1, 1, 1, 1) the overall predictive results slightly improved, due to a better model fit in the over-represented categories. As the model fit in the under-represented categories, however, dropped drastically, the original balancing weights were used to get a more transferable model with equal attention to all response categories. Third, an oversampling approach of under-represented response categories was applied, i.e. drawing with replacement was performed for these categories. While increasing the degree of oversampling, a decrease in predictive quality could be observed due to an over-adaptation to the training data.

As the measure used to evaluate overall predictive performance weights all response categories equally and independently from the occurrence in the test sample (see Section 4.3.3), it is further assumed that the models which perform best on the defined test samples also have the highest predictive quality for the full set of invoices of the same company and the highest transferability to datasets from other companies and markets. Nevertheless, it must be considered that the absolute predictive quality reached on the test samples is clearly not realistic for the full set of invoices from the same company and for other datasets because the existing pre-selection bias cannot fully be compensated by any sampling technique.

4.3 Methodology

4.3.1 Model structure

The basic idea of the modeling approach is to predict the probabilities π_{UP}, π_{UJ}, π_{BI} and π_{NI} that a new invoice belongs to one of the response categories "unperformed services", "unjustified services", "other billing issues" or "no irregularities", respectively. In theory, an invoice could of course belong to more than one response category, for example if a non-covered person uses the insurance card of a relative and a physician performs a medically not necessary x-ray examination. For the sake of simplicity, however, it is assumed that response categories are disjoint and the response probabilities π_{UP}, π_{UJ}, π_{BI} and π_{NI} add up to one. This implies that the class affiliation y_i of invoice i follows a multinomial distribution.

Further, it is assumed that the parameters π_{UP}, π_{UJ}, π_{BI} and π_{NI} of this multinomial distribution depend on historic provider behavior v_P, member behavior v_M and interaction of players/network behavior v_N, as well as several invoice parameters $c_1, \ldots, c_{r_C}$, i.e.

$$\begin{aligned} y_i | c_1, \ldots, c_{r_C}, v_P, v_M, v_N \quad &\sim \quad \text{Mult}((\pi_{\text{UP}i}, \pi_{\text{UJ}i}, \pi_{\text{BI}i}, \pi_{\text{NI}i})) \\ &\quad \text{with } \pi_{\text{NI}i} = 1 - (\pi_{\text{UP}i} + \pi_{\text{UJ}i} + \pi_{\text{BI}i}). \end{aligned} \tag{44}$$

Examples for invoice parameters are the fraud and abuse potential of diagnoses gathered from several markets, or the number of invoice lines. Following the theory of generalized linear models [McCullagh and Nelder, 1989], a multinomial logistic model is used to link the response parameters to the covariates which yields the model equations

$$\begin{aligned} \log\left(\frac{\pi_{ij}}{\pi_{\mathrm{NI}i}}\right) &= \alpha_{j0} + \alpha_{j1}c_{i1} + \ldots + \alpha_{jr_C}c_{ir_C} + \\ &\quad \sum_j v_{Pij}\beta_{Pj} + \sum_j v_{Mij}\beta_{Mj} + \sum_j v_{Nij}\beta_{Nj} \\ &\quad \text{with } j \in \{UP, UJ, BI\}\,. \end{aligned} \tag{45}$$

In this first stage model, the αs are the regression coefficients belonging to the invoice parameters, and the βs the regression coefficients belonging to the historic behavior of players related to invoice i.

As provider behavior, member behavior and network behavior are latent and not directly measurable, analyses are based on a reporting factory indirectly measuring fraudulent and abusive behavior. The reporting factory incorporates both detailed medical knowledge and sound experience on fraudulent and abusive patterns from different markets. In total, 150 reports with focus on provider, member or interaction of players can be used as far as data availability allows it. The prediction results which are presented in Section 4.4 are based on approximately 100 reports/parameters ($r_P = 50$ provider reports, $r_M = 30$ member reports, $r_N = 10$ network reports and $r_C = 10$ reports on invoice properties). Examples are the number of prescriptions per provider, compared to the average of all providers with the same specialty in the same region, the number of doctors an insured member visits relative to the severity of the diagnosis, and the distance between member and provider.

Regarding the goal of the model to reach high predictive quality for new invoices, it is dangerous to directly incorporate the large amount of measures from the reporting factory into the linear predictor of the regression model. Even with variable selection techniques, overfitting can hardly be controlled, especially due to the fact that the learning sample is not representative (see Section 4.2). This assumption is confirmed by the prediction results of corresponding benchmark models applied in the existing data situation (see Sections 4.3.3 and 4.4).

Therefore, provider, member and network behavior are specified as latent variables. Similar to the idea of structural equation models [Goldberger, 1972], the latent variables are estimated in second stage models using observable provider measures $p_1, \ldots, p_{r_P}$, member measures $m_1, \ldots, m_{r_M}$ and network measures $n_1, \ldots, n_{r_N}$ as covariates:

$$
\begin{aligned}
v_{Pij} &= \gamma_{Pj0} + \gamma_{Pj1} p_{i1} + \ldots + \gamma_{Pjr_P} p_{ir_P} + \varepsilon_{Pij} \\
v_{Mij} &= \gamma_{Mj0} + \gamma_{Mj1} m_{i1} + \ldots + \gamma_{Mjr_M} m_{ir_M} + \varepsilon_{Mij} \\
v_{Nij} &= \gamma_{Nj0} + \gamma_{Nj1} n_{i1} + \ldots + \gamma_{Njr_N} n_{ir_N} + \varepsilon_{Nij} \\
&\quad \text{with } j \in \{\text{UP,UJ,BI}\}\,.
\end{aligned} \tag{46}
$$

The γs denote the regression coefficients of the second stage models and the εs the error terms of the models. As target variables of these models, binary indicators are used which determine whether invoice i falls into the category j or not. This means that behavior of players is interpreted as their fraud and abuse potential, which is measurable as a continuous score. Furthermore, it is assumed that this potential is proportional to the probability that the binary indicators equal 1. As class affiliation y_i is already used as response of the first stage model, overfitting is avoided by a) reducing y_i from four categories to a binary indicator and b) assuming a linear relationship between y_i and the report outcomes. Even though linear models are usually not the right choice to predict a probability, the linear setup is used here, because the focus is more on a score with high predictive value for the first stage model than on an interpretable final result. In addition, the linear setup with the assumption of normally distributed εs on the second stage has proven to be numerically more stable than a logistic model. The estimated score values can also be used to rank provider and members by fraud and abuse potential, which is a beneficial side effect.

The model fitting process introduced in Section 4.3.2 is based on an iterative update of first and second stage model parameters. It includes an additional mechanism to control overfitting based on parameter shrinkage in the second stage models.

In summary, these assumptions lead us to the model structure summarized in Figure 30.

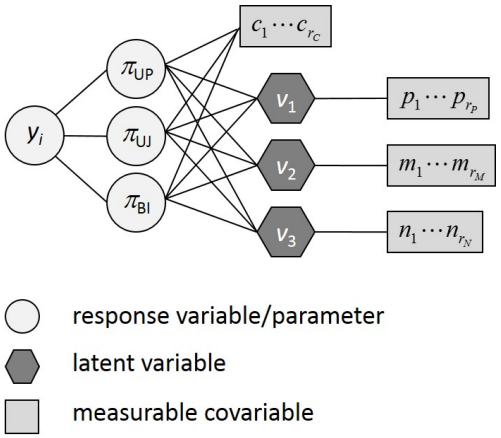

Figure 30: Structure of the proposed latent variable model.

The developed model can be seen as a onedimensional special case of the multidimensional latent variable model proposed by Sammel et al. [1997]. Compared to

other authors who have addressed this class of models, it is here applied to model a multinomial outcome instead of continuous, binary, ordinal or Poisson distributed outcomes [Fahrmeir and Steinert, 2006; Fahrmeir and Raach, 2007; Adebayo et al., 2011].

4.3.2 Model fitting

In order to fit the model parameters $\boldsymbol{\alpha} = (\alpha_{j0}, \ldots, \alpha_{jr_C})$, $\boldsymbol{\beta} = (\beta_{\text{PUP}}, \ldots, \beta_{\text{NBI}})$ and $\boldsymbol{\gamma} = (\gamma_{\text{PUP0}}, \ldots, \gamma_{\text{NBI}r_N})$, a joint fitting algorithm for both stages is applied based on Metropolis-Hastings sampling for each of the subsamples. Final estimates are derived from the bagging approach already mentioned in Section 4.2 by calculating median values over all subsamples. The MCMC sampling approach is a straightforward choice for parameter estimation because it intuitively allows an alternating update of first and second stage model parameters. The algorithm uses the following (simplified) reformulation of the posterior distribution of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$:

$$\begin{aligned} \underbrace{p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} | y, \boldsymbol{v})}_{\text{posterior}} &\propto \underbrace{p(y, \boldsymbol{v} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})}_{\text{likelihood}} \cdot \underbrace{p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})}_{\text{prior}} \\ &= p(y | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \cdot p(\boldsymbol{v} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \cdot p(\boldsymbol{\alpha}, \boldsymbol{\beta} | \boldsymbol{\gamma}) \cdot p(\boldsymbol{\gamma}) \\ &= \underbrace{p(y | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \cdot p(\boldsymbol{\alpha}, \boldsymbol{\beta} | \boldsymbol{\gamma})}_{\text{first stage model}} \cdot \underbrace{p(\boldsymbol{v} | \boldsymbol{\gamma}) \cdot p(\boldsymbol{\gamma})}_{\text{second stage models}} \end{aligned} \tag{47}$$

The likelihood in the first line is split into two parts based on the assumption of conditional independence between $\boldsymbol{v} = (v_{\text{PUP}}, \ldots, v_{\text{NBI}})$ and y given $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

As an additional instrument to control overfitting to the training data, a parameter shrinkage option is included in the algorithm. With the setting described in the following, sampling took on average about 2.2 minutes per subsample without parameter shrinkage and 2.5 minutes per subsample with shrinkage (on a 64 GB RAM working station).

The fitting algorithm starts by drawing the second stage parameters $\boldsymbol{\gamma}$ using the Metropolis-Hastings algorithm [Robert and Casella, 2004]. For the linear models, a hierarchical decomposition model is applied to draw from the posterior distribution of $\boldsymbol{\gamma}$ and the error variance σ^2:

$$\underbrace{p(\boldsymbol{\gamma}, \sigma^2 | v)}_{\text{posterior}} \propto \underbrace{p(v | \boldsymbol{\gamma}, \sigma^2)}_{\text{likelihood}} \cdot \underbrace{p(\boldsymbol{\gamma} | \sigma^2) \cdot p(\sigma^2)}_{\textit{prior}} \tag{48}$$

For the version without parameter shrinkage, Gibbs sampling is used as a special case of the Metropolis-Hastings algorithm (acceptance probability of proposed parameters always equal to one) implemented in the function `MCMCregress` in the R package `MCMCpack` [Martin et al., 2011]. Here, a semi-conjugate prior to the normal likelihood $p(v | \boldsymbol{\gamma}, \sigma^2)$ is used. More precisely, a weakly informative multivariate normal prior distribution is used for $\boldsymbol{\gamma}$:

$$\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}, 0.1\boldsymbol{I}) \tag{49}$$

and an inverse Gamma prior for the error variance σ^2 (with mean 5 and variance 25) which is assumed to be independent from $\boldsymbol{\gamma}$.

For the version with parameter shrinkage, a lasso regression model [Park and Casella, 2008] with reversible jump mechanism [Green, 1995; Troughton and Godsill, 1997] implemented in the `blasso` function in the R package `monomvn` [Gramacy et al., 2007] is applied. The reversible jump algorithm modifies the acceptance probability of the applied Metropolis-Hastings algorithm by introducing moves between parameter spaces of different dimensionality [Troughton and Godsill, 1997].

The lasso penalization is implemented by including the shrinkage parameter λ which controls the degree of parameter shrinkage in the hierarchical model representation. Then, an exponential power distribution is assumed as prior for $\boldsymbol{\gamma}$ which is no longer independent from the error variance σ^2:

$$p(\boldsymbol{\gamma}|\sigma^2) \propto \prod_{l=1}^{r} e^{-\lambda(|\gamma_l|/\sqrt{\sigma^2})} \ . \tag{50}$$

For λ (more precisely, λ^2), a non-informative Jeffrey's hyperprior [Gramacy et al., 2007] is used.

After having updated the second stage models for the first time, the updated parameter vector $\boldsymbol{\gamma}^{(1)}$ is used to calculate first estimations of the latent variable scores $v_{PUP}^{(1)}, \ldots, v_{NBI}^{(1)}$ for all observations in the subsample. Based on these estimations, an update of the first stage model can be performed.

For updating the first stage model, the `MCMCmnl` function in the R package `MCMCpack` [Martin et al., 2011] is used. A multivariate normal distribution with expectation equal to zero and infinite variance for $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$ is assumed, i.e. an improper non-informative prior distribution. First updates $\boldsymbol{\alpha}^{(1)}$ and $\boldsymbol{\beta}^{(1)}$ are received by independent Metropolis-Hastings sampling [Chib et al., 1998]. The Metropolis proposal distribution is centered at the current value of $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$ and has the covariance matrix TCT. T is a diagonal positive definite matrix depending on the tuning parameter that controls the acceptance rate and C is the large sample covariance matrix of the maximum likelihood estimate of $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$.

In each further iteration s ($s = 1, \ldots, S = 250$) of the fitting process, $\boldsymbol{\gamma}^{(s-1)}$ is used on the second stage and $\boldsymbol{\alpha}^{(s-1)}$ and $\boldsymbol{\beta}^{(s-1)}$ on the first stage as initial parameter vectors of the sampling functions. In order to achieve stable convergence of the algorithm, the first 50 draws from each subsample are discarded and the median parameter estimates from iterations 51 to 250 are used as overall estimate for the subsample. Thinning of draws (with thinning parameter 5) was tested, but did not change the parameter estimates significantly. Sampling traces and autocorrelation plots for all parameters show that the sampling algorithm already produces stable estimates in this configuration despite the relatively short burn-in phase and the

relatively small number of draws. As an example, Figure 31 displays sampling trace and autocorrelation plots for γ_{PUP25} (effect of frequency of prescriptions relative to provider specialty and region on score for abusive behavior of providers) based on one subsample.

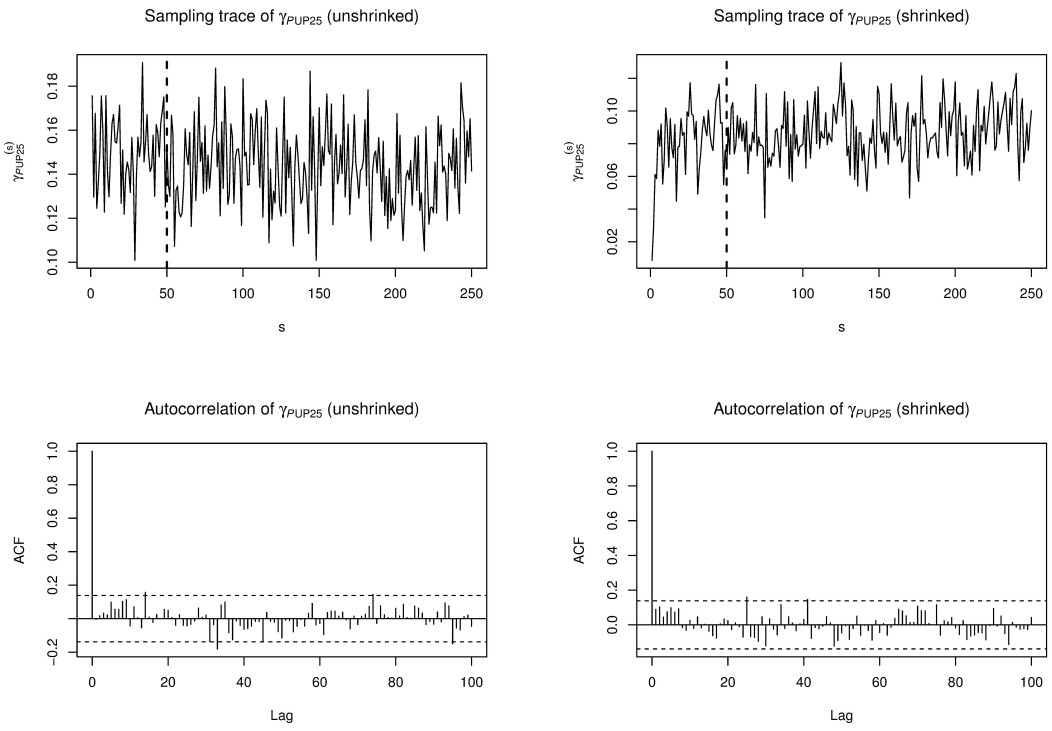

Figure 31: Sampling trace and autocorrelation function of γ_{PUP25} (effect of frequency of prescriptions relative to provider specialty and region on score for abusive behavior of providers) for unshrinked (left-hand side) and shrinked version (right-hand side) of the fitting algorithm.

The prediction result are stabilized by calculating the median of all $\boldsymbol{\alpha}^{(s)}$, $\boldsymbol{\beta}^{(s)}$ and $\boldsymbol{\gamma}^{(s)}$ after burn-in (i.e. $s = 51, \ldots, 250$) over all subsamples denoted by $\boldsymbol{\alpha}_{\text{med}}$, $\boldsymbol{\beta}_{\text{med}}$ and $\boldsymbol{\gamma}_{\text{med}}$. By plugging in these final estimates as well as the observed reporting results and invoice parameters related to a new invoice k into the model equations (45) and (46), a prediction $\boldsymbol{\pi}_k^* = (\pi_{k\text{UP}}^*, \pi_{k\text{UJ}}^*, \pi_{k\text{BI}}^*, \pi_{k\text{NI}}^*)$ of the fraud and abuse probabilities for this invoice is obtained.

Now, for instance, the maximum of $\boldsymbol{\pi}_k^*$ can be used to determine the predicted class y_k^* and compare it with the real class y_k of invoice k in the test sample. Based on this comparison, a misclassification matrix and further measures of predictive performance (see Section 4.4) can be calculated. Of course, Bayesian modeling also allows deriving credibility intervals for the parameter estimates and predicted probabilities. Even though such intervals are a beneficial side result of the fitting algorithm, the focus is rather on the evaluation of median prediction results in this chapter.

An important requirement to the fitting algorithm is that it leads to a reasonable number of reports which have measurable influence on the final predictions. If there are only very few parameters influencing the predictions, the model may be too close to the rule-based system which was applied for pre-selecting the invoices for review. This means that certain patterns will not be recognized by the model, especially if it will be applied to other datasets. If there are too many parameters in the selection model, prediction results may also be poor because multiple correlations will disturb the detection of systematic patterns. From a content perspective, both versions of the algorithm, with shrinkage and without shrinkage, lead to a reasonable number of reports which have measurable influence on the final predictions, where of course, the shrunk version produces sparser models, with some parameters even shrunk to 0. Prediction results for the two test samples introduced in Section 4.2.2 are illustrated in Section 4.4.

In general, the methodology described yields consistent parameter estimates and prediction results, not sensitive to changes in the data input, which is indicated by a relatively small variation of parameter estimates across the subsamples. This stability is a very beneficial property that allows applying the model in other data and market environments (see Section 4.3.4).

4.3.3 Benchmarking

In order to benchmark the predictive performance of the developed Bayesian latent variable model in the existing data situation, it is compared to several other state-of-the-art prediction techniques. As explained, the Bayesian approach uses two stages to predict class probabilities for all response categories. In both steps, the covariates are related to the real response category using different distributional assumptions for the response of both model parts (normal and multinomial). An alternative strategy is to first reduce the dimension of the covariate space without consideration of the target variable and use the resulting factors or principle components as input for the class prediction model. A third possible strategy is to combine variable selection/weighting and class prediction in one step.

