DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 4, 2018, pp. 863–888, http://dx.doi.org/10.2478/JOS-2018-0043

Detecting Reporting Errors in Data from Decentralised Autonomous Administrations with an Application to Hospital Data

Arnout van Delden[1], *Jan van der Laan*[1], *and Annemarie Prins*[2]

Administrative data sources are increasingly used by National Statistical Institutes to compile statistics. These sources may be based on decentralised autonomous administrations, for instance municipalities that deliver data on their inhabitants. One issue that may arise when using these decentralised administrative data is that categorical variables are underreported by some of the data suppliers, for instance to avoid administrative burden. Under certain conditions overreporting may also occur.

When statistical output on changes is estimated from decentralised administrative data, the question may arise whether those changes are affected by shifts in reporting frequencies. For instance, in a case study on hospital data, the values from certain data suppliers may have been affected by changes in reporting frequencies. We present an automatic procedure to detect suspicious data suppliers in decentralised administrative data in which shifts in reporting behaviour are likely to have affected the estimated output. The procedure is based on a predictive mean matching approach, where part of the original data values are replaced by imputed values obtained from a selected reference group. The method is successfully applied to a case study with administrative hospital data.

Key words: Administrative data; measurement errors; predictive mean matching; reporting errors; selective editing.

1. Introduction

Use of administrative data in official statistics offers several advantages over survey data, such as observations for a larger fraction of the target population, reduced data collection costs and lower response burden. Therefore, administrative data is increasingly used by National Statistical Institutes (NSIs) to compile statistics, either as a sole data source or in combination with other sources. Administrative data here refers to data collected by an organisation external to the statistical office for administrative purposes, thus not targeted for use in official statistics (UNECE 2011). When the statistical population, unit and variable definitions coincide with those for the administrative data source, estimation of the statistical output is straightforward. For instance, the total number of persons receiving

[1] Statistics Netherlands, Department of Process Development and Methodology, Henri Faasdreef 312, P.O. Box 24500, 2490 HA The Hague, The Netherlands. Emails: a.vandelden@cbs.nl, and dj.vanderlaan@cbs.nl
[2] Netherlands Institute for Health Services Research (Nivel), P.O.Box 1568, 3500 BN Utrecht, The Netherlands. a.prins@nivel.nl

Acknowledgments: The authors would like to thank the Associate Editor, four anonymous referees, Peter van der Heijden and Peter-Paul de Wolf for their useful comments and suggestions, which have led to a significant improvement of this article. The authors thank Erik van Bracht for drawing the flow chart.

an unemployment benefit is easily derived from the corresponding administrative data. However, when population, unit or variable definitions do not coincide, or when the purpose of the register holder clearly differs from the intended statistical use, methodological issues may arise (Bakker and Daas 2012; Wallgren and Wallgren 2014).

One of the issues that might occur with administrative data is that the registered values differ from the true ones (as defined by the statistical office), resulting in measurement errors. This happens especially with variables that are not of crucial importance to the owner of the data set. For instance, enterprises might register their reported value added tax data as being monthly, whereas in fact it concerns four-week values (Van Delden and Scholtus 2017). The tax office tolerates deviations in monthly values, especially for smaller enterprises, as long as the yearly amount of tax paid is correct. Also, Statistics Netherlands (CBS) uses administrative fire brigade data. Fire brigades need to register the variable "did the fire cause any environmental damage". They underreport any occurrence of environmental damage, because this way they avoid having to register a number of subsequent variables, such as an estimated cost of the environmental damage (Berenschot 2012). From research on questionnaires, it is also known that respondents learn to shorten questionnaire duration by underreporting events (Backor et al. 2007; Shields and To, 2005; Silberstein and Jacobs, 1989). In the case of surveys, there is a lot of literature available on reporting errors, for instance reporting errors might occur when asking sensitive questions, or as a result of socially desirable behaviour (Tourangeau et al. 2010; Tourangeau and Yan 2007). In the case of administrative data, numerous studies on measurement errors have been done (e.g., Groen 2012; Oberski et al. 2017 and references therein), but to the best of our knowledge, the role of the administrative practice of data suppliers on these measurement errors has hardly been given any attention.

Some of the administrative data sets used in official statistics are obtained through decentralised data collection. For example, population data and social benefit data are registered by municipalities. Similar examples concern administrative data sets provided by fire brigades (on fires), by schools (on pupils), by hospitals (on patients), by local authorities (on building activities), by employers (on salary information of employees) and by courts (on legal proceedings). These decentralised administrations will be referred to as "data suppliers" in this article and the corresponding administrative data will be referred to as "decentralised administrative data". Each of these decentralised administrations may have their own administrative practices (Brackstone 1987), resulting in measurement errors that vary with the data supplier. For instance, employers in the Netherlands vary in the intensity of reporting employees' overtime. In surveys, a similar phenomenon occurs with personal interviewing, where interviewer-dependent measurement errors may occur (West and Blom 2017). For instance, homeless respondents reported drug use more frequently in the presence of male interviewers (see West and Blom 2017, 189 and references therein).

From an official statistics point of view, preventing measurement errors in administrative data is desirable, for instance by unifying and improving the "fields" that the administrators have to fill in, or the questions that they have to respond to. Nonetheless, there are at least two obstacles to achieving such improvements in practice. The first obstacle is that local administrations may have different administrative systems (software). This is, for instance, the case with employers reporting salary data for

employees, with hospital data (see Section 2) and with financial administrative data of municipalities. A second obstacle is that local administrations are autonomous and act rather independently of the statistical offices that receive the data. The best that the NSIs can do is to discuss quality issues with them and request improvements; the NSI cannot prescribe any changes to the administrative systems. Before such a discussion can be held, the NSI should have serious indications that measurement errors occur. The present article therefore focusses on the detection of reporting errors in data of decentralised, autonomous administrations.

In order to avoid biased outcomes, NSIs usually correct influential measurement errors in a data editing process. Automatic error correction methods are applied when correct values can be deduced from other variables, or when records are not influential (De Waal et al. 2011). Otherwise, selective manual data editing will be used, and, if needed, respondents are contacted. Selective editing methods aim to identify units with a high risk of influential errors, where an "influential error" is defined as one "that has a considerable effect on the publication figures" (De Waal et al. 2011). Our approach resembles that of selective editing. However, when under- or overreporting of one or more variables occurs in decentralised administrative data, correcting those data may not always be easy. Applying the methodology described in the current article, we are able to detect the data suppliers with measurement errors, but we cannot precisely detect which of the units within a data supplier contain errors.

The data suppliers responsible for those decentralised administrative data will not be able to determine which of the values are incorrect, nor to provide the "correct" values for individual records in the data. The problem is that the correct data either have not been registered, or can only be obtained with considerable effort. Municipalities, for instance, might not be able to identify which students have moved out of their parents' homes and which have not. Hospitals may not be able to see the complete set of diseases of their patients, if not all of them have been registered. In such a situation, the best option is to analyse which of the data suppliers have relatively many measurement errors. Subsequently, one can contact "suspicious" data suppliers and motivate them to improve their administrative processes in order to reduce the number of errors in future data deliveries.

Detecting under- or overreporting in a large number of variables in decentralised administrative data may be especially difficult in the case of level estimates. An option to analyse reporting behaviour in the case of level estimates makes use of a second independent source. In the present article, we aim to detect changes in reporting behaviour between two time periods. More specifically, we aim to develop an automatic procedure to detect suspicious data suppliers in decentralised administrative data in which shifts in reporting behaviour are likely to have affected a targeted change estimate. We apply our method to hospital data.

The remainder of the article is organised as follows. Section 2 gives some background information on the hospital data and the potential reporting errors therein. Section 3 describes the methodology used to select data suppliers with deviating reporting behaviour. How this methodology is applied to the case study is described in Section 4. Section 5 presents the results of the case study. Finally, Section 6 discusses the outcomes.

2. Background of the Case Study

At CBS, Dutch Medical Registration data (LBZ) is used concerning hospital stays of patients. This data set contains patient-related information, such as age and sex, and diagnosis-related variables, such as main diagnosis and comorbidities, which are other diagnoses describing the medical condition of the patient (Elixhauser et al. 1998). These data are compiled by the hospitals to provide a clinical data set that can be used by medical researchers. At each hospital, LBZ data is registered by coders using the administrative data system of the hospital and patient files (Van den Bosch et al. 2010). The LBZ data set is not targeted for use in official statistics, and fulfils the UNECE (2011) description of administrative data. We therefore refer to LBZ data as administrative data in the remainder of this article.

Since 2011, CBS has been responsible for computing the yearly Hospital Standardised Mortality Ratio (HSMR) for Dutch hospitals using LBZ data. The HSMR aims to measure differences in quality of hospital care and its computation was initiated in the United Kingdom by Jarman et al. (1999). Nowadays, it is being computed in a number of countries, such as the United States, Canada, the United Kingdom (Bottle et al. 2011), Australia and the Netherlands. Mortality is taken as a measure of hospital care, since several studies have shown that mortality correlates with quality of hospital care (e.g., Pitches et al. 2007). The HSMR of a hospital is computed as the ratio of observed to expected mortality, normalised to 100 (over all hospitals in a year). The HSMR includes an expected mortality to remove differences between hospitals that are caused by differences in patient populations. The expected mortality is estimated from a logistic regression model that includes a large number of background variables (Israëls et al. 2012).

In the hypothetical situation that we send the same patient to all Dutch hospitals, we would like the hospitals to register the same values for all patient- and disease-related variables. In practice, this is indeed the case in the Netherlands for variables such as age and sex. However, for a number of other variables, differences in reporting frequency were found among hospitals. The largest differences were found for the variables comorbidity and urgency of admission (Jarman, 2008; Pieter et al. 2010; Van der Laan 2013). According to Van den Bosch et al. (2010), reasons for these differences in reporting frequency are time pressure due to a limited number of coders, interpretation differences of the coding rules, and late delivery of patient files. Furthermore, the average number of coders per admission, and consequently, the time typically spent on LBZ registration (Van den Bosch et al. 2010) varies between hospitals. Van der Laan (2013) showed a sharp increase in the average number of reported comorbidities in some hospitals in 2008–2010, where 2010 was the year when the HSMR would become publicly available. Since such a large shift in the patient population of the hospitals in such a short time seems unlikely, this suggests that some hospitals changed their comorbidity reporting (Van der Laan 2013). Note that increased comorbidity reporting – everything else being the same – leads to a decreased HSMR. An increase in the average number of comorbidities by 0.1 led to an estimated HSMR decrease of five points (Van der Laan 2013), implying improved hospital care. This makes the data interesting as a case study. Another quality issue in the hospital data is that hospitals sometimes use the wrong codes (misclassifications) when reporting the main diagnosis or the comorbidities of patients, see for instance Harteloh et al. (2010)

and Quan et al. (2008). Although this is an important quality issue, estimating these misclassifications is beyond the scope of the present article.

