Health Services Research and Managerial Epidemiology
Volume 6: 1-10

DOI: 10.1177/2333392819891005
journals.sagepub.com/home/hme
SAGE

# Estimating Health Cost Repartition Among Diseases in the Presence of Multimorbidity

**Valentin Rousson, PhD[1], Jean-Benoît Rossel, PhD[1], and Yves Eggli, MD, MA, PhD[1]**

**Abstract**

We consider the nontrivial problem of estimating the health cost repartition among different diseases in the common case where the patients may have multiple diseases. To tackle this problem, we propose to use an iterative proportional repartition (IPR) algorithm, a nonparametric method which is simple to understand and to implement, allowing (among other) to avoid negative cost estimates and to retrieve the total health cost by summing up the estimated costs of the different diseases. This method is illustrated with health costs data from Switzerland and is compared in a simulation study with other methods such as linear regression and general linear models. In the case of an additive model without interactions between disease costs, a situation where the truth is clearly defined such that the methods can be compared on an objective basis, the IPR algorithm clearly outperformed the other methods with respect to efficiency of estimation in all the settings considered. In the presence of interactions, the situation is more complex and will deserve further investigation.

**Keywords**

general linear models, health costs, interactions, iterative proportional repartition, linear regression, multimorbidity

## Introduction

Estimating the costs, whether direct or indirect, generated by the different diseases is a recurrent and nontrivial problem that one retrieves at various levels of a health-care system. It is important for choosing the most cost-effective health services. At country level, the distribution of costs by disease is necessary to set priorities, to calibrate prevention programs,[1] to prevent the selection of risks by health insurances,[2] or to understand the causes of an increase in costs, for example.[3] At the level of health-care providers, it allows to compare medical practices, evaluate new technologies,[4] and better control expenses.

One complicate issue is that many patients have multiple diseases at the same time. For instance, 3 quarters of people older than 80 years had sequelae of more than 5 diseases in 2013 in developed countries.[5] It is therefore essential to spread health costs—whether monetary units or years of healthy life lost—among all those diseases.[6,7] Some costs are specific to a disease, for example for drugs, a positive screening test, a tumor biopsy, or a fracture X-ray. Costs are however often difficult to allocate: 1 day of hospitalization can be justified by several diseases at the same time, a consultation with a general practitioner can be due to several simultaneous conditions. Another issue is to ensure that the sum of the costs per diseases per patient gives back the total health costs,[8] avoiding double counting of certain expenses.[9-11]

To achieve a cost repartition among diseases present in a same patient, some authors compare the costs of patients having that disease with those who do not.[7,12] However, one will not retrieve the total health costs using that approach in the presence of interactions, that is, when the cost generated by the simultaneous presence of 2 diseases is on average different from (being superior or inferior to) the sum of the costs of the 2 diseases when present individually. Depending on the method chosen to circumvent this issue, the results may vary a lot.[13] Skirting the difficulty, another option consists in reduce complexity using a classification of patients with only one disease. This option is commonly used with diagnosis-related groups that categorize patients according to a major pathology (or

[1] Center for Primary Care and Public Health (Unisanté), University of Lausanne, Lausanne, Switzerland

Submitted October 25, 2019. Accepted October 25, 2019.

**Corresponding Author:**
Jean-Benoît Rossel, Center for Primary Care and Public Health (Unisanté), University of Lausanne, Lausanne, Switzerland.
Email: jean-benoit.rossel@unisante.ch

operation), possibly creating subcategories to account for the presence or absence of significant comorbidities.[14] The same solution is applied when years of life lost due to premature death are related to a unique "underlying cause of death."[15] This may be acceptable in some simple situations such as cost of hospitalizations related to childbirth, hip prosthesis or appendicitis, or for deaths due to road accidents. Complex situations are nevertheless frequent, especially in developed countries with an increasing life expectancy. Since most of the costs occur among elderly patients with multiple conditions, there is a high risk of underestimating the costs of concomitant diseases, which often appear as comorbidities.[16-19] Finally, various methods of regression could be considered, such as ordinary least squares (OLS) multiple regression or generalized linear models (GLM), which are discussed below.

Alternatively, an iterative proportional repartition (IPR) algorithm has been applied some years ago to share health costs by illness in Switzerland.[20] This nonparametric method could be interesting because it is simple to implement and to understand, although its statistical properties (such as unbiasedness and stability) and a comparison with other methods have not been investigated so far in the literature. The aim of the present article is to fill this gap and to propose and motivate the use of IPR to solve the above-mentioned problem.

Our article is organized as follows. In section 2, we formally introduce the problem. In section 3, we present and discuss different candidate methods to solve that problem, including IPR. In section 4, we apply these methods to health costs data from Switzerland. In section 5, we compare the statistical performances of these methods based on simulated data. Section 6 concludes.

## The Problem

We consider $n$ patients and $p$ potential diseases (eg, $n = 500\,000$ patients and $p = 50$ potential diseases). Let $Y_i$ the (global) health cost spent for patient $i$ (eg, in a given year or during a specific hospital stay, depending on the context considered and of data availability, for $i = 1, \ldots, n$, all these costs being strictly positive). Let $X_{ij}$ a binary variable indicating whether disease $j$ was diagnosed for patient $i$, such that $X_{ij} = 1$ if disease $j$ was diagnosed for patient $i$, and $X_{ij} = 0$ otherwise (for $i = 1, \ldots, n$ and $j = 1, \ldots, p$). In this article, we consider the usual case where all patients have at least one disease, and where a patient can (and generally will) have several diseases (eg, up to 20 diseases). Our first goal is to estimate the mean cost $\mu_j$ of disease $j$, defined over all the patients having that disease, for each disease $j = 1, \ldots, p$. From there, one can obtain the total cost spent for disease $j$ (calculated over the patients in our sample) as $\tau_j = \sum_{i=1}^{n}(X_{ij}\,\mu_j)$, as well as the percentage of the total health cost spent for that disease as $\pi_j = \tau_j / \sum_{k=1}^{p}\tau_k$ (for $j = 1, \ldots, p$), in what follows the cost contribution of disease $j$, such that $\sum_{j=1}^{p}\pi_j = 1$. In other words, we get a health cost repartition among the different diseases, which is our ultimate goal.

