Dissertation
zum Erwerb des Doctor of Philosophy (Ph.D.)
an der Medizinischen Fakultät der
Ludwig-Maximilians-Universität zu München

Statistical problems in the analysis of health claims data: new approaches and applications

vorgelegt von

Christoph Franz Kurz

aus

Dachau

am

20. September 2019

erstellt am

Helmholtz Zentrum München

Supervisor:	Prof. Dr. biol. hom. Rolf Holle
Second Expert:	Prof. Dr. rer. nat. Ulrich Mansmann
Dean:	Prof. Dr. dent. med. Reinhard Hickel
Date of oral defense:	27.1.2020

Contents

List of abbreviations

AIC	Akaike information criterion
FMM	Finite mixture model
GenG	Generalized Gamma distribution
GLM	Generalized linear model
RMSE	Root mean squared error

1 Introductory summary

1.1 Background

The study of health care utilization answers important health care questions. An appropriate management of health care utilization assists to prevent serious health conditions and provides the foundation for more efficient procedures (de Boer et al., 1997). For example, problems with access to care can be indicated by the finding that people with a lower income or who live in certain regions use fewer services (Diehr et al., 1999). However, a high variation among subgroups in rates at which surgery is performed indicates that some individuals may not receive optimal care (Deyo et al., 2010). Therefore, a prediction of total health care costs for a number of individuals remains important, because then the providers can appropriately allocate resources for caring for those people (Bates et al., 2014).

In addition, well-managed health care resource allocation and adherence to recommended care practices tend to reduce the use of emergency rooms and hospitalization, leading to better health outcomes and less expensive care for patients (Hilton et al., 2018).

Currently, the introduction of individual health records and claims data delivered by health care providers and insurers assists in understanding and managing health care utilization (Schneeweiss and Avorn, 2005).

Usually, health claims data include not only information on health services such as the type and location of care, services provided, diagnosis, and procedure codes, but also the individual-level information such as the demographics for thousands of patients. However, utilization data hold several characteristics that foster them challenging to analyze (Griswold et al., 2004; Mihaylova et al., 2011).

Because of the increasing complexity of medical procedures and the focus on evidence-based practice, attention to statistical quality in medical research has risen in recent years. Although randomized trial designs have been employed to evaluate the quality of care and to identify effective interventions in the real world, the empirical basis of public health research is primarily resides on data collected in the observational setting, e.g., in the routine setting of daily practice (Normand, 2008; Clarke et al., 2019). However in routine data, data are often collected for other purposes and are not used experimentally (Powell et al., 2003). Hence, empirical analyses of routine data require statistical tools that can manage the complexity of these data.

This thesis presents three statistical approaches and applications that address problems in analyzing health care utilization using claims data, covered in three different manuscripts. All manuscripts included in this thesis explore and use recently developed statistical methods that may be specifically useful for the analysis of health claims data.

The first approach concerns the unique distributional properties of health care utilization cost data and the problems that accompany accurate modeling and prediction of such data. Many of the problems that typically arise during the analytical process can be solved by the statistical distribution for health care cost data of this approach.

The second approach focuses on discovering of latent subgroups of individuals with distinct utilization patterns. Policymakers, payers, and clinicians seek to improve care and reduce spending in these groups by interventions tailored to subpopulations (Hastings et al., 2014). Here,

a Bayesian mixture regression model for count data identifies distinct subgroups of lung cancer patients with regard to their health as well as treatment, and relationships between other covariates (e.g. age, living in rural areas) and hospital length of stay.

The third approach concentrates on causal inference applications using health claims data. First, the causal effect of bariatric surgery on health care costs was estimated. This estimate derived from a developed and applied Bayesian structural time series model that estimated what costs would have incurred for individuals if they had not received the surgery.

1.2 Accurate modeling and prediction of health care costs

Accurate estimates of health care costs determine the trade offs between medical possibilities, their financial viability, as well as the quality and fairness in any health care system. Health care utilization cost data are usually not normally distributed, as some individuals cause no costs while others cause extremely high costs. In addition, costs cannot be negative. The distribution of costs is called semi-continuous (Min and Agresti, 2002) and poses many problems. For example, common statistical models involving the Gamma or log-normal distributions have difficulty with such a combination of discrete and continuous values due to the significant part of the population with zero costs. Models for these data must be flexible enough to accommodate these features and yet still produce interpretable, policy-relevant results (Mihaylova et al., 2011).

