

Cost-of-illness studies based on massive data: a prevalence-based, top-down regression approach

Björn Stollenwerk · Thomas Welchowski ·
Matthias Vogl · Stephanie Stock

Received: 21 March 2014 / Accepted: 12 January 2015 / Published online: 4 February 2015
© Springer-Verlag Berlin Heidelberg 2015

Abstract Despite the increasing availability of routine data, no analysis method has yet been presented for cost-of-illness (COI) studies based on massive data. We aim, first, to present such a method and, second, to assess the relevance of the associated gain in numerical efficiency. We propose a prevalence-based, top-down regression approach consisting of five steps: aggregating the data; fitting a generalized additive model (GAM); predicting costs via the fitted GAM; comparing predicted costs between prevalent and non-prevalent subjects; and quantifying the stochastic uncertainty via error propagation. To demonstrate the method, it was applied to aggregated data in the context of chronic lung disease to German sickness funds data (from 1999), covering over 7.3 million insured. To assess the gain in numerical efficiency, the computational time of the innovative approach has been compared with

corresponding GAMs applied to simulated individual-level data. Furthermore, the probability of model failure was modeled via logistic regression. Applying the innovative method was reasonably fast (19 min). In contrast, regarding patient-level data, computational time increased disproportionately by sample size. Furthermore, using patient-level data was accompanied by a substantial risk of model failure (about 80 % for 6 million subjects). The gain in computational efficiency of the innovative COI method seems to be of practical relevance. Furthermore, it may yield more precise cost estimates.

Keywords Cost-of-illness · Massive data · Generalized additive models · Error propagation

JEL Classification C1 · C5

B. Stollenwerk (✉) · T. Welchowski · M. Vogl
Institute of Health Economics and Health Care Management,
Helmholtz Zentrum München (GmbH), Ingolstädter Landstraße
1, 85764 Neuherberg, Germany
e-mail: bjoern.stollenwerk@helmholtz-muenchen.de

T. Welchowski
e-mail: welchow@imbie.meb.uni-bonn.de

M. Vogl
e-mail: matthias.vogl@helmholtz-muenchen.de

T. Welchowski
Institut für Medizinische Biometrie, Informatik und
Epidemiologie (IMBIE), Universitätsklinikum Bonn,
Sigmund-Freud-Straße 25, 53105 Bonn, Germany

S. Stock
Institute of Health Economics and Clinical Epidemiology,
University of Cologne, Gleueler Straße 176-178, 50935 Cologne,
Germany
e-mail: stephanie.stock@uk-koeln.de

Introduction

Cost-of-illness (COI) studies are a common type of economic study in the medical literature [1–4]. They are intended to measure either costs per patient or the total costs of a particular disease, including direct, indirect, and intangible costs. The rationale of COI studies has been stated, first, as indicating the potential savings that could be achieved if the target disease was abolished and, second, for prioritization purposes based on the total costs across diseases [2, 5]. Furthermore, findings from COI studies have been used to parameterize decision-analytic models for economic evaluation [6, 7]. However, the relevance of COI studies has also been questioned, as the total amount of expenditure would not reveal anything about how efficiently resources are used in a corresponding area [2, 8–10]. Moreover, preventing a disease is often associated

with a cost increase [2, 9, 11]. Nevertheless, COI studies provide an impression of the economic impact of a disease, and are used to strengthen the relevance of associated research [3].

The methods used for COI studies differ widely. It can be differentiated into incidence-based and prevalence-based approaches [3], top-down and bottom-up approaches [3, 12, 13], studies that sum up all the costs of prevalent subjects, studies that sum up only those costs that are target disease related, matched control studies, and studies based on regression models [1]. Prevalence-based COI studies are much more common than incidence-based approaches [1]. Whereas incidence-based approaches focus on the lifetime costs associated with an incident case, prevalence-based approaches in general calculate (annual) costs that could be avoided if the prevalence of the disease was set to zero [3].

COI regression approaches can also be divided into those in which costs are assumed to be normally distributed [i.e., ordinary least squares (OLS) regression] [1, 14, 15] and those based on right-skewed distributions [16–18]. Simple regression models based on the normal distribution lead to easily interpretable parameters [1, 14]. However, the assumption of normally distributed costs in general does not hold; the (right-skewed) gamma distribution fits much better [16, 19–21]. A disadvantage of the gamma distribution is that it does not allow for zero costs. Although this problem is often negligible and can be solved by adding a small constant [14, 22], two-part models could be used to explicitly model zero costs [14, 23, 24].

The size of the study population used to conduct COI studies differs widely. Based on a systematic review conducted in 2006, the sample size of 365 COI studies ranged from eight subjects to 1.8 million subjects [1]. However, as routine data become more and more available, even cohorts of more than 14 million insured people have been observed [25, 26].

Problems related to the usage of massive data go beyond COI studies. Preliminary work which technically aims at working with large datasets and efficiency has been established in comparable medical topics and other research areas such as astrophysics, data mining, or genetics: e.g., causal inferences on large datasets via propensity score matching [27], estimating cosmological parameters from large data sets [28], efficiency of partitioning a set of objects in databases into homogeneous groups or clusters [29], and computational efficiency of genotype imputation for large data sets [30]. In medicine, public health, health economics, and health services research massive data will be generated for outcome variables such as patient-reported outcome measures (PROMs) in the future [31]. Research on massive data analysis will thus be of pivotal importance [32, 33]. However, techniques to solve computational and statistical pitfalls are not yet developed to full extent. With our

prevalence-based, top-down regression approach we support massive data analysis.