The natural estimate of $\mu_j$ would be $\tilde{\mu}_j = \sum_{i=1}^{n}(X_{ij}Y_{ij}) / \sum_{i=1}^{n}X_{ij}$, where $Y_{ij}$ would be the specific cost of disease $j$ for patient $i$ (for $i = 1, \ldots, n$ and $j = 1, \ldots, p$). The problem is that we only know the global health costs $Y_i = \sum_{j=1}^{p}(X_{ij}Y_{ij})$, not the detail of the $Y_{ij}$. In the next section, we present different methods to get estimates $\hat{\mu}_j$ of the $\mu_j$. For each method, estimates of the $\tau_j$ can then be obtained as $\hat{\tau}_j = \sum_{i=1}^{n}(X_{ij}\,\hat{\mu}_j)$, and estimates of the $\pi_j$ as $\hat{\pi}_j = \hat{\tau}_j / \sum_{k=1}^{p}\hat{\tau}_k$ (for $j = 1, \ldots, p$), such that $\sum_{j=1}^{p}\hat{\pi}_j = 1$.

## Methods

### Linear Regression

One idea to solve this problem would be to consider a regression model with $Y_i$ as the response variable and with the $X_{ij}$ as $p$ binary predictors. In such a model, one decomposes the cost $Y_i$ for individual $i$ as a sum of 2 terms. The first term is the mean cost of those individuals sharing the same diseases (in what follows the same "disease pattern", ie, having the same values of predictors $X_{ij}$) as that patient. The second term is the difference between the cost of a given patient and the mean cost within his or her disease pattern, called a residual. The notation is as follows:

$$Y_i = \text{mean}(Y_i|X_{ij}) + \varepsilon_i. \qquad (1)$$

Equation (1) in its general form is always true since one can always write $A = B + (A - B)$. However, a model should be assumed to describe how the first term (the mean cost) depends on the predictors (the disease pattern), and a probability distribution is needed to describe the second term (which will be typically different within each disease pattern). In a linear model, one assumes that:

$$\text{mean}(Y_i|X_{ij}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}. \qquad (2)$$

A natural option (which we are following throughout this article) would be to remove the "intercept" $\beta_0$ from "model (2)" since a patient without any disease (ie, with $X_{ij} = 0$ for $j = 1, \ldots, p$) would not have any cost. Without intercept, the mean cost of those patients having only disease $j$ is given by $\beta_j$, whereas the mean cost of those patients having more than one disease is obtained by summing up the corresponding $\beta_j$. Under that model, our problem is thus solved by estimating $\mu_j = \beta_j$ for $j = 1, \ldots, p$.

Importantly, the coefficients $\beta_j$ in model (2) can be consistently estimated (ie, without bias and converging to the correct values with an increasing sample size) via estimates $\hat{\beta}_j$ provided by OLS regression, whatever the distribution of the residuals $\varepsilon_i$,[21] yielding consistent estimates $\hat{\mu}_j = \hat{\beta}_j$ of $\mu_j$. Estimation is optimal, however, only if the $\varepsilon_i$ are normally distributed and with the same variance within each disease pattern (homoscedasticity). In practice, this will never be the case, the $\varepsilon_i$ being typically skewed, including outliers and having different variances (heteroscedasticity), and actually quite different distributions in the different disease patterns. This may call for alternative methods to improve estimation. Other

drawbacks of OLS are that it is possible to get negative estimates of $\hat{\mu}_j = \hat{\beta}_j$, and that one does not exactly retrieve the total health cost spent for all the patients by summing up the estimated costs spent for the different diseases, that is, $\sum_{j=1}^{p} \hat{\tau}_j \neq \sum_{i=1}^{n} Y_i$. Let us also recall that classical OLS standard errors might not be valid in the presence of heteroscedasticity. Standard errors can be here obtained via bootstrap, as for the other methods considered in what follows.

Note finally that model (2), without an intercept, describes an "additive" reality. For example, if a first disease alone would cost on average 10 000 CHF and a second disease alone would cost on average 100 CHF, having the 2 diseases simultaneously would cost on average 10 100 CHF. This does not appear quite unreasonable, at least as a first approximation. If not the case, one may consider including interactions in model (2), although with many diseases and interactions, the model might become difficult to fit.[22] However, having interactions in the model necessitates the use of a "marginal" approach (which is described below in the context of a GLM) to achieve a health cost repartition among the different diseases. Alternatively, one may stick with model (2) even in the presence of interactions and study how an interaction (eg, an "extra" cost due to the simultaneous presence of 2 diseases for one patient) will be "allocated" among the different diseases when model (2) is estimated via OLS. This is what we are studying in a companion paper.[23]

## Log-Transformation of the Costs

When a distribution is skewed to the right, a natural (and often successful) approach is to apply a log-transformation. In our problem, one might thus be tempted to consider the above decomposition on the log scale, yielding:

$$\mathrm{Log}(Y_i) = \mathrm{mean}\big(\mathrm{Log}(Y_i)|X_{ij}\big) + \varepsilon_i. \quad (3)$$

In a linear model, one would then assume that:

$$\mathrm{mean}\big(\log(Y_i)|X_{ij}\big) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}. \quad (4)$$

One serious difficulty with model (4) is that it provides estimates of mean log-costs by disease, not of mean costs by disease. Moreover, a "smearing" retransformation to estimate a mean cost from a mean log-cost would strongly rely on a normality assumption of the $\varepsilon_i$ in model (3), and thus on the lognormal distribution of the $Y_i$ within each disease pattern. Even with lognormal distributions, one would need to model the variance (in addition to the mean) of the log-costs in case of heteroscedasticity in model (3). On the other hand, if the distributions are not strictly lognormal, estimates obtained via smearing retransformation will not be consistent.[24] Note also that assuming a lognormal distribution for the specific costs $Y_{ij}$ would not imply a lognormal distribution for the global costs $Y_i$, (but a sum of lognormal distributions which shall not be the same in each disease pattern). In fact, even to determine the distribution of the sum of only 2 lognormal distributions is known to be a difficult problem.[25]

## General Linear Model

The difficulties mentioned in the previous paragraph led recent authors to attempt an estimation of the log of mean costs, rather than of the mean of log costs, via a GLM. In such a model, $Y_i$ is assumed to follow a given parametric distribution, such as a Poisson or a Gamma distribution, where:

$$\mathrm{Log}\big(\mathrm{mean}\big(Y_i|X_{ij}\big)\big) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}. \quad (5)$$

This implies:

$$\mathrm{mean}\big(Y_i|X_{ij}\big) = \exp(\beta_0) \cdot \exp(\beta_1 X_{i1}) \cdot \exp(\beta_2 X_{i2}) \cdot \ldots \cdot \exp\big(\beta_p X_{ip}\big). \quad (6)$$

A serious concern of model (6) is that it describes a multiplicative reality. For example, in a GLM without an intercept, if a first disease alone would cost on average 10 000 CHF and a second disease alone would cost on average 100 CHF, having the 2 diseases simultaneously would cost on average 1 000 000 CHF, which is hard to believe (and things would become even worse with more than 2 diseases).