A popular manner to analyze cost data in the generalized linear models (GLM) framework is the use of two-part models (Duan et al., 1983) that combine a binary model for the dichotomous event of having either zero or positive values with a continuous model for those individuals having positive values. This complements a two-stage decision process, that can be inadequate because the two decisions are not usually made independently (Winkelmann, 2004; Van Ophem, 2011). In constrast, the Tobit model uses a single distribution (Tobin, 1958). This model is based on a zero-truncated Normal distribution but cannot handle excess zeros, i.e. the presence of more zeros in the data than would be expected from the underlying distribution. This linear regression setting assumes constant variance that is inadequate for cost data.

Recent research has primarily focused on developing new models and comparing distributions for the continuous part of the two-part models. For example, Jones et al. (2016) compare several recent developments in parametric and semiparametric regression models, i.e., models that use a finite number of parameters for health care costs, including the generalized Gamma distribution (GenG), Weibull, and exponential distribution. The comparative studies from Basu et al. (2006) and Hill and Miller (2010), study models with either real data, i.e. the true distribution is unknown, or by using simulations. Both analyze positive costs with no emphasis on the zero aspect. The only comparative study considering zero costs has been documented by Buntin and Zaslavsky (2004).

As an alternative, Manuscript 1 (Kurz, 2017) considered a single distribution GLM that can simultaneously model the zeros and continuous positive outcomes for cost data. The number of excess zeros could be arbitrarily high while still providing good support for the positive costs. This model, based on the family of Tweedie densities (Tweedie, 1984), demonstrates many advantages that enable it an ideal candidate for health economic cost data modeling.

The Tweedie family of distributions corresponds to special cases of exponential dispersion

models (Jorgensen, 1997) in which the mean-variance relationship can be flexibly specified. For example, the purely continuous Normal, Gamma and inverse Gaussian distributions are all part of the Tweedie family. However, the class of compound Poisson-Gamma distributions that have positive mass at zero, but are otherwise continuous, are the most relevant subclass for cost data. Because Tweedie distributions also belong to the exponential family of distributions, they can be used in the GLM framework.

Manuscript 1 compared the Tweedie model with the two-part (Binomial/Gamma and Binomial/GenG), the Tobit, and the Poisson models regarding marginal effects (at the means), model fit and prediction error in both Monte Carlo simulation and real data. These are all simple models that are easy to interpret and favored by analysts.

The simulation study assessed both root mean squared error (RMSE), an absolute measure of model quality, and Akaike information criterion (AIC) (Akaike, 1973), a relative measure of model quality, across different settings with low and high correlation (i.e., how much users and non-users differ in their characteristics) and varying the numbers of zeros. If the number of zeros was below 20%, the Tweedie model outperformed the Tobit, the Poisson, and both two-part models in situations with high correlation between users and non-users. When the zero percentage was above 20%, two-part models started to surpass the Tweedie model in both AIC and RMSE.

In a real data application, the AICs of the Tweedie and the two-part Gamma models were almost identical, suggesting a comparable model fit. Yet, the two-part GenG showed slightly superior fit with a lower AIC, but the Tweedie clearly outperformed Poisson and Tobit models.

From these results, the models based on Tweedie distributions provide an interesting alternative for the analysis of semi-continuous health care cost data. Indeed, they remain especially useful when the correlation between users and non-users of health care utilization is high and the proportion of these non-users is low.

1.3 Subgroup identification in health care utilization data

To characterize the utilization behaviors and to investigate the drivers of variations in health care utilization may highlight targeted interventions that can improve disease management and treatment. Clustered data complicates an analysis primarily because no reason exists to assume that observations are statistically independent within a cluster (Normand, 2008). For example, with many public health studies, interventions are assessed on patients who are treated within *practice* settings. When evaluating the quality of health care provided in *ambulatory* treatment settings, all patients in the practice are exposed to the same quality level. Thus, it may be highly likely that the reception of guideline treatment for two patients sampled from the same practice would be more equal than the likelihood of two patients sampled from two different practices. Consequently, identifying subgroups remains an important but challenging task.