In the present study, we focus on the effect of reporting behaviour on estimated changes. In the HSMR case study, for instance, many hours of manual analysis are being spent to clarify whether some hospitals with a changed HSMR have been affected by changes in intensity of comorbidity reporting. Cases of "suspicious results", such as a large change in the average number of comorbidities per hospital stay, are reported by the staff to the data holder, Dutch Hospital Data, which releases the outcomes. Currently, the average number of comorbidities per hospital stay is used as a first simple quality indicator for reporting differences between hospitals. However, applying this simple indicator on previous years showed two serious shortcomings. The first one is that it does not correct for the trend, over all hospitals, in the number of reported comorbidities over time. The second, most serious shortcoming is that it is unclear to what extent changes in the average number of comorbidities per hospital stay affect the estimated HSMR changes of that hospital. The reason is that this effect depends on the patient composition of the hospital. We therefore aim to develop a detection method that overcomes these current shortcomings.

3. Methodology

3.1. Basic Approach

Consider a population U_h of units i ($i = 1, \ldots, N_h$) that are reported in a decentralised administrative data source by data supplier h. Let $\mathcal{Y} = \{y_1, \ldots, y_m, \ldots, y_M\}$ be a set of M binary variables that are prone to under- or overreporting. Further, let the obtained values for the variables y_m for unit i of data supplier h be contained in the vector $\boldsymbol{y}_{hi} = (y_{1hi}, \ldots, y_{mhi}, \ldots, y_{Mhi})^T$. Also, let $\mathcal{Z} = \{z_1, \ldots, z_l, \ldots, z_L\}$ be a set of L covariates (continuous or categorical) for which it is reasonable to assume that they are error-free. Further, let $\boldsymbol{z}_{hi} = (z_{1hi}, \ldots, z_{lhi}, \ldots, z_{Lhi})^T$ be the corresponding vector with the obtained values for unit i of data supplier h. For instance, in the case study, the variables age, sex, socio-economic status and mortality for admissions i of hospital h were considered to be error-free. Further, let θ be the target parameter of interest. In our case study, we have the special situation that we publish a target parameter for each data supplier, denoted by θ_h, but with some minor adaptations our method can also be applied when there is one common target parameter. The target parameter is estimated as $\hat{\theta}_h$, which is a function of the variables y_m and z_l. Throughout the article a hat is used to indicate an estimate.

We aim to compute the effect of under- and overreporting on $\hat{\theta}_h$ for the variables y_m with $m = 1, \ldots, M$. We apply the following four steps to estimate the effect of under- and overreporting (the exact description is given in the next sections):

1. Select a group r of reference suppliers with similar reporting behaviour for the variables y_m. One might use multiple reference groups to analyse the sensitivity of the outcomes to the selected reference group;
2. For the units in the reference group, predict the probability that $y_{mhi} = 1$ given a set of covariates. Use the regression coefficients to predict the probabilities for the nonreference suppliers;

3. Use the predictions of 2) in a predictive mean matching imputation algorithm. If the observed y_{mhi} values of the nonreference suppliers differ significantly from the expected ones, replace them with the reference suppliers' values;
4. Compute the change in the target parameter between two periods for data supplier h as a function of the original $\mathbf{y}_{hi}$ and $\mathbf{z}_{hi}$ values and recompute this change using the imputed values. The difference between those two changes is a measure for the effect of the reporting behaviour of data supplier h on the outcomes.

The four steps are schematically represented in a flow chart, see Figure 1. The details of the steps, for instance the loop over units i in step 3, are explained in the next sections.

3.2. *Select a Reference Group*

We define a model for the variables y_m $(m = 1, \ldots, M)$ to describe the reporting behaviour of the data suppliers. This will be used to select data suppliers with a comparable reporting behaviour. When the intensity of reporting behaviour is expected to be the same for the set of variables y_m one can combine these variables into a single

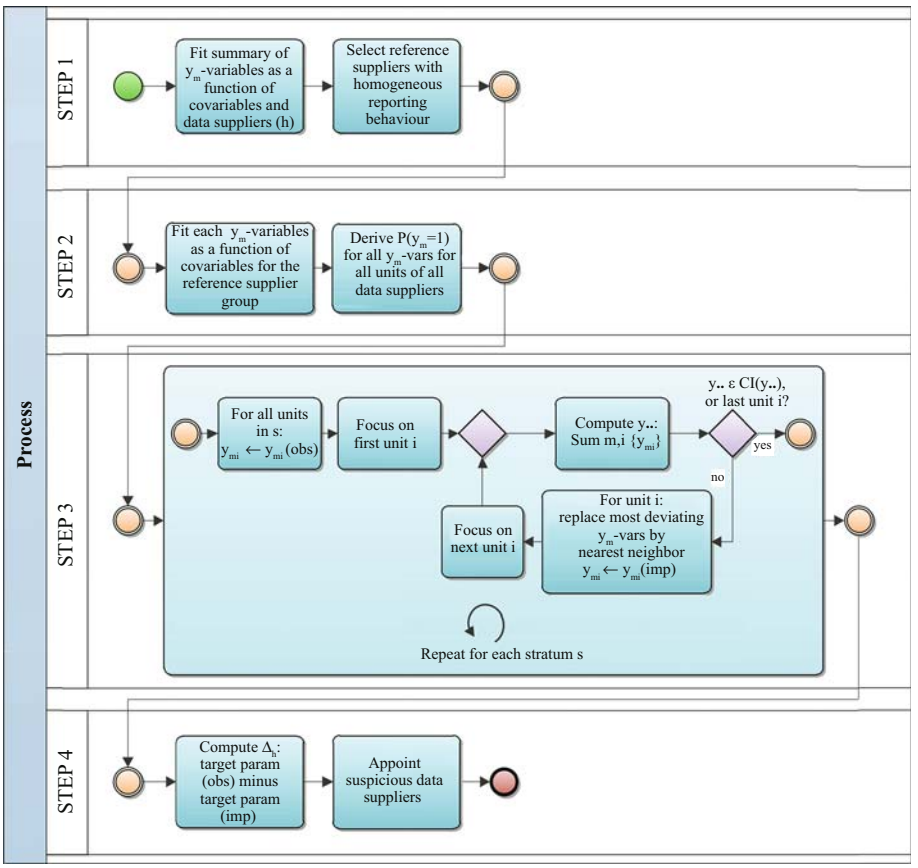

Fig. 1. Flow chart of the four steps of our methodology. The symbols in the chart are simplified compared to the main text; CI stands for confidence interval; the meaning of stratum s is explained in Subsection 3.4.

summary measure. We denote this summary measure as:

$$y_{\bullet hi} = g(\boldsymbol{y}_{hi})$$

where subscript "$\bullet$" denotes that it summarises over a set of variables. A summary variable can be

$$y_{\bullet hi}^{(1)} = 1 - \prod_{m=1}^{M} (1 - y_{mhi}). \tag{1}$$

Thus $y_{\bullet hi}^{(1)}$ equals 0 when $y_{mhi} = 0$ for all y_m variables and it equals 1 otherwise. Alternatively, one might use $y_{\bullet hi}^{(2)} = \sum_{m=1}^{M} y_{mhi}$, which stands for the number of variables with a score of 1. When the variables y_m are not related to each other or when the reporting intensity is expected to vary considerably among the y_m variables, it is better to analyse the effect of reporting behaviour for one variable at a time. In the remainder of this article, we limit ourselves to the analysis of reporting behaviour on a set of variables, because analysing one variable at a time is a special case of this.

In order to assess differences in reporting behaviour among data suppliers, one needs to correct for differences in the population on which they report. We use covariates to capture this population composition. These covariates may coincide with the error-free variables z_l ($l = 1, \ldots, L$) within the administrative data set, but they may also be amended with error-free variables that are not available in the administrative data set at hand. It is important that those variables are error-free to assure unbiased estimates for the supplier effects (the $\hat{\boldsymbol{\gamma}}$ in (2), see below). In the discussion we give some suggestions for the situation that the covariates contain measurement errors. We denote the set of covariates by $x = \{x_1, \ldots, x_k, \ldots, x_K\}$ and their obtained values for unit i of data supplier h are denoted by $\boldsymbol{x}_{hi} = (x_{1hi}, \ldots, x_{khi}, \ldots, x_{Khi})^T$. Let I_{hi} be an indicator variable that is 1 if unit i belongs to data supplier h ($h = 1, \ldots, H$) and 0 otherwise. Let $\boldsymbol{\delta}_{hi}$ be the vector $\boldsymbol{\delta}_{hi} = (I_{1i}, \ldots, I_{Hi})^T$. Further, let $P\left(y_{\bullet hi}^{(1)} = 1\right)$ denote the probability that $y_{\bullet hi}^{(1)} = 1$.

We estimate the data supplier effect on the reporting behaviour using the logistic model:

$$\text{logit}\left\{\hat{P}\left(y_{\bullet hi}^{(1)} = 1 | \boldsymbol{x}_{hi}, \boldsymbol{\delta}_{hi}\right)\right\} = (\boldsymbol{x}_{hi})^T \hat{\boldsymbol{\beta}} + (\boldsymbol{\delta}_{hi})^T \hat{\boldsymbol{\gamma}}, \tag{2}$$

where $\hat{\boldsymbol{\beta}}$ is the vector of estimated regression coefficients concerning the covariates (including the intercept) and $\hat{\boldsymbol{\gamma}} = [\gamma_h]$ is the vector with the estimated data supplier effects. With Equation (2) we assume that there is an overall effect γ_h on a set of binary y_m variables ($m = 1, \ldots, M$) due to the administrative practice of the data supplier. As an alternative to (2) one might estimate the data supplier effect for summary variable $y_{\bullet hi}^{(2)}$ with a simple linear model. When the decentralised data also contains data suppliers that report for a smaller number of units, random effects models might give better estimates of γ_h (see discussion). Note that large γ_h values indicate high reporting levels, whereas small values stand for the opposite.

For (each) reference group r, we aim to select data suppliers with similar γ_h values. Since we are interested in changes of a target parameter ($\hat{\theta}_h$) as affected by shifts in reporting behaviour (between two subsequent periods), we estimate Equation (2) for two subsequent periods and select data suppliers with similar values over two periods. A directly related issue concerns the choice of the group size. This size should, on the one

hand, be small enough to reduce the variability in reporting behaviour within the set, but on the other hand it should be large enough to reliably predict the variables with reporting patterns. See Subsection 4.2 how we operationalised "similar γ_h values" and the group size for the case study.

Note that we regard the computed value for $\hat{P}\left(y^{(1)}_{\bullet hi} = 1 | \boldsymbol{x}_{hi}, \boldsymbol{\delta}_{hi}\right)$ in (2) to be an estimate, although it is derived from administrative data that covers the complete target population. The reason is that we are interested in the reporting behaviour of the data supplier concerning the y_m variables. We regard this reporting behaviour to be an unknown property; the obtained observations can be seen as "input" to monitor this reporting behaviour.