Even if model (6) would hold, one would not have $\mu_j = \beta_j$ as with model (2), and one could not simply use the (maximum likelihood) estimates $\hat{\beta}_j$ of the coefficients $\beta_j$ as estimates of the $\mu_j$. To get such estimates, one possibility would be to estimate the average "marginal cost" which would be saved by removing completely a disease. Such a marginal approach has been used by Moschetti et al.[7], in a context a bit different from ours, who calculated the following quantity:

$$\begin{aligned}\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} &\exp\Big(\hat{\beta}_0\Big) \cdot \exp\Big(\hat{\beta}_1 X_{i1}\Big) \cdot \ldots \\ &\cdot \Big(\exp\Big(\hat{\beta}_j\Big) - 1\Big) \cdot \ldots \cdot \exp\Big(\hat{\beta}_p X_{ip}\Big).\end{aligned} \quad (7)$$

This quantity is the average difference between the global health cost with disease $j$ and the global health cost without disease $j$, calculated over all individuals, while holding the disease status constant for all diseases other than $j$. However, one will not retrieve the total health cost spent for all the patients by summing up the estimated costs spent for the different diseases, the discrepancies being even typically (much) larger than with OLS, as illustrated in our example below. Note also that it is quite possible to get estimates $\hat{\beta}_j$ which are negative for some diseases, implying in turn negative estimates $\hat{\mu}_j$ via equation (7), as it happened in our example below, and as it also happened in Moschetti et al's study.[7]

Another annoying feature of model (6) is that a patient without any disease would cost $\exp(\beta_0)$, which cannot be set to 0 by removing the intercept from the model. In fact, removing the intercept would imply a cost of $\exp(0) = 1$ monetary unit for a patient without any disease, whatever this unit might be. Thus, contrary to model (2), an intercept is needed in model (6) to ensure that the results do not depend on the monetary units considered. Without an intercept in model (6), one would not get the same health cost repartition among the diseases if

one would express the costs in Swiss francs, or in thousands of Swiss francs, for example.

### *Iterative Proportional Repartition*

In the present article, we propose to estimate the mean costs by disease $\mu_j$ via an algorithm yielding an IPR of the global costs $Y_i$ across the different diseases diagnosed for a patient $i$, before averaging the specific costs hence obtained for a disease $j$ over the patients having that disease. Specifically, the IPR algorithm works as follows:

1. Start with some initial estimates $\hat{\mu}_j > 0$ of $\mu_j$ (eg, those estimates provided by OLS if they are all positive, or $\hat{\mu}_j = 1$ for $j = 1, \ldots, p$);
2. Let $\hat{Y}_{ij} = Y_i X_{ij} \hat{\mu}_j / \sum_{k=1}^{p} (X_{ik} \hat{\mu}_k)$ (for $i = 1, \ldots, n$ and $j = 1, \ldots, p$);
3. Update the current estimate of $\mu_j$ as $\hat{\mu}_j = \sum_{i=1}^{n} (X_{ij} \hat{Y}_{ij}) / \sum_{i=1}^{n} X_{ij}$ (for $j = 1, \ldots, p$) and go back to step 2 until a stopping criterion.

In our illustration and simulations below, we used as stopping criterion that the updated estimates of $\mu_j$ differ from the current estimates of $\mu_j$ by less than 1 CHF for all the diseases (ie, for $j = 1, \ldots, p$). As for the other methods, standard errors and confidence intervals can be obtained via bootstrap.

For example, if one would know that a first disease would cost on average 10 000 CHF and a second disease would cost on average 100 CHF, the former being 100 times more expensive than the latter, and if the health costs of 1 patient having the 2 diseases simultaneously would be of 12 120 CHF (which is 2020 CHF more than the sum of the average costs of the 2 diseases, due to an interaction), one would not split the extra amount of 2020 CHF equally among the 2 diseases (ie, one would not consider $10\,000 + 1010 = 11\,010$ CHF for the first and $100 + 1010 = 1110$ CHF for the second disease), but one would allocate it proportionally to the costs of the 2 diseases (yielding $10\,000 + 2000 = 12\,000$ CHF for the first and $100 + 20 = 120$ CHF for the second disease). However, since one does not know the average costs of the diseases in reality, one should proceed iteratively, as described in the algorithm above.

By construction, IPR estimates $\hat{\mu}_j$ of $\mu_j$ will be necessarily strictly positive and one will retrieve the total health spent for all the patients by summing up the estimated costs spent for the different diseases, that is, $\sum_{j=1}^{p} \hat{\tau}_j = \sum_{i=1}^{n} Y_i$ (although the total would no longer be retrieved when applying the IPR estimates to another representative sample of patients from the same population).

Another (slight) advantage of IPR is that, contrary to OLS or GLM, it will not be affected by the potential (numerical) problems due to a possible multicolinearity among the predictors $X_{ij}$. Even in an extreme case of multicolinearity, for example, where 1 patient would have one disease if and only if he or she would have another one disease, IPR could still be calculated, eg, sharing the costs equally among those 2 diseases. Another problematic situation would arise in a case where each disease would be classified either as a "principal" or as a "secondary" disease, such that each patient would have one and only one principal disease (a system which would be similar to diagnosis-related groups, although including in addition secondary diseases). In that situation, including an intercept in a regression model would yield a multicolinearity issue, whereas IPR could still be calculated. A similar issue would also arise in a case where all the patients would have exactly the same number of diseases (eg, if one would consider a population of patients with more than one disease, but considering only the most 2 serious diseases for each patient).

## Illustration

In this section, we illustrate and compare the IPR algorithm with different methods applied to global health costs data from Switzerland, collected on $n = 482\,303$ patients and involving $p = 49$ diseases. These data were representative of the health costs in Switzerland during the year 2006.

The total health costs calculated over all the patients in our sample was of 595 661 616 CHF (Swiss Francs). The distribution of the (global) health costs per patient is shown on the left panel (A) of Figure 1 (transformed on the log scale, base 10), ranging from 1 to 455 200 CHF, with a median cost of 422 CHF and an average cost of 1235 CHF. A total of 1 461 939 diseases have been diagnosed, yielding an average of 3.03 diseases per patient. The distribution of the number of diseases per patient is shown on the right panel (B) of Figure 1. A proportion of 33.2% patients had just 1 disease, 23.1% had 2 diseases, 14.7% had 3 diseases, and 9.1% had 4 diseases, while 1 patient had up to 24 diseases (the maximum number of diseases observed for a single patient). Of course, some diseases were less frequent than other, ranging from 203 to 155 800 occurrences, with a median of 9906 and an average of 29 840 occurrences. Of note, the most expensive diseases occurred (fortunately) not very often, while the Spearman correlation between disease frequency and disease cost was negative (eg, $-0.48$ according to the costs estimated via IPR).