Table 12 summarizes the applied approaches together with concrete prediction techniques used for benchmarking. All benchmark models have been calculated based on the subsampling approach described in Section 4.2.

Approach	Technique	Shortcut
One-stage model with variable selection/weighting	Multinomial model with AIC variable selection	A1
	Random Forest	A2
One-stage model with dimension reduction	Multinomial model based on factor analysis	B1
	Polyclass model based on factor analysis	B2
Two-stage model	Bayesian latent variable model without shrinkage	C1
	Bayesian latent variable model with shrinkage	C2

Table 12: Prediction approaches and techniques applied to the underlying classification problem).

Models A1 and A2 assume a direct relationship between the reporting results and class affiliation of corresponding invoices without any latent variables or factors

in between. Model A1 is a multinomial logit model fitted via Neural Networks as implemented in the function `multinom` of the R package `nnet` [Venables and Ripley, 2002]. To ensure model convergence and avoid overfitting in view of the high number of covariates, this technique is combined with a stepwise (forward) variable selection algorithm based on the AIC criterion. As an alternative direct prediction method, a Random Forest is used which is a standard method for class prediction. The forest is calculated with the R function `randomForest` in the R package of the same name [Breiman, 2001]. In order to give each report the chance to show its predictive value, 15 covariates are randomly chosen for the construction of each of the 100 trees per subsample. Variable importance measures show a good mixture of relevant reports.

Like the Bayesian approach, models B1 and B2 reduce the dimension of the covariate space before performing the final classification. However, no regression models considering the target variable are used here, but a dimension reduction technique for all reporting categories instead. More precisely, a factor analysis is used to conserve as much as possible of the covariance between the reports when reducing the dimension of the covariate space. Considering the underlying number of reports and the form of scree plots, 5 factors are used for provider behavior, 3 factors for member behavior and 2 factors for network behavior. Based on these factors, either a multinomial logit model (like for model A1) or a polyclass model as described by Kooperberg et al. [1997] and Stone et al. [1997] are applied. Polyclass models are related to the MARS (multivariate adaptive regression splines) approach first described by Friedman [1991]. They include a stepwise variable selection approach that also takes into account the potential non-linear impact of continuous influence factors, as well as potential interactions between model covariates. As the preliminary factor analysis reduces the covariate space to a limited number of continuous factors, these properties are supposed to lead to a better prediction compared to model B1. Regarding the results, the MARS algorithm indeed detects both non-linearities and interactions between the calculated factors that improve the classification in the described situation. The polyclass models are implemented in the function `polyclass` of the R package `polspline`.

Models C1 and C2 correspond to the Bayesian approach described in Sections 4.3.1 and 4.3.2. Model C1 represents the fitting algorithm without parameter shrinkage (inverse gamma distribution for $\boldsymbol{\gamma}$) and model C2 represents the version with parameter shrinkage (exponential power distribution for $\boldsymbol{\gamma}$).

The results of these six classification techniques will be compared based on the original test sample test1 and the potentially more realistic test sample test2 (see Figure 29 in Section 4.2). As a measure for comparing the overall predictive quality of the model, the average area under the ROC curves is used for all four response categories $\overline{AUC}$ (see the definition in Appendix C). This criterion is applied, as it considers both type I and type II error for each of the response categories and gives equal importance to all response categories. The latter property is especially important to ensure that the distribution of detected cases will be more balanced in the future. Furthermore, it is assumed that models which perform well with regard to the AUC criterion are more likely to deliver reasonable results in other markets, where there is, for example, a stronger focus on abusive behavior. A

simple adaptation of the criterion to markets with unequal importance of response categories can be carried out by using a weighted average of category-specific AUCs instead of a raw average. A more comprehensive approach based on a business-case logic (i.e. the assignment of costs to each right and wrong decision) can be found in Viaene et al. [2004a]. In addition, the different prediction techniques are compared with regard to some content-specific criteria (see Appendix C) which may be of interest to healthcare payers.

4.3.4 Transferability

Based on the analyses described in this chapter, it is – in general – difficult to draw conclusions for other datasets in the same or other markets due to the pre-selection bias described in Section 4.2.2. Especially, the absolute predictive quality of models (see Section 4.4) is not realistic for other datasets. Based on the applied subsampling approach and the extended test sample (see Section 4.2.2), however, it is assumed that a comparison of models provides an indication with regard to predictive quality for the full set of invoices from the same company and for datasets of other companies. In the following, the focus will be on the transferability of the model to other insurance companies and markets and on a strategy to make use of the insights obtained.

The transferability of fraud and abuse detection systems strongly depends on the health system, the regulatory situation, the business model of the insurance company, the insurance product, the competitive situation in the market, the development status of the market, the cost pressure on providers, data quality and availability, and many other factors. However, the motivation for fraudulent and abusive behavior as well as certain behavioral patterns are often comparable across market boundaries. Therefore, it is assumed that the set of reports, which was developed based on experience from different markets, already provides a significant stand-alone value to most insurance companies worldwide. Of course, the feasibility of reports and the validity of statements strongly depends on the availability and quality of data, especially of diagnoses and procedures in an international standard coding system.

The question whether it is reasonable to also transfer the scoring model and estimated parameters is more difficult. Here, it mainly depends whether a significant number of reviewed cases is available as a learning dataset in the new market. In markets where a large number of own reviewed cases is available, it may be more reasonable to fit an own scoring model, for which the proposed methodology can of course be used. However, in most – even developed – markets, no databases with a significant amount of reviewed cases are available. In this situation, it may still be more efficient to transfer the developed model with the same or modified parameter estimates than to rely on an unsupervised classification technique that is purely focused on outlier detection. To reach the predictive quality that is observed for the dataset on which the model has been developed is certainly unrealistic in this case (see the results in Section 4.4). However, the described Bayesian approach offers some intuitive features which facilitate the transfer in different ways.

In the following, an approach how to transfer the model between datasets of different insurance companies from the same or other markets is presented. The approach is based on the methodology described in Section 4.3.1 and a decomposition of the model input in market-/company-specific and general reports. Even within the same market, it is highly unlikely that the original full set of reports can be calculated, as the feasibility of suggested reports depends on availability of information and the validity of reports on data quality and other company-specific conditions. Therefore, a strategy is needed on how to deal with reports that cannot be calculated or are not reasonable in the new data situation and, vice versa, how to deal with additional reports that can only be calculated in the new data environment. Especially for the transfer to other markets, it is also conceivable that certain reports may have a different impact on provider, member and network score as well as on final probabilities due to different regulatory, product or market conditions, e.g. incentives for providers to perform more outpatient treatments to avoid hospitalizations.

As a general rule of thumb, the transfer of regression coefficients from the original dataset to a new dataset can be considered reasonable if at least 60% of the original set of reports can also be calculated in the new data environment. In any case, it is recommended re-fitting the model on the original dataset based on the intersect of reports that are feasible (and reasonable) in both the original and the new market. This proceeding helps to find surrogate reports for non-available reports in the new data environment.

Reports which are only available in the new market environment can be added to the regression framework as offset in the corresponding model equation. For example, a great distance between provider and member is known to increase the probability of fraud. However, this report cannot be calculated in the original data environment due to the non-availability of member address data. If most other reports can be transferred between the markets, a simple improvement of the scoring model is to use the regression parameters from the original market and add a standardized version of provider-member distance as an offset in the network model when scoring cases in the new market. The increased network score will then directly lead to an increase of the fraud probability in the new data environment.

Reports which exist in both the original and the new market, but are expected to have another impact on final fraud and abuse probabilities based on experience from the target market, can be modified by using alternative (informative) prior distributions for the regression parameters. For example, the expectation of the prior distribution for the corresponding element of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ or $\boldsymbol{\gamma}$ can be increased from 0 to a positive value to weight the report more heavily. Of course, the risk that probability estimation will be dominated by those new reports needs to be controlled in this case. As soon as new cases from the target market are available, the modified prior assumptions will be revised by the Bayesian updating process. As the impact of the prior distribution decreases with increasing number of updates, it is recommended reconsidering the burn-in phase and the number of draws in order to control the impact of the adaptation. In any case, several calibration loops may be necessary.

In general, it is reasonable to re-fit the model on a regular basis (e.g. once per quarter) to identify new trends and adapt the model to behavioral changes. Of course, it is also important to add new reports whenever a new fraudulent or abusive pattern is detected. In this way, the scoring approach can be extended to a self-learning system. If the regulatory conditions allow it, the database of reviewed cases can, of course, also be extended with cases from other insurance companies in the same market or across market boundaries. This is especially helpful in markets where no sufficient number of reviewed cases is available to construct an own scoring model. Here, it is interesting to monitor how the regression coefficients change over time, while the reviewed cases from the own market gain more and more weight.

4.4 Results

In the following, the results of different prediction techniques (see Section 4.3.3) are presented based on the original dataset from the Middle East using the test samples described in Section 4.2.2. As already mentioned in Section 4.2.2, it needs to be considered that both training and test datasets are based on a biased pre-selection of invoices. For this reason, the relative frequency of fraud and abuse cases is very high and the relative frequency of abuse cases compared to fraud cases is comparably low. Also, several other characteristics of the dataset (e.g. age distribution) and the market background are not comparable to most other, especially US and European, markets. Consequently, no generalization of results in terms of absolute predictive quality is possible. The target of the benchmarking is to provide an indication of relative predictive quality for the full set of invoices of the same company and datasets of other insurance companies based on the subsampling approach and the extended test sample described in Section 4.2.2.

First, the overall predictive quality of all applied techniques measured by *AUC* (see Section 4.3.3 and Appendix C) are compared based on both test samples test1 and test2 (see Section 4.2.2). Table 13 shows both the category-specific *AUC*s and the (unweighted) average over all categories. In order to also visualize the results of the analysis, Figure 32 summarizes all ROC curves of the applied classification techniques using the example of test sample test1.

In general, the one-stage approaches with variable selection/weighting (A1 and A2) perform worst in terms of overall predictive quality. The best results in the existing data environment are obtained from the Bayesian latent variable model (C1 and C2), shortly followed by the regression approaches with preliminary dimension reduction (B1 and B2), which confirms the assumption of a latent behavioral effect. In addition, the two-stage approach with both stages relating the covariates to the target variable does not seem to cause an overfitting issue. It also seems to be a reasonable choice here, because *AUCs* are relatively high across all response categories for both test datasets. This indicates that the efficiency of the future investigation process can also be increased with regard to abuse and other billing issues. The dimension reduction approach also yields stable and good results and can be considered for the application in other markets.

test1					
Technique	AUC_{NI}	AUC_{UP}	AUC_{UJ}	AUC_{BI}	$\overline{AUC}$
A1	**0.90**	0.74	0.70	0.75	0.77
A2	0.67	0.60	0.75	0.76	0.70
B1	0.78	0.81	0.77	0.79	0.79
B2	0.74	0.81	0.84	**0.90**	0.82
C1	0.80	**0.83**	**0.92**	0.86	**0.85**
C2	0.80	**0.83**	0.91	0.86	**0.85**
test2					
Technique	AUC_{NI}	AUC_{UP}	AUC_{UJ}	AUC_{BI}	$\overline{AUC}$
A1	**0.88**	0.56	0.60	0.80	0.71
A2	0.64	0.56	0.81	**0.86**	0.72
B1	0.68	0.79	0.80	0.68	0.74
B2	0.72	0.81	0.83	0.78	0.79
C1	0.74	0.87	**0.91**	0.76	0.82
C2	0.76	**0.88**	**0.91**	0.75	**0.83**

Table 13: Category specific and average AUCs for both test samples test1 and test2 (overall best results marked in boldface).

The one-stage AIC selection model (A1) performs best in terms of NI cases, but clearly worse than most other techniques in the identification of fraud and abuse cases. The Random Forest (A2) is the only technique which performs better on the sample with more realistic total fraud and abuse frequency. However, it needs to be considered that compared to all other techniques, the results from the Random Forest strongly depend on tuning parameters and, therefore, are rather unstable. Regarding the techniques with preliminary factor analysis (B1 and B2), it is noticeable that the non-linear polyclass model (B2) performs better in the categories UJ and BI compared to the standard multinomial model (B1). Here, some non-linear effects are observed as well as interactions which clearly improve the predictive quality. On the other hand, it needs to be considered that those non-linear effects may be very market-specific, so that the standard multinomial model may be more stable when the model is applied to another market. The Bayesian latent variable models without and with shrinkage (C1 and C2) deliver very similar results. It is assumed that the effect of variable shrinkage may be more significant for smaller sample sizes and more input parameters.

Beside the overall predictive quality, also some specific performance indicators relevant from the perspective of insurance companies and other healthcare payers are evaluated. For this purpose, the additional category "fraudulent or abusive" (FA) is introduced by summarizing the categories UP, UJ or BI. Table 14 gives an overview of all content-related measures based on the test sample test2 (an explicit definition and explanation of measures can be found in Appendix C).

Technique	TPR_{FA}	RCR_{FA}	PPV_{NI}	AAR_{NI}	PPV_{FA}	$n_{95\%}$
A1	0.69	0.46	**0.99**	0.26	0.13	22
A2	0.76	0.55	0.95	0.27	0.12	24
B1	0.78	0.54	0.84	0.34	0.15	18
B2	**0.81**	0.57	0.84	0.43	**0.17**	**16**
C1	**0.81**	0.61	0.96	0.50	0.16	17
C2	**0.81**	**0.63**	0.96	**0.53**	**0.17**	**16**

Table 14: Content-specific performance measures of all applied classification techniques based on test sample test2 (overall best results marked in boldface).

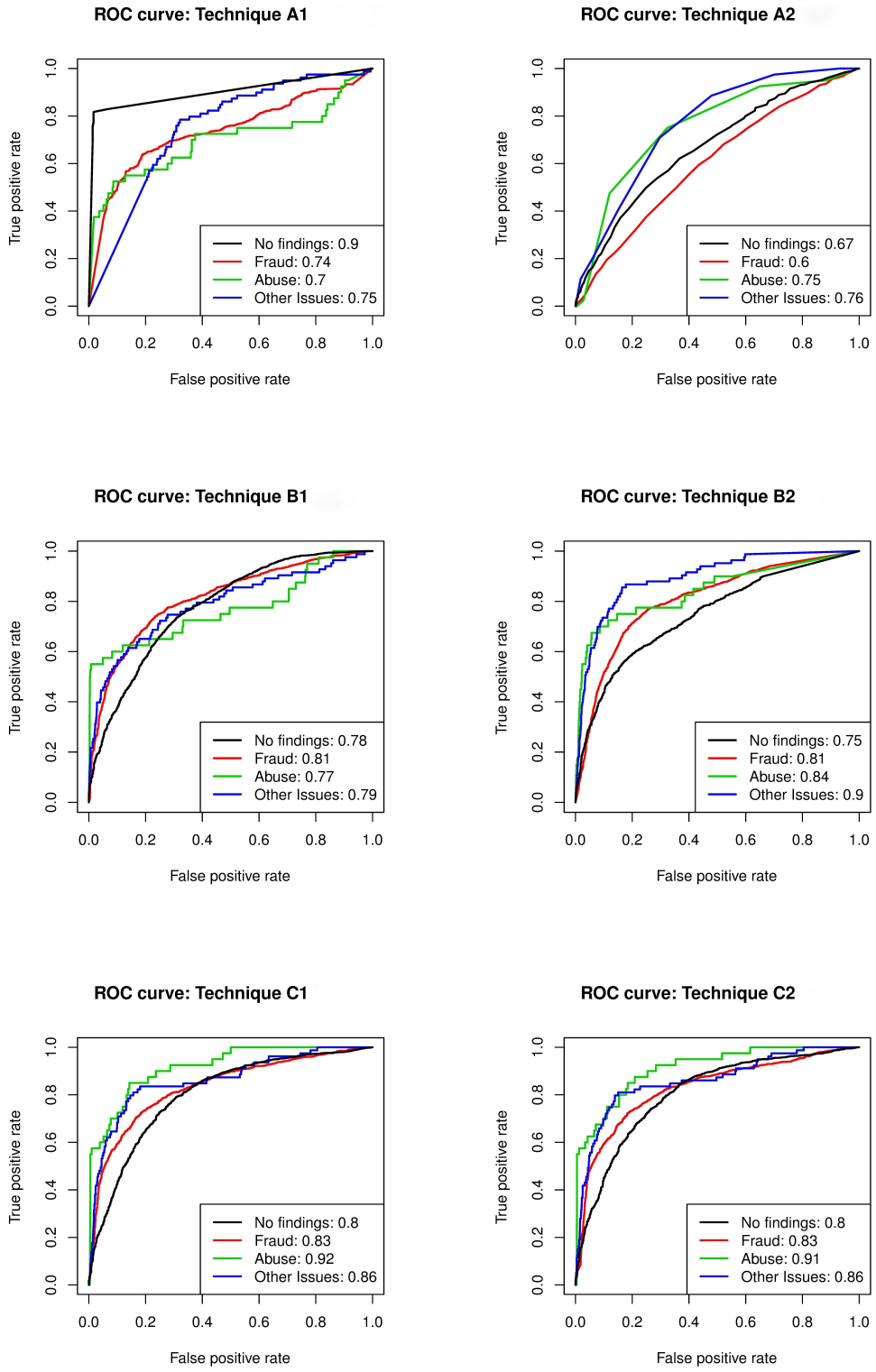

Figure 32: ROC curves for all applied classification techniques based on test sample test1.

The first measure TPR_{FA} gives an indication of the percentage of real fraud and abuse cases that can be identified as such. Also here, it must be considered that the figures are assumed to be substantially lower for the full set of invoices of the same insurance company and for datasets of other insurance companies (from the same or other markets) due to the mentioned pre-selection bias. Regarding the true positive rate based on the test dataset test2, the Bayesian latent variable models perform best together with the non-linear polyclass model based on dimension reduction with an identification rate of 81%.

Also in terms of correct class allocation within the real cases, the Bayesian models deliver the best results with a correct class allocation rate of above 60%. Regarding the precision with which "clean" cases are identified correctly, the AIC selection model (A1) performs best with a precision of 99% (see also the high AUC value for the category NI). However, only about one-fourth of all invoices can be auto-adjudicated based on this classification approach. Here, the Bayesian models are clearly preferable with a similar precision (96%), but the double auto-adjudication potential of about 50%.

Finally, the reliability of a decision to filter out an invoice is analyzed. Here, the one-stage models with preliminary dimension reduction are on the same level as the Bayesian latent variable models with a positive predictive value of fraud and abuse above 15%. This means – at least for the existing data situation – that less than 20 invoices need to be analyzed to uncover one fraud or abuse case with a probability of 95%. Even though these figures only hold for the biased test datasets, the comparison of techniques indicates that the Bayesian models yield the highest savings potential for insurers in terms of operational costs.

Looking at the overall picture, the results are quite similar to the AUC evaluation: the Bayesian models perform best, shortly followed by the dimension reduction models. The content-related measures give a rough indication of the saving potential for insurers and healthcare payers in the existing data situation. Due to the stability of results, it is recommended applying the Bayesian latent variable model and the dimension reduction techniques in similar situations, even though it is assumed that the high absolute predictive quality cannot be generalized. Especially, the results in the abuse category require further validation based on other datasets, due to the small number of underlying cases in the existing situation. The Bayesian approach offers the additional advantages of high stability (see Section 4.3.2) and adaptability based on flexible prior specifications (see Section 4.3.4), which are helpful when transferring the model to other data situations.