3.3. Predict the Variables with Data Supplier Effects

In the second step, we predict the scores for each of the variables y_m ($m = 1, \ldots, M$) for reference group r and judge how well the models fit. A good model fit leads to a better result for the next step: predictive mean matching. Let d be a domain, that is, a category of one variable or a category of a cross-classification of multiple categorical variables. Domains are used when the effect of the covariates on the error-prone y_m variables are expected to vary (over domains). Domains thus capture interactions between the population composition variables with respect to their effect on reporting intensity. Further, let y_{mhdi} be the score for unit i on variable y_m for data supplier h and domain d. Let U_{rd} be the set of units of reference group r within domain d. Denote by $p^{(r)}_{mhdi} = P\left(y^{(r)}_{mhdi} = 1 | \boldsymbol{x}_{hdi}\right)$ the probability that $y_{mhdi} = 1$ for reference group r given the values of a set of covariates. For the set of units $i \in U_{rd}$ we estimate $p^{(r)}_{mhdi}$ by:

$$\text{logit}\left\{\hat{p}^{(r)}_{mhdi}\right\} = (\boldsymbol{x}_{hdi})^T \hat{\boldsymbol{\beta}}^{(r)}_{md} \quad (i \in U_{rd},\, m = 1, \ldots, M) \tag{3}$$

where $\hat{\boldsymbol{\beta}}^{(r)}_{md}$ stands for the estimated regression coefficients that depend on reference group r, variable y_m and domain d. The periods, for instance years, can be included as dummy variables in $\boldsymbol{x}_{hdi}$, which means that the model captures that reporting behaviour for each of the variables y_m may vary with year. Next, also compute $\hat{p}^{(r)}_{mhdi}$ for the nonreference suppliers based on the same regression coefficients $\hat{\boldsymbol{\beta}}^{(r)}_{md}$ in (3). Note that Equation (3), in contrast to Equation (2), does not contain a data supplier effect ($\hat{\boldsymbol{\gamma}}$). The reason is that we wish to model how the comorbidity probabilities $\hat{p}^{(r)}_{mhdi}$ depend on a set of error-free background variables for a set of units $i \in U_{rd}$ that have a similar reporting behaviour.

We used the C-statistic as an evaluation criterion for the predictive validity of the logistic regressions. The C-statistic lies between 0.5 and 1. As a rule of thumb, values of 0.7 to 0.8 indicate an acceptable discrimination and values above 0.9 show an outstanding discrimination (Hosmer and Lemeshow, 2004).

3.4. Predictive Mean Matching

For ease of notation, in the remainder of the article, we will drop the super- and subscripts r, h and d from the notation of the variables, unless we need them to explain the equations. Thus, for instance $\hat{p}^{(r)}_{mhdi}$ will be abbreviated as $\hat{p}_{mi}$.

In order to analyse the effect of reporting behaviour on the target parameter, we did not directly replace the originally observed y_{mi} values by their $\hat{p}_{mi}$ values, for two reasons:

1. the $\hat{p}_{mi}$ values are estimated for each of the variables y_m separately without accounting for their covariances;
2. we wanted to replace the original data only when the existing values differed clearly from their expected values (see below).

Note that in the HSMR case study, the target variable θ is a nonlinear function of the binary variables y_m $(m = 1, \ldots, M)$ and z_l $(l = 1, \ldots, L)$ (see Section 7). Therefore, directly using the $\hat{p}_{mi}$ will not yield the same outcome as using the binary variables themselves.

We use a nearest neighbour hot deck imputation method, whereby the reference suppliers act as donors and the nonreference suppliers as recipients. We use predictive mean matching as our hot deck imputation method (De Waal et al. 2011). We do not impute all units of the nonreference suppliers: our baseline is that we keep the originally supplied data untouched as much as possible, unless there is a large difference between observed and expected values (similar to selective editing).

Before explaining the algorithm, we introduce some additional notation. Let $\mathcal{H}$ denote the full set of data suppliers and let $\mathcal{R}^{(r)}$ denote the set of reference suppliers for reference group r. Thus, $\mathcal{H} \backslash \mathcal{R}^{(r)}$ stands for the group of nonreference suppliers in case of reference group r. The imputation algorithm is repeated for each combination of reference group r, nonreference data supplier h, domain d and period t. We will refer to this combination by "stratum s" and the set of units in a stratum is denoted by U_s and its size by N_s. Within each stratum U_s we will impute the units one by one. After each imputation, we check the difference between observed and expected values to decide whether or not a new unit is to be imputed, see step three below. Let $\ell = 0, 1, \ldots, \mathcal{L}$ (with $\mathcal{L} \leq N_s$) be an index that counts the number of units that have been imputed (so far). Let $\check{y}_{mi}$ denote an imputed value (0 or 1) for variable y_m $(m = 1, \ldots, M)$ of unit i and let $\tilde{y}_{mi}^{(\ell)}$ denote the actual value of unit i when ℓ units have been imputed, that is

$$\tilde{y}_{mi}^{(\ell)} = \begin{cases} y_{mi} \text{ if not imputed, given that } \ell \text{ units have been imputed} \\ \check{y}_{mi} \text{ if imputed, given that } \ell \text{ units have been imputed} \end{cases} \tag{4}$$

The imputation algorithm consists of three steps:

1. For the units of all nonreference data suppliers ($h \in \mathcal{H} \backslash \mathcal{R}^{(r)}$) compute the sum $y_{\bullet i} = \sum_{m=1}^{M} y_{mi}$. Likewise, compute the expected value as $\hat{E}(y_{\bullet i}) = \sum_{m=1}^{M} \hat{p}_{mi}$. Denote its difference by $\hat{\omega}_{\bullet i} = y_{\bullet i} - \hat{E}(y_{\bullet i})$. Additionally, compute $\hat{E}(y_{\bullet\bullet}) = \sum_{i \in U_s} \hat{E}(y_{\bullet i})$, which is the expected total of $y_{\bullet i}$ within stratum s. Thus, in our case study, the total $y_{\bullet\bullet}$ stands for the number of registered comorbidities over all admissions i in nonreference hospital h and main diagnosis d and year t, and the expectation of $y_{\bullet\bullet}$ is determined for each reference group r.
   Let $V(y_{\bullet\bullet})$ denote the variance of $y_{\bullet\bullet}$. Further, let $\mathrm{L}(y_{\bullet\bullet})$ denote the lower and $\mathrm{U}(y_{\bullet\bullet})$ the upper bound of an (approximate) 95%-confidence interval for $\hat{E}(y_{\bullet\bullet})$.

When the stratum size N_s is large we can estimate these bounds by:

$$\hat{\mathrm{L}}(y_{\bullet\bullet}) = \hat{E}(y_{\bullet\bullet}) - 1.96\sqrt{\hat{V}(y_{\bullet\bullet})} \tag{5}$$

$$\hat{\mathrm{U}}(y_{\bullet\bullet}) = \hat{E}(y_{\bullet\bullet}) + 1.96\sqrt{\hat{V}(y_{\bullet\bullet})} \tag{6}$$

We now derive an expression for $V(y_{\bullet\bullet})$. y_{mi} follows a Bernoulli distribution with $E(y_{mi}) = p_{mi}$. Let $E(y_{mi}y_{ni}) = p_{mni}$. We then find $V(y_{\bullet i}) = \sum_{n=1}^{M}\sum_{m=1}^{M}(p_{mni} - p_{mi}p_{ni})$; for $m = n$ we obtain $p_{mmi} = p_{mi}$. Because the y_{mi} variables are independent across units, $V(y_{\bullet\bullet}) = \sum_{i=1}^{N_s}\sum_{n=1}^{M}\sum_{m=1}^{M}(p_{mni} - p_{mi}p_{ni})$. The values that are generated by (3), $\hat{p}_{mi}$, do not account for interactions between the variables, which implies that we use the approximation $\hat{p}_{mni} = \hat{p}_{mi}\hat{p}_{ni}$ for $m \neq n$. This leads to $\hat{V}(y_{\bullet\bullet}) = \sum_{i=1}^{N_s}\sum_{m=1}^{M}\left(\hat{p}_{mi} - \hat{p}_{mi}^2\right)$, which is an approximation of $V(y_{\bullet\bullet})$. When the y_{mi} variables are positively correlated, $V(y_{\bullet\bullet})$ is underestimated. When they are negatively correlated, $V(y_{\bullet\bullet})$ is overestimated.

2. Let u denote a recipient unit that belongs to the nonreference suppliers and let v be a donor unit that belongs to the reference suppliers. We seek a donor v for recipient u such that the sum of the observed values y_{mv} of the donor will be close to the expected sum of y_{mu} for the recipient. Since this expected sum, $\hat{E}(y_{\bullet u})$, follows from the corresponding probabilities, we select a donor by using the Euclidean distance between $\hat{p}_{mu}$ and $\hat{p}_{mv}$: $\sqrt{\sum_{m=1}^{M}(\hat{p}_{mv} - \hat{p}_{mu})^2}$.
3. Within each stratum s for the nonreference data suppliers:
   a. Set $\ell = 0$;
   b. Compute the totals $\tilde{y}_{\bullet\bullet}^{(\ell)} = \sum_{i\in Us}\sum_{m=1}^{M}\tilde{y}_{mi}^{(\ell)}$ of the actual scores (including imputations). If $\tilde{y}_{\bullet\bullet}^{(\ell)} \notin [\hat{\mathrm{L}}(y_{\bullet\bullet}), \hat{\mathrm{U}}(y_{\bullet\bullet})]$ continue with c), otherwise stop.
   c. If $\tilde{y}_{\bullet\bullet}^{(\ell)} > \hat{E}(y_{\bullet\bullet})$ then determine unit u, from the set of units in s that have not been imputed (so far), with the largest value of $|\hat{\omega}_{\bullet u}|$ given that $\hat{\omega}_{\bullet u} > 0$. Likewise, if $\tilde{y}_{\bullet\bullet}^{(\ell)} < \hat{E}(y_{\bullet\bullet})$ then determine unit u with the largest value of $|\hat{\omega}_{\bullet u}|$ given that $\hat{\omega}_{\bullet u} < 0$. Denote this unit by u_0. For recipient u_0, determine the closest donor v_0 according to the Euclidean distance and impute its values $\boldsymbol{y}_{v_0}$. So, the values for all y_m variables from donor v_0 are imputed.
   d. Let $\ell = \ell + 1$ and go to b.