Mean costs $\mu_j$ were estimated by 4 different methods: IPR, OLS, GLM Poisson, and GLM Gamma. All calculations were done/programmed using the statistical software R (version 3.3.3). For IPR, our stopping criterion was met after 21 iterations of the algorithm (which took about 100 seconds on a laptop computer from 2013, with a processor 2 GHz Intel Core i7). Table 1 summarizes the results. For OLS, GLM Poisson, and GLM Gamma, we got, respectively, 4, 8, and 6 negative estimates $\hat{\mu}_j$. For each method, such negative values were set to half of the smallest among the positive values of $\hat{\mu}_j$. As expected, IPR was the only method allowing to retrieve the total health cost from the estimated mean costs, that is, with $\sum_{j=1}^{p} \hat{\tau}_j / \sum_{i=1}^{n} Y_i$ equal to 1, the latter quantity being equal to 1.04, 0.63, and 1.17 for OLS, GLM Poisson, and GLM Gamma, respectively. In particular, the total health cost was

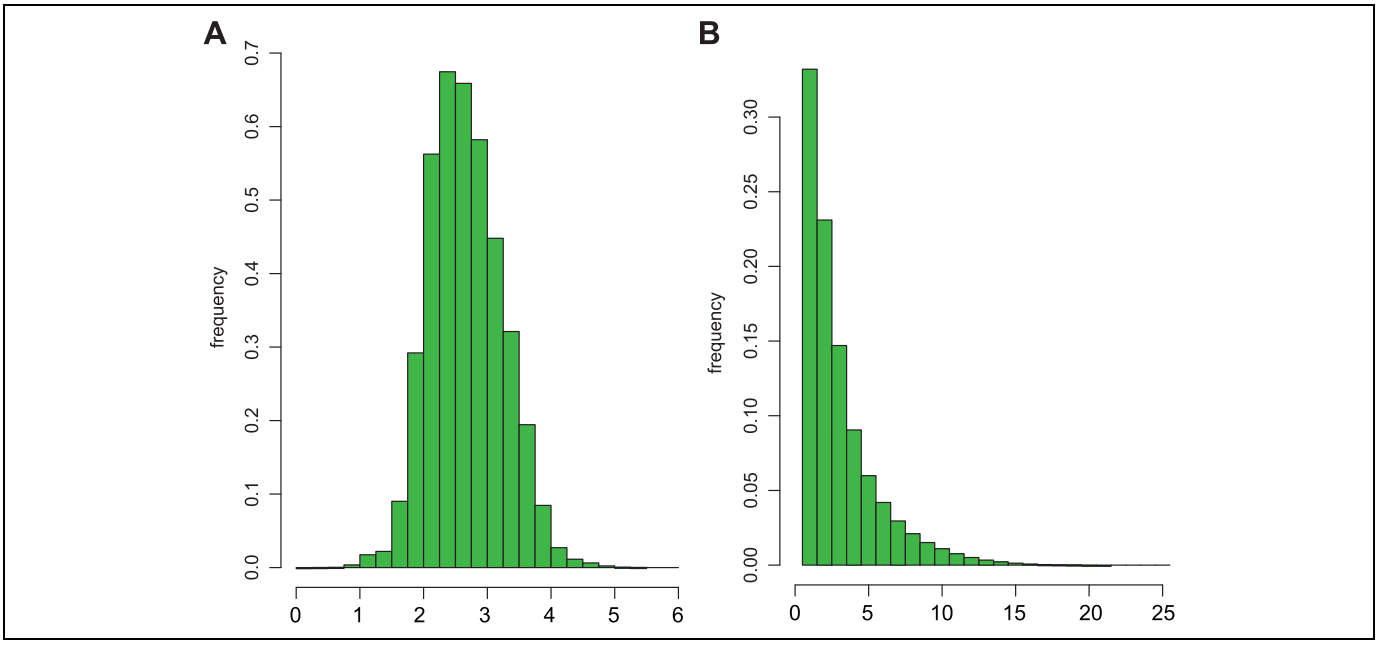


**Figure 1.** Histograms of global health costs (log scale, base 10, left panel) and of numbers of diseases (right panel) calculated from 595 661 616 patients representative of the health costs in Switzerland during the year 2006.

**Table 1.** Comparison of Mean Cost by Disease ($\hat{\mu}_j$) Estimated via Different Methods.

| | OLS | GLM (Poisson) | GLM (Gamma) | IPR |
|---|---|---|---|---|
| % negative mean costs | 4/49 | 8/49 | 6/49 | 0/49 |
| Minimum (CHF) | 28 | 7 | 12 | 87 |
| Maximum (CHF) | 30 330 | 15 730 | 33 070 | 29 410 |
| Median (CHF) | 444 | 246 | 530 | 486 |
| Mean (CHF) | 1991 | 983 | 2239 | 1939 |
| Total health cost (estimation/truth) | 1.04 | 0.63 | 1.17 | 1 |

Abbreviations: GLM generalized linear models; IPR, iterative proportional repartition; OLS, ordinary least squares.

overestimated by 17% using GLM Gamma, while it was underestimated by 37% using GLM Poisson. Summary statistics (minimum, maximum, median, and mean) are also provided in Table 1. We retrieve the fact that estimates were on average much lower when using GLM Poisson, while they were on average higher when using GLM Gamma than when using OLS or IPR.

That the total health cost is not retrieved is not necessarily a problem when the issue is to estimate the cost contributions $\pi_j$ of the different diseases, since the estimates $\hat{\pi}_j$ are standardized to sum to one for each method (and since one knows the true total health cost). Figure 2 shows the relationships between the $\hat{\pi}_j$ obtained via the different methods, together with the corresponding Spearman correlations. One can see that IPR was closer to OLS (Spearman correlation of 0.96) than to GLM Poisson (Spearman correlation of 0.86) or GLM Gamma (Spearman correlation of 0.88). Note also that an absolute difference of more than 1% compared to the $\hat{\pi}_j$ estimated via IPR was observed for 7 (of the 49) diseases when using OLS, and for 11 diseases when using either GLM Poisson or GLM Gamma, the mean absolute difference was 0.4%, 0.7%, and 0.5%, respectively, and the maximum absolute difference was 2.1%, 4.8%, and 3.3%, respectively. Thus, the conclusions which can be drawn regarding health cost repartition among the different diseases in Switzerland in 2006 will partly depend on the method which is used, raising the question of determining which method is best. In Figure 2, 95% confidence intervals for the $\pi_j$ based on 500 bootstrapped resamples are also represented to indicate the variability of the estimates. In general, one can see that IPR yielded the shortest intervals, and thus the most stable estimates. This will be further investigated in the next section via simulations.