A variety of statistical models have been considered for subgroup identification. For example, models based on mixtures of parametric models represent a complicated density as a linear combination of simpler densities and, therefore, identify groups of observations with similar outcomes using unsupervised clustering. These mixture models, also known as latent class models (Böhning and Seidel, 2003; Muthén and Shedden, 1999) or switching models (Frühwirth-Schnatter, 2001), are motivated by the concern that different parts of the response distribution

(e.g., low cost users, high cost users) could be differently affected by covariates. They are widely used (see reviews in McLachlan and Peel (2004) and Titterington et al. (1985)) and often perform better than standard GLMs and the hurdle model (Deb and Trivedi, 1997).

However, fitting mixture models requires the specific number of mixture components (or clusters). Too many clusters will over-fit the data and impair model interpretation while too few will be unable to fully reflect the structure of the data. In the simplest case of mixture modeling, an initial natural hypothesis is posed which subgroups might exhibit different behavior, providing both an initial choice of the number of clusters and even which data points belong to each cluster. However, the problem remains on how to infer the clusters precisely, and derive hypotheses for further study *from* the model. Usually, the number of components will be decided either *ex ante*, by the choice of a convenient and interpretable number such as two or three or *ex post*, by the generation of models with different numbers of components and the manual search for a plausible best fit by the comparison of quantities such as AIC or likelihood ratio (McLachlan and Rathnayake, 2014).

Manuscript 2 (Kurz and Hatfield, 2019) concentrated on the variation in hospital inpatient days among patients diagnosed with lung cancer. To calculate this variation, two implementations of mixture models for zero-inflated count regression were defined and compared: maximum-likelihood-based finite mixture models (FMMs) and parametric Bayesian mixture models. Indeed, an explicit comparison could de determined about the maximum likelihood and parametric Bayesian mixture models for count data. Furthermore, assessment could be achieved about these two approaches' ability to detect the true number of mixture components and estimate component parameters, as well as the practicalities of both approaches could be produced. The model allowed for frequent zero-valued observations.

In summary, the Bayesian mixture modeling allowed the number of clusters to be estimated from the data. Yet, in a simulation study, the selection of the number of clusters in a FMM using model fit statistics such as AIC did not provide a very precise method for determining the true number of clusters. Instead, the posterior clusters probabilities from the Bayesian model were closer to the truth, although slightly overestimated the number of clusters as seen by other authors (Onogi et al., 2011).

Among the claims data set with lung cancer patients, three distinct clusters could be identified using the Bayesian mixture model. The first cluster contained individuals with the fewest hospital days on average; it found many patients undergoing chemotherapy only or undergoing chemotherapy in combination with radiation therapy. This and the lack of surgery, likely indicate that these patients were already in an advanced (metastatic) stage at diagnosis. For these patients, it could be that therapy had a palliative intent with a focus on improving the quality of life. In contrast, patients in clusters 2 and 3 were more likely to have surgery only, surgery and chemotherapy, and the combination of all three treatments. This suggests diagnosis at an earlier stage, and fosters more aggressive treatment.

1.4 Causal inference applications using health claims data

Although not always explicitly stated, the most common goal in public health research involves establishing causation. Causal inference focuses on what would happen to a specific individual

under different intervention options. Interventions can include different treatment options, but also educational or care programs, policy changes, or health promotion campaigns.

Health claims data have not been originally collected for drawing causal estimates from health claims data, because they are observational rather than experimental, and consequently, usually fail to meet most of the assumptions to support a causal conclusion (Shiffrin, 2016). However, health claims data present an exciting opportunity to determine estimates in health economics and health policy research because of their accurate collection of utilization measures. In the absence of a randomized control group, analyzing such data requires the use of new data-adaptive approaches that automatically optimize a confounding control to study causal treatment effects (Schneeweiss, 2018).