3.5. Compute the Imputed Target Parameter

Let superscript 0 denote the parameters that are based on the original input values $\boldsymbol{z}_{hi}$ and $\boldsymbol{y}_{hi}$. Let $\hat{\theta}^{t,0}$ denote the target parameter of interest for period t, based on a function of the original input values $\boldsymbol{z}_{hi}$ and $\boldsymbol{y}_{hi}$ for all data suppliers $h \in \mathcal{H}$ and of the estimated model parameters $\hat{\boldsymbol{\beta}}^0$. Note that this function can be a simple sum of $\boldsymbol{y}_{hi}$ or a complex function, which is the case with the HSMR. We now evaluate the effect of the reporting behaviour of one specific data supplier, h_1, by replacing the original input values for that data supplier by its imputed values. Next, we reestimate the target parameter, denoted by $\hat{\theta}^{t,imp(h_1)}$ based on the imputed values for h_1 and the original values for all other data suppliers. Since we

aim to evaluate estimated changes, we compare the change based on the original input values, $\hat{\theta}^{t,0} - \hat{\theta}^{t-1,0}$, with the imputed version: $\hat{\theta}^{t,imp(h_1)} - \hat{\theta}^{t-1,imp(h_1)}$. We denote its difference by $\hat{\Delta}^{t,t-1(h_1)} = (\hat{\theta}^{t,0} - \hat{\theta}^{t-1,0}) - (\hat{\theta}^{t,imp(h_1)} - \hat{\theta}^{t-1,imp(h_1)})$. We analyse $\hat{\Delta}^{t,t-1(h)}$ for all nonreference data suppliers $h \in \mathcal{H}\backslash\mathcal{R}^{(r)}$. In the case study, we have the special situation that we have a target parameter $\hat{\theta}_h^0$, but the model parameters $\hat{\boldsymbol{\beta}}^0$ continue to be based on all $h \in \mathcal{H}$. This leads to a small modification, which is explained in Subsection 4.2.

Since our method aims to select nonreference data suppliers with extreme values of $\hat{\Delta}^{t,t-1(h)}$, we wish to have a practical rule to appoint which values we consider to be extreme. To that end, we assume that for data suppliers h that are free of under- or overreporting, the values of $\hat{\Delta}^{t,t-1(h)}$ are approximately normally distributed: $\hat{\Delta}^{t,t-1(h)} \sim N\left(0, \sigma^2_{\hat{\Delta}}\right)$. The expected value of $\hat{\Delta}^{t,t-1(h)}$ is taken to be 0, since we expect that the y_m values $(m = 1, \ldots, M)$ of data suppliers without under- or overreporting are not imputed. Further, $\sigma^2_{\hat{\Delta}}$ stands for variation in the outcomes of $\hat{\Delta}^{t,t-1(h)}$ that cannot be explained by the covariates used in the regression. Here, we limit the estimation of $\sigma^2_{\hat{\Delta}}$ to the situation that we have two reference groups r; it can easily be extended to a situation with more reference groups. The situation of a single reference group is treated in the discussion.

Denote the first reference group by A, its corresponding set of data suppliers by $\mathcal{R}^{(A)}$ and its size by $N_{\mathcal{R}^{(A)}}$. Likewise, we denote the second reference group by B with set $\mathcal{R}^{(B)}$ of size $N_{\mathcal{R}^{(B)}}$. When A is selected as the reference group, values of $\hat{\Delta}^{t,t-1(h)}$ are not available for $\mathcal{R}^{(A)}$, since only the values of the nonreference suppliers are imputed, but they are available for $\mathcal{R}^{(B)}$. We consider the variation in $\hat{\Delta}^{t,t-1(h)}$ for $\mathcal{R}^{(B)}$ when A is the reference group as an estimate of $\sigma^2_{\hat{\Delta}}$, since we have selected the set within a reference group to be (more or less) homogeneous in reporting behaviour. We define:

$$s^2_{\hat{\Delta}}(\mathcal{R}^{(B)}|r = A) = \frac{1}{N_{\mathcal{R}^{(B)}} - 1} \sum_{h \in \mathcal{R}^{(B)}} (\hat{\Delta}^{t,t-1(h)} - \hat{\bar{\Delta}}^{t,t-1(B)})^2 \tag{7}$$

with $\hat{\bar{\Delta}}^{t,t-1(B)} = \frac{1}{N_{\mathcal{R}^{(B)}}} \sum_{h \in \mathcal{R}^{(B)}} \hat{\Delta}^{t,t-1(h)}$. Note that in (7) we used the sample mean $\hat{\bar{\Delta}}^{t,t-1(B)}$ with $N_{\mathcal{R}^{(B)}} - 1$ degrees of freedom rather than using the expected value "0". Likewise, we define:

$$s^2_{\hat{\Delta}}(\mathcal{R}^{(A)}|r = B) = \frac{1}{N_{\mathcal{R}^{(A)}} - 1} \sum_{h \in \mathcal{R}^{(A)}} (\hat{\Delta}^{t,t-1(h)} - \hat{\bar{\Delta}}^{t,t-1(A)})^2 \tag{8}$$

We now estimate $\sigma^2_{\hat{\Delta}}$ as the pooled estimate of $s^2_{\hat{\Delta}}(\mathcal{R}^{(B)}|r = A)$ and $s^2_{\hat{\Delta}}(\mathcal{R}^{(A)}|r = B)$:

$$\hat{\sigma}^2_{\hat{\Delta}} = \frac{(N_{\mathcal{R}^{(A)}} - 1) s^2_{\hat{\Delta}}(\mathcal{R}^{(A)}|r = B) + (N_{\mathcal{R}^{(B)}} - 1) s^2_{\hat{\Delta}}(\mathcal{R}^{(B)}|r = A)}{N_{\mathcal{R}^{(A)}} + N_{\mathcal{R}^{(B)}} - 2} \tag{9}$$

Using $\hat{\sigma}^2_{\hat{\Delta}}$, we construct an (approximate) 95%-confidence interval for $\hat{\Delta}^{t,t-1(h)}$ as $0 \pm 1.96\sqrt{\hat{\sigma}^2_{\hat{\Delta}}}$. Data suppliers outside this confidence interval are considered to have a deviating reporting behaviour.

4. Application to the Case Study

The method described in the previous section was applied to the HSMR case study, which is calculated from the LBZ data. The HSMR is limited to hospital stays, also called inpatient admissions. Day admissions are excluded, since they are usually non-life-threatening. As was mentioned in the introduction, there are strong indications that some of the variables in this data set are affected by reporting differences between the hospitals, which affects the output. The next subsections describe the administrative data and how our method is applied to the HSMR case study. The method of calculating the HSMR is given in the Appendix. Section 5 describes the results.

4.1. LBZ Data

We used LBZ data for the consecutive years 2011 and 2012 with a total of 1,221,414 inpatient admissions. We wanted to use data near 2010, which was announced to be the first year when the HSMR would become publicly available. We found clear shifts in intensity of comorbidity reporting by some of the hospitals a few years before and after that period. Although the HSMR model (see Section 7, Appendix) is normally estimated over three years, we did not include 2013 or more recent years as most hospitals switched to a new coding system (ICD10) in 2013, which might affect the results (Van der Laan et al. 2015). In total, 83 hospitals provided LBZ data for both 2011 and 2012. Four of these hospitals submitted data of poor quality (an incomplete data set). One hospital was very specialised with only a few main diagnoses. We excluded those five hospitals and used a net population of 78 hospitals in the analysis.

4.2. Analysis of the Reporting Effects

The target parameter θ_h is in this case the HSMR of hospital h (see Appendix and Van der Laan et al. 2015 for how this parameter is calculated). The HSMR is the ratio between the observed mortality and the expected mortality. The expected mortality is calculated using a logistic regression model for mortality at patient level using background properties such as age, sex and comorbidities. We studied the effect of reporting intensity on the comorbidity variables. These comorbidity variables were grouped into 17 Charlson groups (see Section 7, Appendix). So in the case study, the variable y_m ($m = 1, \ldots, M$) stands for one of the 17 Charlson groups for admission i of hospital h. We now describe how we applied Subsections 3.2–3.5 to our case study.

4.2.1. First Step

We asked experts which hospitals they thought were expected to have correct comorbidity reporting, but they were unable to answer the question. Therefore, we decided to select *two* reference groups, to be able to determine the sensitivity of the outcomes for the choice of the reference group. We selected a "middle group" representing average levels of comorbidity reporting and a "top group" representing high levels of comorbidity reporting.

To summarise the variables y_m we used $y^{(1)}_{\bullet hi}$, defined in Equation (1), which has a value 1 when at least one comorbidity code is given to admission i and 0 otherwise. We used the

logistic regression model in (2) to model $y^{(1)}_{\bullet hi}$ as a function of the hospital effect γ_h and of a set of diagnoses- and patient-related variables that are given in Table 1 in the Appendix (column "select reference hospitals"). The logistic regression was applied to 2011 and 2012 separately. Within a year, it was applied to all admissions of all hospitals. This means that we did not let the hospital effect vary with main diagnosis, but it did vary with year. We did so, because we were interested in estimating the overall hospital effect on reporting behaviour and because previous experience showed that hospitals may vary their reporting behaviour from one year to the next. Note that in Equation (2) we used a fixed hospital effect γ_h. Prins (2016) has also computed outcomes where the hospital effect was included as a random effect within a multilevel model, which yielded near-identical results.

We wanted to select two reference groups that were homogeneous in their reporting behaviour for two subsequent years (2011, 2012). Thus, we wanted the γ_h values not to vary too much from one year to the next. We computed the difference $d_h^{t,t-1} = \gamma_h^t - \gamma_h^{t-1}$ of the hospital effects, with $t = 2012$ and its variance: $V\left(d_h^{t,t-1}\right) = \frac{1}{H-1}\sum_{h=1}^{H}\left(d_h^{t,t-1} - \bar{d}_h^{t,t-1}\right)^2$, with $\bar{d}_h^{t,t-1} = \frac{1}{H}\sum_{h=1}^{H} d_h^{t,t-1}$. Next, we computed an (approximate) 80% confidence interval according to $\bar{d}_h^{t,t-1} \pm 1.28\sqrt{V\left(d_h^{t,t-1}\right)}$, based on a Normal distribution. Hospitals with $d_h^{t,t-1}$ values outside this interval were excluded from the reference group. For the "middle reference group" we selected the hospitals within the 80% confidence interval whose absolute γ_h^{2012} values were closest to 0. For the "top reference group" we selected the hospitals within the 80% confidence interval with the largest γ_h^{2012} values. A reference group size of 15 (we investigated 10, 15 and 20) was found to be the smallest group size for which the models could be reasonably accurately

Table 1. Variables used in the various computations.