## Simulations

In this section, we present the results of simulations where cost data have been generated according to the simulation design which is presented below. To get a realistic scenario, in particular in terms of disease pattern frequencies, it is inspired from the Swiss health cost data described in the previous section.

We considered $n = 100\,000$ patients and $p = 49$ diseases. In our basic simulation design, we have (randomly) selected 100 000 lines $(X_{i1}, \cdots, X_{ip})$ from the matrix of the $X_{ij}$ observed in the Swiss health data, such that disease frequencies, and more generally disease pattern frequencies, resembled those observed in the Swiss health data (some combinations of diseases being more frequent than others). For each individual

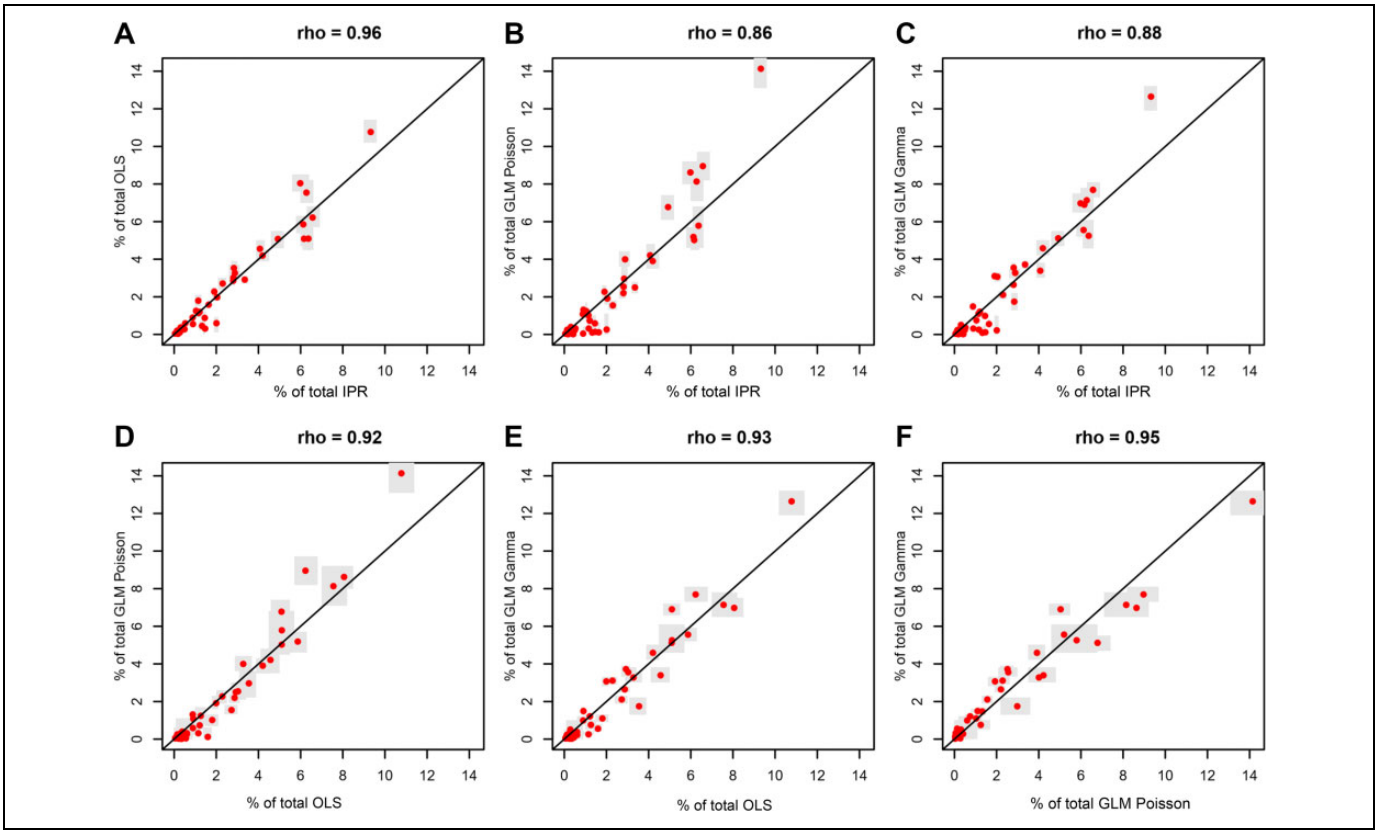


**Figure 2.** Scatterplots, together with Spearman (rho) correlations, comparing the estimates of percentages of the total health costs ($\hat{\pi}_j$) spent for the 49 diseases obtained by 4 different methods, each panel comparing 2 different methods (IPR, OLS, GLM Poisson, and GLM Gamma, each point representing 1 disease, an identity line being added as a reference line), for the Swiss health cost data. The length and width of the gray-shaded rectangles represent 95% confidence intervals (for the corresponding methods) obtained from 500 bootstrapped resamples. GLM indicates generalized linear models; IPR, iterative proportional repartition; OLS, ordinary least squares. Panel A: IPR vs. OLS, Panel B: IPR vs. GLM Poisson, Panel C: IPR vs. GLM Gamma, Panel D: OLS vs. GLM Poisson, Panel E: OLS vs. GLM Gamma, Panel F: GLM Poisson vs. GLM Gamma.