Manuscript 3 (Kurz et al., 2019) used claims data from the largest health insurance provider in Germany to estimate the causal effect of bariatric surgery on health care costs. Surgical measures to combat obesity are very effective in terms of weight loss, recovery from diabetes, and improvement in cardiovascular risk factors but their effect on health care utilization remains unclear. The economic aspects of bariatric surgery should be an important policy question, because being overweight and obesity pose high economic costs to health care providers (Yates et al., 2016; Tsai et al., 2011). This is particularly important in Germany because the absorption of bariatric surgical procedures expenses is not currently included in the statutory health insurance standard benefit catalogue. For individual cases, however, the interventions can be requested and funded by the patient's insurance fund. This might be the reason why the number of bariatric procedures in Germany is significantly lower than in neighboring nations. The frequency of operations for morbid obesity is currently 9 per 100,000 adults in Germany; in other European countries it is many times higher (e.g., Sweden: 77; France: 57; Belgium: 108) (Angrisani et al., 2015).

Due to the absence of a control group in the provided data, none of the traditional methods to establish causal estimates such as regression control and matching (Stuart, 2010) could be used. Instead, Manuscript 3 employed a Bayesian structural forecasting model to construct a synthetic control group. These methods have been shown to be useful in the analysis of intervention effects through time-series data in the absence of a randomized controlled trial (Bouttell et al., 2018). Traditional synthetic control methods usually involve the construction of a weighted combination of groups used as controls, to which the treatment group is compared.

In contrast, the Bayesian structural model estimates the model on the pre-treatment period using Gibbs sampling and then iterated each sampling trajectory forward using the estimated parameters to construct the post-intervention counterfactual. This is essentially a forecasting method trained on pre-treatment outcomes to construct a post-intervention counterfactual, i.e. what would have happened to individuals who did not receive bariatric surgery. This approach has the advantage that it does not require a set of dedicated control units, i.e., untreated individuals, and instead can use any sort of related time series to predict the counterfactual.

Using this method, Manuscript 3 found that bariatric surgery was associated with a cost reduction in pharmaceuticals and physician services, but also with rising costs for inpatient care. In total, health care costs increased slightly after bariatric surgery.

1.5 Conclusion and outlook

In conclusion, the current doctoral thesis highlights the new statistical challenges and opportunities that occur with analyzing health care utilization using health claims data. In Manuscript 1, a new statistical distribution was introduced that has the potential to more accurately model cost data. It showed many advantages over current methods for predicting and analyzing costs in health care settings. In future work, this could be extended to cost-effectiveness analyses in which the distributional aspects of the data also play an important role. Manuscript 2 focused on identifying different inpatient utilization patterns with the aim of improving health outcomes and care quality while reducing health care spending in groups with substantial heterogeneity. The presented Bayesian approach could be extended with more sophisticated priors to potentially use infinite mixture models. Also, a variational Bayesian approach to lower the computational burden might be considered. Manuscript 3 estimated the effect of bariatric surgery on health care costs using a synthetic control group estimated by a Bayesian structural model. This remains an important health policy question that needs further investigation with an increased number of patients and possibly a set of untreated controls to verify the results obtained in this manuscript.

2 Published manuscripts

1. Christoph Kurz. Tweedie distributions for fitting semicontinuous health care utilization cost data. *BMC Medical Research Methodology*, 17:171, 2017, doi: 10.1186/s12874-017-0445-y.

2. Christoph Kurz, Laura Hatfield. Identifying and interpreting subgroups in health care utilization data with count mixture regression models. *Statistics in Medicine*, 2019, 1-13, doi: 10.1002/sim.8307.

3. Christoph Kurz, Martin Rehm, Rolf Holle, Christina Teuner, Michael Laxy, Larissa Schwarzkopf. The effect of bariatric surgery on health care costs: a synthetic control approach using Bayesian structural time series. *Health Economics*, 2019, doi: 10.1002/hec.3941.

2.1 Manuscript 1

Christoph Kurz. Tweedie distributions for fitting semicontinuous health care utilization cost data. *BMC Medical Research Methodology*, 17:171, 2017, doi: 10.1186/s12874-017-0445-y.

2.2 Manuscript 2

Christoph Kurz, Laura Hatfield. Identifying and interpreting subgroups in health care utilization data with count mixture regression models. *Statistics in Medicine*, 2019, 1-13, doi: 10.1002/sim.8307.

