ACTA UNIVERSITATIS AGRICULTURAE ET SILVICULTURAE MENDELIANAE BRUNENSIS

Volume 65 183 Number 5, 2017
https://doi.org/10.11118/actaun201765051779

ECONOMIC ASPECTS OF THE MISSING DATA PROBLEM – THE CASE OF THE PATIENT REGISTRY

Hatice Uenal[1], David Hampel[1]

[1]Department of Statistics and Operation Analysis, Mendel University in Brno, Zemědělská 1, 613 00 Brno, Czech Republic

Abstract

UENAL HATICE, HAMPEL DAVID. 2017. Economic Aspects of the Missing Data Problem – the Case of the Patient Registry. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis,* 65(5): 1779–1791.

Registries are indispensable in medical studies and provide the basis for reliable study results for research questions. Depending on the purpose of use, a high quality of data is a prerequisite. However, with increasing registry quality, costs also increase accordingly. Considering these time and cost factors, this work is an attempt to estimate the cost advantages of applying statistical tools to existing registry data, including quality evaluation. Results for quality analysis showed that there are unquestionable savings of millions in study costs by reducing the time horizon and saving on average € 523,126 for every reduced year. Replacing additionally the over 25 % missing data in some variables, data quality was immensely improved. To conclude, our findings showed dearly the importance of data quality and statistical input in avoiding biased conclusions due to incomplete data.

Keywords: Benford law, data source quality, missing-at-random mechanism, missing data problem, reducing study costs

INTRODUCTION

Registries play an irreplaceable role in medical studies, where they are essential for the reliability of study results for arbitrary research questions. The quality of data in registries may have extreme effects on the significance and meaningfulness of results. Therefore, in particular in data for sensitive uses, such as pharmaceutical data, the highest reliability is required. Data from patients with any disease are collected until enough data is available to achieve reliable study results. Depending on the research question and the percentage of missing key values, in some cases volunteers are even interviewed after 10 years and data collected, until the registry has a high degree of reliability. However, with increasing registry quality, the costs increase accordingly.

In a data matrix (or a registry), data concerning specific subjects share specified characteristics or information. These for example can be patients with a particular disease (patient registry), customer answers to a product (market research), or a sales related registry (selling strategy). A registry is an important tool for investigating and evaluating specific outcomes of a research question. A registry can be seen as a list of data entries in which each entry represents an individual case or person with defined components of specific information. However a data matrix can be an abstract or a sample of a registry, where representativeness is assumed (e.g. study data). Depending on the aim of a study, data are not always collected in the same way and in this case the completeness of missing data is crucial to high quality as well as correct data acquisition without any manipulation for reliable study outcomes. Especially in pharmaceutical studies, only high quality data registries are trusted (Rothenbacher *et al.*, 2015).

As a consequence, study periods are often prolonged with the aim to extend the records in order to gain more complete data sets, see Uenal *et al.* (2014a). However, these data sets might already have been of good quality, and only missing a few small pieces of information (e.g. the patient has not stated

certain questionnaire items of high importance or clinical measurements were not completed). Thus, these are pieces of information that can be estimated statistically and need not to be collected in a new study, which would be expensive.

This means, that theoretically well-planned registries would guarantee high quality, but are never achieved to a lack of the funding that would be necessary for continuing the registry-based study and compensating for missing data and cases. In these situations, statistical input could solve the problem. In order to evaluate this, distributions are tested (Benford distribution and original distribution of the variables). A simulated low-quality registry created from the same source will be compared to measure the quality evaluation as well as improvements and cost differences. For example Glicklich *et al.* (2014) set up guidelines concerning registry quality ("Registries for Evaluating Patient Outcomes: A User's Guide"), but these suggestions do not cover the data quality and manipulations (intended or unintended) themselves.

The objective of this work is to investigate the cost advantages by comparing the quality of statistical input to costs. We estimate the cost advantages of applying statistical tools to existing registry data and evaluate the quality as well as the reliability for possible study results (instead of obtaining new data by prolonging the study period). Further we measure the quality of data (low to high) using Benford's Law (BL). We use a published data matrix containing gastric cancer patients from Ma *et al.* (2015).

Data Quality Problem

Although the overall problem of data matrix quality is discussed in several publications, they mainly focus on the issues of incomplete and missing data in a registry. If they deal with cost effectiveness, then it is only examined from a general financial point of view. Ward discusses the advantages of cost probability by using statistical methods but relates them to the fields of accounting and product costs (Ward, 1968). The same is attempted by Yewdall *et al.* (1969). Zhang *et al.* (2007) discuss cost-effectiveness by using estimation of data instead of carrying out expensive research. Still his approach does not deal with the aspect of quality; quality loss is only mentioned as a problem in all these approaches. Bankhofer and Praxmeier (1998) discuss the missing data problem in market research, but their main focus is on the reasons for missing data. Dinh and Zhou (2006) as well as Bansal *et al.* (2008) describe missing data in cost sensitive areas in the business and health sectors, but they do not examine the quality aspect.

However, there is no publication yet that combines all of these aspects with regard to improving data quality and saving costs (Silvia *et al.*, 2014; Enders, 2010; Lewis, 2008). Thus what is missing in these publications is an analysis of the cost savings when using statistical methods to evaluate quality instead of trying to improve low quality data by prolonging the study and collecting more data, which is a very expensive and time-consuming process.