$i = 1, \cdots, n$ and each disease $j = 1, \cdots, p$, the specific cost $Y_{ij}$ was generated according to a lognormal distribution with true mean cost $\mu_j$, taken as mean cost $\hat{\mu}_j$ estimated via IPR (see the previous section), and with a (within-disease) coefficient of variation (CV = standard deviation/mean) of 2. Global costs were then obtained as $Y_i = \sum_{j=1}^{p}(X_{ij}Y_{ij})$ (for $i = 1, \ldots, n$). To simplify the presentation, diseases were sorted according to their mean costs, $\mu_1$ denoting the mean cost from the cheapest disease, $\mu_p$ the mean cost from the most expensive disease. We have then considered various variants from this basic simulation design, resulting in the 10 following simulation settings:

1. [Basic design or BD] As just described.
2. [GLM mean costs] Same as BD, except that true mean costs $\mu_j$ were taken as mean costs $\hat{\mu}_j$ estimated via GLM Gamma (see the previous section).
3. [Small CV] Same as BD, except that the CV was set to 0.5.
4. [CV decreasing with mean cost] Same as BD, except that the CV was taken different for each disease, being $2 - 1.5(j-1)/(p-1)$ for disease $j$.
5. [CV increasing with mean cost] Same as BD, except that the CV was taken different for each disease, being $0.5 + 1.5(j-1)/(p-1)$ for disease $j$.
6. [Gamma distribution] Same as BD, except that the specific costs were generated according to a Gamma (instead of a lognormal) distribution.
7. [Frequency decreasing with mean cost] Same as BD, except that the columns $(X_{1j}, \cdots, X_{nj})'$ from the matrix of the $X_{ij}$ have been reordered, such that disease frequency was monotonically decreasing from the cheapest to the most expensive disease.
8. [Frequency increasing with mean cost] Same as BD, except that the columns $(X_{1j}, \cdots, X_{nj})'$ from the matrix of the $X_{ij}$ have been reordered, such that disease frequency was monotonically increasing from the cheapest to the most expensive disease.
9. [Uniform frequencies] Same as BD, except that the $p$ elements $X_{ij}$ of each line $(X_{i1}, \cdots, X_{ip})$ from the matrix of the $X_{ij}$ have been randomly swapped (independently for $i = 1, \ldots, n$), such that disease distribution was uniform.

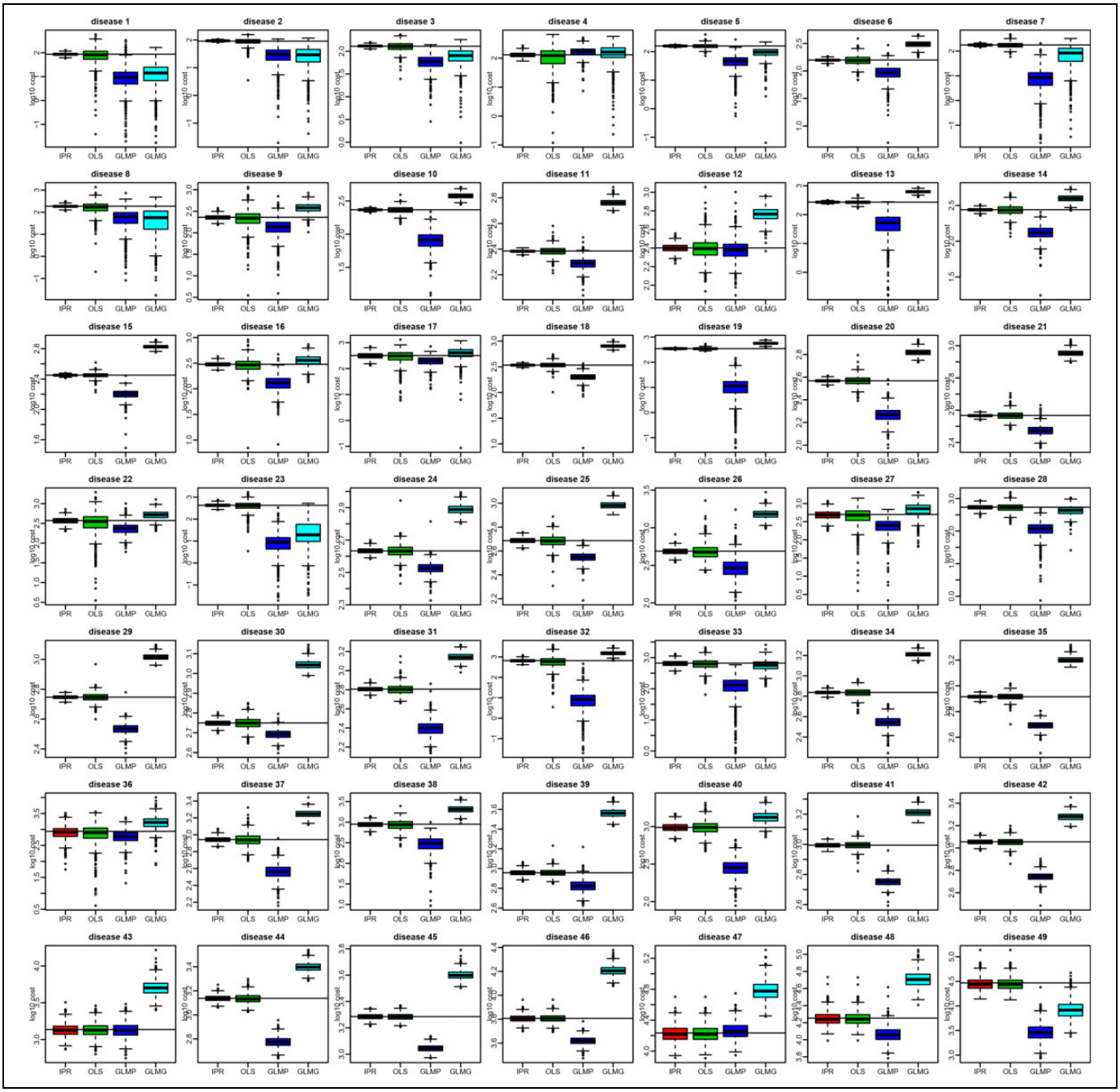


**Figure 3.** Boxplots of the estimates of mean costs by disease ($\hat{\mu}_j$, log scale, base 10) for the 49 diseases obtained by 4 different methods (IPR, OLS, GLM Poisson, and GLM Gamma, each boxplot representing 1 disease) over 1000 simulated data sets of health costs for 100 000 patients generated from our basic design. Horizontal lines represent the true mean costs from which data have been generated. GLM indicates generalized linear models; IPR, iterative proportional repartition; OLS, ordinary least squares.

10. [More diseases] Same as the former design but with twice more diseases per patient (or set back at 24 in case of more than 24 diseases), resulting in an average of 6.04 (instead of 3.04) diseases per patient.

We generated $N = 1000$ data sets under each of these 10 settings. In each generated data set $I = 1, \ldots, N$, we got estimates $\hat{\mu}_j = \hat{\mu}_j(I)$ of the mean costs by disease $\mu_j$ obtained via either IPR, OLS, GLM Poisson, or GLM Gamma (the few negative estimates being treated as in the previous section), as well as the corresponding estimates $\hat{\pi}_j = \hat{\pi}_j(I)$ of the cost contributions by disease $\pi_j$ (for $j = 1, \ldots, p$). For IPR, our stopping criterion was met after less than 150 iterations (on average, after 32 iterations) for all generated data sets under any setting. Figure 3 shows the boxplots of the 1000 estimates obtained for each method and each disease under the

basic design (on the log scale, base 10). One can see that IPR and OLS yielded estimates which were mostly unbiased, the latter with a higher variability (ie, a worse accuracy) than the former, while GLM Poisson and GLM Gamma yielded biased estimates. Although this bias was mostly negative for GLM Poisson and mostly positive for GLM Gamma, it was not of the same amplitude (and not even in the same direction!) for each disease, making a bias correction problematic.