4.5 Summary and Outlook

This chapter transfers the idea of Bayesian latent variable modeling known from other contexts, such as social science [Fahrmeir and Raach, 2007; Fahrmeir and Steinert, 2006], to the problem of fraud and abuse detection. To handle the challenges of this specific problem, a Bayesian latent variable model for multinomial response variables was developed. The idea behind the approach is that the report

results do not directly affect the class affiliation of an invoice, but indirectly via latent variables which summarize behavioral aspects of each reporting perspective. In this way, the full information potential of the underlying expert system shall be exploited. The developed fitting algorithm involves a lasso parameter shrinkage option intended to control overfitting to the training sample. To further increase the transferability of the model, a subsampling approach is used to balance the skewed class distribution in training data.

The prediction results indicate that the introduced Bayesian approach improves the relative predictive quality compared to several benchmarking techniques. Therefore, it is assumed that the methodology can successfully be applied to the full set of invoices of the same insurance company and to datasets of other insurance companies. However, the observed absolute predictive quality is not realistic due to a pre-selection bias affecting training and test data. Especially those approaches which directly use the report results as input do not exploit the full potential of the data in terms of predictive quality. Regression techniques with preliminary dimension reduction (factor analysis) yield a predictive quality similar to the Bayesian approach. As these techniques are computationally faster and based on standard routines of analytical software packages, they may be preferred in practical situations, especially in the case of very large datasets.

A general advantage of the Bayesian approach is its adaptability to other data and market environments. The expected influence of new and already existing reports can be adjusted by corresponding (informative) prior specification. In situations where no or only few reviewed cases are available as learning data, the transfer of the model including (modified) parameter estimates is assumed to be more efficient than the application of an unsupervised system only focused on outlier detection. Prerequisite for such transfers are comparable data availability and quality in the original and the new market. On the other hand, specification of wrong informative prior distributions may decrease the predictive quality in case of small training datasets which is a drawback compared to frequentist approaches.

The developed model can be applied in real time based on a score card approach. This means that the model does not need to be re-fitted for every new invoice, but only on a monthly or quarterly basis. This frequency is assumed to be sufficient to keep pace with the dynamic adaptation of behavior of providers and insured persons. The high stability of the described modeling approach and the prediction results (i.e. low dependency on data input) ensures the validity of the scoring for this period. Another quality of the Bayesian approach is its high interpretability. In particular, it allows backtracking of a high probability to those reports which caused it. This gives the investigator a starting point for further evaluation and allows a more targeted investigation process.

The efficiency of the claims adjudication process can further be increased by establishing specific investigation units which are trained to take appropriate action based on the scoring results. It is also important to note that usually only a small percentage of the saving potential with regard to fraudulent and abusive behavior is related to direct recovery. The larger part is assumed to arise from a measurable

deterrent effect as well as a better position of the insurance company in provider network negotiations.

Statements and conclusions presented in this chapter mainly refer to the existing data situation. Therefore, the subject of further research will be to test the (absolute) predictive performance and stability of the model based on other insurance or healthcare datasets. Based on validated results, it further needs to be assessed from an economic perspective if the increase in predictive quality can be translated into cost savings which justify the implementation effort. Especially interesting is whether the shrinkage option will lead to a stronger differentiation of predictive quality in other datasets. Also, further benchmarking techniques, like a recent boosting approach for multi-class problems with high dimensional covariate space suggested by Zahid and Tutz [2013], may be considered. A potential extension of the model is the inclusion of a spatial term in the linear predictors of first stage models [Adebayo et al., 2011], as fraudulent and abusive behavior is known to be strongly dependent on geographic location. Besides, it would be interesting to evaluate whether more latent variables can improve the predictive quality, e.g. the allocation of provider reports to several behavioral aspects represented by own latent variables/first-stage models.

Bayesian latent variable models may have many other applications in the health insurance industry. For instance, Section 5 proposes a similar Bayesian latent variable approach to optimize the risk adjustment in the related context of analyzing medical outcome quality.

5 Provider Quality Measurement

Beside the financial behavior of medical providers, insufficient medical treatment quality is a major cost factor for healthcare payers and may lead to serious consequences for patients. Section 5 illustrates a model-based approach that allows a fair scoring of medical providers based on historic performance in specific medical treatments, especially surgeries. Goal of the statistical approach that estimates a complication score is to perform a risk adjustment considering the medical risks of the providers' patients. In order to account for the fact that different complications of a surgery may have different related risk factors, a latent variable model fitted by Bayesian inference techniques is applied and benchmarked with other statistical approaches. Based on the statistical scoring approach, an objective and transparent provider quality ranking is derived that can be used by patients to choose the best doctor for planned surgeries, by insurance companies to improve their steering of provider networks and by regulators to impose fair pay-for-performance systems controlling healthcare costs.

5.1 Introduction

The treatment quality in medical surgeries is a factor strongly influencing total treatment costs and patients quality of life. The costs of treating occurring complications can easily exceed the costs of the original surgery and cause long-term health issues [Dimick et al., 2003, 2006]. Therefore, it is beneficial for patients and health insurers to identify those treating doctors which reduce the risk of complications as far as possible. For this purpose, some health insurers already offer their insureds rankings of top experts/hospitals for specific (predictable) surgeries based on a retrospective quality score. At the same time, a risk-adjusted and, therefore, fair performance scoring forms the foundation of regulatory pay-for-performance systems which encourage high outcome quality and lead to a cost control in healthcare [Porter and Teisberg, 2006].

An important challenge to create such quality score is to define an adequate risk adjustment which does not discriminate treating doctors with more severe cases and vice versa [Powell et al., 2003]. This means that the score needs to be based on a comprehensive risk analysis of historic cases. Also, the ideal treating doctor for one patient might not be the right choice for another patient with different clinical history. So, an individualization of a doctor ranking to the individual (medical) situation of the patient is preferable.

In principle, the methodology described in this chapter can be applied to every kind of medical treatment with measurable risk factors on patient side (like age or comorbidities) and measurable complications. From an insurance perspective, it is preferable to concentrate on prevalent and cost-intensive surgeries. Besides, steering possibilities are limited for emergency treatments. Therefore, it is recommended to focus on predictable treatments, like cataract removal, keratomileusis (eye lasering), (planned) ceasarian section, cruciate ligament surgery, knee/hip replacement and

lithotripsy. In the following, the example of cataract surgery is used to illustrate the suggested analytical approach. Section 5.2 gives an overview of the underlying dataset and relevant literature on medical outcome quality research.

The modeling approach yields a risk-adequate and individualized quality scoring of treating doctors (or hospitals) with regard to specific surgeries. This scoring shall both support patients in their doctor's/hospital's choice as well as allow insurance companies to partner with best performing doctors/hospitals and improve their steering of provider networks. Therefore, the main focus lies on the reduction of medical risks for future patients and of follow-up costs for the insurer.

To ensure the focus on medical risk adjustment, the modeling approach only considers patient related risk factors (like sociodemographic factors, clinical history or comorbidities). Other factors (like service quality or geographical proximity to the patient) which may also influence the choice of the medical provider could – from a technical perspective – easily be integrated in the modeling. However, such criteria may bias the medical risk adjustment and not lead to best medical and financial outcomes. It is therefore recommended that these criteria are considered separately (e.g. by applying filters in the final ranking).

Foundation of the statistical modeling is a comprehensive medical analysis of potential complications, their reasons, consequences and related risk factors. As the resulting provider ranking shall not only consider one aspect of medical treatment quality, but cover the risks of all important potential complications of the surgery, a comprehensive complication score is used as target variable of the model. To account for the fact that the risk for each potential complication might be driven by different risk factors (and interactions of risk factors), a Bayesian latent variable model is used which allows for more flexibility than traditional regression models (see Section 5.3.1).

Beside the risk adjustment/modeling approach itself, Section 5.3 also describes an intuitive way to translate the model outcomes in a risk-adjusted provider ranking. Based on a simple clustering approach, this ranking can be individualized for single patients who can determine the best treating doctor according to their individual risk parameters. Another side result of the modeling approach is the possibility to predict expected costs of complications which can be used as input for pre-authorization decisions or case reserving in health insurance. To increase the predictive power and, therefore, the transferability of the model re-sampling and shrinkage techniques are applied.

Finally, several risk adjustment techniques are compared with regard to different measures of predictive quality to identify the best approach (see Sections 5.3.4 and 5.4). Predictive quality is particularly important, if an individualized provider scoring for a new patient shall be provided. Section 5.4 also analyzes the stability of the approach and provides further side results.

5.2 Background

In the following, further background information on the statistical assessment of medical treatment quality is provided. Section 5.2.1 gives an overview of related literature and carves out the innovative aspect of the developed risk adjustment approach described in Section 5.3. Additionally, Section 5.2.2 summarizes the most important aspects on the data foundation that is required to perform a meaningful analysis of medical quality in surgical procedures.

5.2.1 Literature

Many, especially US authors address the topic of medical quality assessment from a pay-for-performance perspective. They often report controlled trials analyzing if pay-for-performance providers deliver better outcomes than control providers (see for example Lindenauer et al. [2007]; Bardach et al. [2013]; Kirschner et al. [2013]; Shih et al. [2014]). Often, the main focus of these articles is on the efficacy of the pay-for-performance system, less on the performance of single doctors or hospitals, like in the analyses described in this chapter. Therefore, the following literature overview will mainly concentrate on articles addressing the application of statistical methods for quality assessment and risk adjustment of single medical providers.

Statistical approaches are applied in medical quality assessment to quantify differences in medical outcomes, perform risk adjustment with regard to the providers' case-mix and predict (costs of) adverse outcomes. Depending on the underlying question, methods range from simple t-tests/ANOVAs comparing two or more groups of providers in terms of a defined outcome measure [Bardach et al., 2013] until (advanced) regression models [Cohen et al., 2013] and machine learning approaches [Zheng et al., 2015]. For example, Zheng et al. [2015] predict the risk of re-admissions using different machine learning techniques (e.g. random forests, neural networks and support vector machines).

Typical outcome measures are risk-adjusted mortality (within a defined timeframe), re-operation, re-admission, occurrence of (specific) complications, costs of surgery (incl. follow-up costs) and length of stay [Thomas and Hofer, 1999; Elmallah et al., 2015; Cohen et al., 2013; Hobson et al., 2015; Lukasiewicz et al., 2016]. Some authors also describe (weighted) composite measures which summarize different endpoints [Shwartz et al., 2008]. Other authors discuss the question, if single endpoints can be used as surrogate measure for overall quality. For example, Press et al. [2013] found that (risk-adjusted) re-admission rates alone are not sufficient to measure overall hospital quality.

Another differentiation in medical quality studies is the question, if quality scores are allocated to single treating doctors or institutions (e.g. practices/hospitals). Especially in the analysis of hospital quality, analyses are often not only focused on one specific procedure, but consider the performance in groups of procedures related to a provider specialty or even in all procedures offered by the provider [Cohen et al., 2013]. As in the case of surgical procedures medical outcomes strongly depend on

the (experience of) the surgeon [Birkmeyer et al., 2013], it is preferable to derive a clinician specific score in this case. For example, Birkmeyer et al. [2013] analyze the impact of the surgeon on medical outcome quality in bariatric surgery and report significant differences with regard to post-operative complications, re-operation and re-admission rates. An aggregation of single clinicians' scores to a hospital score is, of course, always possible.

Cohen et al. [2013] describe a statistical risk adjustment approach based on logistic regression models similar to the approach described in Section 5.3. They also derive a quality score based on a ratio between observed and expected complications. Difference to the modeling approach described in this chapter is that they build separate models for each potential complication. Other interesting components of their work are also the introduction of a random effect for the hospital, a shrinkage concept that allows to also evaluate providers with small number of observations and the calculation of confidence intervals for the quality score. Such confidence intervals allow to judge if a provider is significantly better (or worse) than the market also considering the number of underlying cases.

Also Bayesian techniques have already been applied in medical quality assessment. For example, Shwartz et al. [2008] use a Bayesian latent variable model to calculate a composite measure of hospital quality based on different quality indicators.

For the specific example of cataract surgery which is used to illustrate the mentioned risk adjustment approach, there exists a lot of literature on potential complications and risk-factors Chan et al. [2010]; Patalano [2016]; Powe et al. [1994]. Statistical analyses are mainly applied to identify risk factors for specific complications, like endophthalmitis [Li et al., 2004] and retinal detachment [Tielsch et al., 1996].

In a targeted research of scientific literature, no articles could be found that measure medical outcome quality in cataract surgery based on a comprehensive complication score. Also, no latent variable approach for risk adjustment that allows for different impacts of risk factors per complication has yet been applied in the context of cataract surgery and other surgical procedures. Other application examples for Bayesian latent variable models can be found in Section 4 where a similar model is applied in the context of fraud and abuse detection.

5.2.2 Data Environment

Like the described statistical analyses in all previous chapters, the analysis of medical provider quality is also based on insurance (policy and claims) data. Both for the analysis of disease management programs and fraud and abuse detection the availability of high-quality medical information is a plus. In the context of medical quality assessment, however, meaningful results are only possible based on a significant amount of data with high medical coding quality (diagnoses and procedures) [Powell et al., 2003]. Also, the number of observed surgeries meeting the inclusion criteria drives the reliability of resulting provider scores. Additional medical information/clinical data, like results of lab tests or diagnostic tests, of course, improve the risk adjustment capability of statistical models.

In the case described in this chapter, the analysis is based on more than 20,000 cataract surgeries (CPT codes 66982, 66983, 66984) and more than 60 surgeons which have performed at least 25 surgeries captured in the database. For model fitting all observations have been used, the resulting provider rankings only include those providers with more than 25 surgeries to reduce the influence of single outliers. Even though, some authors provide approaches to also rank providers with less observations, like the shrinkage approach described by Cohen et al. [2013], conclusions based on a small number of cases bear the danger of a ranking driven by single (outlying) events.

In the data used for the analysis, medical coding quality with regard to diagnosis and procedure (incl. drug) coding is high and the information is available over a period of about 6 years. In total, 17 measurable complications, like re-operation, retinal detachment and endophthalmitis, are considered and summarized in different complication scores (see Section 5.3.1). For risk adjustment, socio-demographic (age, gender) and non-cataract related risk factors (general health score) as well as 20 indication-related risk factors/co-morbidities (e.g. presence of glaucoma, diabetes, cardiovascular disease, strong short-/long-sightedness) are included in the model.

5.3 Methodology

In this methodology section, a detailed description of the Bayesian latent variable model used for risk adjustment in provider scoring for cataract surgery will be given. Section 5.3.1 illustrates the model assumptions and structure of the developed latent variable model. In Section 5.3.2, the iterative fitting algorithm using parameter shrinkage and re-sampling techniques to increase the predictive power of the model are outlined. The question how to translate model estimates into a provider scoring/ranking is addressed in Section 5.3.3. Here, also a concept for the individualization of the ranking to the specific medical situation of a patient is introduced. Furthermore, the capability of the model to provide individual predictions of complication costs is discussed. Finally, Section 5.3.4 summarizes the applied benchmarking concept in preparation of the results overview given in Section 5.4.

5.3.1 Model structure

As mentioned in the introduction, goal of the analysis is not to score providers based on a single aspect or a surrogate measure for performance (like the re-admission rate [Press et al., 2013]). Instead, the scoring shall be based on a comprehensive complication score which summarizes the most relevant potential complications of a specific surgery. Therefore, we use the complication score R as target variable of our modeling approach. The realizations r_i ($i = 1, \ldots, n$, $n = 20{,}666$) are defined as weighted sum of q ($q = 17$) different complications indicators y_{ij}

$$r_i = \sum_{j=1}^{q} y_{ij} \cdot w_j \quad (i = 1, \ldots, n; j = 1, \ldots, q) \tag{51}$$

where y_{ij} is equal to one if the jth complication occurs in surgery i and else equal to zero. The complication-specific weights w_j are defined from different perspectives:

- From a cost perspective (i.e. the perspective of a healthcare payer), it is straightforward to directly use the costs of the complication as weight w_j. In theory, one could use the costs c_{ij} of complication j in case i as weight w_{ij}. However, it is difficult based on insurance claims data to exactly determine the costs for a complication, especially if complications are defined as occurrence of a specific diagnoses, as costs are related to procedures and not diagnosis. To balance this uncertainty, it is suggested to use the average costs $\bar{c}_j$ of complication j which are determined by the difference of the average claims costs of surgeries where complication j occurs (as only complication) and the average claims costs of surgeries without complications. The weight w_j can then be defined as absolute average complication costs $w_j = \bar{c}_j$ or as relative average complication costs $w_j = \bar{c}_j / \sum_{j=1}^{p} \bar{c}_j$. Even though the relative version may have some numerical advantages, the absolute version is used in the following due to its better interpretability (see Section 5.3.4). The target variable based on $w_j = \bar{c}_j$ is denoted in the following as R_1.
- As the first complication score has a multi-modal distribution (see Figure 33) which is an undesirable property for the target variable of (standard) regression models, a second cost-based complication score is defined. For this second complication score (denoted as R_2), a discrete weight function is used: $w_j = 1$ for low-cost complications, $w_j = 2$ for medium-cost complications and $w_j = 3$ for high-cost complications. Allocation to the cost categories is based on scientific literature (as far as available) and medical expert judgement.
- The third complication score (denoted as R_3) represents the patients' perspective and accounts for severity of the complication, i.e. the impairment of the patient's quality of life. Like for the second complication score, a discrete weight function is used: $w_j = 1$ for mild/short-term complications, $w_j = 2$ for medium/mid-term complications and $w_j = 3$ for severe/long-term complications. It is suggested to use the related models for all applications scenarios where future patients are offered a provider ranking to support their doctor's choice. As the similar shape of the histograms of R_2 and R_3 indicates (see Figure 33), both complication scores are highly correlated (correlation greater 0.9). Therefore, also the cost perspective is reflected by the severity-based score.

An intuitive way to perform risk adjustment based on the defined target variables is to fit a multiple (linear) regression model using the risk factors $X_1, \ldots, X_p$ ($p = 23$) described in Section 5.2.2 as covariates which yields the regression equation:

$$\mathbb{E}(r_i | x_{i1}, \ldots, x_{ip}) = \gamma_0 + \gamma_1 x_{i1} + \ldots + \gamma_p x_{ip}. \tag{52}$$

In order to account for the non-negative values of the complication scores a logarithmic transformation of the target variable has been tested ($\log(r_i + 1)$ to account for the large number of zeros). However, the related model yields worse results in

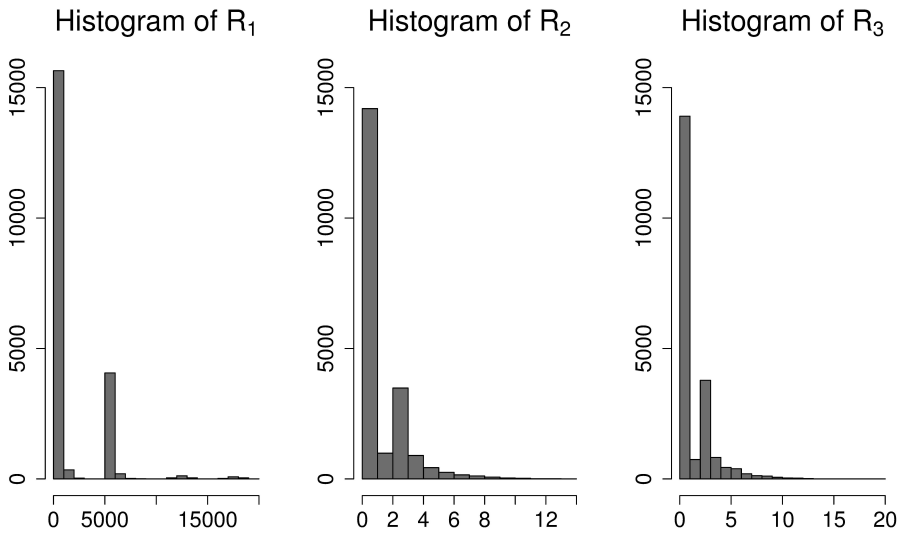

Figure 33: Histograms of the three defined complication scores.

terms of prediction and is not considered further. The linear model approach is used as a baseline benchmark (see Section 5.3.4) for the performance of the more sophisticated models described in the following. The major disadvantage of this approach is, that it implicitly assumes the same impact of each risk factor for each complication. Also, there might be different interaction effects of risk factors on different complications which cannot be accounted for.