Analyzing massive data is computationally intensive and requires further considerations when choosing a research design. Analyzing only a random sample would be associated with a loss of information and could be considered inefficient. However, traditional approaches may fail for numerical reasons when analyzing the data; for example, the numerical demand may increase exponentially with the number of observations [34, 35]. Furthermore, model assumptions of traditional approaches might be oversimplifying, as massive data allow much more precise estimates. For example, a linear or log-linear relationship between age and costs would not make efficient use of the data [36].

The objective of this study is, first, to present a new COI approach to analyze massive routine data, which can be categorized as a prevalence-based, top-down regression approach, accounting for right-skewed data. Second, we aim to assess this novel approach in terms of computational efficiency. This approach has not yet been discussed explicitly; however, it has already been applied in the context of breast cancer and coronary artery disease [4, 37].

Methods

The methodological approach

To perform COI studies based on massive data, which we define as data sets with a huge number of observations, we propose an approach that consists of the following steps: specification of the regression equation; data aggregation; estimating the regression model parameters; predicting costs for diseased and non-diseased subjects (recycled predictions); comparing costs between diseased and non-

The cost-of-illness data aggregation and regression approach

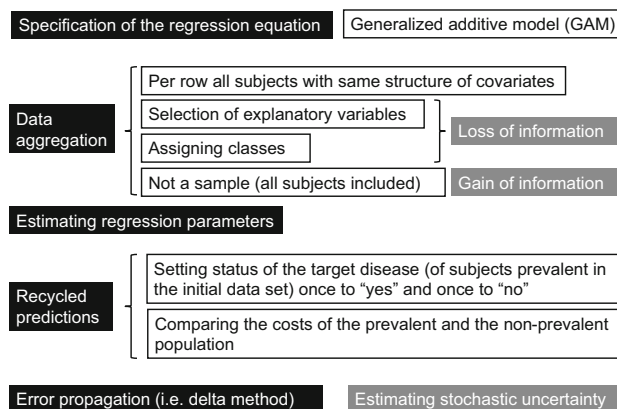


Fig. 1 Steps of the cost-of-illness data aggregation and regression approach

diseased subjects; and applying error propagation to construct uncertainty estimates of attributable cost estimates (Fig. 1).

Specification of the regression equation

As stated above, the assumption of a linear or a log-linear association between explanatory variables and costs may be too simplified, if massive data are available for quantification. A way to incorporate smooth terms into regression models is given by generalized additive models (GAMs) [36]. GAMs model smooth terms via splines, which are piecewise defined numeric functions. The purpose of splines is to model smooth curves without any a priori specification of a shape or a parametric structure. Splines are flexible and follow the course of the data points provided. Originally splines passed through all points provided. Regression splines are similar as a regression line in the linear model: they pass through a cloud of points, but not all points are lying on the line provided. Regression splines are fitted based on a penalty term, which avoids curves to wiggle too much. This means smoothness is balanced against a close fit to the data. Thin plate regression splines are a special and sophisticated type of splines, which fulfill some optimality criteria. In contrast to other types of splines, thin plate splines can smooth several variables simultaneously, e.g., they can provide a two-dimensional smoothing [38]. Just as in generalized linear models (GLMs), GAMs allow the response variable (i.e., the costs) to follow distributions other than the normal distribution [20, 39]. When cost data are used as the dependent variable, typically a gamma distribution and a log-link function are applied [16, 19–21].

Data aggregation

GAMs are numerically extensive, and fitting these might cause numerical problems with respect to massive data. Thus, we propose, first, to aggregate the massive data, in such a manner that the loss of relevant information is kept small and, second, to fit the GAM to the aggregated dataset. In the unaggregated data set we assume the columns to represent variables and the rows to present observations. In the aggregated dataset, the columns still represent the variables, but each row represents multiple subjects. Additionally to the original variables a count variable will be added to the aggregated data set, specifying how many subjects are represented by each row. Before aggregating the data, the explanatory variables have to be chosen and classified. For example, classifying the age into 1-year age groups keeps the loss of information negligibly small. After all chosen explanatory variables have been classified, one can determine the average costs for each unique

combination of covariates. Furthermore, if the explanatory variable has originally been measured on a continuous scale, one may choose to maintain the continuous scale by assigning a single value (e.g., the class mean) to each class of this variable.

One of the explanatory variables should correspond to whether the target disease is present (i.e., 1 = ‘diseased’, 0 = ‘not diseased’). In a prevalence-based approach, this enables us to use the predictions of our regression model for prevalent subjects also for the non-prevalent subjects in a second step (recycled predictions). Therefore, we set the prevalence status of diseased subjects to ‘not diseased’, when estimating attributable costs. When choosing the further explanatory variables, one has to ensure that none of them is causally affected by the target disease, as this could lead to biased results [3, 40]. In contrast, the main relevant confounders should be included [1]. As the number of rows in the aggregated dataset increases with the number of selected explanatory variables, and the number of categories per variable, the choice of both variables has to be made very carefully within regression approaches for massive data.

Estimating the model parameters

After aggregating the data, the regression model has to be fitted (i.e., a GAM with response variable costs, log-link function, and gamma-distributed response variable). The number of subjects represented by each row serves as the weight. Even though this model will be quite accurate in predicting costs, the parameters of this model will not be easily interpreted. This is due to applying a log-link function, incorporating smooth terms and potentially including interactions of explanatory variables into the regression equation. Models, which are too simple, given the available data, may yield easily interpretable parameters, but are imprecise and do not make an efficient use of the data. More sophisticated models, like the one we present, make efficient use of the data, but the parameters cannot be easily interpreted. Incorporating multiple smooth terms and interactions, like in our case, make it impossible to interpret the parameters directly. However, the study results do not necessarily need to be extracted from the regression parameters. In the current case, we present the results graphically: In the figure “Chronic lung disease attributable costs” we present results, which are easily to grasp but cover the complexity of the real data structure.