In order to not overpenalize the GLM methods, which were thus underestimating (GLM Poisson) or overestimating (GLM Gamma) the total health cost, and since our ultimate goal is an estimation of the health cost repartition among the different diseases, we compare below the methods with respect to bias and accuracy of the $\hat{\pi}_j$ (instead of the $\hat{\mu}_j$). Under each setting and for each disease $j = 1, \ldots p$, methods can be compared via classical criteria such as bias, standard deviation (ie, variability), and root mean square error (RMSE) defined as:

$$B(j) = \frac{1}{N}\sum_{I=1}^{N} \hat{\pi}_j(I) - \pi_j$$

$$SD(j) = \sqrt{\frac{1}{N}\sum_{I=1}^{N} \hat{\pi}_j^2(I) - \left(\frac{1}{N}\sum_{I=1}^{N} \hat{\pi}_j(I)\right)^2}$$

$$\mathrm{RMSE}(j) = \sqrt{B^2(j) + SD^2(j)} = \sqrt{\frac{1}{N}\sum_{I=1}^{N} \left(\hat{\pi}_j(I) - \pi_j\right)^2}$$

An alternative (close in spirit) to RMSE is the mean absolute error defined as:

$$\mathrm{MAE}(j) = \frac{1}{N}\sum_{I=1}^{N} |\hat{\pi}_j(I) - \pi_j|.$$

This quantity is interpretable as the expected absolute error of cost contribution of disease $j$. Those $\mathrm{MAE}(j)$ quantities could then be averaged over all the diseases, yielding what we shall call the overall absolute error (OAE), a useful summary of the global performance of a method, obtained as:

$$\mathrm{OAE} = \frac{1}{p}\sum_{j=1}^{p} \mathrm{MAE}(j).$$

Figure 4 summarizes the performance of the methods in terms of OAE under the 10 simulation designs described above. More complete results involving the other criteria above are available in the supplementary material. Results were pretty clear-cut with IPR consistently outperforming the other 3 methods. Second best was OLS with an OAE between 1.5 and 4.7 times higher than for IPR depending on the simulation setting, whereas it was between 5.6 and 22.9 times higher for GLM Poisson, and between 4.5 and 19.4 times higher for GLM Gamma. Various additional simulations from still different settings confirmed these results.

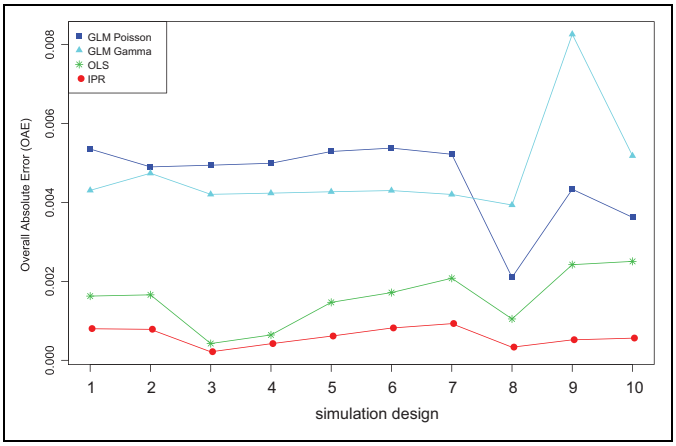


**Figure 4.** Overall absolute errors (OAE) of estimates of percentages of the total health costs ($\hat{\pi}_j$) averaged over the 49 diseases obtained by 4 different methods (IPR, OLS, GLM Poisson, and GLM Gamma) over 1000 simulated data sets of health costs for 100 000 patients generated from the 10 simulation designs described in section 5. GLM indicates generalized linear models; IPR, iterative proportional repartition; OLS, ordinary least squares.

## Conclusions

Estimating mean cost and cost contribution by disease from global health costs is a challenging and nonstandard statistical problem. One possible approach would be to consider a regression model. While an additive model certainly represents a reasonable (first) approximation of the reality, a classical method of estimation such as OLS is not optimal in presence of asymmetric and heteroscedastic distributions. On the other hand, a multiplicative model such as GLM appears quite unrealistic and provided (not surprisingly) bad results in the context of an additive model even using a marginal approach. Finally, we have introduced the IPR method, which is nonparametric, simple to understand and to implement, and which turned out to consistently outperform the other methods considered in our simulation study.

An obvious (in fact inevitable) limitation of our simulation study is that we have simulated global health costs from a model without interactions. However, this is only under that setting that the health cost repartition among the diseases is clearly defined such that one unambiguously knows the "truth." In that context, it becomes possible to compare the different methods on an objective basis.

In presence of interactions (which in practice will be the rule rather than the exception), it is no longer clear how an extra (or an economy of) cost due to the simultaneous presence of several diseases for 1 patient *should be* allocated among those diseases. We are actually studying this far-reaching question in a companion paper in the simple case of 2 diseases. It turns out that the different methods will allocate this extra cost differently among the diseases such that each method will estimate a different (true) health cost repartition. In such a context, it is thus not relevant to compare the performance of the methods via the usual statistical

criteria, such as those used in the previous section. In other words, in presence of interactions, the choice of a method should not be based on the efficiency of estimation, but on the relevance of what is estimated, such that a comparison of methods via simulations would become obsolete.

As mentioned in our method section, IPR is allocating the extra (or the economy of) cost due to the simultaneous presence of 2 diseases for 1 patient proportionally to the average cost of those 2 diseases, which again seems natural to us, but which is not achieved using another method (see our companion paper).

Other advantages of the IPR method are that it retrieves by construction the total health cost by summing up the estimated costs spent for the different diseases (whether interactions are present or not), and that it does not face neither the potential issues of multicolinearity nor the problem of the negative estimates, whereas other methods should be updated accordingly (in some more or less artisanal way) to accommodate these issues. This is why we see IPR as the natural method to achieve a health cost repartition and we would like to encourage its use.

We end up by mentioning 2 (related) open problems raised by the reviewers for interested readers. One would be to study which (simple) objective function is formally minimized using IPR. Another one would be to prove mathematically that the IPR algorithm does always converge, which we were not able to demonstrate, despite the fact that our stopping criterion was met in 100% of our simulations.