A high quality registry is essentially a complete registry, i.e. without any missing information and a high percentage of completeness (assumption that all patients were collected), where data has been collected consistently for all patients to international validated standards (Glicklich *et al.*, 2012 and 2014; Nagel *et al.*, 2012 and 2013; Uenal *et al.*, 2014a). Accordingly, a data matrix is assumed that has complete observations. Missing data and human coding mistakes as well as data manipulation possibilities should be conscientiously managed from the beginning of data collection. Confirmation of high quality, data accuracy and completeness is expected.

Here the problem arises that despite the careful planning of a study and its duration as well as well-organised data collection to international standards, doubts concerning the quality of data can remain. The reason for this is the lack of awareness that there is the possibility of investigating quality (Tam *et al.*, 2007; de Vocht and Kromhout, 2012; Judge and Schechter, 2009; Spencer, 1985).

Aside from the completeness of a data matrix, the aspect of data quality and data collection is also hugely relevant. Spencer discusses the decision-theory of different scenarios concerning the level of data quality. While he constructs different hypothesis and cases, they all rely on the assumption that data quality is needed at that level where "the benefit minus the cost is largest", see Spencer (1985), p. 565. A number of publications use the BL stated in Benford (1938) for exploring distributions and data set applicability in different contexts. Leemis *et al.* (2000) confirm the relevance of Benford in survival distribution research, while Judge and Schechter (2009) discuss the usability for biased conclusions in survey data and the Archambault and Archambault (2011) explore the BL in management area.

According to Tam *et al.* (2007), BL is also usable for data fraud detection. A deeper explanation in the field of data quality detection in medical studies applying the BL is described by de Vocht and Kromhout (2012). Wang (1996) as well as Lee and Strong (2003) discuss the importance of data quality and the consequences of poor data quality. Cappiello *et al.* (2003) examine the relevance of time for data quality. Lee (2003) emphasises the importance of context variables for data quality. In contrast to Tam *et al.* (2007), Cecchini *et al.* (2010) examine fraud detection in data sources by using the Beneish Law, see Beneish (1997) and (1999).

Maritz (2003) describes the importance of data management systems as an organisational resource and in handling archives. Esteva *et al.* (2013) discuss practical ways of data mining and working with large archives.

In the case of the missing data of observed patients in a data set to be further analysed, the ideal case would be that the data are missing completely at random (MCAR) and that they lie in an appropriate range. More than 25% missing relevant data in a registry of all data entries could also be a basis for biased study results, when ignoring or not handling them appropriately (Uenal *et al.*, 2014b).

When ensuring a high quality registry it is important to make sure that the data collection complies with international standards (Glicklich *et al.*, 2012 and 2014; Nagel *et al.*, 2012; Nagel *et al.*, 2013; Rothenbacher *et al.*, 2015) to avoid bias already in the phase of data collection and that a high level of registry coverage is guaranteed (Uenal *et al.*, 2014b; Bishop, 2007; Williams, 2002).

MATERIALS AND METHODS

Example Registry Data Set: Solitary Lymph Node Metastasis in Gastric Cancer

The example used of registry data of gastric cancer patients is a data set of a study conducted at the Cancer Center of Sun Yat-Sen University in South China. The aim of the study was to examine the clinical significance of risk factors towards solitary lymph node metastasis (SLM) in gastric carcinoma (Ma *et al.*, 2015). In total, 385 patients with gastric carcinoma who had a D2 lymphadenectomy were included into the research. To validate the study results, data from the Sun Yat-Sen University Hospital were compared.

The retrospective research took place from July 2000 to July 2012. The researchers used clinicopathological data of gastric cancer patients. In total 82 (22.3 %) patients were observed with SLM and 303 (78.7 %) patients with NLM, respectively. Follow-up visits were made at an average of 52.3 (with SD equal 25.9) months. In their study, they could show that SLM is an independent risk factor for gastric cancer and there was no survival difference between the two SLM groups (skip / no skip). We will work with following variables: CRP (C-reactive protein) level, CEA (carcinoembryonic antigen) level, CA-199 level and CA-724 level (the name CA is derived from commercial carbohydrate antigen sets). These variables belong to so-called tissue-specific tumor markers obtained from blood. Further medical details can be studied in Ma *et al.* (2015). The volume and contents of the data set is a good basis for investigating the quality as well as further scenarios of situations (for instance: the accuracy of missing data estimation, reduced study duration in case-control studies and registry quality evaluation).

Methods to Detect Registry Data Quality

There exist several methods to detect the data quality of a study (or registry). Researchers use data of their study for further analysis and deal with the subjects of quality in data collection phase as well as when working with collected data. Unfortunately, there are more factors to consider than proper data collection. Awareness of other factors seriously affecting the study results of any research question is not taken seriously enough. Intended or unintended mistakes often remain undetected in studies. The more sensitive the purpose of the data use (health, medical and pharmaceutical research, etc.), the more important is the quality, see Spencer (1985) and Rothenbacher *et al.* (2015).