Variable (no of classes[1])	HSMR model	Select reference hospitals	Predict each comorbidity (15)
Age (5-year classes)	x	x	x (5 knot spline)
Comorbidity group (17)	x		
Hospital (78)		x	
Main diagnosis (50)	*	x	*
Medical specialty (44)			x
(bi-) Month of admission (6)	x	x	x
Re-admission (2)	x	x	x
Reason of admission (3)			x
Sex (2)	x	x	x
Severity main diagnosis (9)	x	x	x
Social-economic status (6)	x	x	x
Source of admission (3)	x	x	x
Type of hospital (2)			x
Urgency (2)	x	x	x
Year of discharge (2)	x	*	x

x: variable included as independent variable in the regression; *: class for which the regression is run separately.
[1] an explanation of the categories can be found in Israëls et al. (2012) and van der Laan (2013)

estimated (all categories reasonably filled; few categories in the recipient group that were not present in the donor group).

4.2.2. Second Step

The second step in the case study was to predict each of the comorbidity variables y_m as a logistic function of patient- and disease-related variables, according to Equation (3). Because the occurrence of comorbidities varied greatly with main diagnosis, the model was fitted separately for each main diagnosis. Thus, main diagnosis represents domain d in Equation (3). For instance, Charlson group 15 (HIV) occurs mainly with main diagnosis 38 (non-Hodgkins lymphoma), see Van der Laan et al. (2015). The patient- and disease-related variables used in this second step are given in the final column of Table 1 in the Appendix. In addition to the covariates that we used to select the reference hospitals, we added medical specialty, type of hospital and reason of admission. For the regressions, two sets of comorbidity groups were very similar and therefore combined: Charlson comorbidity group 17 (Severe liver disease) was combined with Charlson comorbidity group 9 (Liver disease) and Charlson comorbidity group 11 (Diabetes complications) was combined with Charlson comorbidity group 10 (Diabetes), leading to a total of 15 groups (see Table 1, final column). Charlson groups 17 and 11 occur only rarely, and a preliminary analysis showed that merging these comorbidity groups had only a minor effect on the HSMR outcomes (Israëls et al. 2012).

4.2.3. Third Step

The third step was to apply the imputation algorithm. To estimate $V(y_{\bullet\bullet})$ we assumed that the probabilities p_{mi} for most of the variables are very small. We assumed that $\hat{p}_{mi} - \hat{p}_{mi}^2 \approx \hat{p}_{mi}$. Recall from Subsection 3.4 that $\hat{V}(y_{\bullet\bullet}) = \sum_{i=1}^{N_s} \sum_{m=1}^{M} (\hat{p}_{mi} - \hat{p}_{mi}^2)$. We now approximate this variance by $\hat{V}(y_{\bullet\bullet}) \approx \sum_{i=1}^{N_s} \sum_{m=1}^{M} \hat{p}_{mi} = \hat{E}(y_{\bullet\bullet})$. Note that the same variance would have been obtained by assuming that the data are Poisson-distributed.

In the case study, we used the Euclidean distance between the probabilities $\hat{p}_{mu}$ of recipient u and $\hat{p}_{mv}$ of donor v. We explain why we use a distance function based on probabilities in Subsection 3.4 (step 2 of the imputation algorithm). In preliminary computations, we also performed the imputation algorithm using the Euclidean distance between the logit of the probabilities, which resulted in near identical results.

4.2.4. Fourth Step

The fourth and final step was to compute the HSMR for each of the hospitals, according to Equations (10) and (11) in the Appendix, based on the observed and the imputed comorbidity scores. Let $\hat{\theta}_h^{t,0}$ denote the HSMR based on the original input values for hospital h and year t and let $\hat{\theta}_h^{t,imp}$ be its imputed version. $\hat{\theta}_h^{t,0}$ is estimated according to the logistic regression for the HSMR given in (11). The imputed HSMR for a specific data supplier h_1, $\hat{\theta}_{h_1}^{t,imp}$, is defined in Subsection 3.5 as the outcome of (11), based on the values $\tilde{y}_{h_1 mi}$ and the original values for the other input variables, combined with the original values for all other data suppliers $h \neq h_1$. Since the data set is changed only slightly, we assume that we can ignore the change in the regression coefficients when we impute

only one data supplier. Therefore, we use the original regression coefficients when calculating $\hat{\theta}_h^{t,imp}$.

Finally, we compared the HSMR change based on the original comorbidity values, $\hat{\theta}_h^{2012,0} - \hat{\theta}_h^{2011,0}$, with the change based on the imputed version: $\hat{\theta}_h^{2012,imp} - \hat{\theta}_h^{2011,imp}$. We denote its difference by $\hat{\Delta}_h^{2012,2011} = \left(\hat{\theta}_h^{2012,0} - \hat{\theta}_h^{2011,0}\right) - \left(\hat{\theta}_h^{2012,imp} - \hat{\theta}_h^{2011,imp}\right)$.

5. Results

5.1. Selection of Reference Group

The hospital effects in 2012 $\left(\gamma_h^{2012}\right)$ and the differences in those hospital effects over the two years $\left(\gamma_h^{2012} - \gamma_h^{2011}\right)$ are given in Figure 2. The γ_h^{2012} values ranged from -2.57 to 1.29. These values are the logarithm of the log-odds of the probabilities of reporting at least one comorbidity per hospital admission. These probabilities can be found by $\hat{P}\left(y_{\bullet hi}^{(1)} = 1 | \boldsymbol{x}_{hi}, \boldsymbol{\delta}_{hi}\right) = \frac{1}{1+\exp\left\{-(\boldsymbol{x}_{hi})^T \hat{\boldsymbol{\beta}}_d - (\boldsymbol{\delta}_{hi})^T \hat{\boldsymbol{\gamma}}\right\}}$ from Equation (2). Values of the logistic regression are given as the difference to a reference category (in case of categorical variables). So, for a patient admission that matches the reference category, the probability of having at least one comorbidity in 2012 among the hospitals ranged from $\frac{1}{1+\exp\{2.57\}} = 0.22$ to $\frac{1}{1+\exp\{-1.29\}} = 0.93$. These results indicate that there was considerable variation

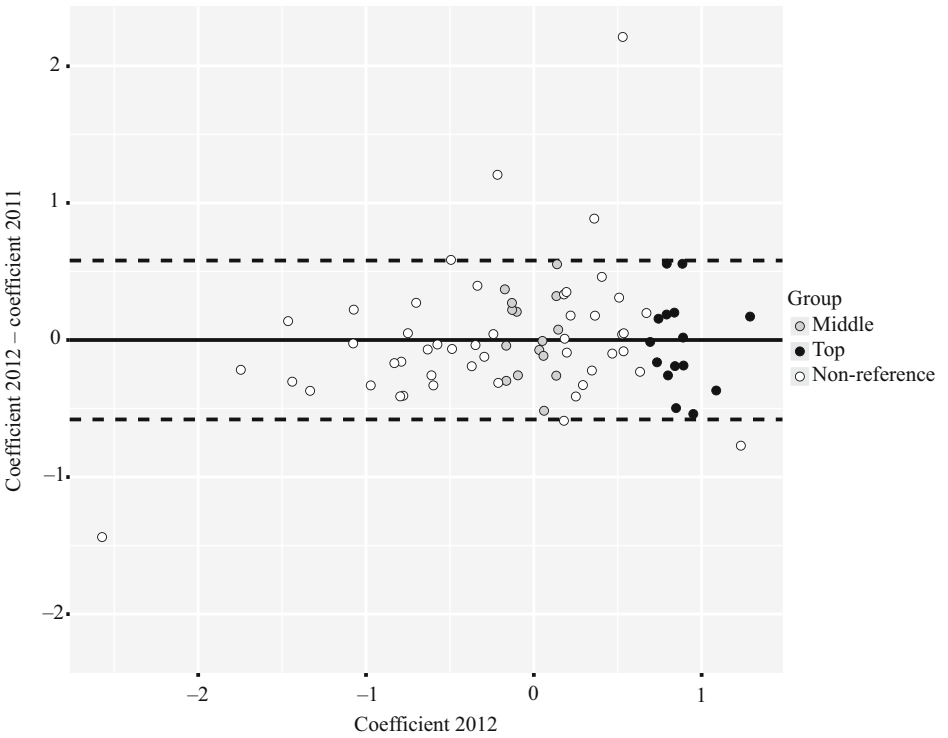

Fig. 2. The difference in the hospital effects (γ_h) of 2012 minus 2011 versus the hospital effects of 2012.

among the hospitals in intensity of comorbidity reporting, after correcting for differences in patient and diagnosis characteristics.

Given a group size of 15 units, the grey points in Figure 2 show the middle reference group and the black points show the top reference group. A standard deviation of 0.452 was found for $d_h^{2012,2011}$ leading to an 80% margin of $\pm$ 0.580. Using these margins as an additional selection criterion implied that one hospital was excluded from the top reference group (with $\gamma_h^{2011} = 2.00$ and $\gamma_h^{2012} = 1.23$) and no hospital was excluded from the middle group (the extreme value shown in Figure 2 with 0.179 in 2012 had order position 16).

5.2. *Prediction of the Incidence of the Charlson Groups*

The fit of the predicted probabilities $\left(\hat{p}_{mi};\ m = 1, \ldots, M\right)$ based on the admissions of the two reference groups according to (3) varied slightly between the different Charlson groups. The averages of the C-statistic for the middle and top groups were relatively small for Charlson group 6 (0.78, 0.71) and 10 (0.74, 0.72) whereas they were large for Charlson group 5 (0.89, 0.90), 8 (0.89, 0.89) and 9 (0.91, 0.87). Overall, the fraction of "main diagnosis $\times$ Charlson group" combinations with a C-statistic of at least 0.7 was 0.92 for the middle group and 0.86 for the top group. Since values of 0.7 and higher indicate an acceptable fit (see Subsection 3.3) we considered the results of the C-statistics to be sufficient to use the predicted probabilities ($\hat{p}_{mi}$) for predictive mean matching.

5.3. *Results of the Predictive Mean Matching*

Figure 3 displays for each hospital the distribution of the fraction-imputed records after applying the imputation algorithm across the 50 main diagnosis groups and the two years for both reference groups. We computed the average and third quantile per hospital of this distribution over diagnosis groups and years. The average per hospital was at most 0.16 (hospital 78) in case of the middle reference group and 0.24 (hospital 48) in case of the top reference group. Furthermore, the third quantile of this distribution was at most 0.24 (hospital 78) for the middle reference group and 0.34 (hospital 48) for the top reference group. These findings clearly show that only a limited number of records for each hospital were imputed, which is in line with the imputation approach that we intended (see Subsection 3.4). The minimum value of these averages over the set of recipient hospitals was 0.003 (middle reference group) and 0.015 (top reference group). This implies that for all recipient hospitals at least some records were imputed.

We also computed the average fraction of imputed records per main diagnosis group over the set of the recipient hospitals and the two years (not shown). For the middle reference group, the three smallest average fractions were 0.007, 0.010 and 0.015 and the three largest ones were 0.0977, 0.134 and 0.141. In the top reference group, the three smallest average fractions were 0.0208, 0.0255 and 0.0278 and the three largest ones were 0.181, 0.202 and 0.224. There were a few "main diagnosis $\times$ hospital" combinations where all admissions were imputed. This was not always for the same hospital or for the same diagnosis.