Therefore, it is proposed to use a more complex model structure involving the latent variables $V_1, \ldots, V_q$ that can be interpreted as risk for the occurrence of complication j. The underlying assumption is that the complication score R does not directly depend on the risk factors $X_1, \ldots, X_p$, but indirectly via the latent variables $V_1, \ldots, V_q$ (see Figure 34).

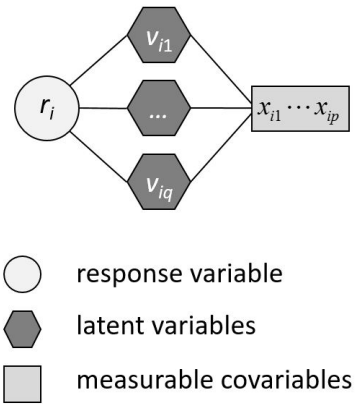

Figure 34: Model structure of latent variable model for risk adjustment.

Similar as for the multinomial latent variable model for fraud and abuse detection presented in Section 4, the indirect impact of covariates on the target variable is reflected in a two-stage model setup.

The first stage model involves the latent variables in the linear predictor:

$$\mathbb{E}(r_i|v_{i1},\ldots,v_{iq}) = \alpha_0 + \alpha_1 v_{i1} + \ldots + \alpha_q v_{iq}. \tag{53}$$

The realizations of the latent variables are estimated in second stage models of the form

$$\begin{aligned} g(\mathbb{E}(v_{i1}|x_{i1},\ldots,x_{ip})) &= \beta_{01} + \beta_{11}x_{i1} + \ldots + \beta_{p1}x_{ip} \\ &\vdots \\ g(\mathbb{E}(v_{iq}|x_{i1},\ldots,x_{ip})) &= \beta_{0q} + \beta_{1q}x_{i1} + \ldots + \beta_{pq}x_{ip}. \end{aligned} \tag{54}$$

As target variables of the second stage models the binary complication indicators y_{ij} are used instead of the unknown latent variables v_{ij}. Therefore, it is an intuitive choice to assume a binomial distribution for $Y_1,\ldots,Y_q$ and use a logistic link function to model the probability $\mathbb{P}(Y_j = 1)$ of complication j as latent risk score (i.e. $g(\cdot) = \text{logit}(\cdot)$). As a second option, one can also argue that the latent variables represent a continuous risk score which does not necessarily need to range between zero and one. In this case, the second stage models are again linear models with binary target variables (i.e. $g(\cdot) = \text{id}$). The latter approach has the advantage that it is clearly more stable, as no separability problems can occur due to a large number of binary covariates (occurrence of risk factors: yes/no). Another theoretical advantage is that the regression estimates for the latent variables which are used as covariates of the first stage model rather follow a normal distribution than in the binomial setup.

To account for the most important medical correlations between the risk factors, manually selected (based on medical expert judgement) interaction terms are additionally considered in the second stage models. Similarly, a moderate number of interaction terms between different latent variables are included in the first stage models to respect the correlation of different complications. Finally, an iterative estimation algorithm is applied which alternatingly updates parameters of first and second stage models (see 5.3.2) to take into account the dependency of first and second stage target variables.

This latent variable approach strongly increases the flexibility of the risk adjustment. However, it is quite complex and requires many parameters. In order to anyway receive a model that allows extrapolation to new observations and has adequate predictive power, re-sampling and shrinkage techniques are applied (see Section 5.3.2). In addition, it was tested if a simpler model which summarizes complications with similar risk factors and uses a smaller number of latent variables can improve the predictive power. However, the measured predictive quality of the related models is (slightly) worse, so that this approach is not described in more detail. This finding also underlines the need to separately model the impact of the risk factors on each complication.

With the choice of a GLM-based approach, it is implicitly assumed that all surgical cases are independent. Of course, this does not hold in general, because some

surgeries are performed by the same doctors and in the same hospitals. Therefore, an intuitive approach to determine a risk-adjusted hospital or clinician score is the fitting of mixed models with clinician- and/or hospital-specific random effects. The estimates of the related random parameters can directly be interpreted as quality indicators for hospitals or clinicians. The scoring approach described in Section 5.3.3, however, is based on the determination of the clinician (or hospital) effect by calculating an observed-to-expected ratio. The applied GLMs purely focus on eliminating the effects of medical risk factors. As the presence of provider-specific random effects naturally leads to a decrease of the effect size of medical risk factors and, therefore, less risk adjustment, we consciously neglect the dependency of single observations in the modeling. Also, the mixed models tested lead to a significantly worse prediction of defined complication scores than the corresponding models without random effects.

5.3.2 Model fitting

The model parameters $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_q)$ and $\boldsymbol{\beta} = (\beta_{01}, \ldots, \beta_{pq})$ are calculated by iterative sampling from first and second stage models using different Metropolis-Hastings techniques (see also the fitting algorithm for the fraud and abuse detection model in Section 4.2.2). The implemented algorithm uses the following (simplified) reformulation of the joint posterior distribution of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$:

$$\begin{aligned} \underbrace{p(\boldsymbol{\alpha}, \boldsymbol{\beta} | r, \boldsymbol{v})}_{\text{posterior}} &\propto \underbrace{p(r, \boldsymbol{v} | \boldsymbol{\alpha}, \boldsymbol{\beta})}_{\text{likelihood}} \cdot \underbrace{p(\boldsymbol{\alpha}, \boldsymbol{\beta})}_{\text{prior}} \\ &= p(r | \boldsymbol{\alpha}, \boldsymbol{\beta}) \cdot p(\boldsymbol{v} | \boldsymbol{\alpha}, \boldsymbol{\beta}) \cdot p(\boldsymbol{\alpha} | \boldsymbol{\beta}) \cdot p(\boldsymbol{\beta}) \\ &= \underbrace{p(r | \boldsymbol{\alpha}, \boldsymbol{\beta}) \cdot p(\boldsymbol{\alpha} | \boldsymbol{\beta})}_{\text{first stage model}} \cdot \underbrace{p(\boldsymbol{v} | \boldsymbol{\beta}) \cdot p(\boldsymbol{\beta})}_{\text{second stage models}} \end{aligned} \tag{55}$$

For the decomposition of the likelihood, conditional independence between $\boldsymbol{v} = (v_1, \ldots, v_q)$ and r given the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is assumed.

Each iteration s ($s = 1, \ldots, S$) of the fitting algorithm starts with drawing one realization of $\boldsymbol{\beta}^{(s)}$ from all q second stage models. Depending on the distributional assumption for the latent variables (normal or binomial likelihood) and the activation of parameter shrinkage, different prior distributions are used for $p(\boldsymbol{\beta})$.

For the version without parameter shrinkage, the functions `MCMClogit` (for a binomial likelihood) and `MCMCregress` (for a normal likelihood) in the R package `MCMCpack` are applied [Martin et al., 2011]. In both cases, a weakly-informative multivariate normal prior distribution is used for the corresponding subvector $\boldsymbol{\beta}_j = (\beta_{0j}, \ldots, \beta_{pj})$ of $\boldsymbol{\beta}$, i.e. $p(\boldsymbol{\beta}_j) \sim \mathrm{N}(\mathbf{0}, 0.1\boldsymbol{I})$. If a normal distribution is assumed for the latent variables, inverse Gamma distributions (with mean 5 and variance 25) are used for the additional parameters σ_j^2 and Gibbs-sampling (acceptance probability of proposed parameters always equal to one) is applied. In the binomial case, the Metropolis proposal distribution is centered at the current value of $\boldsymbol{\beta}_j$ and has the

variance-covariance matrix $\boldsymbol{V} = \boldsymbol{T}(10\boldsymbol{I} + \boldsymbol{C}^{-1})^{-1}\boldsymbol{T}$. $\boldsymbol{T}$ is a diagonal positive definite matrix depending on the tuning parameter that controls the acceptance rate and $\boldsymbol{C}$ is the large sample covariance matrix of the maximum likelihood estimate of $\boldsymbol{\beta}_j$.

For the version with parameter shrinkage and a binomial distribution assumption, the function `logit.spike` in the R package `BoomSpikeSlab` is used [Chipman et al., 2001]. Idea of the related "spike-and-slab" shrinkage approach [George and McCulloch, 1997] is that the spike prior $p(\boldsymbol{\gamma}_j)$ controls the number of non-zero regression coefficients and, therefore, the degree of the shrinkage. $\boldsymbol{\gamma}_j = (\gamma_{1j}, \dots \gamma_{pj})$ is a vector of inclusion indicators for each of which a Bernoulli distribution with parameter $\pi = 0.2$ is assumed as spike prior. This means, that each second stage model is reduced to approximately $p/5$ (most relevant) risk factors. The slab prior is the prior distribution of the regression coefficients $p(\boldsymbol{\beta}_j|\boldsymbol{\gamma}_j)$. The slab prior is assumed to follow a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{V}$ which depends on the Fisher information available in the data and further hyperparameters [Chipman et al., 2001]. The dimensions of mean vector and covariance matrix are determined by the number of inclusion indicators equal to one.

Finally, the version with parameter shrinkage and a normal distribution assumption is implemented based on the function `blasso` in the R package `monomvn` [Gramacy et al., 2007]. Here, we apply the same lasso shrinkage approach with reversible jump mechanism [Green, 1995; Troughton and Godsill, 1997] which has already been used for the second stage models in the fraud and abuse context (see Section 4.2.2). According to (50), an exponential power distribution is assumed for $p(\boldsymbol{\beta}_j|\sigma_j^2)$ with a non-informative Jeffrey's hyperprior for the shrinkage parameters λ_j [Gramacy et al., 2007].

After having updated $\boldsymbol{\beta}$, the resulting estimates of the latent variables $\hat{\boldsymbol{v}}_1, \dots, \hat{\boldsymbol{v}}_q$ are plugged in the first stage model. For the update of the parameters $\boldsymbol{\alpha}$ of the first stage model, the same Gibbs-sampling approach is used like for the second stage model with normal distribution assumption and no shrinkage.

In total, this alternating update of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is repeated $S = 250$ times. Based on the analysis of different sampling traces, a burn-in phase of 50 iterations and thinning ratio of 1 (i.e. no thinning) are applied. Consequently, estimates $\hat{\boldsymbol{r}}$ of the three complication scores defined in Section 5.3.1 are received by averaging over the sampling results $\hat{\boldsymbol{r}}^{(51)}, \dots, \hat{\boldsymbol{r}}^{(250)}$. In addition to the described shrinkage options, a bagging [Breiman, 1996] approach is applied to increase predictive power and transferability of the model. This means that 50 bootstrap samples are drawn from the full dataset on each of which the different fitting algorithms are applied. By averaging over the results of all 50 bootstrap samples final estimates of the complication scores $\hat{\boldsymbol{r}}_1$, $\hat{\boldsymbol{r}}_2$ and $\hat{\boldsymbol{r}}_3$ are calculated. The following Section 5.3.4 illustrates how the estimates of the complication scores are translated into risk-adjusted provider scores. A comparison of the different latent variable models and additional benchmarking models with regard to predictive quality can be found in Section 5.4.

The execution of the described fitting algorithm is relatively time-consuming and strongly depends on the underlying computing power and the distributional assumption for the second stage models. On a working station with 8 GB RAM, estimation

time for all 50 bootstrap samples ranged between 2.5 and 5 hours. Naturally, the Gibbs sampler for the Gaussian option without shrinkage is faster than the three other options. From an application perspective, the performance of the algorithm is not that crucial as the provider ranking is quite stable and does not drastically change with each new case. So, if the model is used by a health insurer, a monthly or quarterly update appears to be sufficient. The prediction of complication costs and the provisioning of an individualized provider ranking for future patients can be realized by a scorecard approach which does not require a new run of the fitting algorithm.

5.3.3 Model outcomes

The idea behind the calculation of a provider-specific quality score is to compare the actual performance of the provider with the expected performance under consideration of the medical risk of the providers' case-mix. The medical quality score (QS) is calculated as provider-specific expected to observed ratio of the defined complication scores, e.g. for a provider A:

$$\mathrm{QS}_A = \frac{\sum_{i \in \mathcal{A}} \hat{r}_i}{\sum_{i \in \mathcal{A}} r_i}. \tag{56}$$

Here, $\mathcal{A}$ is an index set, identifying all cases of provider A from the full set of n cases.

The scores QS can be interpreted in the way that a provider with QS > 1 performs better than the average of all providers on a comparable case-mix. Vice versa, a QS < 1 indicates below average quality. The mean (or median) QS based on all bootstrap samples can be used as ranking criterion for a risk-adjusted provider ranking. Based on the Bayesian fitting algorithm described in Section 5.3.2, it is also possible to evaluate the credibility of the calculated QS. A Bayesian 95%-credibility interval can be formed by calculating the 2.5%- and 97.5%-quantile based on all bootstrap samples. If the lower bound of the interval is greater than 1, the provider performs significantly better than the average provider. If the upper bound is smaller than 1, the provider performs significantly worse. These results can be used by insurance companies for a transparent network management incentivizing or penalizing providers.

Even though the model is calculated based on all available cases in the data, it can be very misleading to include providers with a very small number of cases in a ranking based on the described quality score. Due to the high proportion of uncertainty in the prediction and the influence of single outliers, it is recommended to exclude all providers with less than 25 cases from the ranking. Approaches which correct for the small number of cases of single providers, like the shrinkage approach of Cohen et al. [2013], can certainly not fully prevent the risk of wrong conclusions. If insurance companies provide such quality ranking as a service to their insured persons, the exclusion of providers below the defined threshold is of course a disadvantage for

these providers. On the other hand, this may be a wanted effect, as the experience of surgeons is an important quality criterion, especially for more complex surgeries, like cataract surgery [Chan et al., 2010].

If the number of observed cases is large enough, an individualization of the ranking considering individual risk factors of the patient is possible. For example, an unsupervised K-means clustering could be applied to split all patients that underwent the treatment in the past into K clusters based on all observed risk factors. The quality score for provider A and cluster k ($k = 1, \ldots, K$) is then calculated as

$$\mathrm{QS}_{A,k} = \frac{\sum_{i \in \mathcal{A} \cap k} \hat{r}_i}{\sum_{i \in \mathcal{A} \cap k} r_i}. \tag{57}$$

Based on all scores QS_k an individualized provider ranking for all patients belonging to cluster k is possible. Future patients can be allocated to one of the K clusters based on known risk factors and will receive an individualized recommendation of best surgeons/providers. Of course, also for an individualized ranking a lower limit of performed treatments (in the corresponding cluster) should be applied to avoid biased results by single outliers. In order not to exclude too many providers K should be chosen relatively small and depending on the number of observed cases. In the data used for developing the described approach with about 20,000 underlying cataract surgeries, $K = 3$ and a lower boundary of 25 cases per provider and cluster turned out to be a reasonable choice.

The described scoring approach can be applied on clinician or on institution/hospital level. Based on the assumption that the treating physician has the larger influence on the medical outcome than the hospital where the treatment is performed, it is recommended to calculate the score on clinician-level if underlying insurance claims data contain this information. In countries with very heterogeneous hospital quality standards and hygiene conditions a hospital based score may be more meaningful.

It is important to note that this approach makes the simplifying assumption that all uncertainty in the medical outcome is related to the provider/clinician including the unexplained medical risk (see also Section 5.3.4). Therefore, the fairness of the scoring strongly depends on the quality of risk adjustment and the explanatory/predictive power of the underlying model. In this regard, each analytical scoring approach has clear limits. If the score is calculated on clinician level, there are a lot of exogenous factors that can not or not directly be influenced by the treating physician, like the performance of assisting staff or the hygiene conditions in the operation theater. A random effect for the hospital may correct this potential bias. On the other hand, it is the duty of the treating physician to ensure an optimal environment for the treatment. From a patient and also from an insurance perspective, it is therefore meaningful to also link this exogenous risk factors to the clinician.

A useful side result of the modeling approach (using complication score R_1) is that it yields a prediction $r^*_{1,i'}$ for the complication costs of a future patient i' based on the known risk factors. Considering the absolute predictive power of all prediction models (see Section 5.4) which indicates a large percentage of unexplained future

risk, a reliable prediction on individual level seems difficult. However, the results may still give a better indication of expected complication costs than just assuming an average over all cases independent of the underlying risk. From an insurance perspective, this indication can be used for pre-authorization decisions and more reliable case reserving. For example, the insurability can be increased by covering previously uncovered surgeries with good cost prognosis. On the actuarial side, case reserves can be adjusted, especially in the case of surgeries with high predicted complication risk, to avoid under-reserving.

5.3.4 Benchmarking

The complication scores described in Section 5.3.1 quantify the unobservable risk of the treatment. As illustrated in Figure 35, it is assumed that this treatment risk consists of the provider related risk (which is measured by the quality score) and the medical/patient related risk. Therefore, a fair provider scoring requires a risk adjustment approach which eliminates the underlying medical risk as far as possible, as the described provider quality score is otherwise biased by the unexplained medical risk. Only in the (theoretical) optimal case that the statistical model explains/predicts the complete medical risk, the described quality score exactly measures the provider related risk. Therefore, the provider scoring and ranking is as more precise as higher the predictive power of the risk adjustment model (see Figure 35). In addition, the predictive power of the model also measures the transferability of the model to new observations needed for pre-authorization and case reserving purposes as described in Section 5.3.3.

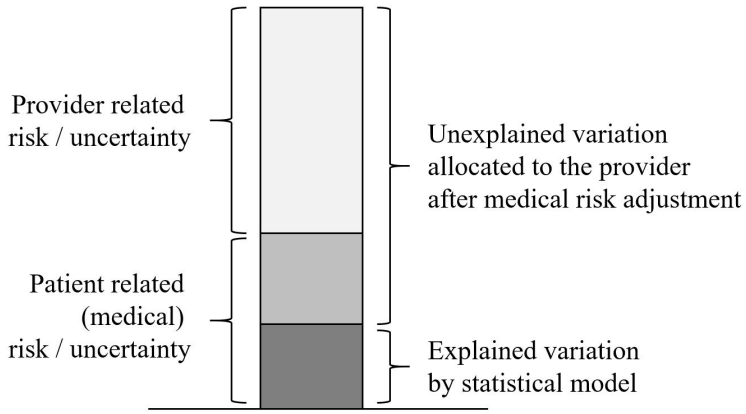

Figure 35: Schematic visualization of uncertainty in medical quality assessment.

From these reasons, the described latent variable approaches are compared to other risk adjustment techniques based on different measures of predictive power for continuous outcomes (like the described complications scores): the mean predictive squared error (MPSE), the mean absolute error (MPAE) and the predictive R-Squared (R^{2^*}, based on the correlation measures of Pearson and Spearman) as defined in Appendix A. To control overfitting to the training data, we calculate

these measures based on the out-of-bag-samples resulting from the bootstrapping approach described in Section 5.3.2.

As benchmarks to the four described latent variable models, we use

- a (one-stage) linear model which does not allow for covariate effects that vary between different complications and
- a Random Forest which naturally accounts for interactions between different risk factors and is known to deliver optimal prediction results in similar situations.