Recycled predictions

The measure of interest of a COI analysis is the disease attributable costs. Attributable costs have been calculated for the whole population (or a subpopulation), but also per prevalent subject. Disease attributable costs were estimated

by predicting costs via the fitted regression model, first for prevalent subjects and then for non-prevalent subjects. However, the first prediction corresponds to the rows in the aggregated data frame that refer to target disease-prevalent subjects. The second prediction is based on ‘recycled predictions’, i.e., the data frame of the first prediction is used, but the prevalence status is set to ‘no’ (i.e., 0 = ‘not diseased’). The sum of the first cost prediction corresponds to the costs of all prevalent subjects, and the sum of the second prediction corresponds to the hypothetical costs among the same subjects given that the target disease was eliminated. The difference between both sums corresponds to the disease attributable costs. The goal of the method is to estimate the costs that are due to a target disease. In the study population, however, many factors affect the costs. To calculate the costs that are due to the target disease, some adjustment has to take place. We know the population characteristics of the subjects who are diseased. The healthy population, however, has different characteristics. The method of “recycled predictions” is a method for controlling these population characteristics. A hypothetical population is generated that has the same population characteristics as the diseased population. The only difference between these two populations is the target disease status. By predicting for both of these populations the costs and by calculating the difference, the disease attributable costs can be estimated.

Error propagation

So far, the target disease attributable costs have been estimated per age group and gender. However, the stochastic uncertainty of this estimate still needs to be quantified. This can be done via Gauss’s error propagation law [37]. Basically, this is done using the standard error (SE) of each single cost prediction and their covariance matrix. These values are provided by standard statistical software [36, 41]. Also, the uncertainty of the prevalence of disease is considered. The single SEs, taking into account their covariance matrix, can be transformed into a SE estimate of the target parameter. Details regarding the application of Gauss’s error propagation law have already been described elsewhere at some length [37].

Application example of the cost-of-illness approach

To apply the COI method for massive data, we used data from four major German sickness funds [25, 26, 37, 42]. The data represent 7.3 million insured people; the base year was 1999. We approximated this number by dividing the observed days of insurance by 365 [4, 37]. Due to health insurance changes of few patients we cannot provide an

exact number for the insured people represented by the dataset. The variables of the dataset are the health expenditures (i.e., the total costs resulting from hospital stays, medication spending and sickness benefit; hospital costs reported include all costs for inpatient care, i.e., physician costs, medication costs, general costs for hospital stay and nursing care) per day of insurance, the age and gender of subjects and the prevalence of seven chronic diseases [i.e., hypertension, diabetes mellitus, heart failure (HF), coronary heart disease, breast cancer, stroke and chronic lung disease; chronic lung disease was defined as asthma and chronic obstructive pulmonary disease (COPD)]. In the application example, the variables included were given, and not explicitly selected for the purpose of this application. Costs were converted from ‘Deutsche Mark’ (DM) to euros (exchange rate 1 euro = 1.95583 DM). The dataset has already been highly aggregated (9,517 rows), and reports the overall days of insurance per unique combination of age, gender and comorbidities. It has been aggregated, for data protection purposes, in part by the sickness funds themselves and in part by the Institute of Health Economics and Clinical Epidemiology, University of Cologne. The data has originally been collected for a survey report to assess a forthcoming health care reform in Germany.

In our application example, we estimate the age- and gender-specific attributable costs of chronic lung disease. The response variable of the GAM is the average costs per day of insurance. A log-link function was applied. Smooth terms were the age and the interactions of age with gender and with each chronic disease. Thin plate regression splines were used to represent smooth terms [38]. Covariates of the regression models were age, gender, and the seven chronic diseases. Furthermore, pairwise interactions were included. For convergence reasons, no gamma distribution was assumed, but the quasi likelihood was maximized [36, 37]. The number of subjects represented by each row was used as weights within regression analysis.

Attributable costs for the whole population of Germany in 1999 were derived by multiplying age- and gender-specific costs by the corresponding population size, as supplied by the German Federal Statistical Office [43], and by the age- and gender-specific prevalence of chronic lung disease. All confidence intervals were calculated based on Gauss’s error propagation law.

Whereas fitting the GAM is only one step in the method we propose, the figure of the excess costs is the main output, i.e., the result which method produces. The fitted GAM is used to predict costs for a population with vs. a population without the disease. Subsequently, the difference between these two cost predictions is calculated (in our case separately for each age-gender-group). This difference is the output of the method, which we displayed graphically.

Simulation study to assess computational efficiency

To assess whether the gain in computational efficiency resulting from aggregation is of practical relevance, we performed a simulation study. As the dataset used for the application example is available only in an aggregated form, we first simulated an individual-level dataset from the original dataset. The number of subjects for each unique combination of age, gender and comorbidities (i.e., for each row of the aggregated data frame) was estimated based on the observed number of days of insurance. This was done by dividing the number of days of insurance by 365 and taking the ceiling of this value. Individual costs were simulated via the gamma distribution. The parameters of the gamma distribution were approximated based on the standard deviance and the expected value. The average cost estimates (i.e., the expected value used for gamma simulation) were predicted via the fitted GAM. Furthermore, SEs of the predicted cost estimates were supplied by the GAM. However, as the standard deviation (SD) required for simulating costs on an individual level exceeds the SE (the SE refers to the precision of the expected value), a further assumption was required. For simplicity, as more detailed information was missing, the relationship between the SE and the SD in the context of a single sample's mean was applied (see one sample *t* test) [44]. In other words, the SD was approximated by multiplying the SE by the square root of the number of subjects represented by the average estimate. Based on these assumptions, an individual patient-level data set representing 7.3 million subjects was simulated.