In this work, we only refer to the BL and how to handle missing data and its reliability. The CARE method is an alternative possibility, when several data sources are provided. While the CARE method would require data sets where a case is uniquely identified (e.g. ID) and the same case ID naturally occurs in other data sources, so one could estimate the total number of missing cases in a registry without any population parameter information, but only by intersection information (set theory). This method is appropriate when a uniquely labelled case can be re-identified in the other data sources. Then the completeness of a registry can also be estimated regarding quality questions of new implemented registry studies, when completeness is doubtful due to hidden cases or not discoverable (i.e. low survival of cases of a rare disease, see Uenal *et al.* 2014a).

Every method, especially in certain situations, has advantages as well as disadvantages. Referring to a single data matrix and observation of several variables, the method cannot be applied, since we do not have varied data sources. In this case, the BL is a good tool to investigate data quality and to detect human coding mistakes, imprecise answers or poorly designed questionnaires (Bredl *et al.*, 2008; Benford, 1938), missing observations (i.e. complete case analysis and its influence on possible biased study results) Uenal *et al.* (2014b), "black numbers" regarding hidden observations or even fraud (i.e. finances) Cecchini *et al.* (2010). The easily applied law uses the first digits of numbers of any kind (decimal numbers, medical measurements, financial data), from 1 to 9.

In combination with missing data (MD) in the data matrix, the BL still can be applied by replacing the missing values first by multiple imputation or an appropriate missing data method, see Uenal *et al.* (2014b). At the same time, the reliability of automatically replaced values is evaluated using significance tests such as the Chi-squared to test for differences between the distributions, the BF distribution and the distribution of the same variable of interest with the imputed values.

Further methods have been suggested by others, such as multivariate analysis or cluster and discriminant analysis (Bredl *et al.*, 2008), which are not further described.

The Benford Approach to Determining Data Quality

The BL underlay the observation that the digits 1–9 (each as an initial digit) are not equally

distributed. According to BL, a lower digit is more likely to occur at the beginning than a higher digit. The astronomer and mathematician "Simon Newcomb" first observed the phenomena in 1881, when he noticed that first pages of logarithm tables were more intensively used than other pages (Fewster, 2009; de Vocht and Kromhout, 2012). "Benford's Law" was born in 1938 after years of oblivion, when Benford described the digits observation in his paper analysing various data sets, see Benford (1938). Proof was produced by the American mathematician Theodore Hill who also came up with the practical application; see Hill (1995a, 1995b).

The BL as well as the underlying distribution can be applied to random data sets and any type of data sets with any variable of interest. Manipulated data sets are affected by natural human coding mistakes, mistakes in the data collection stage, transfer errors, fraud or the handling of missing data (whether deleted complete observations or replaced inappropriately). Such data differ from the natural Benford distribution resulting in low data quality at the same time (de Vocht and Kromhout, 2012). The method even helps in registry quality questions, determining the first digits and detecting the total size of the missing observations. Although the BL is evasive, the law has been widely applied by many researchers, but at the same time it has still not been satisfactorily explained, see Fewster (2009).

Generally for the k first digits, consider a set of data, which is Benford distributed, then the probability that a digit d with a basis B in the n'th place from the beginning is as follows:

$$p_n(d) = \sum_{k=B^{n-2}}^{B^{n-1}-1} \log_B\left(1+\frac{1}{kB+d}\right).$$

For the first digits, the equation of the probability that first digit is d reduces to

$$p(d) = \log_{10}\left(\frac{d+1}{d}\right) = \log_{10}(d+1) - \log_{10}(d) =$$

$$= \log_{10}\left(1+\frac{1}{d}\right), d = 1,\ldots,9.$$

Which means in more detail that numbers in a variable that is following BL, when the leading digit $d \in 1, \ldots, 9$ occurs with a probability $p(d)$.

One of the special advantages in the properties of BL is that the BL can be applied to any kind of data. In the case of decimal values, the next possible digit is used. The law is valid in any field of research where numbers in a variable can be summarised by their first digit. Unfortunately, the method does not function for binary or category data, see Judge and Schechter (2009).

The evaluation of data quality for each of the variables "CRP", "CEA", "CA-199" and "CA-724" is estimated by their extent of divergence from the BL rates and the observed rates. Then the common Chi-squared goodness of fit significance test is conducted to investigate the level on BL-accordance:

$$\chi^2 = \sum_{d=1}^{9} \frac{(obs_d - bl_d)^2}{bl_d},$$

where obs_d is the observed counts of the digits d and bl_d the expected BL counts. Depending on a significance level of 0.10, 0.05 or 0.01, the cut-off values are known as $\chi^2 = 13.36$, 15.51 or 20.09 and considered significant, when χ^2 is then the critical threshold for the χ^2 test with 8 degrees of freedom (Judge and Schechter, 2009). It is important to note that distribution of the χ^2 test statistics is asymptotic χ^2 distribution only. Observed counts should be large enough for reasonable results; in the context of our task it is necessary to check this assumption especially in the case of a short study period.

Handling Missing Data

Missing values are a common problem in medical research and may be a considerable source of bias when ignored or not handled appropriately, see Uenal *et al.* (2014b). Especially in quality questions, decisions on handling MD are at the same time decisions on potential financial consequences.