Finally, we would like to underline that our problem of estimating the repartition of health cost among the diseases is based on a classification system (into $p$ diseases) which is not necessarily straightforward to define. One issue is that some diseases might be the consequences of other. For instance, an impaired renal function is a frequent consequence of a congestive heart failure. In that case, one may wish to entirely allocate the cost of the former to the latter. This can be solved by an appropriate choice of classification system (for example, by occulting the former disease in those cases or by considering a "new" disease which would be the combination of the 2). Another point is that it might be desirable to consider several episodes of illnesses per patient to take into account time diseases overlap. Of course, changing the classification system or the number of episodes per patient will change the health cost repartition whatever the method which is used.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## ORCID iD

Jean-Benoît Rossel https://orcid.org/0000-0001-5803-3233

## Supplemental Material

Supplemental material for this article is available online.

## References

1. Reynales-Shigematsu LM, Campuzano-Rincón JC, Sesma-Vásquez S, et al. Costs of medical care for acute myocardial infarction attributable to tobacco consumption. *Arch Med Res*. 2006;37(7):871-879.
2. Juhnke C, Bethge S, Mühlbacher AC. A review on methods of risk adjustment and their use in integrated healthcare systems. *Int J Integr Care*. 2016;16(4):4.
3. McPhail SM. Multimorbidity in chronic disease: impact on health care resources and costs. *Risk Manag Healthc Policy*. 2016;9: 143-156.
4. Gagnon MP, Desmartis M, Poder T, Witteman W. Effects and repercussions of local/hospital-based health technology assessment (HTA): a systematic review. *Syst Rev*. 2014;3: 129.
5. Global Burden of Disease Study 2013 Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the global burden of disease study 2013. *Lancet*. 2015;386(9995): 743-800.
6. Nurmagambetov T, Kuwahara R, Garbe P. The economic burden of asthma in the United States, 2008-2013. *Ann Am Thorac Soc*. 2018;15(3):348-356.
7. Moschetti K, Zabrodina V, Wangmo T, et al. The determinants of individual health care expenditures in prison: evidence from Switzerland. *BMC Health Serv Res*. 2018;18:160.
8. Eggli Y, Halfon P, Chikhi M, Nguyen L, Decollogny A, Weissbaum F. Analyse des prestations prises en charge par l'assurance maladie obligatoire des soins. Cadre conceptuel et étude de faisabilité centrée sur trois pathologies: cancer, diabète et affections mentales. Bern, Switzerland: OFSP; 2016.
9. OECD. *Health Divison, Estension of work on Expenditure by Disease, Age and Gender*. OECD, Paris; 2013. https://www.oecd.org/els/health-systems/Extension-of-work-on-expenditure-by-disease-age-and-gender_Final-Report.pdf. Accessed October 25, 2019.
10. Trogdon JG, Finkelstein EA, Hoerger TJ. Use of econometric models to estimate expenditure shares. *Health Serv Res*. 2008; 43(4):1442-1452.
11. Teo WS, Tan WS, Chong WF, et al. The economic burden of chronic obstructive pulmonary disease. *Respirology*. 2012; 17(1):120-126.
12. Nurmagambetov T, Kuwahara R, Garbe P. The economic burden of asthma in the United States, 2008-2013. *Ann Am Thorac Soc*. 2018;15(3):348-456.
13. Ward MM, Javitz HS, Smith WM, Bakst A. A comparison of three approaches for attributing hospitalizations to specific diseases in cost analyses. *Int J Technol Assess Health Care*. 2000; 16(1):125-136.
14. Fetter RB, Freeman JL. Diagnosis related groups: product line management within hospitals. *Acad Manage Rev*. 1986;11(1):41-54.
15. WHO. *International Statistical Classificcation of diseases and related Health Problems.* Tenth revision. Volume 1. Geneva, Switzerland: WHO; 1992.

16. Hodgson TA, Cohen AJ. Medical care expenditures for diabetes, its chronic complications, and its comorbidities. *Prev Med*. 1999; 29(3):173-186.
17. Hodgson TA, Cai L. Medical Care expenditures for hypertension, its complications, and its comorbidities. *Med Care*. 2001;39(6): 599-615.
18. Yang L, Sung HY, Mao Z, Hu TW, Rao K. Economic costs attributable to smoking in China: update and an 8-year comparison, 2000-2008. *Tob Control*. 2011;20(4):266-272.
19. Barnett SB, Nurmagambetov TA. Costs of asthma in the United States: 2002-2007. *J Allergy Clin Immunol*. 2011;127(1):145-152.
20. Eggli Y, Halfon P, Seker E. Estimation of the profitability of chronic diseases prevention through routinely available data in Switzerland. Cost savings achieved by the prevention of chronic diseases estimated on the basis of routinely available data. Bern, Switzerland: OFSP; 2012 (expert report).
21. Aitkin AC. On least squares and linear combination of observations. *Proceed Roy Soci Edinbur*. 1935;55:42-48.
22. Manning W, Mullahy J. Estimating log models: to transform or not to transform? *J Health Eco*. 2001;20:461-494.
23. Rossel JB, Rousson V, Eggli Y. Statistical methods for allocating disease costs in the presence of interactions. In preparation.
24. Jones M, Lomas J, Moore P, Rice N. A quasi-Monte Carlo comparison ORF developments in parametric and semi-parametric regression methods for heavy tailed and non-normal data: with an application to healthcare costs. *J R Stat Soc Ser A Stat Soc*. 2016;179(4):951-974.
25. Beaulieu NC, Abu-Dayya AA, McLane PJ. Estimating the distribution of a sum of independent lognormal random variables. *IEEE Transact Comm*. 1995;43(12):2869.

## Author Biographies

**Valentin Rousson** obtained a PhD in Statistics at the University of Neuchâtel in 1998. He then worked as a postdoc and as a scientific collaborator at the Australian National University in Canberra, and at the University of Zurich, before joining the University of Lausanne and Unisanté, where he is Associate Professor in Biostatistics since 2007.

**Jean-Benoît Rossel** obtained a PhD in Mathematics at the Ecole polytechnique fédérale de Lausanne (EPFL) in 2013. He has been working as a statistician at Unisanté since 2014, where he split his time between statistical research and statistical consultations to medical doctors.

**Yves Eggli** obtained a PhD in Health Services Management at the University of Montréal in 2002, after a doctorate in Medicine and a Master of Arts at the University of Lausanne. He has expertise in health management, planning, cost controlling, resource allocation, patients' classification systems, and quality assessment.