When a certain category of the covariates occurred in the recipient set that was absent in the donor set, the $\hat{p}_{mi}$ probabilities could not be predicted and those records were excluded

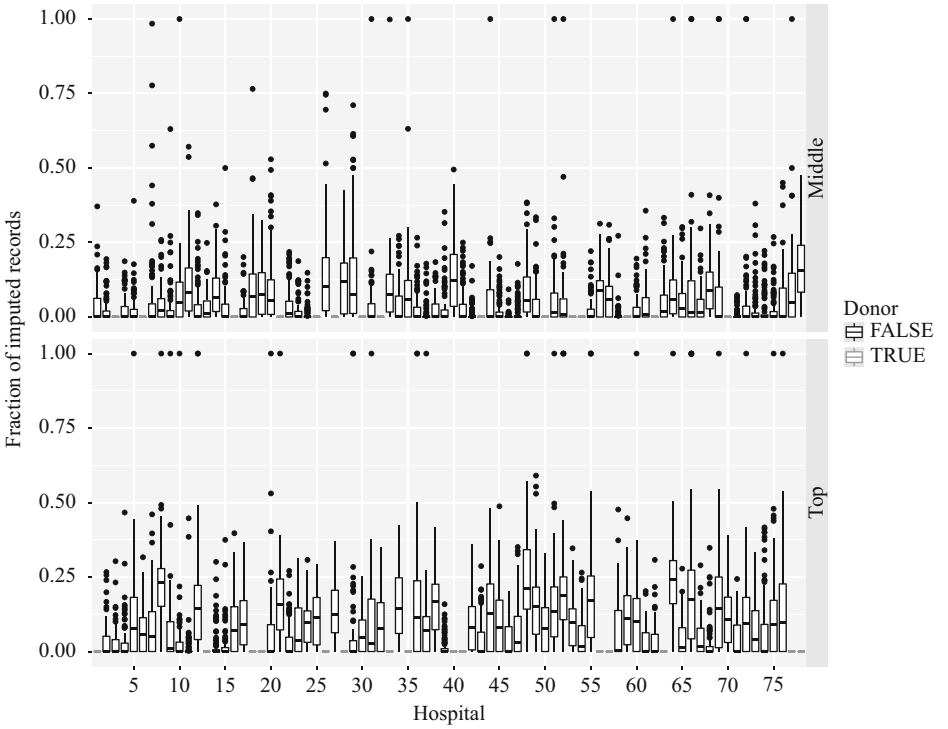

Fig. 3. Boxplot of the fraction of imputed records (distribution over 50 main diagnoses and both years) for each hospital with the middle (upper panel) or the top reference group (lower panel).

from the imputation algorithm. We computed the average, median, standard deviation, maximum and minimum fraction of units without a predicted comorbidity group score per main diagnosis. These values were 0.13, 0.070, 0.18, 0.93, 0.0058 for the middle and 0.080, 0.039, 0.14, 0.79, 0.0017 for the top reference group. In both reference groups, the median fraction of units without predicted scores was small, but in each reference group there were a few main diagnoses with a large fraction. The reason for this larger fraction was that those main diagnoses occurred mainly in certain categories of the patient- and diagnosis-related variables that were (by accident) absent in the reference group.

5.4. Computation of the Imputed HSMR

Production staff at CBS are interested in knowing to what extent the original HSMR development represents a change in the quality of hospital care or whether it results from a change in intensity of comorbidity reporting. The fraction of reported Charlson groups per admission for a given year is denoted by $\bar{y}_h$ and defined as $\bar{y}_h = \frac{1}{N_h M} \sum_{i \in Uh} \sum_{m=1}^{M} y_{mhi}$, where N_h stands for the number of admissions for hospital h. The development of $\bar{y}_h$ from 2011 to 2012 is denoted by $\bar{y}_h^{2012,2011}$, with $\bar{y}_h^{2012,2011} = \bar{y}_h^{2012} - \bar{y}_h^{2011}$. In Figure 4 we plotted $\hat{\Delta}_h^{2012,2011}$ against $\bar{y}_h^{2012,2011}$ and we fitted a simple linear regression through the data. We tested whether the slope differed from zero, under the assumption that the residuals are independent and identically distributed. We found a slope of -42.3 for the middle

reference group and of -55.7 for the top reference group at a p-value < 0.001. Recall that an increase in comorbidity reporting – everything else being the same – leads to a decrease in the HSMR. The latter represents an improvement in the quality of hospital care. The regression results imply that an increase in "the fraction of reported Charlson groups" $\left(\bar{y}_h^{2012,2011}\right)$ of 0.1 leads to an HSMR development which reduced by 4.2 points (middle group) or 5.5. point (top reference group). So, the hospitals that are plotted in the bottom-right of Figure 4 are hospitals with a large increase in comorbidity reporting from 2011 to 2012. It concerns hospitals where the original HSMR development $\left(\hat{\theta}_h^{2012,0} - \hat{\theta}_h^{2011,0}\right)$ is lower than the imputed one $\left(\hat{\theta}_h^{2012,imp} - \hat{\theta}_h^{2011,imp}\right)$, suggesting that their original improvement in the quality of hospital care was partly due to reporting effects.

Figure 5 shows that the $\hat{\Delta}_h^{2012,2011}$-values of the nonreference hospitals (circles) with the middle reference group were clearly related to those of the top group, with a correlation of 0.91. We found an estimated variance $s^2_{\hat{\Delta}}$ of 11.57 when the middle group was used as the reference group and of 15.24 when the top group was used as reference group. This resulted in a pooled variance of 13.41 and an estimated 95%-confidence interval of ± 7.17 index points. We thus found five hospitals with significant reporting effects. Three hospitals in the bottom-left of Figure 5 have a negative value for $\hat{\Delta}_h^{2012,2011}$ and in Figure 4

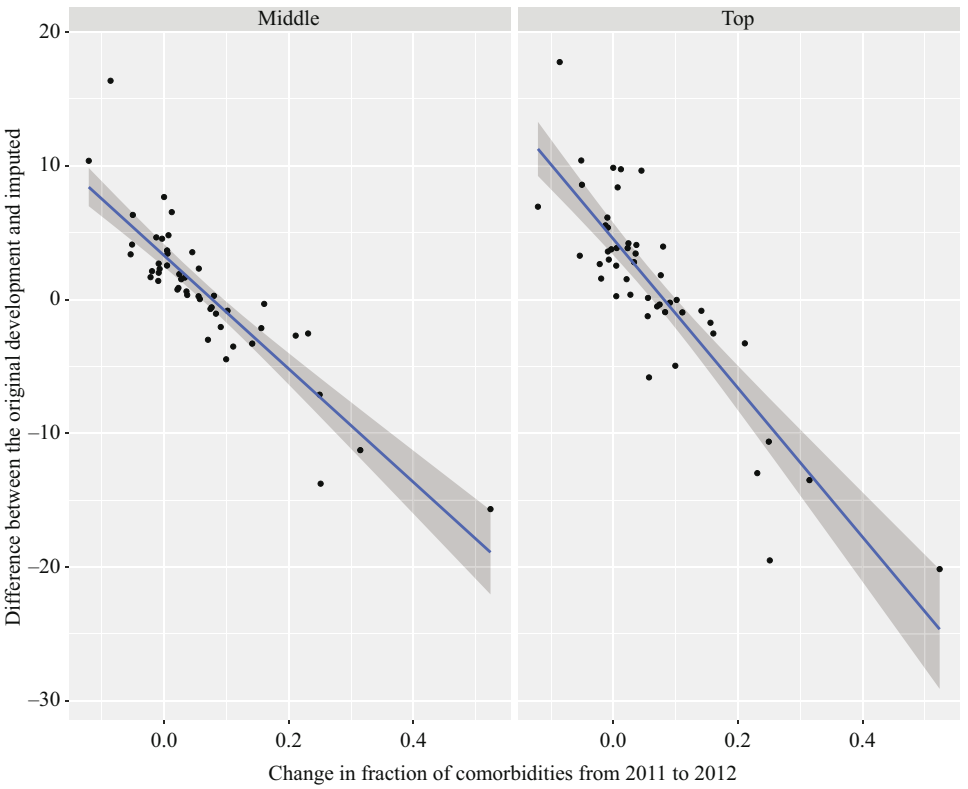

Fig. 4. Difference between the original and the imputed HSMR development as a function of the change in the fraction of comorbidities (2012 minus 2011) for both reference groups. Shaded areas represent the 95% confidence intervals of the linear regressions.

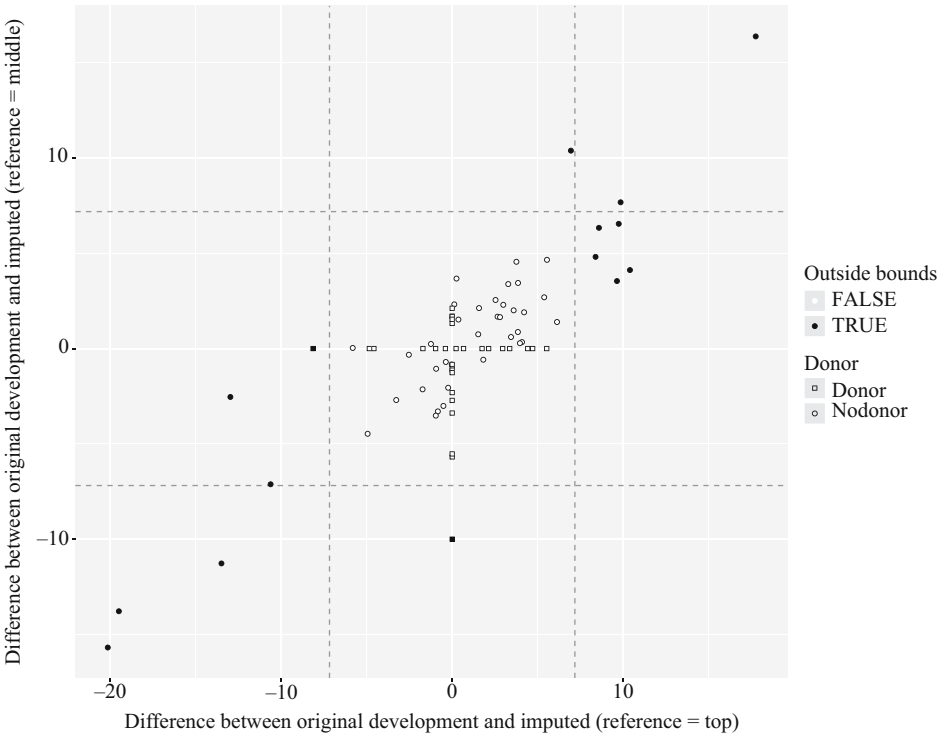

Fig. 5. Change (2012 – 2011) of the original minus that of the imputed HSMR $\left(\hat{\Delta}_h^{2012,2011}\right)$ *for the middle group versus the top group of reference hospitals. Broken lines represent confidence bounds based on* $\hat{\sigma}^2_{\hat{\Delta}}$.

we can see that it concerns three hospitals with a large increase in comorbidity reporting, suggesting that their original change in hospital care was "too positive". Two hospitals in the top-right have a positive value for $\hat{\Delta}_h^{2012,2011}$ suggesting that their original change in hospital care was "too negative". These two hospitals had a very low value of original comorbidity reporting in both years (not shown). These five hospitals are the suspicious hospitals to be contacted.