The goal was to assess to what extent the gain in numerical efficiency via the aggregation process is of practical relevance. For this purpose, we randomly drew (with replacement) datasets from the simulated individual-level data frame with a total number of observations ranging from 100,000 to 7 million subjects. The grid of the sample sizes was narrower in the range from 100,000 observations to 1 million observations (i.e., increments of 100,000 observations) and wider in the range from 1 to 7 million observations (i.e., increments of 1 million observations). For each given sample size, 10 data frames were randomly sampled. To each of the sampled data frames, a regression model was fitted with the same regression equation as in the application example. The time needed to fit each regression model was measured. Furthermore, as fitting the regression model sometimes failed for numerical reasons, we performed a logistic generalized additive regression with the probability of failure as the response and the sample size (smooth term, thin plate regression splines) as the explanatory variable. For analysis, we used a desktop personal computer (PC), Windows 7 (64-bit) with 8 processor kernels and 16 gigabytes (GB) of RAM. Furthermore, the statistical software R (version 3.1.0) was

used, including the R-package 'mgcv' (version 1.7-29) to fit GAMs [36, 41]. Even though we applied the R function 'gam' to fit the aggregated data regression model (9,517 rows), we applied the function 'bam' to the individual-level datasets, as 'bam' is computationally faster for big data (bam = 'GAMs for very large datasets'). However, 'bam' has not been applied to the aggregated dataset, as the algorithm is not stable for datasets with 'few' observations.

Results

Cost-of-illness of chronic lung disease in Germany

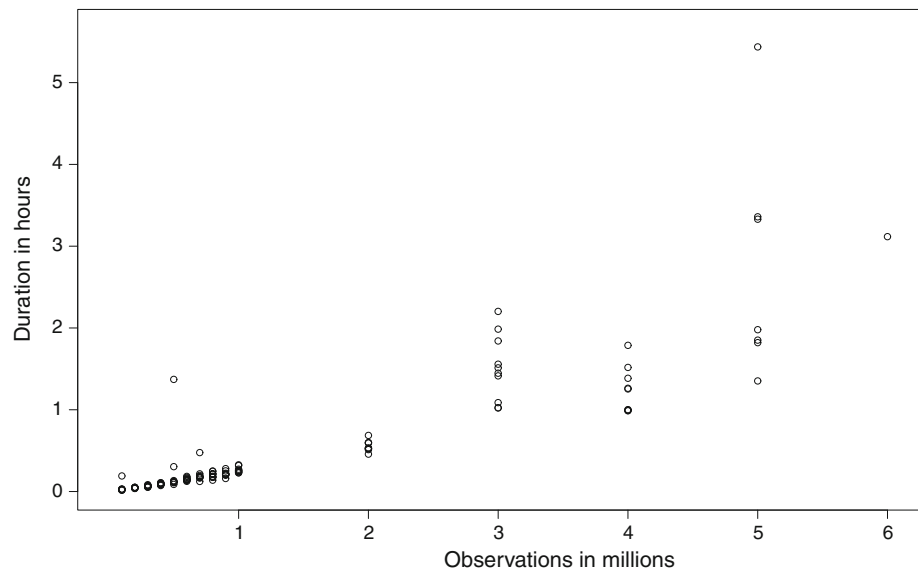
The age- and gender-specific costs of COPD and asthma-related chronic lung disease are displayed in Fig. 2. The coefficients of the regression model are presented in Table 1. The course of attributable costs differs significantly from a linear or a log-linear relationship. The stochastic uncertainty is reasonably small, which can be explained by the huge amount of data incorporated into the analysis. There are four peaks in attributable costs of COPD/asthma: in the first years after birth, in the early 20 s, in the 60 s, and in the 80 s. Taking into account the age and gender distribution of Germany from 1999, the quantified annual costs of chronic lung disease amount to 6.78 million euros (SE = 0.01 million euros).

A simple linear or log-linear relationship does not capture the course of the attributable costs. This can be concluded (a) from the significance values of the smooth terms in the regression model: the smooth terms are highly significant; and (b) from the figure illustrating the "chronic lung disease attributable costs": this figure includes very narrow confidence bands. A straight line or a line which is straight on the log scale would not stay within the confidence bands at all.

Computational performance of the COI method for massive data

The computational time for fitting the aggregated data regression model was 6.2 min plus 12.5 min for calculating the point estimates and applying the error propagation approach. The computational time required for fitting the individual-level data regression model increased disproportionately more by sample size (Fig. 3). Fluctuation of time used increased with the size of the dataset, as more runs fail with an increasing number of observations. For 5 million observations, it amounted on average to 2.8 h. The increase in computational time was accompanied by an increasing failure probability (Fig. 4). For 3 million observations, one in ten regression approaches failed, for 5 million observations, three approaches failed, for 6 million

Fig. 2 Computational time needed to fit the person-level data regression model (given the model fitting did not fail)



observations, nine approaches failed, and for 7 million observations, all ten approaches failed.