Table 15 summarizes all applied risk adjustment models at a glance.

Shortcut	Model Type	Distribution assumption for second stage	Shrinkage
lm	Linear model	–	no
lvmb	Latent variable model	binomial	no
lvmg	Latent variable model	gaussian	no
lvmsb	Latent variable model	binomial	yes
lvmsg	Latent variable model	gaussian	yes
rf	Random forest	–	–

Table 15: Overview of applied risk adjustment approaches.

5.4 Results

The results tables and figures shown in the following are referring to complication score R_1, i.e. the complication risk is represented directly by the sum of average complication costs (see Section 5.3.1). Due to the high correlations between the scores R_1, R_2 and R_3 described in Section 5.3.1, the results based on the scores R_2 and R_3 hardly deviate from the shown results and do not change the derived conclusions.

The results given in Figure 36 show the measures of predictive quality defined in Section 5.3.4 based on the 50 out-of-bag-samples resulting from the described bootstrapping approach. The measures show quite consistent results with regards to the risk adjustment capability of the displayed approaches (see Table 15).

If the predictive R-Squared is assumed to represent the share of explained variance in future complication risk, the results given in Figure 36 show that the best prediction models explain about 20% of future complication risks. As probably not all remaining 80% are related to the provider, the influence of medical risks can obviously not fully be eliminated. However, there are clear differences with regards to the predictive power of the different risk adjustment approaches. The developed latent variable models perform better than the Random Forest and clearly better than the simple linear model across all applied measures of predictive power. Therefore, it is recommended to base a provider scoring and ranking on the developed latent variable approach.

Within the latent variable models the models with Gaussian distribution assumption for the second stage models perform slightly better than the binomial models. The

shrinkage approach minimally increases the predictive power. The best model is the latent variable model with Gaussian distribution assumption for the second stage models and activated shrinkage option. Considering the clearly longer runtime of the shrinkage models to the Gaussian model without shrinkage, the simpler algorithm may be preferred in practice.

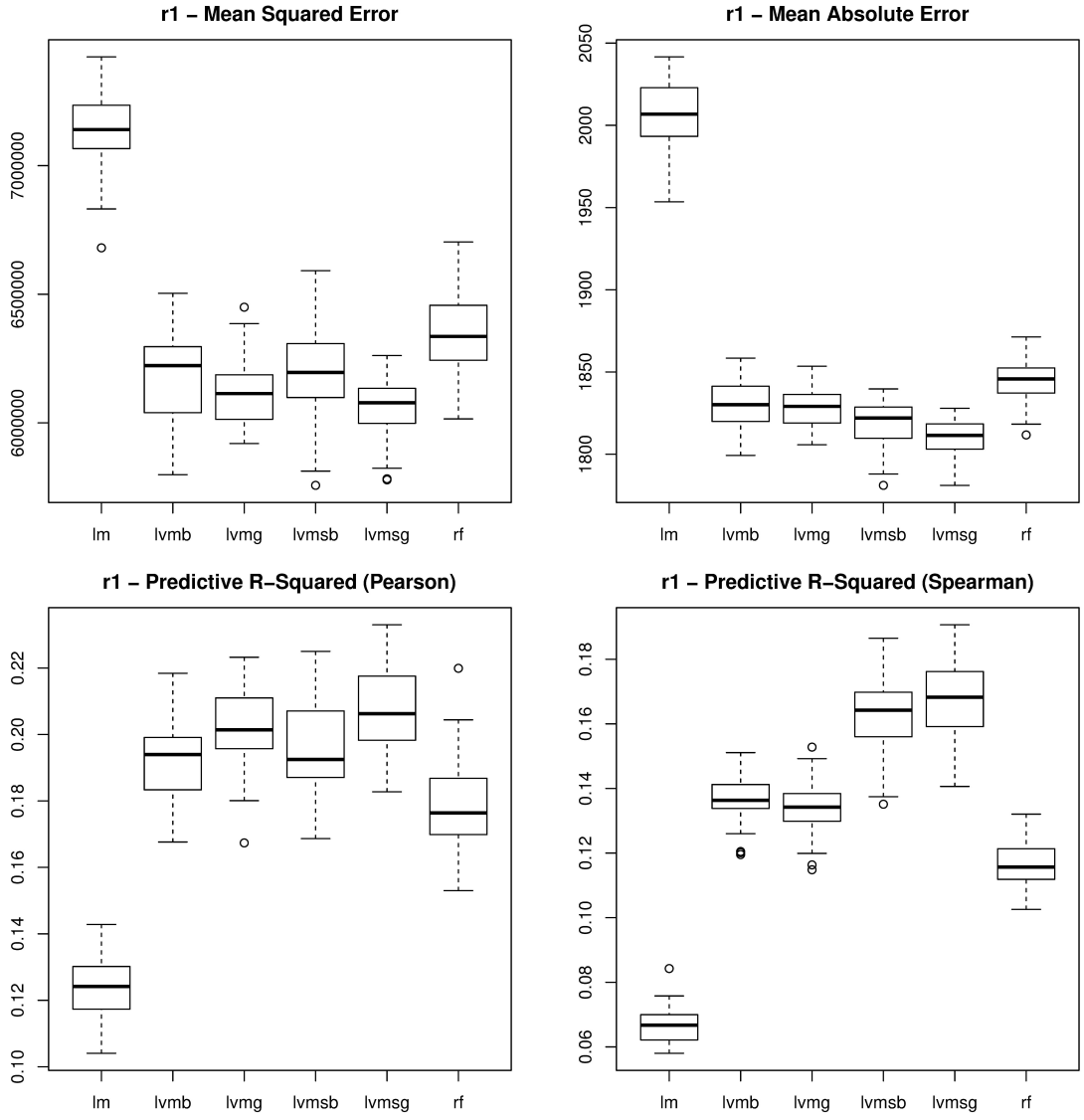

Figure 36: Comparison of different risk adjustment approaches based on different predictive quality measures for complication score R_1 (model shortcuts explained in Table 15).

Figure 37 compares the provider ranking based on the best model (lvmsg) in terms of predictive power

- to a raw complication ranking without risk adjustment ordering providers directly based on observed complication scores (raw),
- to the provider rankings based on the benchmark models (lm and rf) and

- to the provider rankings based on the other latent variable models (lvmb, lvmg and lvmsb).

The graphs show that no or insufficient risk adjustment may lead to completely different provider rankings. For example, there is one provider on rank 11 in the raw ranking and only on rank 60 (out of 63 providers with more than 25 cases) in the risk-adjusted ranking. Recommending this provider to patients can lead to serious consequences for the patient and unnecessary high complication costs for the insurance company. Also, using a simple linear model for risk adjustment leads to rank differences of up to 22 in the analyzed cataract data. Accounting for the fact that the best model explains about twice the amount of complication risk, the simple linear model approach does not yield optimal results. The other risk-adjusted rankings including the Random Forest approach lead to comparably similar rankings, especially among the best and worst providers. Therefore, the results show that it is beneficial to use more complex risk adjustment models for assessment of medical outcomes, the concrete approach does not change the outcome significantly.

Compared to the Random Forest, the latent variable models offer more transparency as single effects (high values of single latent variables) can easily be interpreted and explained to providers, e.g. in network negotiations. In this way, insurance companies can demonstrate concrete deficits to providers based on model outcomes which enables a targeted quality improvement process on provider side.

Finally, the individualized provider rankings based on the K-means clustering approach described in Section 5.3.3 are considered. Table 16 shows the correlation matrix of the provider rankings in the $K = 3$ clusters built based on all 20 risk factors. Only those providers which had more than 25 cases in each of the clusters (21 providers) have been considered for the analysis. The rankings in cluster 1 and 3 are nearly identical (and could be merged in this specific situation), however, the lower correlations between cluster 2 and 3 as well as 1 and 3 reveal that an individualized ranking makes sense for the existing data set. Compared to cluster 1 and 3, cluster 2 represents mainly age-related complications and co-morbidities which indicates that elderly patients (with affected risk factors) should rather rely on the corresponding ranking.

Correlation	Cluster 1	Cluster 2	Cluster 3
Cluster 1	1.00	0.75	0.98
Cluster 2	0.75	1.00	0.76
Cluster 3	0.98	0.76	1.00

Table 16: Correlations of provider rankings in different medical risk clusters (based on complication score R_1).

5.5 Summary and Outlook

Section 5 introduces a Bayesian latent variable model for risk adjustment in the evaluation of medical outcome quality. The results summarized in Section 5.4 show that more complex risk adjustment approaches are required to receive a provider

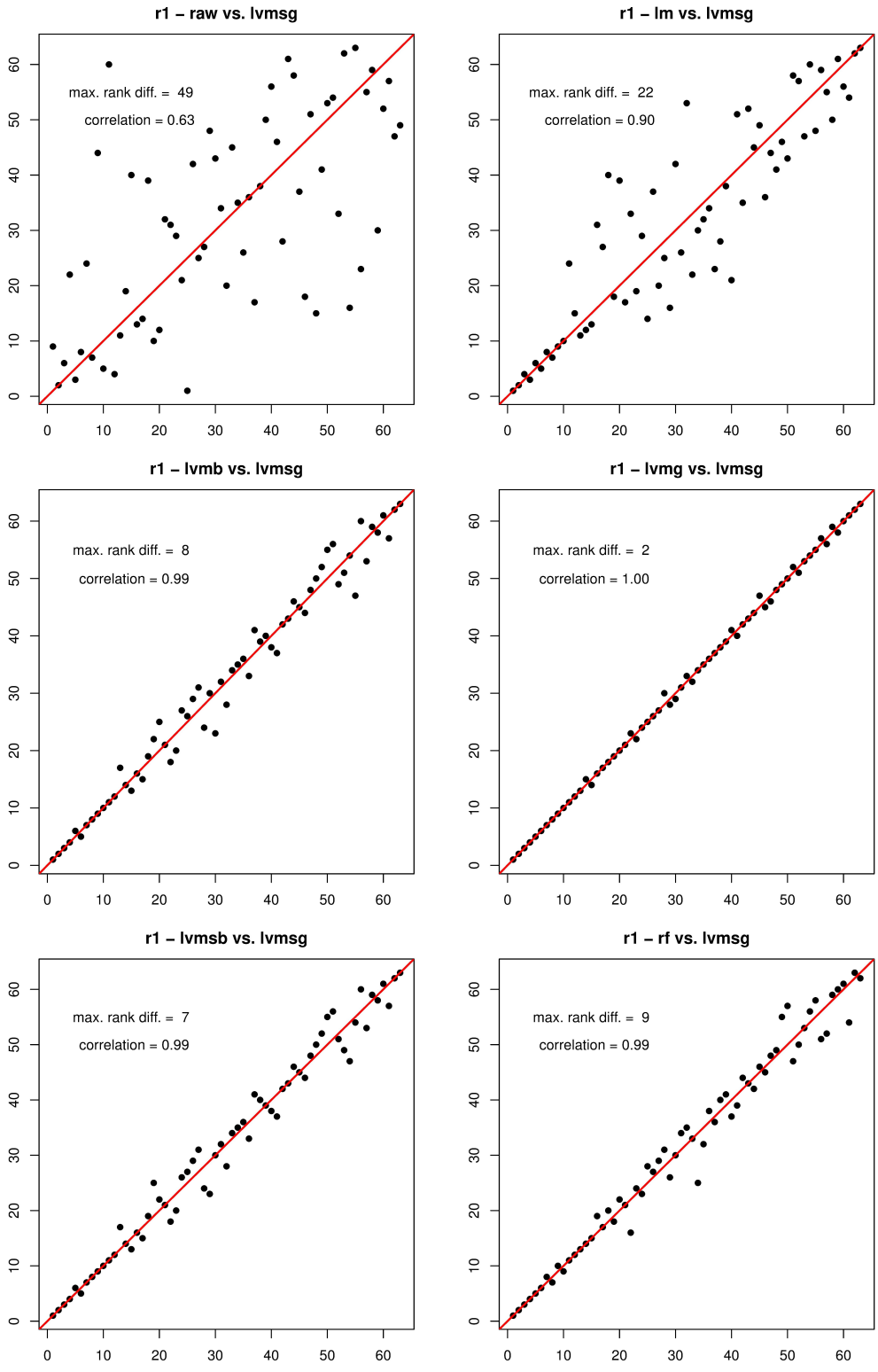

Figure 37: Rank correlations between different provider quality scores (raw, lm, lvmb, lvmg, lvmsb and rf) and best approach (lvmsg) in terms of predictive quality for complication score R_1 (model shortcuts explained in Table 15).

scoring which eliminates as much as possible of the patient-related medical risk. The developed Bayesian latent variable models yield high predictive power and outperform different benchmarking techniques including a Random Forest. In addition, they provide transparency towards providers by easily interpretable effects which may reveal concrete quality deficits. The described individualization approach allows to tailor the provider ranking to the individual medical situation of patients and, therefore, enables the choice of the best doctor for each patient. In this way, severe and expensive complications can be avoided.

From an insurance perspective, the provided results can be used for the optimization of patient steering. The publication of (individualized) medical quality rankings enriches the service offering of the insurance company and can be combined with further information, like geographical proximity between provider and patient. Further, the individual prediction of complication costs supports pre-authorization and case reserving decisions. Similarly like the results of the fraud and abuse analysis in Section 4, insufficient medical quality outcomes can be used in network management, e.g. for negotiations on discounts or exclusions from preferred networks. The introduced risk-adjusted quality scores can – applied on a wider scale – also form the basis for a transparent regulatory pay-for-performance system encouraging an overall improvement of medical treatment quality.

The described latent variable approach can be applied for the assessment of every medical treatment in which the performance of the provider strongly influences the occurrence of complications, like cataract surgery. Prerequisite for the implementation of an informative provider scoring/ranking is a large enough database of observed cases which contains detailed medical information on the patients' history before and after the treatment. Usually, insurance claims data with high quality medical coding (diagnoses and procedures) observed over a period of at least 5 years meet these requirements in the case of prevalent surgeries. The described individualization of results to the medical situation of (future) patients requires a large number of providers who frequently perform the corresponding treatment.

Transferring the methodology to other (predictable) surgical treatments and confirming the results of this study will be subject to future research. Other possible applications scenarios are, for example, ceasarian section, cruciate ligament surgery or (total) knee and hip replacement. From a statistical perspective, it will be interesting to evaluate if the shrinkage option will improve the predictive power of the models more strongly if more potential risk-factors (and interactions) are available as model input.

Considering the absolute predictive power of the best risk adjustment models (see Section 5.4), analytical models can naturally not fully explain/predict the treatment related risk. These results may improve with the enrichment of (insurance) data by additional information like clinical data (e.g. results of lab tests) or data from health trackers (e.g. activity levels of patients). Also, this information will not lead to a complete predictability of medical outcomes, but it will be interesting to observe the extent of improvement related to these upcoming trends.

6 Success Factors and Future Trends

What are the Main Conclusions from this Thesis?

The creation of statistical outcomes which provide value to health insurance companies is a challenging task and can go terribly wrong if important principles are not respected. Wrong results or wrong interpretations of results can lead to the disqualification of the whole subject and, therefore, analysts and researchers have a strong responsibility in this regard. In the following, the major observations and conclusions from the academic research carried out in relation with this thesis are summarized.

Using all relevant information is key!

In general, it is a clear advantage in statistical modeling to have more covariates available. Models which incorporate all relevant information, especially medical information in the claims sector, nearly always outperform traditional insurance approaches only based on age and gender. However, the inclusion of hundreds of parameters without controlling correlations does not improve statistical outcomes, in particular not predictions. Especially, when working with large portfolios, not all significant parameters are also relevant. Therefore, targeted methods of variable selection (as described in Section 2 and 3) or parameter shrinkage (as described in Section 4 and 5) are required to control model complexity and avoid overfitting. In times of increasing "Big Data", techniques which are able to filter the relevant information are more important than ever.

Subject-matter expertise matters!

The blind application of statistical techniques does usually not lead to optimal outcomes. Many analyses described in the previous chapters show that only the combination of insurance knowledge and statistical know-how leads to optimal outcomes. Especially medical information needs to be structured and clustered in an adequate way to reach the best predictions. For example, diagnoses codes require some context specific categorization to efficiently use this information for statistical models. Also, the ex-ante definition of known causal correlations and incorporation of this knowledge clearly improves model results. Such prior knowledge is less relevant in machine learning models, because they are supposed to automatically identify this information from training data. This proceeding, however, bears the danger of overfitting as, for example, the comparison of regression techniques and machine learning models in Section 2 shows.

One size does not fit all!

Standard prediction techniques, like Random Forests, deliver meaningful results in many situations. However, both Sections 4 and 5 show that tailored algorithms which are designed to reflect certain known data characteristics can still improve the results. For example in Section 5, the medical hypothesis that different complications depend on different risk factors (and the resulting introduction of separate latent variables) clearly improves the outcomes in medical quality assessment. This

is interesting from an academic perspective, but especially if statistical techniques are applied to large portfolios, even small improvements of predictive power can also cause a significant difference in financial impact. On the other hand, the development of such tailored algorithms is, of course, also related to efforts and costs which carefully have to be weighed against the benefits. In terms of cost-efficiency the transferability of solutions plays an important role. Sections 3 and 4 demonstrate how statistical techniques can also support health insurance companies in this regard. In particular, the Bayesian approach to specify prior knowledge on fraud and abuse patterns allows insurance companies from other markets to use existing results, but still to benefit from a highly customizable solution.

More complex is not always better!

In large insurance companies who have stable portfolios and capture data in a high quality for many years, more sophisticated statistical techniques have a clear value. As especially the study of disease management candidate selection shows (see Section 2), for small and strongly growing portfolios, also comparably simple (regression) techniques provide stable and reliable results. Additional complexity can make sense, but requires careful analysis and understanding if the additional features also provide sustainable value. Especially for prediction purposes, sparser models have proven in many situations to be at least equivalent with more complex alternatives. Combining this statement with the previous one on tailored solutions, it can be noticed that usually those models with "targeted complexity" deliver the best outcomes.

How to tap the Economic Potential of Statistical Insights?

Even if all those statistical best practice rules – which certainly not only hold for the analysis of health insurance data – are considered and optimal statistical outcomes are generated, the realization of savings or upsides is not guaranteed. In order to tap the full economic potential of statistical insights, several other prerequisites need to be fulfilled. Two of these prerequisites to successfully apply Statistics in health insurance are adequate processes around the statistical analysis and an analytical mindset across the company.

The business loop illustrated in Figure 38 shows what additional measures are required before and after analyzing data. In economy like in academic research, everything starts with a business/research question which needs to be defined involving all relevant stakeholders. In the examples provided in this thesis, claims and network managers, but also IT managers have been involved, and an initial strategy to reach a defined (financial) goal has commonly been defined. Based on this starting point, all relevant data (internal and eventually external) need to be collected and made available for the analysis. After comprehensive statistical analyses, the important business integration process starts. This involves technical integration, but especially the common definition of targeted business interactions as consequence of the statistical results. For example in fraud and abuse detection, claims investigators need to understand the meaning of a high fraud probability and require explanations

from which data patterns this high probability arises. Only if this understanding is given an efficient follow-up investigation or targeted network negotiations can be carried out. In this regard, transparent statistical models ("white-box" approaches) have the clear advantage that the results can be interpreted more easily, also by non-experts. After a certain test period, it needs to be evaluated if the defined interactions lead to the achievement of the initially defined (financial) goal. Usually, multiple adaptations of models and processes – again involving all parties – are needed to fully exploit the potential of statistical analyses.