There are several MD approaches which have to be distinguished. Imputation of missing values should increase the quality of a registry, since deleting registry observations with missing values could manipulate the representativeness of patient data, Uenal *et al.* (2014b). Before applying an algorithm, one has to distinguish between the MD mechanisms. The "Missing Completely At Random" (MCAR) mechanism is suitable when the probability of missing data is unrelated to covariates and the values of the target variable. "Missing At Random" (MAR) is useful when the probability of missing data could be related to covariates, but not to values of the target variable. "Not Missing At Random" (NMAR) describes both, the probability of missing data relates to unobserved values of the target variable even after control for covariates. It is useful for example if a person suffering from depression does not respond to questions on his mental health, see Uenal *et al.* (2014b).

Economic Aspects: Quality Improvements and Cost Differences

Quality improvements leading to high data quality by applying BL considering initial data collection criteria are evaluated, see Glicklich *et al.* (2014). Any improvement in the data (patient data matrix) compared to the quality at the beginning is considered and after statistical methods have been employed describing the extent of improvement between the original quality and the improved quality. The cost differences are calculated then based on typical monthly study costs.

To keep a main overview of the costs K, creating a cost function with influencing variables, we keep mainly variable costs and fixed costs. Generally, the fixed costs are defined monthly having the same

costs every month. Variable costs are defined as changing costs (cannot be summarised monthly by the same amount). The cost function is thus investigated including all cost changing factors in the whole variating interval (in our case several kink points). Therefore, the function we set for cost saving investigation is different to proportional cost functions (only one influencing variable with a constant slope) or the fixed cost function. We propose the following scheme of cost function:

$$K = F_G + t\sum_{i=1}^{10} F_{M_i} + \sum_{i=1}^{4}\sum_{j=1}^{t} x_{i,j},$$

where $x_{1,j}$ means Working students or assistants in month j (total monthly working hours in €), $x_{2,j}$ means Medician in month j (probands; total monthly attended probands in €), $x_{3,j}$ means Medical management costs in month j (total probands monthly in €), $x_{4,j}$ means Management costs – study nurse in month j (total probands monthly in €) and t is number of months of study duration. The variables $x_{1,j},\ldots,x_{4,j}$ are observed monthly. Further, F_{G} means study-specific purchases and reserves and F_{M_i} covers total consumables and the monthly fixed costs, where F_{M_1} relates to Professors, F_{M_2} to Scientists, F_{M_3} to Data Managers, F_{M_4} to Disease specialised medicians, F_{M_5} to Medical assistants, F_{M_6} to Disease specialised assistants, F_{M_7} to Cost of materials (including medical and biological), F_{M_8} to Total consumables (probands), F_{M_9} to Other operating expenses and $F_{M_{10}}$ to Management costs of the project. The created function is applicable to any study design. For instant, in a typical case-control study, healthy probands are also collected to compare healthy people with diseased patients 1:n (matching 1 patient to n probands). Considering possible study designs, possible cost advantages are investigated more intensively. For a given t we wish to estimate cost function of the form

$$K = \beta_0 + \beta_1 x_1 + \ldots + \beta_4 x_4 + \beta_5 F_M,$$

where $F_M = \sum_{i=1}^{10} F_{M_i}$.

The cost analysis as well as the cost function were estimated using statistical software R.

RESULTS

Original data set characteristics

Concerning the study of Ma *et al.* (2015), several advantages, disadvantages and complications in the stage of data collection and the study itself were published and confirmed by the underlying data matrix (for summary information about the data see Tab. I). The quality of data completeness on the one hand can be seen in "action required", since only patients who fulfilled the full criteria were included and patients without follow-up data were excluded. On the other hand, there are several MD in the analysis, which can influence data quality and thus the study results as well as the finances. Although there was some missing data, the researchers included and investigated in all conscience their data. A control group was used for data validation. The small number of cases and the high rate of MD in several variables led to complications resulting in a quality that possibly can

I: *Summary of current situation of data quality criteria*

Quality Criteria	Comments
Data Completeness	Only patients who fulfilled the criteria were included; patients without complete follow-up data were excluded; The total number of missing including excluded cases is unknown. No analysis of completeness.
Missing Data	MD in the variables CRP, CEA, CA-199 and CA-724 and in 2 of them more than 25% missing
Covariate Collection	Investigation of all risk factors
Data validation	Comparison group
Data manipulations (intended/unintended) and accuracy	No analysis of data quality

II: *Descriptive statistics of the raw data*

Variable	Missing data (%)	MIN	MAX	MEAN	MEDIAN	SD
CRP	101 (26.2)	2	19	9.0	6.0	8.9
CEA	16 (4.2)	1	995	674.4	777.5	299.4
CA-199	58 (15.1)	5	11620	2855.9	743.0	4361.9
CA-724	110 (28.6)	1	1169	767.6	776.0	192.5

bias results and expand finances. In the subsections we will investigate if the exclusion of persons and if missing values influenced the data quality as well as the study costs.

For all variables, MD was observed with 26.2% for the CRP, 4.2% for the CEA, 15.1% for the CA 199 – and 28.6% for the CA-724 variable. The stated missing values are intended to be multiple imputed in the next sections and compared by their data quality using BL, when ignored. The current quality of MD (over 25% of missing values in 2 variables) signals in advance the possible influence of suffering data quality. The first impression of the variables including the level of missing data is described in Tab. II.