6. Discussion

We presented a method to detect under- and overreporting by data suppliers for decentralised administrative data in case of change estimates. With our approach, we estimated the impact of correlated measurement errors within a data supplier on the target outcomes. We successfully applied the method to administrative hospital data to detect hospitals that show large changes in reporting of the comorbidities of their patients. Previous studies have also found reporting differences among hospitals (Jarman 2008; Van der Laan 2013) but they were unable to estimate the impact of the reporting intensity on the outcomes. With the current method we expect to reduce the number of hours spent on manual data analysis. Moreover, we can contact the suspicious data suppliers, in order to improve the accuracy of future administrative data deliveries. The question remains how to proceed with the output based on the current data delivery. When reporting errors of a

variable that are widespread and severe, one might decide not to publish the outcomes that are based on this variable. When it concerns only a limiting number of data suppliers, then one might set those cases to "missing" and use a robust estimation, a weighting model or an imputation model to correct for it. In the special case of the HSMR, the output is at the level of the data supplier. When the quality of data delivered by a certain supplier is insufficient, one might publish a remark along with the outcomes, decide not to publish the output of that supplier, or exclude that data supplier from the data set, depending on the severity of the errors.

Our method is developed for a situation with decentralised administrative data, where it is possible to detect differences in reporting behaviour among data suppliers, but it is not possible to exactly pinpoint which units within the suspicious data supplier have measurement errors. This is opposed to the situation described in Van Delden and Scholtus (2017), where reporting patterns at unit level are detected with deterministic rules. That concerned turnover patterns derived from reported value added tax data. Our method requires that the total set of data suppliers $\mathcal{H}$ is large enough to set aside a group $\mathcal{R}$ that can act as reference suppliers.

A number of points are to be addressed before our method can be applied in statistical production and before it can be applied to forms of decentralised administrative data other than hospital data. First, tooling should be developed to enable analysts to perform the four steps in Subsection 3.1. Second, some practical guidance is needed in treating the decentralised data structure in the parameter estimates. Third, a practical application of our method would be enhanced by relaxing some of the assumptions and conditions of the currently reported method, because they may not hold in practice. Fourth, for application of the method to other decentralised data than the hospital data, it would be very useful to extend the method in terms of the types of variables it concerns, the types of errors treated and the forms of output. In the next four paragraphs we elaborate on the second, third and fourth point.

The decentralised, hierarchical, data structure needs to be accounted for in the estimation of the regression coefficients in step 1. In practical applications, one has to choose whether to treat the data supplier effects as random effects within a multilevel model or as fixed effects. In a data set where at least some of the data suppliers have a limited number of units per data supplier, we would prefer to model the data supplier effects as random effects, since one can then make use of the shrinkage factor (Efron and Morris 1975). In the HSMR case study, we had a large number of units for each data supplier. We found that treating the hospital effects as random or fixed effects yielded near-identical estimates, whereas the convergence of the latter model was much faster than that of the multi-level model, in line with Kim et al. (2013).

An example of a useful relaxation concerns the imputation algorithm (step 3). The current procedure does not account for the actually reported scores $y_{mi} = 1$. We propose the following refinement. Let $\mathcal{Y}_v$ be the set with $y_{mv} = 1$ (with $m = 1, \ldots, M$), for recipient v and let $\mathcal{Y}_u$ be the corresponding set for donor u. If the size of $\mathcal{Y}_v$ is smaller than expected, thus $y_{\bullet v} < \hat{E}(y_{\bullet v})$, which is an indication for underreporting, then it might be reasonable to assume that any reported values $y_{mv} = 1$ are correct and replacing them in the imputation algorithm by zeros should be avoided. In that case, one might limit the set of donors to those for which it holds that the observed set is a subset of the donor set:

$\mathcal{Y}_u \supset \mathcal{Y}_v$. Conversely, if the size of $\mathcal{Y}_v$ is larger than expected, thus $y_{\bullet v} > \hat{E}(y_{\bullet v})$, which is an indication of overreporting, a donor u could be selected such that the donor set is a subset of the observed set: $\mathcal{Y}_u \subset \mathcal{Y}_v$. This refinement is only feasible when the donor set is large enough.

Another example of a relaxing a condition of the reported method concerns the estimation of the residual variance $\sigma^2_{\hat{\Delta}}$. In our article the estimation of $\sigma^2_{\hat{\Delta}}$ is based on multiple reference groups, but the question remains how this variance can be estimated with a single reference group. The latter situation might occur when a reference group is appointed by experts. This residual variance stems from four error sources. The first is that not all covariates explaining the y_m variables ($m = 1, \ldots, M$) in Equation (3) might in fact be available, so there is unexplained variance. A second, related, cause is uncertainty in the imputation procedure due to uncertainty in the regression coefficients and in appointing the nearest neighbour. A third error source is the presence of random reporting errors in the y_m variables among the data suppliers in the reference group. The extent of this error source might be investigated by letting two or more administrators independently register the same cases. The final issue is that we are interested in capturing the reporting behaviour of data suppliers, whereas data from a single data supplier can be seen as just one realisation of an (unknown) distribution. Using multiple years of data from the same supplier might help to analyse the extent of this error source. When all four error sources are quantified, one might apply a repeated sampling procedure to estimate their effect on the variance $\sigma^2_{\hat{\Delta}}$. Possibly, a multiple imputation approach, originating from Rubin (1978, 1987), is useful in this context. Using that approach, we then aim to draw multiple versions of the regression coefficients of Equation (3) that capture the combined effect of the four error sources. Next, multiple versions of the matching algorithm and of $\hat{\Delta}^{t,t-1(h)}$ are obtained. It needs to be tested whether this approach yields good results.

Before our method can be applied to forms of decentralised administrative data other than the hospital data, research is needed on adaptation and extension of the method to other types of variables, errors and output forms. First, we will give two examples of potential other applications and then we will go into those adaptations and extensions. CBS has municipalities' administrative data on inhabitants' receipt of social benefits. It not only concerns social benefit data, but also additional information such as fraud occurrence, estimated fraud values, and training activities to find a job. Different municipalities have different reporting intensities, especially concerning the additional information. Suppose our aim is to detect changes in the intensity of fraud activities. We can then use covariates such as received social benefit, age, profession, social economic status, current duration of unemployment and so on (we have a social statistical database with many potentially useful variables). We could compute the (expected) changes in fraud intensity per municipality after applying steps 1–3, compare this with the original changes and select the suspicious municipalities. Likewise, we could detect underreporting of the occurrence of environmental damage reported by fire brigades using covariates like type of building, type of surrounding, presence of chemicals and so on.

It would be useful to apply the method to new examples to find out whether it works, and which adaptations or extensions are needed. We have foreseen some of those

adaptations and extensions already. A first small adaptation to the method can be done when one applies the approach to a *single* binary variable (representing reporting behaviour) at a time, rather than to a *set* of variables. Then, one could replace the predictive mean matching step by drawing a binary value from a Bernoulli distribution for each unit in the data set using the estimated probabilities. Second, the method could be extended by handling *continuous* variables with reporting errors in a selective group of data suppliers. That requires a robust way of estimating the data supplier effects (γ_h) especially in the case of large measurement errors (Rousseeuw and Leroy 1987). Third, it would be useful to develop an analysis procedure that combines the detection of under- and overreporting in classification variables with that of misclassifications. A fourth extension would be to increase the level of detail: in addition to analysing effects at data supplier level, one could analyse effects in underlying domains. Those underlying domains could, in fact, represent administrative agencies underlying the formal data suppliers, for instance clinics within large hospitals or establishments within large schools. When the reference group is selected at the more detailed domain level one may have to find a procedure to deal with a limited number of units per domain. A fifth extension is to address the effect of reporting behaviour on *level* estimates. This requires a subset of data suppliers for which we are certain that they are reporting correctly. One way to do this is to use expert knowledge to obtain such a set. It is a point of future research whether there are other possibilities for such an analysis.

7. Appendix: Computation of the HSMR

The target parameter θ_h, denoting the HSMR for hospital h, is computed as follows. Let O_{hd} be the observed mortality for main diagnosis d of hospital h and let E_{hd} be the corresponding expected mortality based on the patient population. Further, let θ_{hd} be standardised mortality ratio (SMR) for the set of units U_{hd} within main diagnosis d of hospital h. The SMR is an indicator for the quality of hospital care per main diagnosis. θ_{hd} is given by

$$\theta_{hd} = 100\frac{O_{hd}}{E_{hd}} \tag{10}$$

with $O_{hd} = \sum_{i \in U_{hd}} D_{hdi}$ and $E_{hd} = \sum_{i \in U_{hd}} E_{hdi}$, where D_{hdi} is a variable that equals 1 when the patient died during hospital admission i and 0 otherwise. The expected mortality E_{hdi} for admission i is estimated from a logistic regression with patient- and diagnosis-related variables as covariates. Hospital-related variables are left out of the model, such as the number of doctors per bed, because these are directly related to the quality of hospital care that the HSMR tries to measure. This logistic regression is fitted for each main diagnosis separately. Recall that the patient- and diagnosis-related covariates are split up into an error-free and an error-prone part. Let $\boldsymbol{z}_{hdi} = \left(z_{1hdi}, \ldots, z_{lhdi}, \ldots, z_{Lhdi}\right)^T$ denote the L-vector of error-free covariates (including the intercept), $\boldsymbol{y}_{hdi} = \left(y_{1hdi}, \ldots, y_{mhdi}, \ldots, y_{Mhdi}\right)^T$ denote the M-vector of error-prone covariates and $\boldsymbol{\beta}_d$ denote the vector of regression coefficients for the joint covariates

vector. $\hat{E}_{hdi}$ is given by:

$$\hat{E}_{hdi} = \hat{P}(D_{hdi} = 1|\boldsymbol{x}_{hdi}) = \frac{1}{1 + \exp\{-[(\boldsymbol{z}_{hdi})^T, (\boldsymbol{y}_{hdi})^T]\hat{\boldsymbol{\beta}}_d\}} \quad (11)$$

The patient- and diagnosis-related variables are given in Table 1.