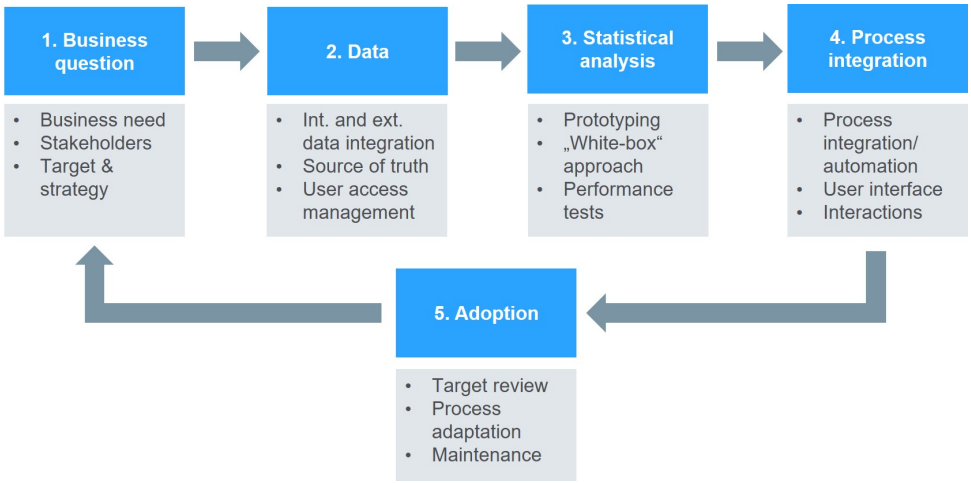

Figure 38: Business loop ensuring the transformation of statistical insights into business value.

Beside optimized processes, another requirement for leveraging the value of Statistics in health insurance is a company-wide analytical mindset. This mindset starts from the top management, but affects all employees of the company. For example, in reporting, a democratization of data, i.e. access for all staff to understandable self-service reports, is required. The implementation of new statistical applications needs a sponsor on top-management level who can make financial decisions on a sound decision basis. As the creation of insights is a try-and-error-process which not always (directly) leads to measurable impacts, also a certain belief in data-based steering is required which does not stop with the first failure. Finally, data analysts need to have a support function for other departments and must not be perceived as control body to ensure a close cooperation with various specialist departments.

If all (or most of these) conditions are fulfilled, the implementation of statistical models can lead to significant financial impacts. Already today, there are companies which cut about 3% of their claims costs based on efficient statistical fraud and abuse reporting or create upsides of 7% in premium revenues by targeted churn prevention models. Together with the growing awareness and the promoting factors described in the introduction (see Section 1), statistical methods have best opportunities to further gain importance in health insurance.

How will the Creation of Statistical Insights change in the Future?

As outlined in this thesis, the (time-consuming) tailoring of statistical methods, especially the process of structuring covariate information and incorporating subject-matter knowledge, optimizes statistical outcomes. But will this conclusion also hold in the future? Many prestigious scientists and futurologists predict that the process of statistical insight generation will drastically change in the next years. There are many examples from other industries where artificial intelligence and deep learning techniques lead to an automation of data analysis already today. The principle is simple: as long as enough data information is available as an input, self-learning algorithms are able to decrypt even most complex patterns in the data on their own. Therefore, it is a valid question, if the automation of statistical insight generation will become a game changer and lead to a loss of importance of content knowledge.

Indeed, machine learning techniques, like neural networks, which form the foundation of artificial intelligence, already provide solid results as shown in this thesis. Even though, human intervention is still needed here, for example in the tuning of Random Forests, an automation of this optimization process is thinkable. Intelligent grid search algorithms for parameter tuning have already been developed. As described in the introduction also more data will certainly be available in health insurance and the relevance of external data will grow. Figure 39 illustrates this assumption and stresses the particular importance of structured external data which are generated from dedicated information extraction algorithms, like text, image or voice mining. Especially in developed markets, the shift from an internal focus to a combination of internal and external data has already begun.

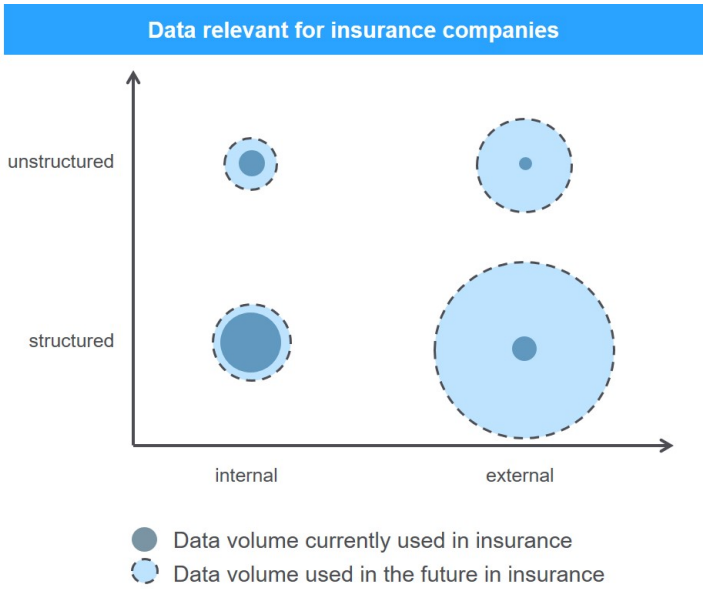

Figure 39: Expected shift in relevant data in health insurance.

On the other hand, a lot of health insurance companies are still struggling with the handling of their internal data. Other companies who have already addressed the

topic, have stopped "Big Data" initiatives due to disappointing business outcomes or invincible regulatory hurdles, in particular the special protection of health insurance data. For example, the European Union strongly tightens its data protection rules in a basic ordinance valid from 2018. Other international standards are likely to follow. Also, the complexity of health insurance systems as well as the high influence of chance on human health impose high demands to automated data analysis and artificial intelligence.

On the long run, it is assumed that these hurdles will definitely delay the process of automated insight generation in health insurance. However, it is highly questionable if they will prevent that these technologies are entering also the health insurance industry. Until then, it's definitely still worth investing time and resources in intelligent statistical methods and algorithms.

A Predictive Measures in DMP Candidate Selection

In this appendix, two kinds of predictive measures are introduced that permit a comparison of cost prediction techniques with regard to an efficient DMP selection:

a) measures for the accuracy of a prediction model that indirectly assess how many of the patients with the highest saving potential can be identified,

b) measures for the sorting capacity of a prediction model that directly examine the same question.

The latter group of measures may be more suitable to measure the financial benefit of DMP selection methods. For comparing the general ability of an approach to optimize the selection of DMP participants by predicting claimed amounts, the measures in group a) are equally important.

a) Accuracy measures

Two predictive measures quantifying the prediction error are the mean predictive squared error (MPSE) and the mean predictive absolute error (MPAE) defined as

$$\begin{aligned} \text{MPSE} &:= \frac{1}{n}\sum_{i=1}^{n}(y_i^* - y_{o,i})^2 \quad \text{and} \\ \text{MPAE} &:= \frac{1}{n}\sum_{i=1}^{n}|y_i^* - y_{o,i}|\,. \end{aligned} \tag{58}$$

Both measures analyze the differences between predicted costs $\boldsymbol{y}^* = (y_1^*, \ldots, y_n^*)'$ and actually observed costs $\boldsymbol{y}_o = (y_{o,1}, \ldots, y_{o,n})'$. The MPSE is based on a quadratic loss function. This means that predictions that avoid extreme discrepancies between $\boldsymbol{y}^*$ and $\boldsymbol{y}_o$ are rated best. By contrast, the MPAE that is based on an absolute loss function favors predictions that are good "on average". The MPSE assures precise predictions for high-cost members. This is very relevant for DMP selection, because not recognizing a member who will – without preventive interaction – produce exploding medical costs in the near future means that there is no possibility to realize the individual saving potential related to the DMP.

A disadvantage of both MPSE and MPAE is that these measures are not normed like, for example, the coefficient of determination or model R-squared that measures goodness-of-fit in linear models and ranges between 0 and 1. Hence, it is desirable to define a normed measure that is bound to a limited interval of possible values and measures the predictive quality of a model. For this purpose, the so-called predictive R-squared R^{2^*} is defined according to the formulation of the model R-squared (that assesses the squared linear correlation between $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$) in order to measure the squared linear correlation between $\boldsymbol{y}^*$ and $\boldsymbol{y}_o$:

$$R^{2^*} := \frac{\sum_{i=1}^{n}(y_i^* - \bar{y}^*)(y_{o,i} - \bar{y}_o)}{\sqrt{\sum_{i=1}^{n}(y_i^* - \bar{y}^*)^2}\sqrt{\sum_{i=1}^{n}(y_{o,i} - \bar{y}_o)^2}}. \tag{59}$$

$\bar{y}^*$ and $\bar{y}_o$ denote the arithmetic means of $\boldsymbol{y}^*$ and $\boldsymbol{y}_o$, respectively. Unlike the model R-squared, the predictive R-squared does not measure goodness-of-fit and cannot be interpreted as percentage of explained variance or deviance in the classical sense, since the decomposition of variance or deviance [Davison, 2003; Tutz and Fahrmeir, 2001] that holds for estimated values $\hat{\boldsymbol{y}}$ does not hold for predicted values $\hat{\boldsymbol{y}}^*$. However, R^{2^*} is also bound to the interval $[0; 1]$ with values closer to one indicating a higher predictive quality. Therefore, it gives an indication of which percentage of future costs can be explained by the model.

b) Sorting capacity measures

Two measures directly characterizing the sorting capacity of a prediction model are the Spearman rank correlation coefficient R_{Sp} and the area under the "matching curve" AUC_m.

The Spearman or rank correlation coefficient R_{Sp} measures the monotonic correlation between $\boldsymbol{y}^*$ and $\boldsymbol{y}_o$:

$$R_{\text{Sp}} = \frac{\sum_{i=1}^{n}(\text{rank}(y_i^*) - \overline{\text{rank}}(\boldsymbol{y}^*))(\text{rank}(y_{o,i}) - \overline{\text{rank}}(\boldsymbol{y}_o))}{\sqrt{\sum_{i=1}^{n}(\text{rank}(y_i^*) - \overline{\text{rank}}(\boldsymbol{y}^*))^2}\sqrt{\sum_{i=1}^{n}(\text{rank}(y_{o,i}) - \overline{\text{rank}}(\boldsymbol{y}_o))^2}}. \tag{60}$$

where $\overline{\text{rank}}(\cdot)$ denotes the average rank of the respective cost vector. R_{Sp} ranges between -1 and $+1$ where values close to $+1$ indicate a high positive monotonic correlation meaning that predicted and observed claimed amounts are similarly ordered. For DMP selection, this is a desirable property.

The idea of the matching curve m is derived from the concept of the ROC (receiver operating characteristic) curve that is used to assess the predictive quality of binary regression models [Pearce and Ferrier, 2000]. $m(i)$ is defined as the percentage of those i members with the highest observed values who can also be found among the i members with the highest predicted values where i ranges between 1 and n:

$$m(i) := \frac{1}{i}\sum_{j=1}^{i}\mathbb{1}(c_{o,(j)} \in c_{(1)}^*, \ldots, c_{(i)}^*), \quad i = 1, \ldots, n. \tag{61}$$

In this definition, $c_{o,(1)}, \ldots, c_{o,(n)}$ represents the vector of member IDs sorted in descending order by the corresponding observed claimed amounts $\boldsymbol{y}_o$. In parallel, $c_{(1)}^*, \ldots, c_{(n)}^*$ denotes the vector of member codes sorted in descending order by the corresponding predicted claimed amounts $\boldsymbol{y}^*$. $\mathbb{1}(\cdot)$ is an indicator function that is equal to 1 if the condition in brackets is fulfilled and 0 if not. Thus, $m(i)$ indicates the percentage of matching member IDs among the first i elements of the vectors

$c_{o,(1)}, \ldots, c_{o,(n)}$ and $c^*_{(1)}, \ldots, c^*_{(n)}$. The area under the matching curve AUC_m is obtained by calculating $\frac{1}{n} \sum_{i=1}^{n} m(i)$. The maximum AUC_m is 1, which occurs if the members have the same order in respect of observed and predicted values.

Figure 40 shows an example of a matching curve $m(i)$ for the developed predictive model (evaluated on the grid $i = 50, 100, \ldots, 9{,}150$) and the expected matching curve of a randomly ordered sample. In the context of DMP selection, it is particularly interesting to compare values of different matching curves in the high-cost region in order to measure which percentage of members who actually have the highest claimed amounts can be identified by the prediction approach. This is why $m(i)$ is also called the identification or hit ratio. Based on the identification ratio and some experience-driven assumptions on the average saving potential per cost group, a health insurer can easily compare the potential overall savings of different selection methods.

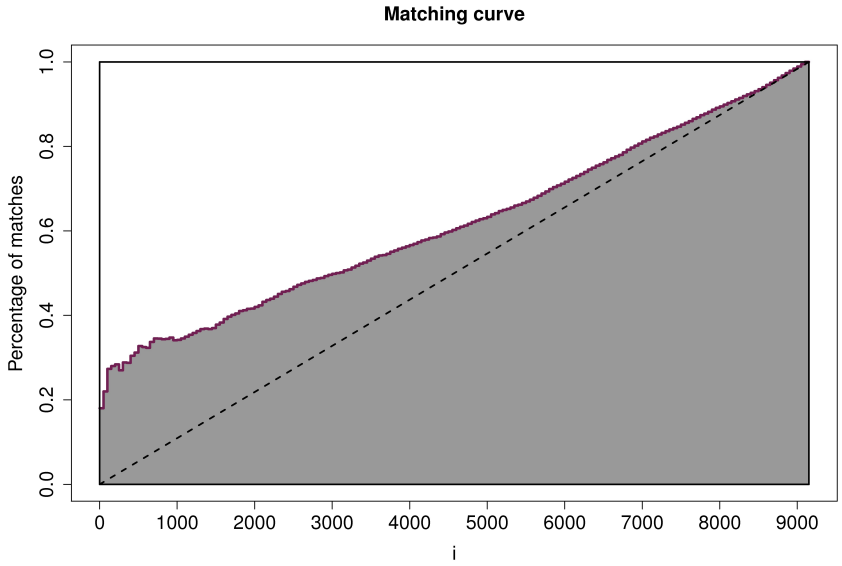

Figure 40: Matching curve m of the developed GLM (solid line) with area under curve AUC_m (gray) and expected matching curve of a randomly ordered sample (dashed line).

B Estimation of Covariances in DMP Evaluation

In this appendix, an estimate for the covariances $\mathrm{Cov}(\hat{\bar{d}}_\theta, \hat{\bar{d}}_{\theta'})$ with $\theta < \theta'$ is defined that can be plugged into Equation (38) defined in Section 4.4:

$$\begin{aligned}\widehat{\mathrm{Cov}}(\hat{\bar{d}}_\theta, \hat{\bar{d}}_{\theta'}) &= \widehat{\mathrm{Cov}}(S_\theta - P_\theta, S_{\theta'} - P_{\theta'}) \\ &= \underbrace{\widehat{\mathrm{Cov}}(S_\theta, S_{\theta'})}_{\sigma_S} - \underbrace{\widehat{\mathrm{Cov}}(S_\theta, P_{\theta'})}_{\sigma_{SP}} - \\ &\quad \underbrace{\widehat{\mathrm{Cov}}(P_\theta, S_{\theta'})}_{\sigma_{PS}} + \underbrace{\widehat{\mathrm{Cov}}(P_\theta, P_{\theta'})}_{\sigma_P}.\end{aligned} \tag{62}$$

The components σ_S, σ_{SP}, σ_{PS} and σ_P are estimated by empirical covariance estimates using pairwise complete observations. This means, for example, that for estimating the covariance between participants in program year $\theta = 1$ and participants in program year $\theta' = 2$ only those $|\mathcal{R}'_\theta|$ participants that have been in the program for at least 2 years are used. This results in

$$\begin{aligned}\hat{\sigma}_S &= \frac{1}{|\mathcal{R}_{\theta'}|}\left(\frac{1}{|\mathcal{R}_{\theta'}| - 1}\right. \\ &\quad \left.\sum_{r \in \mathcal{R}_{\theta'}} (y_{s(p_r),\theta} - \bar{y}_{s(p_r),\theta})(y_{s(p_r),\theta'} - \bar{y}_{s(p_r),\theta'})\right),\end{aligned} \tag{63}$$

$$\begin{aligned}\hat{\sigma}_{SP} &= \frac{1}{|\mathcal{R}_{\theta'}|}\left(\frac{1}{|\mathcal{R}_{\theta'}| - 1}\right. \\ &\quad \left.\sum_{r \in \mathcal{R}_{\theta'}} (y_{s(p_r),\theta} - \bar{y}_{s(p_r),\theta})(y_{p_r,\theta'} - \bar{y}_{p_r,\theta'})\right), \\ \hat{\sigma}_{PS} &= \frac{1}{|\mathcal{R}_{\theta'}|}\left(\frac{1}{|\mathcal{R}_{\theta'}| - 1}\right. \\ &\quad \left.\sum_{r \in \mathcal{R}_{\theta'}} (y_{p_r,\theta} - \bar{y}_{p_r,\theta})(y_{s(p_r),\theta'} - \bar{y}_{s(p_r),\theta'})\right), \\ \hat{\sigma}_P &= \frac{1}{|\mathcal{R}_{\theta'}|}\left(\frac{1}{|\mathcal{R}_{\theta'}| - 1}\right. \\ &\quad \left.\sum_{r \in \mathcal{R}_{\theta'}} (y_{p_r,\theta} - \bar{y}_{p_r,\theta})(y_{p_r,\theta'} - \bar{y}_{p_r,\theta'})\right).\end{aligned} \tag{64}$$

For these estimates, the same assumptions as for estimates (35), (36) and (37) are made, i.e. the dependencies between the "virtual twins" are neglected as outlined in Section 4.4.

C Predictive Measures in Fraud and Abuse Detection

As described in Section 4.3, the average of all category-specific AUCs is applied as criterion for the predictive quality of the different classification techniques. In the following, this and different other criteria frequently used in the classification context will be defined and explained. On this basis, some additional measures for the predictive quality of fraud and abuse detection models are introduced which may be relevant for insurance companies and healthcare payers.

A simple criterion for predictive quality is the misclassification rate m calculated based on the misclassification matrix M:

$$M = (f_{ij})_{ij},\ i,j \in \{\mathrm{NI}, \mathrm{UP}, \mathrm{UJ}, \mathrm{BI}\}, \quad \text{with } f_{ij} := \frac{1}{n_{\text{test}}} \sum_{k=1}^{n_{\text{test}}} \mathbb{1}(y_k = i \wedge y_k^* = j) \tag{65}$$

where y_k is the real class affiliation of invoice k and y_k^* the predicted class affiliation ($k = 1, \ldots, n_{\text{test}}$). The misclassification rate is defined as:

$$m = 1 - \sum_i f_{ii} \quad \text{with } i \in \{\mathrm{NI}, \mathrm{UP}, \mathrm{UJ}, \mathrm{BI}\}. \tag{66}$$

The raw misclassification rate assigns a higher weight to more frequent response categories in the test dataset. This is an undesirable property, as a model is needed that performs equally well in all response categories and can also be applied in other markets.