Analysing the first digits, data quality is evaluated in detail by stars (poor data quality is assumed when at least in one digit no star was achieved).

Ignoring the missing values of 26.2% in the CRP-variable, there are no anomalies in the observed frequencies obs_d from the BL expected frequencies bl_d. After computing in detail for every digit, the χ^2 test resulted in not significant neither at 5% significance ($\chi^2 = 4.2 < 15.51$) nor at 10% significance, ($\chi^2 = 4.2 < 13.36$). Summarising the given data quality points for all digits, really good quality is assumed and even the 26.2% missing data did not affect the overall data quality.

Concerning CEA, the Benford approach yielded that the data quality is clearly more action required as for the CRP-variable. Ignoring the missing values, part of the Benford frequencies differed extremely. For digit 2, $\chi^2 = 15.781$ and so is significantly different from the BL distribution (5% level). After detailed computing of χ^2 for every digit the test result summed up for all 9 digits shows significant differences from BL distribution at 1% level ($\chi^2 > 20.09$), what means that really poor data quality is counted in this case. A detailed analysis however shows that aside from digit 2, the other digits seem to be within range. The overall data quality gain therefore is still 1 point.

In CA-199, very poor data quality is observed. A detailed view of each digit showed that digit 6 especially was expected to be lower than observed, where the other digits did not yield a significant χ^2, the χ^2 test was even highly significant with $\chi^2 > 20.09$. Part of the Benford frequencies differed extremely. Compared to the Benford distribution, the distribution of the original data was highly significant. Summed up the squared differences for all digits, the overall χ^2 – value yielded 51.703 and was highly significant considering that the cut-off value for the 1% significance level is $\chi^2 > 20.09$.

CA-724 did not show any anomalies detailed for each digit, but summed up for all digits, the overall χ^2 test was highly significant at the 1% level with ($\chi^2 = 55.35 > 20.09$) . The variable has 28.6% MD, which could contribute to the result.

Handling Missing Data – original data set

It is important to use the comprehensive decision-making algorithm to find an appropriate MD method and in order to estimate the missing values. For the variables "CRP", "CEA", "CA-199" and "CA-724", ignorable missing data is assumed, since no systematic missing data was observed. For all four variables, at least MAR as a MD mechanism is suitable. Since the variables are of a continuous scale level, the algorithm led to the left side of the decision tree. For both MD pattern (monotone and arbitrary), the MCMC method (Markov Chain Monte Carlo) is suggested. MDs were therefore imputed by the MCMC method using NORM 2.03 software (Schafer *et al.*, 1999).

The investigation of the Benford approach with imputed values yields that the variable CRP still had good data quality. Considering the 26.2% of replaced MD, summing up χ^2 – values for all 9 digits, χ^2 was over 2 times higher, but still was not significant even at a 10% significance level. In comparison with the Benford distribution, the distribution of the original imputed data was as before detailed for every digit as being of very good quality. Especially for this variable, the comparison from the Benford signals a low influence of missing vales or even missing patients in the study. Although data was collected over a study period of 6 years, the quality seemed to be stable constantly over the time concerning the variable CRP. The digit 1 was observed 5 times more with imputed MD. This could be a signal of missing patients who did not meet the inclusion criteria.

In CEA, the investigation of the Benford approach yields the same results as with 4.2% MD. After the missing values were replaced, Benford frequencies again yield significantly less data for digit 2. Accordingly χ^2 for digit 2 is significant with $\chi^2 > 15.51$. Summed up for all digits, $\chi^2 = 38.83 > 20.09$ and so was highly significant. The poor quality signals a high influence of missing patients. Probably this is at least partly the influence of patients who were excluded from the study, but not the influence of MD as we see.

The data quality in CA-199 only improved slightly, but in total the quality remains very poor signalling that besides missing values other factors could have influenced the overall quality considering also the representativeness. Comparing the Benford distribution from the original imputed data distribution, digit 6 again showed significant anomalies of $\chi^2 = 19.165 > 15.51$. The overall χ^2 summed up for all digits was again highly significant with $\chi^2 = 49.066 > 20.09$.

For CA-724 an immense improvement in data quality was observed. After the 28.6% MD were replaced, the χ^2 statistical value (summed up for all digits) was reduced greatly from $\chi^2 = 55.35$ to $\chi^2 = 21.60$. The value is of high significance at 1% ($\chi^2 > 20.09$), but the 5% significance level limit value came a lot closer. Nor did a detailed analysis of each digit's χ^2 yield any anomalies. The variable showed a high degree of influence of quality caused by missing data. Other factors, such as human coding or data entry manipulation (intended or

unintended) or missing patients in a study due to exclusion criteria still influence data quality. The fact of "handling MD" shows especially in this case the importance of correct handling, which allows a good chance to completely improve its data quality examining a reduced study period. Properties of all variables after imputation using this method were the same as before imputation.