Discussion

In this paper, we have presented a COI method for massive data, and applied it in the context of chronic lung disease. Furthermore, we have assessed the practical relevance of the computational efficiency gain of this method. Up to now, no method has yet been presented for COI studies based on massive data. The smooth terms included in the regression models were highly significant, and the age-specific costs per gender group differed significantly from a simple model structure that would have been achieved via, for example, a linear or a log-linear relationship. The average attributable costs of men and women have different shapes and are even crossing. This completely different curve is significant, as we can conclude from the narrow confidence bands (Fig. 3). Furthermore, applying extensive GAMs to massive data was associated with a disproportionately high time increase, and even more problematic, with a high probability of computational failure. Thus, applying current COI methods to massive data would be oversimplifying, would lead to computational problems or would require the drawing of a subsample.

With respect to the application example, the attributable costs per patient were similar to previous studies, considering differences in severity of stage, included cost components and differences in health system [45–50]. The peak in younger years is due to asthma and corresponds with the prevalence of asthma in children and adolescents. There are few studies regarding the prevalence of COPD. One is the international BOLD study (burden of obstructive lung

disease), which assessed the prevalence of COPD after the age of 40. The results of the cost distribution in our study correspond with the prevalence distribution of COPD in the BOLD study with men having a higher prevalence compared with women and the highest prevalence between the ages of 65 and 84 years [51]. The decrease in attributable costs between the ages of 60 and 70 years can be explained by the consideration of sickness benefit and by the fact that many people retire within this age range.

The approach that we present is a prevalence-based approach. Even though prevalence-based approaches have been stated to be most suitable for assessing the current economic burden of a disease [3], it might be desirable to use massive data to conduct incidence-based analyses. However, this would require longitudinal data, and the transferability of the principle of fitting GAMs to aggregated data in the context of incidence-based COI studies still needs to be shown.

Furthermore, in our application example, we referred to health expenditures from a sickness fund perspective. However, a complete COI study includes the overall direct, indirect, and intangible costs [1, 3]. Indirect costs such as loss of earnings are partially not covered by our approach from a sickness fund perspective and intangible costs such as loss in quality of life are not covered at all. If, as in our case, the data cover costs only partially, further data sources have to be analyzed to quantify the overall costs associated with a disease.

A limitation of the presented approach is that data might get lost due to aggregation. In our particular case, this has not been a problem, as the individual-level data have already been classified, and 1-year age groups have been used. However, the big advantage of not losing participants due to aggregation is reduced by classifying (e.g.,

Table 1 Estimates of the fitted generalized additive model to predict health expenditures

	Parameter estimate	Standard error	<i>t</i> statistic	<i>p</i> value
Intercept	0.72	0.01	114.1	<0.001
Female	0.11	0.16	0.7	0.499
Diabetes	4.82	0.67	7.2	<0.001
CHD	2.63	0.90	2.9	0.003
Hypertension	0.00 [§]	0.00 [§]	NA	NA
Asthma	0.00 [§]	0.00 [§]	NA	NA
HF	-2.30	0.60	-3.8	<0.001
BC	0.00 [§]	0.00 [§]	NA	NA
Stroke	0.00 [§]	0.00 [§]	NA	NA
Female × diabetes	0.05	0.01	4.4	<0.001
Female × CHD	-0.08	0.01	-8	<0.001
Female × hypertension	-0.09	0.01	-7.8	<0.001
Female × asthma	-0.07	0.01	-6.7	<0.001
Female × HF	-0.10	0.01	-8.7	<0.001
Female × BC	0.27	0.14	1.9	0.054
Female × stroke	0.15	0.01	10.1	<0.001
Diabetes × CHD	-0.06	0.01	-5.4	<0.001
Diabetes × hypertension	-0.44	0.02	-24.8	<0.001
Diabetes × asthma	-0.14	0.01	-10.1	<0.001
Diabetes × HF	-0.04	0.02	-2.7	0.007
Diabetes × BC	-0.47	0.04	-11.5	<0.001
Diabetes × stroke	-0.23	0.02	-14.1	<0.001
CHD × hypertension	-0.10	0.02	-6.1	<0.001
CHD × asthma	-0.14	0.01	-11.3	<0.001
CHD × HF	0.00	0.01	0.1	0.914
CHD × breast cancer	-0.37	0.04	-9.4	<0.001
CHD × stroke	-0.29	0.02	-19.0	<0.001
Hypertension × asthma	-0.34	0.02	-21.1	<0.001
Hypertension × HF	-0.11	0.02	-5.5	<0.001
Hypertension × BC	-0.40	0.04	-11.2	<0.001
Hypertension × stroke	-0.42	0.02	-17.6	<0.001
Asthma × HF	-0.04	0.02	-2.3	0.022
Asthma × BC	-0.35	0.04	-9.8	<0.001
Asthma × stroke	-0.25	0.02	-12.4	<0.001
HF × BC	-0.15	0.04	-4.0	<0.001
HF × stroke	-0.27	0.02	-12.5	<0.001
BC × stroke	-0.87	0.06	-13.6	<0.001
	Estimated <i>df</i>	Estimated rank	<i>F</i>	<i>p</i> value
<i>Approximate significance of smooth terms</i>				
Age	9.0	9	1,035	<0.001
Age × female	8.9	9	187	<0.001
Age × diabetes	8.9	9	203	<0.001
Age × CHD	7.4	8	110	<0.001
Age × hypertension	10.0	10	593	<0.001
Age × asthma	9.9	10	866	<0.001
Age × HF	8.4	9	34	<0.001
Age × BC	6.7	8	100	<0.001
Age × stroke	9.6	10	1,014	<0.001

CHD coronary heart disease, HF heart failure, BC breast cancer, NA not available, *df* degrees of freedom

[§] Parameters were set to zero for identification reasons (identification problems, i.e., columns in the design matrix that could be represented as linear transformation of the remaining columns, occurred due to the various interactions included into the regression model)