Let $\mathcal{D}$ be the set of main diagnoses that are included in the computation of the HSMR. We included 50 out of 200 main diagnoses in the HSMR computation. These 50 diagnoses comprised about 80 per cent of all hospital admissions (Israëls et al. 2012). Further, let θ_h be the HSMR of hospital h, which is computed by $\theta_h = \sum_{d\in\mathcal{D}}\theta_{hd}$.

The SMR in (10) and HSMR θ_h are estimated using the observed mortality relative to the estimated expected mortality according to (11). The comorbidities are not directly used in Equation (11) as covariates. Instead, they are transformed into 17 binary variables. Each binary variable stands for a group of related diseases according to the classification of the so called Charlson Index (Charlson et al., 1987). This binary variable is 1 when one or more comorbidities are registered for that specific class of the Charlson index and 0 otherwise.

When the number of admissions for a certain class within the patient- or diagnosis-related variables was smaller than 50, classes were merged. This was done to avoid that the standard errors of the regressions became too large. The procedure for merging classes can be found in Van der Laan et al. (2015).

8. References

Backor, K., S. Golde, and N. Nie. 2007. "Estimating Survey Fatigue in Time Use Study." Paper presented at the 29th Annual Conference of the International Association of Time Use Research, 17–19 October 2007, Washington, DC, U.S.A. Available at http://www.atususers.umd.edu/wip2/papers_i2007/Backor.pdf (accessed October 2018).

Bakker, B.F.M. and P.J.H. Daas. 2012. *Methodological Challenges of Register-based Research*. Statistica Neerlandica, 66: 2–7. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00505.x.

Berenschot. 2012. Inventarisatie informatiebehoefte brandweerstatistiek. Eindrapport (in Dutch). Available at https://www.wodc.nl/binaries/2131-volledige-tekst_tcm28-72208.pdf (accessed February 2018).

Bottle, A., B. Jarman, and P. Aylin. 2011. "Hospital Standardized Mortality Ratios: Sensitivity Analyses on the Impact of Coding." *Health Services Research* 46: 1741–1761. Doi: http://dx.doi.org/10.1111/j.1475-6773.2011.01295.x.

Brackstone, G.J. 1987. "Issues in the Use of Administrative Records for Statistical Purposes." *Survey Methodology* 13: 29–43. Available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1987001/article/14467-eng.pdf?st=DEZo9O3B (accessed October 2018).

Charlson, M.E., P. Pompei, K.L. Ales, and R. MacKenzie. 1987. "A New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation." *Journal of Chronic Diseased* 40: 373–383. Doi: http://dx.doi.org/10.1016/0021-9681(87)90171-8.

De Waal, T., J. Pannekoek, and S. Scholtus. 2011. "Handbook of Statistical Data Editing and Imputation." New York: John Wiley and Sons.

Efron, B. and C.N. Morris. 1975. "Data Analysis using Stein's Estimator and Its Generalizations." *Journal of the American Statistical Association* 74: 311–319. Doi: http://dx.doi.org/10.1080/01621459.1975.10479864.

Elixhauser, A., C. Steiner, D.R. Harris, and R.M. Coffey. 1998. "Comorbidity Measures for Use with Administrative Data." *Medical Care* 36: 8–27. Doi: http://dx.doi.org/10.1097/00005650-199801000-00004.

Groen, J.A. 2012. "Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures." *Journal of Official Statistics* 28: 173–198. Available at http://www.sverigeisiffror.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/sources-of-error-in-survey-and-administrative-data-the-importance-of-reporting-procedures.pdf (accessed October 2018).

Harteloh, P., K. de Bruin, and J. Kardaun. 2010. "The Reliability of Cause-of-death Coding in The Netherlands." *The European Journal of Epidemiology* 25: 531–538. Doi: http://dx.doi.org/10.1007/s10654-010-9445-5.

Hosmer, D.W. and S. Lemeshow. 2004. Applied Logistic Regression. New York: John Wiley and Sons.

Israëls, A., J. van der Laan, J. van der Akker-Ploemacher, and A. de Bruin. 2012. HSMR 2011: Methodological report. Technical report, Statistics Netherlands. Available at https://www.cbs.nl/NR/rdonlyres/E7EC3032-B244-4566-947D-543B8AAE6E4A/0/2012hsmr2011methoderapport.pdf (accessed February 2018).

Jarman, B. 2008. "In Defence of the Hospital Standardised Mortality Ratio." *Healthcare Papers* 8: 37–41. Doi: http://dx.doi.org/10.12927/hcpap.2008.19974.

Jarman, B., S. Gault, B. Alves, A. Hider, S. Dolan, A. Cook, B. Hurwitz, and L.I. Iezzoni. 1999. "Explaining Differences in English Hospital Death Rates Using Routinely Collected Data." *Biomedicial Journal (BMJ)* 318: 1515–1520. Doi: http://dx.doi.org/10.1136/bmj.318.7197.1515.

Kim, Y., Y-K. Choi, and S. Emery. 2013. "Logistic Regression with Multiple Random Effects: A Simulation Study of Estimation Methods and Statistical Packages." *The American Statistician* 67: 171–182. Doi: http://dx.doi.org/10.1080/00031305.2013.817357.

Oberski, D.L., A. Kirchner, S. Eckman, and F. Kreuter. 2017. "Evaluating the Quality of Survey and Administrative Data with Generalized Multitrait-Multimethod Models." *Journal of the American Statistical Association*. Available at http://dx.doi.org/10.1080/01621459.2017.1302338 (accessed February 2018).

Pieter, D., R.B. Kool, and G.P. Westert. 2010. Nederlands Tijdschrift voor Geneeskunde, 154, A2186 [in Dutch]. Available at https://www.ntvg.nl/artikelen/beperkte-invloed-gegevensregistratie-op-gestandaardiseerd-ziekenhuissterftecijfer-hsmr/volledig (accessed February 2018).

Pitches, D.W., M.A. Mohammed, and R.J. Lilford. 2007. "What Is the Empirical Evidence That Hospitals with Higher-Risk Adjusted Mortality Rates Provide Poorer Quality Care? A Systematic Review of the Literature." *BMC Health Services Research* 7: 91–98. Doi: http://dx.doi.org/10.1186/1472-6963-7-91.

Prins, M.J. 2016. *The Effect of Coding Practice on the Hospital Standardised Mortality Ratio, Master Thesis*. Utrecht University. (available upon request).

Quan, H., B. Li, L.D. Saunders, G.A. Parsons, C.I. Nilsson, A. Alibhai, and W.A. Ghali. 2008. "Assessing Validity of ICD-9-CM and ICD-10 Administrative Data in Recording Clinical Conditions in a Unique Dually Coded Database." *Health Services Research* 43: 1424–1441. Available at http://onlinelibrary.wiley.com/doi/10.1111/j.1475-6773.2007.00822.x/full (accessed February 2018).

Rousseeuw, P.J. and A.M. Leroy. 1987. *Robust Regression and Outlier Detection*. New York: John Wiley and Sons.

Rubin, D.B. 1978. *Multiple Imputations in Sample Surveys – a Phenomenological Bayesian Approach to Nonresponse. Proceedings of the Section on Survey Research Methods*. American Statistical Association. Available at http://ww2.amstat.org/sections/srms/proceedings/papers/1978_004.pdf (accessed February 2018).

Rubin, D.B. 1987. *Multiple Imputation for Non-response in Surveys*. New York: John Wiley and Sons.

Shields, J. and N. To. 2005. "Learning to Say No: Conditioned Underreporting in an Expenditure Survey." Paper presented at the American Association for Public Opinion Research Annual Conference, 12–15 May 2005, Miami Beach, U.S.A. Available at http://ww2.amstat.org/sections/srms/proceedings/y2005/Files/JSM2005-000432.pdf (accessed October 2018).

Silberstein, A.R. and C.A. Jacobs. 1989. Symptoms of Repeated Interview Effects in the Consumer Expenditure Survey. In *Panel Surveys*, edited by D. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh, 289–303. New York: John Wiley and Sons.

Tourangeau, R., R.M. Groves, and C. Redline. 2010. "Sensitive Topics and Reluctant Respondents: Demonstrating a Link between Nonresponse Bias and Measurement Error." *Public Opinion Quarterly* 74(3): 413–432. Doi: http://dx.doi.org/10.1093/poq/nfq004.

Tourangeau, R. and T. Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133(5): 859–883. Doi: http://dx.doi.org/10.1037/0033-2909.133.5.859.

United Nations Economic Commission for Europe. 2011. *Using Administrative and Secondary Sources for Official Statistics: a Handbook of Principles and Practices*. New York and Geneva: United Nations. Available at http://www.unece.org/fileadmin/DAM/stats/publications/Using_Administrative_Sources_Final_for_web.pdf (accessed February 2018).

Van Delden, A. and S. Scholtus. 2017. "Correspondence between survey and admin data on quarterly turnover." CBS Discussion Paper 2017-03. Available at https://www.cbs.nl/en-gb/background/2017/07/correspondence-between-survey-and-admin-data-on-quarterly-turnover (accessed February 2018).

Van den Bosch, W.F., J. Silberbusch, K.J. Roozendaal, and C. Wagner. 2010. Variatie in codering Patiëntengegevens beïnvloedt gestandaardiseerd ziekenhuissterftecijfer (HSMR). Nederlands Tijdschrift voor Geneeskunde, 154 A1189 [in Dutch]. Available at https://www.ntvg.nl/artikelen/variatie-codering-patiëntengegevens-beïnvloedt-gestandaardiseerd-ziekenhuissterftecijfer/volledig (accessed February 2018).

Van der Laan, J. 2013. *Quality of the Dutch Medical Registration (LMR) for the calculation of the Hospital Standardised Mortality Ratio*. Discussion Paper. Statistics

Netherlands. Available at https://www.cbs.nl/NR/rdonlyres/6290A0A8-4CC9-4DBF-AF0B-A3C6742EEA89/0/201308x10pub.pdf (accessed February 2018).

Van der Laan, J., A. de Bruin, J. van den Akker-Ploemacher, C. Penning, and F. Pijpers. 2015. *HSMR 2014: Methodological Report. Technical Report*. Statistics Netherlands. Available at https://www.cbs.nl/NR/rdonlyres/2AFF4E96-C02F-4BB0-97A9-035D17DF1104/0/2015hsmrmethodologicalreport2014.pdf (accessed February 2018).

Wallgren, A. and B. Wallgren. 2014. *Register-based Statistics. Statistical Methods for Administrative Data* (2nd edition). New York: John Wiley and Sons.

West, B.T. and A.G. Blom. 2017. "Explaining Interviewing Effects: A Research Synthesis." *Journal of Survey Statistics and Methodology* 5: 175–211. Doi: http://dx.doi.org/10.1093/jssam/smw024.

Received June 2017
Revised April 2018
Accepted May 2018