Handling Missing Data – reduced study period

Using the same data set presented in the previous section with imputed missing data means that by using the appropriate MD method (MCMC chosen by the comprehensive decision-making algorithm) for the variables "CRP", "CEA", "CA-199" and "CA-724", an increased level of data quality is observed as the study period is reduced. Reduction of the study period in years means reducing at the same time missing or unrecorded patients and improves the representativeness of data. To demonstrate in detail changing key values, results are presented for a study period of three years (half of the observations with $n = 193$ patients) assuming that the original data set with 385 patients were recorded chronologically.

In the variable CRP, the investigation of the Benford approach with imputed values reducing the study period to three years yielded indeed still good data quality as expected. Summing up all 9 digits' χ^2 values, χ^2 was higher, but still not significant even at a 5% significance level. The comparison from the Benford signals still a low influence of missing vales or even missing patients in the study concerning the variable CRP. Although data was collected over a study period of 6 years, the quality seemed to be stable even at half of the study period. The digit 1 was observed more frequently. This could be a signal of still missing patients (inclusion/exclusion criteria).

To demonstrate in detail changing key values, results are presented for the years of the study period (half of the observations with $N = 193$ patients) assuming that the original data set with 385 patients were recorded chronologically.

Concerning CEA, a study period of 6 years yielded the same results as for only three years. Benford frequencies again had less data for digit 2, but improved accordingly from 16.26 to $\chi^2 = 12.99$ for digit 2. Summed up for all digits, $\chi^2 = 3.61 < 20.09$ and still was highly significant. The poor quality signals the still high influence of missing patients. Although the reduction of the study period did not yield as high an improvement as expected, the quality of data shows an improvement in the detailed investigation of the first digit.

For CA-199, the data quality improved greatly, but in total the quality remains very poor, confirming the assumption of other factors that could have influenced the overall quality considering also the representativeness. Comparing the Benford distribution from the original, digit 6 is highlighted again with significant anomalies of $\chi^2 = 25.30 > 15.51$ therefore earning no star. The overall χ^2 summed up for all digits improved to $\chi^2 = 39.15 > 20.09$ instead of $\chi^2 = 49.07$, but both χ^2 – values still remain highly significant on halving the study period.

As previously assumed, the variable CA-724 signalled a good chance to improve quality immensely. The investigation of the Benford approach with imputed values confirms the assumption as expected. The χ^2 statistical value (summed up for all digits) reduced greatly from $\chi^2 = 55.35$ to $\chi^2 = 21.60$ after replacing 28.6% MD and again improved greatly after reducing the study period to three years in total to $\chi^2 = 9.83 < 13.36$. This value is not significant anymore resulting even very good data quality. The quality of the variable was influenced possibly by a high degree of missing data and possibly also by many missing patients or other factors, such as data entry manipulation (intended or unintended). The fact "handling MD" and possible profits from cost savings in reducing the period show accordingly the great need for correct data handling given the additional cost saving advantages.

Economic aspects: Quality improvements and the cost differences

According to the results, the data quality shows good quality after reducing the study period, assuming that the underlying study can be shortened. By shortening the study period up to the half of the costs could be saved.

III: *Comparison of significance testing with original and imputed missing values (study period: 1, 2, 3, 4 and 6 years).*

χ^2 (study period)	Original Data				Imputed Data			
	CRP	**CEA**	**CA-199**	**CA-724**	**CRP**	**CEA**	**CA-199**	**CA-724**
χ^2 (6 years)	**4.19*** **	**35.61**	**51.70**	**55.35**	**10.17*** **	**38.83**	**49.07**	**21.60**
χ^2 (4 years)	6.14***	27.64	36.79	33.70	10.70***	31.20	35.08	12.69***
χ^2 (3 years)	**4.92*** **	**30.17**	**41.91**	**23.79**	**14.11** **	**33.61**	**39.15**	**9.83*** **
χ^2 (2 years)	3.48***	17.04*	17.30*	15.11**	6.76***	19.11*	16.64*	6.93***
χ^2 (1 year)	5.07***	2.23***	8.72***	10.13***	9.37***	2.62***	6.74***	2.41***

***Data quality is very good ($\chi^2 < 13.36$ => not significant at 10%-significance level). **Data quality is good ($\chi^2 < 15.51$ => not significant at 5%-significance level). *Data quality is moderate ($\chi^2 < 20.09$ => not significant at 1%-significance level). No star: very poor.

Average study-related costs are defined by experience in real life studies and are set as constants or variables to measure the cost differences in order to investigate cost advantages (or disadvantages). The costs defined are suggested as generally replaceable with any reasonable values; nevertheless the expenditures and outlays in Tab. IV are reliable and coming from real experience.

The costs are calculated depending on the study design considering costs per patient and if needed additionally per proband. Regarding the variable x_1, let the total number of observed patients be p in the first month and the working hours per working student be 2 hours per patient (e.g. study specific preparations and materials), then the first value of the variable (first month) is €150 (€15/h × 2 h × p; p = 5). In a study design, where 1 patient is matched by 2 healthy people, we have €450 for the first value ((€15/h × 2 h × p) × 2 + €15/h × 2 h × p; p = 5). Considering the variable x_2, the first value for the first month is €1800 ((€120 × p) × 2 + €120 × p) for the total number of patients attended by the medicians (study specific medical treatments), analogously for x_3 and x_4 with €1500 and €1050. Summarising for one month, F_M includes in total €53,800. Regarding the variable costs, the total amount of monthly expenditures depends on the number of hours worked as well as across the board costs per patient (and probands). Because of signals that patients are not included in the study, additional costs were included considering the observed number of patients in the study period of 6 years. The monthly number of patients was assigned for the cost function randomly between 5 to 10 patients in a month set by observed experience. All surrounding costs were then included per assigned patient.