Fig. 3 Chronic lung disease attributable costs (i.e., excess costs) in 1999 euros

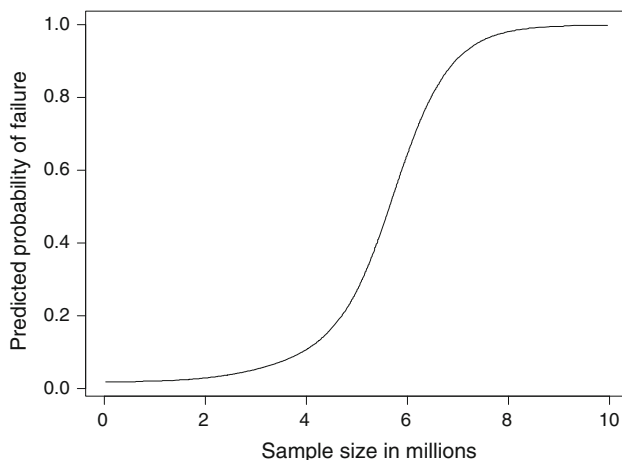
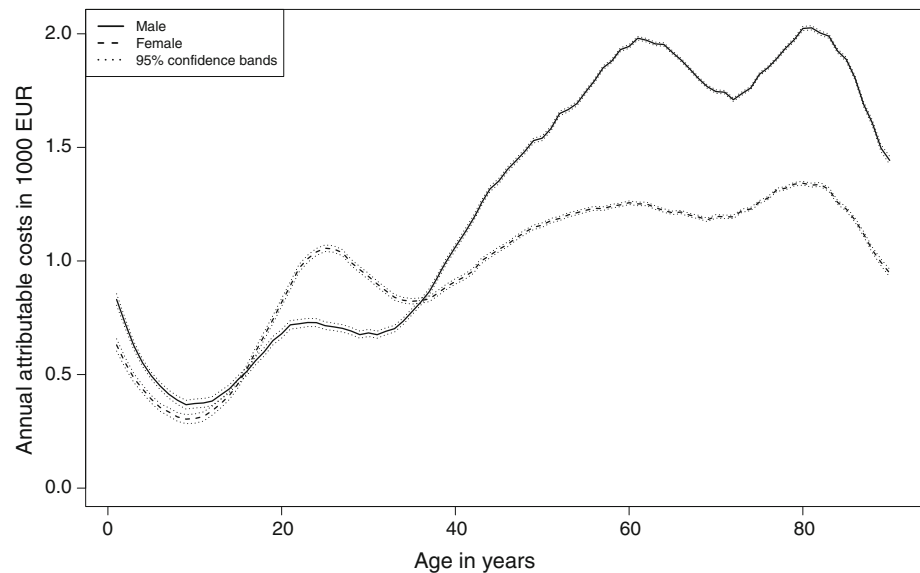


Fig. 4 Probability of failure for fitting the individual-level data regression model

information gets lost when transforming continuous variables into ordinal variables) and choosing variables for aggregation. Furthermore, it might be questioned whether it is really necessary to include all individuals, or whether performing, for example, a propensity score matching would not lead to sufficient results.

In our methodological approach, we chose GAMs as regression models, which are based on GLMs. We made this choice because GAMs include smooth terms. However, there is a wide range of approaches that can be used for skewed outcome data. Manning and Mullahy, for example, compared log models, GLMs and OLS models with log-transformed response variables, but none of these options has been found to be best under all conditions [52]. Furthermore, the generalized gamma (GMM) distribution was found to yield potentially more robust results than

GLMs with gamma distribution [53, 54]. However, coefficients of log-link GAMs with smooth terms cannot be interpreted. They are meant to be used for cost prediction. Without smooth terms, exponentiated coefficients could be interpreted as relative changes compared to the base category.

Finally, in the R package “mgcv” the optimization is based on direct optimization and on an approximation of the penalized likelihood function (i.e., backfitting via the Fisher scoring). This results in a high dimensional system of equations. When the Fisher information matrix needs to be calculated based on huge matrices, numerical problems might result from the coordination of the numeric derivatives. Furthermore, this process is based on inverting matrices. As there are a high number of predictors, this may yield numerical instabilities.

Regarding the application example, there are several potential sources of bias that should be mentioned. First, the data may not be representative for Germany, as the insured subjects only represent the four corresponding sickness funds. Second, there might be some bias due to the methods used to classify the prevalence status of the seven chronic diseases [4, 37]. In consequence, the supplied standard errors only represent the stochastic uncertainty. The true uncertainty, due to the presence of bias, is probably reasonably larger. However, this is a limitation that holds for all COI studies, which should not preclude efforts to optimize the precision of cost estimates. Third, in the current approach, the ICD-9 code 493 and the ACT code RO3 were used to define the population of subjects with COPD and asthma, which probably underestimates the prevalent subjects with chronic lung disease and thus also the overall costs. Furthermore, several cost components

were not included in the analysis. In particular, physician costs, medical advice, and indirect costs were not included.

In future analyses, the impact of the number of categories per variable and the number of variables itself on the possibility of numerical failure and the length of computational time should be analyzed to get a broader view on abilities and limitations of regression approaches for COI analysis with massive data. In conclusion, we provide an innovative method to conduct COI studies based on massive data. This method may yield more precise cost estimates and also improves computational efficiency.