2.3 Manuscript 3

Christoph Kurz, Martin Rehm, Rolf Holle, Christina Teuner, Michael Laxy, Larissa Schwarzkopf. The effect of bariatric surgery on health care costs: a synthetic control approach using Bayesian structural time series. *Health Economics*, 2019, doi: 10.1002/hec.3941.

3 Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory*, pages 267–281. Akademiai Kiado, Budapest.

Angrisani, L., Santonicola, A., Iovino, P., Formisano, G., Buchwald, H., and Scopinaro, N. (2015). Bariatric surgery worldwide 2013. *Obesity Surgery*, 25(10):1822–1832.

Basu, A., Arondekar, B. V., and Rathouz, P. J. (2006). Scale of interest versus scale of estimation: comparing alternative estimators for the incremental costs of a comorbidity. *Health Economics*, 15(10):1091–1107.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., and Escobar, G. (2014). Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, 33(7):1123–1131.

Böhning, D. and Seidel, W. (2003). Editorial: Recent developments in mixture models. *Computational Statistics & Data Analysis*, 41(3-4):349–357.

Bouttell, J., Craig, P., Lewsey, J., Robinson, M., and Popham, F. (2018). Synthetic control methodology as a tool for evaluating population-level health interventions. *Journal of Epidemiology and Community Health*, pages 2000–2017.

Buntin, M. B. and Zaslavsky, A. M. (2004). Too much ado about two-part models and transformation?: Comparing methods of modeling medicare expenditures. *Journal of Health Economics*, 23(3):525–542.

Clarke, G. M., Conti, S., Wolters, A. T., and Steventon, A. (2019). Evaluating the impact of healthcare interventions using routine data. *British Medical Journal*, 365:l2239.

de Boer, A. G., Wijker, W., and de Haes, H. C. (1997). Predictors of health care utilization in the chronically ill: a review of the literature. *Health Policy*, 42(2):101–115.

Deb, P. and Trivedi, P. (1997). Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics*, 12(3):313–336.

Deyo, R. A., Mirza, S. K., Martin, B. I., Kreuter, W., Goodman, D. C., and Jarvik, J. G. (2010). Trends, major medical complications, and charges associated with surgery for lumbar spinal stenosis in older adults. *Jama*, 303(13):1259–1265.

Diehr, P., Yanez, D., Ash, A., Hornbrook, M., and Lin, D. (1999). Methods for analyzing health care utilization and costs. *Annual Review of Public Health*, 20(1):125–144.

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

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 96(453):194–209.

Griswold, M., Parmigiani, G., Potosky, A., and Lipscomb, J. (2004). Analyzing health care costs: a comparison of statistical methods motivated by medicare colorectal cancer charges. *Biostatistics*, 1(1):1–23.

Hastings, S. N., Whitson, H. E., Sloane, R., Landerman, L. R., Horney, C., and Johnson, K. S. (2014). Using the past to predict the future: latent class analysis of patterns of health service use of older adults in the emergency department. *Journal of the American Geriatrics Society*, 62(4):711–715.

Hill, S. C. and Miller, G. E. (2010). Health expenditure estimation and functional form: applications of the generalized gamma and extended estimating equations models. *Health Economics*, 19(5):608–627.

Hilton, R. P., Zheng, Y., and Serban, N. (2018). Modeling heterogeneity in healthcare utilization using massive medical claims data. *Journal of the American Statistical Association*, 113(521):111–121.

Jones, A. M., Lomas, J., Moore, P. T., and Rice, N. (2016). A quasi-monte-carlo comparison of parametric and semiparametric regression methods for heavy-tailed and non-normal data: An application to healthcare costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(4):951–974.

Jorgensen, B. (1997). *The theory of dispersion models*. CRC Press.

Kurz, C. F. (2017). Tweedie distributions for fitting semicontinuous health care utilization cost data. *BMC Medical Research Methodology*, 17(1):171.

Kurz, C. F. and Hatfield, L. A. (2019). Identifying and interpreting subgroups in health care utilization data with count mixture regression models. *Statistics in Medicine*, 0(0):1–13.

Kurz, C. F., Rehm, M., Holle, R., Teuner, C., Laxy, M., and Schwarzkopf, L. (2019). The effect of bariatric surgery on health care costs: a synthetic control approach using Bayesian structural time series. *Health Economics*, 0(0):1–22.