The multivariate cost function with the influencing independent variables $x_1, \ldots, x_4$ resulting in total costs K as a dependent variable was reduced to only two summarised independent variables and has the following form:

$$K = 13.000 + 10.67\, x_1 + F_M.$$

IV: *Summary of study-related costs (per month).*

Outlay*	Estimated costs** (Euros) on average per month	Comments
Professor	6000	At least 1 scientist
Scientists	5000	At least 1 scientist
Data manager	5000	At least 1 data manager
Working students or assistants	100 hours × 15	
Medicians and Medical employees		
Medician (probands)	120 × proband	At least 1 medician
(Disease specialised) medician	5000	At least 1 medician
Technical administrative employees		
Medical assistants	4000	At least 1 assistant (per example study nurse)
(Specialised) Medical assistants	4000	At least 1 assistant (per example in the laboratory)
Properly direct costs		
Cost of materials (including medical and biological)	1500	food, medical care needs, devices and other materials
Consumables of probands		
Total Consumables	2000	Pre-analytics, shipping, laboratory, pharmacy economy needs, MRI, analytics virology, microbiology, pathology, transfer and other expenditures
Ground costs (study begin)		
Total Consumables	3000	Travel, meetings, publications and other expenditures
Other operating expenses	1000	Administrative requirements, CEO supplies, maintenance, other ordinary expenses
Other outlays		
Management costs – project	2500	
Management costs – medicians	120 × proband	per proband
Management costs – study nurse	70 × proband	per proband

*Pre-calculation design of medical projects (an extract by the Clinical University of Frankfurt, Germany). **Suggested costs on average set by experience (Brutto (€); case-control study, University Ulm).

The fitted parameter estimate yields the same results generating 1999 bootstrap replicates of the regression coefficients (see Tab. V).

According to all independent variables, the costs reduce from about €3,138,760 (6 years study period) to €535,240 (one year study period). After ensuring statistically that data quality is reliable for reducing study period for 2 or 3 years, what reducing also the number of excluded patients in the study, study costs reduces still to €1,055,560 or €1,582,600. This is about half of the original costs; it is visible that savings due to improved data quality are significant, and study results of any research question still have high quality.

V: *Summary of bootstrap statistics: Fitting the cost function.*

Fitted	Original	Bias*	SE*	Med*	Skew*	Kurtosis*
Intercept	13000	7.7489e-10	8.6162e-10	13000	0.047942	0.31166
V_1	10.67	–1.1724e-13	2.0352e-13	10.67	0.324086	3.17978
F_M	1	1.6653e-15	3.8472e-15	1	–0.356306	0.24296

*Bootstrapped.

VI: *Total summed up fixed and variable costs (one to six years)*

Variable Costs	Number of months	MIN	MAX	MEAN	MEDIAN	SD
K (6 years)	**72 months**	**53800**	**3138760**	**1600426.67**	**1603960**	**907313.78**
x_1	*12 months*	450	50040	25633.75	25965	14431.58
x_2	*12 months*	1800	200160	102535	103860	57726.32
x_3	*12 months*	1500	166800	85445.83	86550	48105.26
x_4	*12 months*	1050	116760	59812.08	60585	33673.68
F_M	*12 months*	36000	2592000	1314000	1314000	753424.18
K (4 years)	**48 months**	**53800**	**2108680**	**1080840**	**1077880**	**609629.09**
x_1	*12 months*	450	34470	17422.5	17145	9904.64
x_2	*12 months*	1800	137880	69690	68580	39618.58
x_3	*12 months*	1500	138000	58075	57150	33015.48
x_4	*12 months*	1050	80430	40652.5	40005	23110.84
F_M	*12 months*	36000	1728000	882000	882000	504000
K (3 years)	**36 months**	**53800**	**1582600**	**819266.67**	**821080**	**457997.65**
x_1	*12 months*	450	25650	13150	13320	7381.87
x_2	*12 months*	1800	102600	52600	53280	29527.5
x_3	*12 months*	1500	138000	43833.33	44400	24606.25
x_4	*12 months*	1050	59850	30683.33	31080	17224.37
F_M	*12 months*	36000	1296000	666000	666000	379283.54
K (2 years)	**24 months**	**53800**	**1055560**	**558720**	**558040**	**308102.39**
x_1	*12 months*	450	16740	8973.75	8910	5023.52
x_2	*12 months*	1800	66960	35895	35640	20094.06
x_3	*12 months*	1500	138000	29912.5	29700	16745.05
x_4	*12 months*	1050	39060	20938.75	20790	11721.54
F_M	*12 months*	36000	864000	450000	450000	254558.44
K (1 year)	**12 months**	**53800**	**535240**	**296920**	**297400**	**158004.86**
x_1	*12 months*	450	8460	4680	4725	2646.84
x_2	*12 months*	1800	33840	18720	18900	10587.36
x_3	*12 months*	1500	138000	15600	15750	8822.8
x_4	*12 months*	1050	19740	10920	11025	6175.97
F_M	*12 months*	36000	432000	234000	234000	129799.85

CONCLUSION

Summarising the main results, the data quality examined by BL conformed to assumptions. After initial observation of over 25% of MD in some variables, it seemed almost impossible to consider reduced study duration. Too many observations had MD. One assumed a need to prolong the study period to avoid unreliable results. The consequence was incurring over half of the really needed study costs. With this work, a detailed investigation could show that the quality of data and thus the study results were even improved, shortening over half of the duration and costs, too.