Based on the misclassification matrix, several other prediction criteria can be derived for each response category i ($i \in \{\mathrm{NI}, \mathrm{UP}, \mathrm{UJ}, \mathrm{BI}\}$). In the following, those criteria which we consider the most relevant in the fraud and abuse context are listed:

a) The *sensitivity* or *true positive rate*

$$\begin{aligned} TPR_i &= \frac{\sum_{k=1}^{n_{\text{test}}} \mathbb{1}(y_k = i \wedge y_k^* = i)}{\sum_{k=1}^{n_{\text{test}}} \mathbb{1}(y_k = i)} \\ &= \frac{f_{ii}}{\sum_j f_{ij}} \quad \text{with } j \in \{\mathrm{NI}, \mathrm{UP}, \mathrm{UJ}, \mathrm{BI}\} \end{aligned} \tag{67}$$

describes the relative frequency of allocating an invoice to the right category given the real category is i.

b) The *specificity* or *true negative rate*

$$\begin{aligned} TNR_i &= \frac{\sum_{k=1}^{n_{\text{test}}} \mathbb{1}(y_k \neq i \wedge y_k^* = i)}{\sum_{k=1}^{n_{\text{test}}} \mathbb{1}(y_k \neq i)} \\ &= \frac{\sum_{r \neq i} f_{ri}}{\sum_{r \neq i} \sum_j f_{rj}} \quad \text{with } j \in \{\mathrm{NI}, \mathrm{UP}, \mathrm{UJ}, \mathrm{BI}\} \end{aligned} \tag{68}$$

describes the relative frequency of allocating an invoice to another category than i given the real category is not i.

c) The *precision* or *positive predictive value*

$$\begin{aligned} PPV_i &= \frac{\sum_{k=1}^{n_{\text{test}}} \mathbb{1}(y_k = i \wedge y_k^* = i)}{\sum_{k=1}^{n_{\text{test}}} \mathbb{1}(y_k^* = i)} \\ &= \frac{f_{ii}}{\sum_r f_{ri}} \quad \text{with } r \in \{\text{NI}, \text{UP}, \text{UJ}, \text{BI}\} \end{aligned} \tag{69}$$

describes the relative frequency that an invoice really belongs to category i if the predicted category is i.

Criterion b) measures the type I error of the classification technique, i.e. the probability that an invoice will be allocated to one of the fraud and abuse categories even if it is not fraudulent or abusive. Criterion a) measures the type II error, i.e. the probability of allocating an invoice to category NI even though it is fraudulent or abusive. Adequate quality criteria for classification techniques need to control both type I and II errors, i.e. consider both criteria a) and b).

An intuitive combination of criteria a) and b) in binary classification problems is the ROC (receiver operating characteristic) curve [Pearce and Ferrier, 2000]. It relates the false positive rate (1–true negative rate) on the x-axis to the true positive rate on the y-axis, depending on the cut-point in terms of predicted probability from which an invoice is allocated to category i (see Figure 41). An optimal classifier would reach a sensitivity of 1 and at the same the time a specificity of 1 independent of the chosen cut-point. Hence, the area under the curve (AUC) is 1 in this case. A random classifier which randomly allocates invoices to one of the response categories reaches an expected AUC of 0.5.

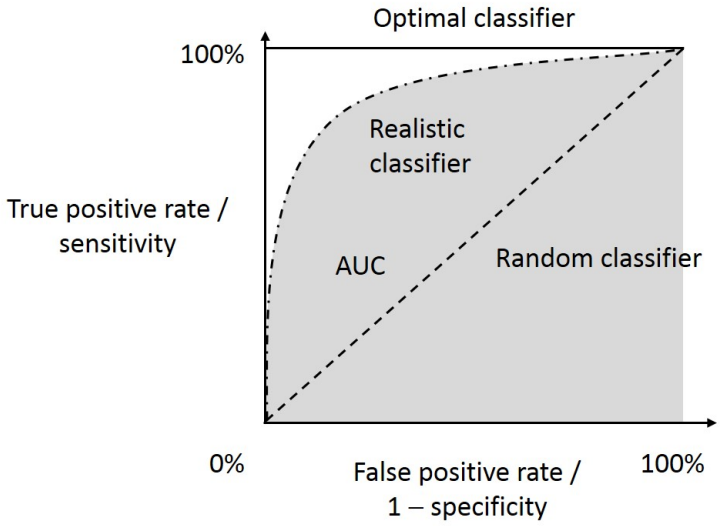

Figure 41: Concept of ROC curve for binary classifiers.

In this way, the AUC can be used to compare different classifiers. Since there exist more than two response categories in the described classification problem, the

ROC curves are separately visualized for each response category based on binary variables, indicating whether the real category was correctly predicted or not. In this way, an AUC is obtained for each response category and an average $\overline{AUC}$ over all categories can be calculated to receive an overall quality criterion. Hence, a measure is obtained which assigns equal weights to all response categories regardless of the occurrence in the test sample, and considers both type I and type II errors for each response category. Of course, a weighted average of AUCs can be used if sensitivity and specificity for one response category seem more important than for others in a specific context.

For the definition of some additional context-related measures, the misclassification matrix is broken down to the categories "no irregularities" (NI) and "fraudulent or abusive" (FA, i.e. one of the categories UP, UJ or BI).

One relevant measure may be how many of the real fraud and abuse cases can be identified as such. Based on the definitions above, this corresponds to the true positive rate of all fraud and abuse cases TPR_{FA}:

$$TPR_{FA} = \frac{\sum_{k=1}^{n_{\text{test}}} \mathbb{1}(y_k = FA \wedge y_k^* = FA)}{\sum_{k=1}^{n_{\text{test}}} \mathbb{1}(y_k = FA)}. \tag{70}$$

For the further investigation process, the percentage of all real cases which were allocated to the right subcategory (right category rate, RCR_{FA}) is also interesting:

$$RCR_{FA} = \frac{f_{UP,UP} + f_{UJ,UJ} + f_{BI,BI}}{\sum_j f_{UP,j} + f_{UJ,j} + f_{BI,j}} \quad \text{with } j \in \{\text{NI}, \text{UP}, \text{UJ}, \text{BI}\}. \tag{71}$$

Another important quantity may be the reliability of a decision that an invoice is "clean" represented by the positive predictive value of category NI:

$$PPV_{NI} = \frac{\sum_{k=1}^{n_{\text{test}}} \mathbb{1}(y_k = NI \wedge y_k^* = NI)}{\sum_{k=1}^{n_{\text{test}}} \mathbb{1}(y_k^* = NI)}. \tag{72}$$

If this figure is sufficiently high, it may make sense to automate the adjudication of all invoices with predicted category NI (and to only perform a small number of random checks of these invoices). The auto-adjudication rate AAR_{NI} is then equal to the amount of invoices which are allocated to the category NI by the classification technique:

$$AAR_{NI} = \sum_i f_{i,NI} \quad \text{with } i \in \{\text{NI}, \text{UP}, \text{UJ}, \text{BI}\}. \tag{73}$$

Finally, an interesting quantity in terms of efficiency of the fraud and abuse detection algorithm is the probability that a filtered case is really a fraud and abuse case. This probability corresponds to the positive predictive value of category FA PPV_{FA}:

$$PPV_{FA} = \frac{\sum_{k=1}^{n_{\text{test}}} \mathbb{1}(y_k = FA \wedge y_k^* = FA)}{\sum_{k=1}^{n_{\text{test}}} \mathbb{1}(y_k^* = FA)}. \tag{74}$$

From a business perspective, this quantity can be used to derive the number of invoices that need to be reviewed to find at least one fraud and abuse case (e.g. with a probability of 95%, $n_{95\%}$):

$$n_{95\%} = \frac{\log(1 - 0.95)}{\log(1 - PPV_{FA})}. \tag{75}$$

References

A. Abadie, D. Drukker, J. L. Herr, and G. W. Imbens. Implementing matching estimators for average treatment effects in Stata. *The Stata Journal*, 4(3):290–311, 2004.

S. B. Adebayo, L. Fahrmeir, C. Seiler, and C. Heumann. Geoadditive Latent Variable Modeling of Count Data on Multiple Sexual Partnering in Nigeria. *Biometrics*, 67:620–628, 2011.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

K. Antonio and J. Beirlant. Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics*, 40(1):58–76, 2007.

K. D. Aral, H. A. Güvenir, İ. Sabuncuoğlu, and A. R. Akar. A prescription fraud detection model. *Computer Methods and Programs in Biomedicine*, 106:37–46, 2012.

N. S. Bardach, J. J. Wang, S. F. De Leon, S. C. Shih, W. J. Boscardin, L. E. Goldman, and R. A. Dudley. Effect of pay-for-performance incentives on quality of care in small practices with electronic health records: a randomized trial. *Journal of the American Medical Association*, 310(10):1051–1059, 2013.

A. Bayerstadler, F. Benstetter, C. Heumann, and F. Winter. A predictive modeling approach to increasing the economic effectiveness of disease management programs. *Health Care Management Science*, 17(3):284–301, 2014. available at http://link.springer.com/article/10.1007/s10729-013-9246-y.

A. Bayerstadler, L. van Dijk, and F. Winter. Bayesian multinomial latent variable modeling for fraud and abuse detection in health insurance. *Insurance: Mathematics and Economics*, 71:244–252, 2016. available at http://www.sciencedirect.com/science/article/pii/S0167668715302845.

T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions (1683-1775)*, pages 370–418, 1763.

C. Belitz, A. Brezger, T. Kneib, and S. Lang. *BayesX - Software for Bayesian Inference in Structured Additive Regression Models, Version 2.0.1*, 2009. erhältlich unter: http://www.stat.uni-muenchen.de/~bayesx.

G. D. Berg and S. Wadhwa. Diabetes disease management results in Hispanic Medicaid patients. *Journal of Health Care for the Poor and Underserved*, 20(2):432–443, 2009.

G. D. Berg, S. Wadhwa, and A. E. Johnson. A Matched-Cohort Study of Health Services Utilization and Financial Outcomes for a Heart Failure Disease-Management Program in Elderly Patients. *Journal of the American Geriatrics Society*, 52(10): 16551661, 2004.

L. Bermúdez, J. M. Pérez, M. Ayuso, E. Gómez, and F. J. Vázquez. A bayesian dichotomous model with asymmetric link for fraud in insurance. *Insurance: Mathematics and Economics*, 42(2):779–786, 2008.

D. M. Berwick and A. D. Hackbarth. Eliminating Waste in US Health Care. *Journal of the American Medical Association*, 307:1513–1516, 2012.

S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland. Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50:602–613, 2011.

J. Billings and T. Mijanovich. Improving the Management of Care for High-Cost Medicaid Patients. *Health Affairs*, 26(6):1643–1655, 2007.

J. D. Birkmeyer, J. F. Finks, A. O'Reilly, M. Oerline, A. M. Carlin, A. R. Nunn, J. Dimick, M. Banerjee, and N. J. O. Birkmeyer. Surgical skill and complication rates after bariatric surgery. *New England Journal of Medicine*, 369(15):1434–1442, 2013.

A. Blackman, B. Lahiri, W. Pizer, M. Rivera Planter, and C. Muñoz Piña. Voluntary environmental regulation in developing countries: Mexico Clean Industry Program. *Journal of Environmental Economics and Management*, 60(3):182–192, 2010.

D. K. Blough, C. W. Madden, and M. C. Hornbrook. Modeling risk using generalized linear models. *Journal of Health Economics*, 18:153–171, 1999.

T. Bodenheimer, K. Lorig, H. Holman, and K. Grumbach. Patient Self-management of Chronic Disease in Primary Care. *Journal of the American Medical Association*, 288(19):2469–2475, 2002a.

T. Bodenheimer, E. H. Wagner, and K. Grumbach. Improving Primary Care for Patients with Chronic Illness. *Journal of the American Medical Association*, 288 (15):1909–1914, 2002b.

R. J. Bolton and D. J. Hand. Statistical fraud detection: A review. *Statistical science*, 17:235–255, 2002.

S. Boriah, V. Chandola, and V. Kumar. Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the eighth SIAM International Conference on Data Mining*, pages 243–254, 2008.

G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. Wiley, 2013.

L. Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

R. Brown, B. Pham, and O. de Vel. Design of a digital forensics image mining system. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 395–404. Springer, 2005.

M. B. Buntin and A. M. Zaslavsky. Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures. *Journal of Health Economics*, 23:525–542, 2004.

R. S. Busch. *Healthcare Fraud: Auditing and Detection Guide*. John Wiley & Sons, 2008.

A. C. Cameron and P. K. Trivedi. *Regression Analysis of Count Data*. Cambridge University Press, New York, 1998.

E. Chan, O. A. R. Mahroo, and D. J. Spalton. Complications of cataract surgery. *Clinical and Experimental Optometry*, 93(6):379–389, 2010.

S. Chib, E. Greenberg, and Y. Chen. MCMC Methods for Fitting and Comparing Multinomial Response Models. *Economics Working Paper Archive, Econometrics*, 9802001, 1998.

H. Chipman, E. I. George, R. E. McCulloch, M. Clyde, D. P. Foster, and R. A. Stine. The practical implementation of Bayesian model selection. *Lecture Notes-Monograph Series*, pages 65–134, 2001.

M. Coca-Perraillon. Local and Global Optimal Propensity Score Matching. In *SAS Global Forum 2007: Statistics and Data Analytics*. Harvard Medical School, Boston, 2007.

W. G. Cochran. Matching in analytical studies. *American Journal of Public Health*, 43:684–691, 1953.

M. E. Cohen, C. Y. Ko, K. Y. Bilimoria, L. Zhou, K. Huffman, X. Wang, Y. Liu, K. Kraemer, X. Meng, and R. Merkow. Optimizing ACS NSQIP modeling for evaluation of surgical quality and risk: patient risk adjustment, procedure mix adjustment, shrinkage adjustment, and surgical focus. *Journal of the American College of Surgeons*, 217(2):336–346, 2013.

M. S. Conti. Effect of disease management programs on Medicaid costs. Master's thesis, University of Notre Dame, 2011.

R. B. D'Agostino, S. Grundy, L. M. Sullivan, and P. Wilson. Validation of the Framingham Coronary Heart Disease Prediction Scores: Results of a Multiple Ethnic Groups Investigation. *Journal of the American Statistical Association*, 286(2):180–187, 2001.

A. C. Davison. *Statistical Models*. Cambridge University Press, New York, 2003.

P. De Jong and G. Z. Heller. *Generalized Linear Models for Insurance Data*. Cambridge University Press, 2008.

R. H. Dehejia and S. Wahba. Propensity Score-Matching Methods for Nonexperimental Causal Studies. *The Review of Economics and Statistics*, 84(1):151–161, 2002.

A. E. Dembe and L. I. Boden. Moral Hazard: A Question of Morality? *New Solutions*, 10(3):257–279, 2000.

P. Diehr, D. Yanez, M. Ash, A. Hornbrook, and D. Y. Lin. Methods for analyzing health care utilization and costs. *Annual Review of Public Health*, 20:125–144, 1999.

J. B. Dimick, P. J. Pronovost, J. A. Cowan, and P. A. Lipsett. Complications and costs after high-risk surgery: Where should we focus quality improvement initiatives? *Journal of the American College of Surgeons*, 196(5):671–678, 2003.

J. B. Dimick, W. B. Weeks, R. J. Karia, S. Das, and D. A. Campbell. Who pays for poor surgical quality? Building a business case for quality improvement. *Journal of the American College of Surgeons*, 202(6):933–937, 2006.

H. G. Dove and I. Duncan. *An Introduction to Care Management Interventions and their Implications for Actuaries*. Society of Actuaries, 2004.

P. Dua and S. Bais. Supervised Learning Methods for Fraud Detection in Healthcare Insurance. In *Machine Learning in Healthcare Informatics*, pages 261–285. Springer, 2014.

N. Duan, W. G. Manning, C. N. Morris, and J. P. Newhouse. A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics*, 1(2):115–126, 1983.

T. Ekina, F. Leva, F. Ruggeri, and R. Soyer. Application of bayesian methods in detection of healthcare fraud. *Chemical Engineering Transaction*, 33, 2013.

R. K. Elmallah, J. J. Cherian, H. Amin, J. J. Jauregui, T. P. Pierce, and M. A. Mont. Readmission rates in patients who underwent total hip arthroplasty. *Surgical Technology International*, 27:215–217, 2015.

E. J. Emanuel and L. L. Emanuel. The economics of dying – the illusion of cost savings at the end of life. *New England Journal of Medicine*, 330(8):540–544, 1994.

D. Esposito, R. Brown, A. Chen, J. Schore, and R. Shapiro. Impacts of a Disease Management Program for Dually Eligible Beneficiaries. *Healthcare Financing Review*, 30(1):27–45, 2008.

European Union Commission. Study on corruption in the healthcare sector. http://www.ehfcn.org/images/EHFCN/Documents/EHFCN_ECORYS_20131219_study_on_corruption_in_the_healthcare_sector_en.pdf, October 2013.

J. E. Ezekiel and V. R. Fuchs. The Perfect Storm of Overutilization. *Journal of the American Medical Association*, 299:2789–2791, 2008.

L. Fahrmeir and T. Kneib. *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*. Oxford University Press, 2010.

L. Fahrmeir and A. Raach. A Bayesian Semiparametric Latent Variable Model for Mixed Responses. *Psychometrika*, 72:327–346, 2007.

L. Fahrmeir and S. Steinert. A geoadditive Bayesian latent variable model for Poisson indicators. Technical report, Discussion paper Sonderforschungsbereich 386 der Ludwig-Maximilians-Universität München, 2006.

L. Fahrmeir, A. Hamerle, and G. Tutz. *Multivariate Statisische Verfahren*. de Gruyter, Berlin, 1996.

A. J. Fairfax, C. D. Lambert, and A. Leatham. Systemic embolism in chronic sinoatrial disorder. *New England Journal of Medicine*, 295(4):190–192, 1976.

M. Fishbein and M. C. Yzer. Using Theory to Design Effective Health Behavior Interventions. *Communication Theory*, 13(2):164–183, 2003.

L. Francis. Neural Networks demystified. Technical report, Casualty Actuarial Society Forum, 2001. available at http://casualtyactuarialsociety.com/pubs/forum/01wforum/01wf253.pdf.

L. Francis. Martian Chronicles: Is MARS better than Neural Networks? Technical report, Casualty Actuarial Society Forum, 2003. available at http://casualtyactuarialsociety.com/pubs/forum/03wforum/03wf027.pdf.

R. Freeman, K. M. Lybecker, and D. W. Taylor. The Effectiveness of Disease Management Programs in the Medicaid Population. Technical report, The Cameron Institute, 2011. available at http://cameroninstitute.com.

E. W. Frees and E. A. Valdez. Hierarchical Insurance Claims Modeling. *Journal of the American Statistical Association*, 103(484):1457–1469, 2008.

E. W. Frees, V. R. Young, and Y. Luo. A longitudinal data analysis interpretation of credibility models. *Insurance: Mathematics and Economics*, 24:229–247, 1999.

A. A. Freitag. *Data mining and knowledge discovery with evolutionary algorithms*. Springer Verlag, 2002.

J. H. Friedman. Multivariate Adaptive Regression Splines. *Annals of Statistics*, 19 (1):1–141, 1991.

C. F. Gauss. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum*. Perthes and Besser, 1809.

J. Gee and M. Button. The financial cost of healthcare fraud 2014. Technical report, BDO LLP, 2014.

E. I. George and R. E. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, pages 339–373, 1997.

A. Giordano, S. Scalvini, E. Zanelli, U. Corrà, G. L. Longobardi, V.A. Ricci, P. Baiardi, and F. Glisenti. Multicenter randomised trial on home-based telemanagement to prevent hospital readmission of patients with chronic heart failure. *International Journal of Cardiology*, 131(2):192–199, 2009.

R. Z. Goetzel, R. J. Ozminkowski, V. G. Villagra, and J. Duffy. Return on investment in disease management: a review. *Health Care Financing Review*, 26(4):1–19, 2005.