McLachlan, G. and Peel, D. (2004). *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley.

McLachlan, G. and Rathnayake, S. (2014). On the number of components in a Gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(5):341–355.

Mihaylova, B., Briggs, A., O'Hagan, A., and Thompson, S. G. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health Economics*, 20(8):897–916.

Min, Y. and Agresti, A. (2002). Modeling nonnegative data with clumping at zero: a survey. *Journal of the Iranian Statistical Society*, 1(1):7–33.

Muthén, B. and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, 55(2):463–469.

Normand, S.-L. T. (2008). Some old and some new statistical tools for outcomes research. *Circulation*, 118(8):872–884.

Onogi, A., Nurimoto, M., and Morita, M. (2011). Characterization of a Bayesian genetic clustering algorithm based on a dirichlet process prior and comparison among Bayesian clustering methods. *BMC Bioinformatics*, 12(1):263.

Powell, A., Davies, H., and Thomson, R. (2003). Using routine comparative data to assess the quality of health care: understanding and avoiding common pitfalls. *BMJ Quality & Safety*, 12(2):122–128.

Schneeweiss, S. (2018). Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. *Clinical Epidemiology*, 10:771.

Schneeweiss, S. and Avorn, J. (2005). A review of uses of health care utilization databases for epidemiologic research on therapeutics. *Journal of Clinical Epidemiology*, 58(4):323–337.

Shiffrin, R. M. (2016). Drawing causal inference from big data. *Proceedings of the National Academy of Sciences*, 113(27):7308–7309.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Applied section. Wiley.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26(1):24–36.

Tsai, A. G., Williamson, D. F., and Glick, H. A. (2011). Direct medical cost of overweight and obesity in the usa: a quantitative systematic review. *Obesity Reviews*, 12(1):50–61.

Tweedie, M. C. (1984). An index which distinguishes between some important exponential families. In *Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference*, volume 579, pages 579–604.

Van Ophem, H. (2011). The frequency of visiting a doctor: is the decision to go independent of the frequency? *Journal of Applied Econometrics*, 26(5):872–879.

Winkelmann, R. (2004). Health care reform and the number of doctor visits—an econometric analysis. *Journal of Applied Econometrics*, 19(4):455–472.

Yates, N., Teuner, C. M., Hunger, M., Holle, R., Stark, R., Laxy, M., Hauner, H., Peters, A., and Wolfenstetter, S. B. (2016). The economic burden of obesity in Germany: results from the population-based KORA studies. *Obesity Facts*, 9(6):397–409.

4 Acknowledgements

Many people contributed to the development of this dissertation. I would like to take this opportunity to thank everyone who has supported me on this journey.

My special thanks goes to my supervisor Rolf Holle for his excellent guidance and support in theoretical and practical matters and constructive criticism in all phases of this work. It was always a pleasure to learn from him.

I would also like to thank my thesis advisory committee and all co-authors who contributed to the individual manuscripts. Their ideas and suggestions improved their scientific quality. I especially thank Laura Hatfield for her excellent advice and thoroughness.

In addition, I would also like to thank former and current colleagues for always having open doors at the Institute of Health Economics and Health Care Management. I am extraordinarily grateful to Michael, Julia, Manuel, Florian (x2), Bogi, Johanna, Karl, Renée, Sara, Christian, Sebastian, Christina and Adriana. The excellent working atmosphere is also a merit of the institute director Reiner Leidl.

Last but not least, I would like to thank my family for their support.

I hereby declare that the submitted thesis, entitled

Statistical problems in the analysis of health claims data: new approaches and applications,
is my own work. I have only used the sources indicated and have not made unauthorised use of the services of a third party. Where the work of others has been quoted or reproduced, the source is always given. I further declare that the submitted thesis or parts thereof have not been presented as part of an examination degree to any other university.

München, January 28, 2020

Christoph Kurz
(Signature doctoral candidate)

I hereby declare that the electronic version of the submitted thesis, entitled

Statistical problems in the analysis of health claims data: new approaches and applications,
is congruent with the printed version in both content and format.

München, 28. Januar 2020

Christoph Kurz
(Signature doctoral candidate)