Acknowledgments Financial support for this study was provided by the Helmholtz Zentrum München, German Research Centre for Environmental Health (HMGU) and the Institute of Health Economics and Clinical Epidemiology, University of Cologne, Germany. The funding agreement enabled the authors to design the study, interpret the data, and write and publish the manuscript. The following authors are employed by the sponsors: Björn Stollenwerk (HMGU), Stephanie Stock (University of Cologne). We thank Heather Hynd for proofreading the manuscript.

References

- Akobundu, E., Ju, J., Blatt, L., Mullins, C.D.: Cost-of-illness studies: a review of current methods. *Pharmacoeconomics* **24**(9), 869–890 (2006)
- Byford, S., Torgerson, D.J., Raftery, J.: Economic note: cost of illness studies. *BMJ* **320**(7245), 1335 (2000)
- Larg, A., Moss, J.R.: Cost-of-illness studies: a guide to critical evaluation. *Pharmacoeconomics* **29**(8), 653–671 (2011)
- Gruber, E.V., Stock, S., Stollenwerk, B.: Breast cancer attributable costs in Germany: a top-down approach based on sickness funds data. *PLoS One* **7**(12), e51312 (2012)
- Ament, A., Evers, S.: Cost of illness studies in health care: a comparison of two cases. *Health Policy* **26**(1), 29–42 (1993)
- Stollenwerk, B., Gandjour, A., Lungen, M., Siebert, U.: Accounting for increased non-target-disease-specific mortality in decision-analytic screening models for economic evaluation. *Eur. J. Health Econ.* (2012). doi:10.1007/s10198-012-0454-z
- Stollenwerk, B., Gerber, A., Lauterbach, K.W., Siebert, U.: The German coronary artery disease risk screening model: development, validation, and application of a decision-analytic model for coronary artery disease prevention with statins. *Med. Decis. Making* **29**(5), 619–633 (2009)
- Shiell, A., Gerard, K., Donaldson, C.: Cost of illness studies: an aid to decision-making? *Health Policy* **8**, 317–323 (1987)
- Wiseman, V., Mooney, G.: Burden of illness estimates for priority setting: a debate revisited. *Health Policy* **43**(3), 243–251 (1998)
- Reuter, P.: What drug policies cost: estimating government drug policy expenditures. *Addiction* **101**(3), 315–322 (2006)
- Shenoy, A.U., Aljutaifi, M., Stollenwerk, B.: Limited economic evidence of carotid artery stenosis diagnosis and treatment: a systematic review. *Eur. J. Vasc. Endovasc. Surg.* **44**(5), 505–513 (2012)
- Liu, J.L., Maniadakis, N., Gray, A., Rayner, M.: The economic burden of coronary heart disease in the UK. *Heart* **88**(6), 597–603 (2002)
- Hodgson, T.A., Meiners, M.R.: Cost-of-illness methodology: a guide to current practices and procedures. *Milbank. Mem. Fund. Q. Health Soc.* **60**(3), 429–462 (1982)
- Andersen, C.K., Andersen, K., Kragh-Sorensen, P.: Cost function estimation: the choice of a model to apply to dementia. *Health Econ.* **9**(5), 397–409 (2000)
- Andersen, C.K., Lauridsen, J., Andersen, K., Kragh-Sorensen, P.: Cost of dementia: impact of disease progression estimated in longitudinal data. *Scand. J. Public Health* **31**(2), 119–125 (2003)
- Maetzel, A., Li, L.C., Pencharz, J., Tomlinson, G., Bombardier, C.: The economic burden associated with osteoarthritis, rheumatoid arthritis, and hypertension: a comparative study. *Ann. Rheum. Dis.* **63**(4), 395–401 (2004)
- Penberthy, L.T., Towne, A., Garnett, L.K., Perlin, J.B., DeLorenzo, R.J.: Estimating the economic burden of status epilepticus to the health care system. *Seizure* **14**(1), 46–51 (2005)
- Perencevich, E.N., Sands, K.E., Cosgrove, S.E., Guadagnoli, E., Meara, E., Platt, R.: Health and economic impact of surgical site infections diagnosed after hospital discharge. *Emerg. Infect. Dis.* **9**(2), 196–203 (2003)
- Bassi, A., Dodd, S., Williamson, P., Bodger, K.: Cost of illness of inflammatory bowel disease in the UK: a single centre retrospective study. *Gut* **53**(10), 1471–1478 (2004)
- Dobson, A.J.: An introduction to generalized linear models. Chapman and Hall/CRC, London (2002)
- Wenig, C.M.: The impact of BMI on direct costs in children and adolescents: empirical findings for the German Healthcare System based on the KiGGS-study. *Eur. J. Health Econ.* **13**(1), 39–50 (2012)
- van Ruitten- Molken, M.P., van Doorslaer, E.K., van Vliet, R.C.: Statistical analysis of cost outcomes in a randomized controlled clinical trial. *Health Econ.* **3**(5), 333–345 (1994)
- Menn, P., Heinrich, J., Huber, R.M., Jorres, R.A., John, J., Karasch, S., Peters, A., Schulz, H., Holle, R.: Direct medical costs of COPD: an excess cost approach based on two population-based studies. *Respir. Med.* **106**(4), 540–548 (2012)
- Mihaylova, B., Briggs, A., O'Hagan, A., Thompson, S.G.: Review of statistical methods for analysing healthcare resources and costs. *Health Econ.* **20**(8), 897–916 (2010)
- Stock, S., Redaelli, M., Luengen, M., Wendland, G., Civello, D., Lauterbach, K.W.: Asthma: prevalence and cost of illness. *Eur. Respir. J.* **25**(1), 47–53 (2005)
- Stock, S.A., Redaelli, M., Wendland, G., Civello, D., Lauterbach, K.W.: Diabetes—prevalence and cost of illness in Germany: a study evaluating data from the statutory health insurance in Germany. *Diabet. Med.* **23**(3), 299–305 (2006)
- Rubin, D.B.: Estimating causal effects from large data sets using propensity scores. *Ann. Intern. Med.* **127**(8 Pt 2), 757–763 (1997)
- Tegmark, M., Taylor, A.N., Heavens, A.F.: Karhunen-Loève eigenvalue problems in cosmology: how should we tackle large data sets? *Astrophys. J.* **480**(1), 22–35 (1997)
- Huang, Z.: Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Min. Knowl. Disc.* **2**(3), 283–304 (1998)
- Browning, B.L., Browning, S.R.: A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**(2), 210–223 (2009). doi:10.1016/j.ajhg.2009.01.005
- Department of Health: Guidance on the Routine Collection of Patient Reported Outcome Measures (PROMs). http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_092625.pdf (2010). Accessed 12 Jan 2014
- Murdoch, T.B., Detsky, A.S.: The inevitable application of big data to health care. *JAMA* **309**(13), 1351–1352 (2013). doi:10.1001/jama.2013.393
- Schneeweiss, S.: Learning from big health care data. *N. Engl. J. Med.* **370**(23), 2161–2163 (2014). doi:10.1056/NEJMp1401111

34. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R.: *Advances in knowledge discovery and data mining*. MIT Press, Cambridge (1996)
35. Witten, I.H., Frank, E., Hall, M.A.: *Data mining: practical machine learning tools and techniques*, 3rd edn. Morgan Kaufmann Publishers/Elsevier, Burlington, MA (2011)
36. Wood, S.N.: *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, London (2006)
37. Stollenwerk, B., Stock, S., Siebert, U., Lauterbach, K.W., Holle, R.: Uncertainty assessment of input parameters for economic evaluation: Gauss's error propagation, an alternative to established methods. *Med. Decis. Making* **30**(3), 304–313 (2010)
38. Wood, S.N.: Thin plate regression splines. *J. R. Stat. Soc. B* **65**(1), 95–114 (2003)
39. Blough, D.K., Madden, C.W., Hornbrook, M.C.: Modeling risk using generalized linear models. *J. Health Econ.* **18**(2), 153–171 (1999)
40. Mullahy, J.: Econometric modeling of health care costs and expenditures: a survey of analytical issues and related policy considerations. *Med. Care.* **47**(7 Suppl 1), S104–S108 (2009)
41. R Development Core Team: *R: a language and environment for statistical computing*. In: *R Foundation for Statistical Computing*, Vienna (2012)
42. Stock, S.A., Stollenwerk, B., Redaelli, M., Civello, D., Lauterbach, K.W.: Sex differences in treatment patterns of six chronic diseases: an analysis from the German statutory health insurance. *J. Womens Health (Larchmt)* **17**(3), 343–354 (2008)
43. Statistisches Bundesamt: *Bevölkerung Deutschlands bis 2060: 12. koordinierte Bevölkerungsvorausberechnung*. DESTATIS, Wiesbaden (2009)
44. Zolman, J.F.: *Biostatistics*. Oxford University Press, Oxford (1993)
45. Miravittles, M., Murio, C., Guerrero, T., Gisbert, R.: Costs of chronic bronchitis and COPD: a 1-year follow-up study. *Chest* **123**(3), 784–791 (2003)
46. van Rutten- Molken, M.P., Postma, M.J., Joore, M.A., Van Genugten, M.L., Leidl, R., Jager, J.C.: Current and future medical costs of asthma and chronic obstructive pulmonary disease in the Netherlands. *Respir. Med.* **93**(11), 779–787 (1999)
47. Nielsen, R., Johannessen, A., Omenaas, E.R., Bakke, P.S., Askildsen, J.E., Gulsvik, A.: Excessive costs of COPD in ever-smokers: a longitudinal community study. *Respir. Med.* **105**(3), 485–493 (2011)
48. Koleva, D., Motterlini, N., Banfi, P., Garattini, L.: Healthcare costs of COPD in Italian referral centres: a prospective study. *Respir. Med.* **101**(11), 2312–2320 (2007)
49. van Rutten- Molken, M.P., Feenstra, T.L.: The burden of asthma and chronic obstructive pulmonary disease: data from the Netherlands. *Pharmacoeconomics* **19**(Suppl 2), 1–6 (2001)
50. Ungar, W.J., Coyte, P.C., Chapman, K.R., MacKeigan, L.: The patient level cost of asthma in adults in south central Ontario. *Pharmacy Medication Monitoring Program Advisory Board. Can. Respir. J.* **5**(6), 463–471 (1998)
51. Buist, A.S., McBurnie, M.A., Vollmer, W.M., Gillespie, S., Burney, P., Mannino, D.M., Menezes, A.M., Sullivan, S.D., Lee, T.A., Weiss, K.B., Jensen, R.L., Marks, G.B., Gulsvik, A., Nizankowska-Mogilnicka, E.: International variation in the prevalence of COPD (the BOLD Study): a population-based prevalence study. *Lancet* **370**(9589), 741–750 (2007)
52. Manning, W.G., Mullahy, J.: Estimating log models: to transform or not to transform? *J. Health Econ.* **20**(4), 461–494 (2001)
53. Basu, A., Rathouz, P.J.: Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics* **6**(1), 93–109 (2005)
54. Manning, W.G., Basu, A., Mullahy, J.: Generalized modeling approaches to risk adjustment of skewed outcomes data. *J. Health Econ.* **24**(3), 465–488 (2005)