A. S. Goldberger. Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society*, pages 979–1001, 1972.

P. I. Good. *Introduction to Statistics Through Resampling Methods and R/S-PLUS*. Wiley Interscience, 2005.

R. B. Gramacy, J. H. Lee, and R. Silva. On estimating covariances between many assets with histories of highly variable length. *arXiv preprint arXiv:0710.5837*, 2007.

S. W. Grant, A. D. Grayson, J. Zacharias, M. J. R. Dalrymple-Hay, P. D. Waterworth, and B. Bridgewater. What is the impact of endoscopic vein harvesting on clinical outcomes following coronary artery bypass graft surgery? *Heart*, 98(1): 60–64, 2012.

P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82:711–732, 1995.

G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, 1982.

S. Guo and M. Fraser. *Propensity Score Analysis: Statistical Methods and Applications*. Thousand Oaks, CA: SAGE Publications, Inc., 2010.

S. Haberman and A. E. Renshaw. Actuarial Applications of Generalized Linear Models. In D. Hand and S. Jacka, editors, *Statistics in Finance*, chapter 3, pages 41–65. Arnold, E., 1998.

H. He, J. Wang, W. Graco, and S. Hawkins. Application of neural networks to detection of medical fraud. *Expert Systems with Applications*, 13:329–336, 1997.

C. Henderson, M. Knapp, J.-L. Fernández, J. Beecham, S. P. Hirani, M. Cartwright, Rixon L., M. Beynon, A. Rogers, P. Bower, H. Doll, R. Fitzpatrick, A. Steventon, M. Bardsley, J. Hendy, and S. P. Newman. Cost effectiveness of telehealth for patients with long term conditions (Whole Systems Demonstrator telehealth questionnaire study): nested economic evaluation in a pragmatic, cluster randomised controlled trial. *BMJ*, 346, 2013.

K. Hirano and G. W. Imbens. Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. *Health Services and Outcomes Research Methodology*, 2:259–278, 2001.

C. Hobson, T. Ozrazgat-Baslanti, A. Kuxhausen, P. Thottakkara, P. A. Efron, F. A. Moore, L. L. Moldawer, M. S. Segal, and A. Bihorac. Cost and mortality associated with postoperative acute kidney injury. *Annals of Surgery*, 261(6):1207–1214, 2015.

K. Hollenbeck. *On the use of administrative data for workforce development program evaluation*. US Department of Labor, Employment and Training Administration, 2005.

C. Holton. Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem. *Decision Support Systems*, 46:853–864, 2009.

S. M. Iacus, G. King, and G. Porro. Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Analysis*, 20(1):1–24, 2012.

IDF. International Diabetes Federation. http://atlas.idf-bxl.org/content/economic-impacts-diabetes and http://www.idf.org/node/23640, 2011.

G. W. Imbens and A. Abadie. *Simple and Bias-Corrected Matching Estimators for Average Treatment Effects*. National Bureau of Economic Research, 2002.

S. C. Inglis, R. A. Clark, F. A. McAlister, J. Ball, C. Lewinter, D. Cullington, S. Stewart, and J. G. F. Cleland. Structured telephone support or telemonitoring programmes for patients with chronic heart failure. Cochrane Database Syst Rev 2010;8:CD007228, 2010.

Y. Jin, R. M. Rejesus, and B. B. Little. Binary choice models for rare events data: a crop insurance fraud application. *Applied Economics*, 37:841–848, 2005.

H. Joudaki, A. Rashidian, B. Minaei-Bidgoli, M. Mahmoodi, B. Geraili, M. Nasiri, and M. Arab. Using data mining to detect health care fraud and abuse: A review of literature. *Global Journal of Health Science*, 7:194–202, 2014.

E. Kirkos, C. Spathis, and Y. Manolopoulos. Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, 32: 995–1003, 2007.

K. Kirschner, J. Braspenning, R. P. Akkermans, J. E. A. Jacobs, and R. Grol. Assessment of a pay-for-performance program in primary care designed by target users. *Family Practice*, 30(2):161–171, 2013.

I. Kolyshkina, S. S. W. Wong, and S. Lim. Enhancing Generalised Linear Models with Data Mining. Discussion Paper, Casualty Actuarial Society, Arlington, Virginia, 2004. available at http://www.casact.org/pubs/dpp/dpp04/04dpp279.pdf.

C. Kooperberg, S. Bose, and C. J. Stone. Polychotomous regression. *Journal of the American Statistical Association*, 92(437):117–127, 1997.

D. Kralik, T. Koch, K. Price, and N. Howard. Chronic illness self-management: taking action to create order. *Journal of Clinical Nursing*, 13:259–267, 2004.

R. LaBelle, G. Stoddart, and T. Rice. A re-examination of the meaning and importance of supplier-induced demand. *Journal of Health Economics*, 13(3):347–368, 1994.

L. M. Lamers. A risk-adjuster for capitation payments based on the use of prescribed drugs. *Medical Care*, 37:824–830, 1999.

L. M. Lamers. AIC and BIC – Comparisons of Assumptions and Performance. *Sociological Methods Research*, 33(2):188–229, 2004.

S. Leatherman, D. Berwick, D. Iles, L. S. Lewin, F. Davidoff, T. Nolan, and M. Bisognano. The Business Case for Quality: Case Studies and an Analysis. *Health Affairs*, 22(2):17–30, 2003.

A. M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. Firmin Didot, Paris, 1805.

J. P. Leigh and J. F. Fries. Health Habits, Health Care Use and Costs in a Sample of Retirees. *Inquiry*, 1:44–54, 1992.

J. Li, N. Morlet, J. Q. Ng, J. B. Semmens, and M. W. Knuiman. Significant non-surgical risk factors for endophthalmitis after cataract surgery: EPSWA fourth report. *Investigative Ophthalmology & Visual Science*, 45(5):1321–1328, 2004.

J. Li, K.-Y. Huang, J. Jin, and J. Shi. A survey on statistical methods for health care fraud detection. *Health care management science*, 11:275–287, 2008.

K. Y. Liang and S. Zeger. GEE estimators. *Biometrika*, 73(1):13–22, 1986.

A. Linden, J. L. Adams, and N. Roberts. Using Propensity Scores to Construct Comparable Control Groups for Disease Management Program Evaluation. *Disease Management and Health Outcomes*, 13(2):107–115, 2005.

A. Linden, J. L. Adams, and N. Roberts. Strengthening the case for disease management effectiveness: un-hiding the hidden bias. *Journal of Evaluation in Clinical Practice*, 12(2):140–147, 2006.

P. K. Lindenauer, D. Remus, S. Roman, M. B. Rothberg, E. M. Benjamin, A. Ma, and D. W. Bratzler. Public reporting and pay for performance in hospital quality improvement. *New England Journal of Medicine*, 356(5):486–496, 2007.

K. R. Lorig, P. Ritter, A. L. Stewart, D. S. Sobel, B. W. Brown, A. Bandura, V. M. Gonzalez, D. D. Laurent, and H. R. Holman. Chronic Disease Self-Management Program: 2-Year Health Status and Health Care Utilization Outcomes. *Medical Care*, 39(11):1217–1223, 2001.

A. M. Lukasiewicz, R. A. Grant, B. A. Basques, M. L. Webb, A. M. Samuel, and J. N. Grauer. Patient factors associated with 30-day morbidity, mortality, and length of stay after surgery for subdural hematoma: a study of the American College of Surgeons National Surgical Quality Improvement Program. *Journal of Neurosurgery*, 124(3):760–766, 2016.

D. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

P. C. Mahalanobis. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):4955, 1936.

J. A. Major and D. R. Riedinger. EFD: A Hybrid Knowledge/Statistical-Based System for the Detection of Fraud. *Journal of Risk and Insurance*, 69:309–324, 2002.

W. G. Manning. *The costs of poor health habits*. Harvard University Press, 1991.

W. G. Manning. The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics*, 17:283–295, 1998.

W. G. Manning and J. Mullahy. Estimating log models: to transform or not to transform? *Journal of Health Economics*, 20:461–494, 2001.

A. D. Martin, K. M. Quinn, and J. H. Park. MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software*, 42:1–21, 2011.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall / CRC, 1989.

C. E. McCulloch and S. R. Searle. *Generalized, Linear, and Mixed Models*. Wiley, New York, 2001.

S. Mehmud and R. Winkelman. A Comparative Analysis of Claims-Based Tools for Health Risk Assessment. Technical report, Society of Actuaries, 2007. available at http://www.soa.org/files/pdf/risk-assessmentc.pdf.

J. Meyer and B. M. Smith. Chronic Disease Management: Evidence of Predictable Savings. Technical report, Health Management Associates, 2008. available at http://www.idph.state.ia.us/hcr_committees/common/pdf/clinicians/savings_report.pdf.

A. Miksch, G. Laux, D. Ose, S. Joos, S. Campbell, B. Riens, and J. Szecsenyi. Is there a survival benefit within a German primary care-based disease management program? *American Journal of Managed Care*, 16(1):49–54, 2010.

A. J. Miller. *Subset Selection in Regression*. Chapman and Hall, New York, 1990.

J. Mullahy. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics*, 17:247–281, 1998.

R. M. Musal. Two models to investigate medicare fraud within unsupervised databases. *Expert Systems with Applications*, 37:8628–8633, 2010.

J. P. Newhouse, W. G. Manning, E. B. Keeler, and E. M. Sloss. Adjusting capitation rates using objective health measures and prior utilization. *Health Care Financing Review*, 10(3):41–54, 1989.

E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50:559–569, 2011.

G. Ninot, G. Moullec, M. C. Picot, A. Jaussent, M. Hayot, M. Desplan, J. F. Brun, J. Mercier, and C. Prefaut. Cost-saving effect of supervised exercise associated to COPD self-management education program. *Respiratory Medicine*, 105:377–385, 2011.

R. Nugent. Chronic Diseases in Developing Countries: Health and Economic Burdens. *Annual of the New York Academy of Sciences*, 1136:70–79, 2008.

A. A. Okunade and V. N. R. Murthy. Technology as a 'major driver' of health care costs: a cointegration analysis of the Newhouse conjecture. *Journal of Health Economics*, 21(1):147–159, 2002.

S. T. Parente, B. Schulte, A. Jost, T. Sullivan, and A. Klindworth. Assessment of Predictive Modeling for Identifying Fraud within the Medicare Program. *Health Management, Policy and Innovation*, 1:8–37, 2012.

T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103:681–686, 2008. http://www.stat.ufl.edu/~casella/Papers/Lasso.pdf.

V. J. Patalano. The risks and benefits of cataract surgery. Technical report, Massachusetts Eye and Ear Infirmary, 2016. http://www.djo.harvard.edu/site.php?url=/patients/pi/408.

J. Pearce and S. Ferrier. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modeling*, 133:225–245, 2000.

Y. Peng, G. Kou, A. Sabatka, J. Matza, Z. Chen, D. Khazanchi, and Y. Shi. Application of classification methods to individual disability income insurance fraud detection. In *Computational Science–ICCS 2007*, pages 852–858. Springer, 2007.

C. Phua, V. Lee, K. Smith, and R. Gayler. A comprehensive survey of data mining-based fraud detection research. *Artifical Intelligence Review*, pages 1–14, 2005.

N. Porta, C. Bonet, and E. Cobo. Discordance between reported intention-to-treat and per protocol analyses. *Journal of Clinical Epidemiology*, 60(7):663–669, 2007.

M. E. Porter and E. O. Teisberg. *Redefining Health Care: Creating value-based Competition on Results*. Boston, 2006.

N. R. Powe, O. D. Schein, S. C. Gieser, J. M. Tielsch, R. Luthra, J. Javitt, and E. P. Steinberg. Synthesis of the literature on visual acuity and complications following cataract extraction with intraocular lens implantation. *Archives of Ophthalmology*, 112(2):239–252, 1994.

A. E. Powell, H. T. O. Davies, and R. G. Thomson. Using routine comparative data to assess the quality of health care: understanding and avoiding common pitfalls. *Quality and safety in health care*, 12(2):122–128, 2003.

C. A. Powers, C. M. Meyer, M. C. Roebuck, and B. Vaziri. Predictive modeling of total healthcare costs using pharmacy claims data: A comparison of alternative econometric cost modeling techniques. *Medical Care*, 43(11):1065–1072, 2005.

M. J. Press, D. P. Scanlon, A. M. Ryan, J. Zhu, A. S. Navathe, J. N. Mittler, and K. G. Volpp. Limits of readmission rates in measuring hospital quality suggest the need for added metrics. *Health Affairs*, 32(6):1083–1091, 2013.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL http://www.R-project.org. ISBN 3-900051-07-0.

B. Riegel, B. Carlson, D. Glaser, and P. Hoagland. Which Patients with Heart Failure Respond Best to Multidisciplinary Disease Management? *Journal of Cardiac Failure*, 6(4):290–299, 2000.

B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 2004.

D. B. Rubin. Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies. *Journal of the American Statistical Association*, 74(366a):318–328, 1979.

D. B. Rubin. *Matched Sampling for Causal Effects*. Cambridge University Press, 2006.

Y. Sahin and E. Duman. Detecting credit card fraud by ANN and logistic regression. In *2011 International Symposium on Innovations in Intelligent Systems and Applications (INISTA)*, pages 315–319. IEEE, 2011a.

Y. Sahin and E. Duman. Detecting credit card fraud by decision trees and support vector machines. In *International MultiConference of Engineers and Computer Scientists*, volume 1, 2011b.

M. D. Sammel, L. M. Ryan, and J. M. Legler. Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 59:667–678, 1997.

G. E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2): 461–464, 1978.

T. Shih, L. H. Nicholas, J. R. Thumma, J. D. Birkmeyer, and J. B. Dimick. Does pay-for-performance improve surgical outcomes? an evaluation of phase 2 of the premier hospital quality incentive demonstration. *Annals of Surgery*, 259(4):677, 2014.

H. Shin, H. Park, J. Lee, and W. C. Jhee. A scoring model to detect abusive billing patterns in health insurance claims. *Expert Systems with Applications*, 39: 7441–7450, 2012.

M. Shwartz, J. Ren, E. A. Peköz, X. Wang, A. B. Cohen, and J. D. Restuccia. Estimating a composite measure of hospital quality from the Hospital Compare database: differences when using a Bayesian hierarchical latent variable model versus denominator-based weights. *Medical Care*, 46:778–785, 2008.

J. Sidorov, R. Shull, J. Tomcavage, S. Girolami, N. Lawton, and R. Harris. Does Diabetes Disease Management Save Money and Improve Outcomes? *Diabetes Care*, 25(4):684–689, 2002.

A. Skrondal and S. Rabe-Hesketh. *Generalized Latent Variable Modeling*. Chapman & Hall / CRC, 2004.

S. Smith, J. P. Newhouse, and M. S. Freeland. Income, Insurance, And Technology: Why Does Health Spending Outpace Economic Growth? *Health Affairs*, 28(5): 1276–1284, 2009.

S. Stock, A. Drabik, G. Büscher, C. Graf, W. Ullrich, A. Gerber, K. W. Lauterbach, and M. Lüngen. German Diabetes Management Programs Improve Quality Of Care And Curb Costs. *Health Affairs*, 29(12):2197–2205, 2010.

C. J. Stone, M. Hansen, C. Kooperberg, and Y. K. Truong. The use of polynomial splines and their tensor products in extended linear modeling (with discussion). *Annals of Statistics*, 25:1371–1470, 1997.

L. Šubelj, Š. Furlan, and M. Bajec. An expert system for detecting automobile insurance fraud using social network analysis. *Expert Systems with Applications*, 38:1039–1052, 2011.

J. W. Thomas and T. P. Hofer. Accuracy of risk-adjusted mortality rate as a measure of hospital quality of care. *Medical Care*, 37(1):83–92, 1999.

B. Thompson. *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. American Psychological Association, Washington DC, 2004.

D. Thornton, R. M. Mueller, P. Schoutsen, and J. van Hillegersberg. Predicting Healthcare Fraud in Medicaid: A Multidimensional Data Model and Analysis Techniques for Fraud Detection. *Procedia Technology*, 9:1252–1264, 2013.

J. M. Tielsch, M. W. Legro, S. D. Cassard, O. D. Schein, J. C. Javitt, A. E. Singer, E. B. Bass, and E. P. Steinberg. Risk factors for retinal detachment after cataract surgery: a population-based case-control study. *Ophthalmology*, 103(10):1537–1545, 1996.

P. T. Troughton and S. J. Godsill. A reversible jump sampler for autoregressive time series, employing full conditionals to achieve efficient model space moves. CUED/F-INFENG/TR. 304, University of Cambridge: Department of Engineering, 1997.

G. Tutz. *Analyse kategorialer Daten*. Oldenbourg Verlag, 2000.

G. Tutz and L. Fahrmeir. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New York, 2001.

C. J. vanVonno, R. J. Ozminkowski, M. W. Smith, E. G. Thomas, D. Kelley, R. Goetzel, G. D. Berg, S. K. Jain, and D. R. Walker. Evaluation of Savings and Return on Investment for a Pilot Congestive Heart Failure Disease Management Program Offered to Federal Employees. *Disease Management*, 8(6):346–360, 2005.

P. J. Veazie, W. G. Manning, and R. L. Kane. Improving risk adjustment for medicare capitated reimbursement using nonlinear models. *Medical Care*, 41(6): 741–752, 2003.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, fourth edition edition, 2002.

S. Viaene, R. A. Derrig, B. Baesens, and G. Dedene. A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection. *Journal of Risk and Insurance*, 69:373–421, 2002.

S. Viaene, R. A. Derrig, and G. Dedene. Cost-sensitive learning and decision making for Massachusetts PIP claim fraud data. *International journal of intelligent systems*, 19:1197–1215, 2004a.

S. Viaene, R. A. Derrig, and G. Dedene. A case study of applying boosting Naive Bayes to claim fraud diagnosis. *Knowledge and Data Engineering, IEEE Transactions on*, 16(5):612–620, 2004b.

S. Viaene, G. Dedene, and R. A. Derrig. Auto claim fraud detection using bayesian learning neural networks. *Expert Systems with Applications*, 29:653–666, 2005.

V. G. Villagra and A. Tamim. Effectiveness Of A Disease Management Program For Patients With Diabetes. *Health Affairs*, 23(4):255–266, 2004.

B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos. Class imbalance, redux. In *2011 IEEE 11th International Conference on Data Mining (ICDM)*, pages 754–763. IEEE, 2011.

R. W. M. Wedderburn. Quasi-Likelihood Functions, Generalized Linear Models and the Gauss-Newton method. *Environmental Research*, 104:402–409, 1974.

A. R. Wells, B. Hamar, C. Bradley, W. M. Gandy, P. L. Harrison, J. A. Sidney, C. R. Coberley, E. Y. Rula, and J. E. Pope. Exploring Robust Methods for Evaluating Treatment and Comparison Groups in Chronic Care Management Programs. *Population Health Management*, 16(1):35–45, 2013.

S. S. Wilks. On the distribution of statistics in samples from a normal population of two variables with matching sampling of one variable. *Metron*, pages 87–126, 9.

W.-S. Yang and S.-Y. Hwang. A process-mining framework for the detection of healthcare fraud and abuse. *Expert Systems with Applications*, 31:56–68, 2006.

K. W. Yau, A. H. Lee, and A. S. K. Ng. A zero-augmented gamma mixed model for longitudinal data with many zeros. *Australian & and New Zealand Journal of Statistics*, 44(2):177–183, 2002.

F. M. Zahid and G. Tutz. Multinomial logit models with implicit variable selection. *Advances in Data Analysis and Classification*, 7(4):393–416, 2013.

B. Zheng, J. Zhang, S. W. Yoon, S. S. Lam, M. Khasawneh, and S. Poranki. Predictive modeling of hospital readmissions using metaheuristics and data mining. *Expert Systems with Applications*, 42(20):7110–7120, 2015.