Analysing the study originally performed by Ma *et al.* (2015) from the year 2000 to 2012 with 385 patients, the examined data quality suffered in reliability. Excluding patients from the original data matrix even signals additionally that the representativeness of the observed data suffers. Handling MD correctly by statistical imputation methods as well as realising the financial advantages means saving time and costs and at the same time improving data towards reliable representativeness and study outcomes. Results regarding quality analysis show that there are clear savings of millions in study costs by reducing the time horizon.

Missing values were observed in all investigated variables. By replacing MD of over 25% MD in some variables (26.2% for the CRP, 4.2% for the CEA, 15.1% for the CA-199, and 28.6% for the CA-724), the data quality was immensely improved, except for one variable (CRP still showed in the beginning good quality, even without MD imputing of 26.2% missing data). Conversely the other variables with moderate to poor data quality were all summarised as improved after statistical input.

In Euros, you save in average €523,126.70 for every reduced year. Compared to a six-year study period, the cost reduction to a one-year period is about 83%. Although the cost reduction is enormous, a study period of only one year is not reliable and the data quality even has the tendency to become worse due to the short period of time. Reducing by more than four years, over 66% of costs are saved, but unlike increasing further there is a positive effect on quality. Reducing by three years still has a cost saving effect of 50% which is roughly the expected half of the costs.

The time reduction is proportional to the cost reduction. This fact seems even more important when the previously stated results even have the side effect of a data quality improvement.

The quality of the original data varies from variable to variable. While the variable CRP seems to be very stable before and after the improvement process, the other variables show some difficulties in the BL deviation comparison. The BL deviation was expected especially in the variable CA-199 digit 6 significantly less than observed evaluating. But replacing MD in a second step, the same digit shows significant results assuming good quality for the digit 6 in the deviation. However summing up all the digits, the overall evaluation differs for the variable, since χ^2 is still greater than 20.09 (significant at a 1% significance level).

Although MD methods are not able to bring all the data back, the knowledge that every poor data registry can be solved ex-post is important. However, the study period should not be reduced too much. There should be a reasonable balance and emphasis between data quality and the study period.

CARE based methods could give further information about investigating the number of excluded patients as a possibility to confirm signals of data quality suffering. To avoid a certain bias which was observed in several studies, the log-linear model controls the dependencies of multiple data sources as an additional advantage of this approach, see Chao *et al.* (2001).

Because there is not any information on the exact number of monthly observed (or not observed) patients, patients (per month) were randomly assigned (Poisson deviation) for the cost function between 5 to 10 patients (assumed by experience). The cost function includes additional possible probands causing costs (independently from the study data source Ma *et al.* (2015) matching two assumed probands to randomly assigned patients – 1:2 matching.

Registry quality depends on several modalities including the researcher's working behaviour itself, which all influence the total end costs, see Tab. I. Suggestions regarding how to plan and conduct a study are essential regarding consideration of hidden costs, see Nagel *et al.* (2012) and (2013). The basic rule is generally besides a well-chosen study design a good detailed plan in advance to keep surprising costs within an acceptable range. This includes standardised study-specific procedures, carefully handling data collection in a standardised manner for all study participants. Also following the international quality standards is essential in the stage of data collection process and earlier in the planning phase. One of these points is controlling questionnaire items and responses from the beginning. This allows – when needed – to recall or re-ask for items not responded to. Up to two weeks recalling not responding participants are much cheaper than excluding patients completely at the end of the study.

In the study of Ma *et al.* (2015), the total number of unknown missing patients seems to be so high that the BL analysis even signalled anomalies (not only caused by missing values). A percentage of these excluded patients possibly could be solved by providing an extract of all important variables anonymised when no improved consent is available or with the help of the state (declaration of

clearance, ethical committee; comparable the successful implemented ALS study), see Nagel *et al.* (2012) and (2013).

Our findings concerning the study Ma *et al.* (2015) show clearly the importance of data quality and analysis in order to take possible study results with highest carefulness to avoid biased conclusions due to incomplete data.

Regarding MD, the decision for the chosen MD method is another important point concerning increased study costs caused by impaired data quality. The decision-making MD algorithm helps to determine a reliable MD mechanism and to consider the scale level of the target variable. Many methods are very complex and the imputing of a large fraction is difficult. On the other hand they are important since a complete case analysis often leads to the loss of much observed data. The number of different algorithms and the fact that no software combines all the existing ones increases the difficulty of finding the right approach. Therefore the decision algorithm makes possible orientation in different MD situations.

In this research, at least the MAR mechanism was assumed and the MCMC method was chosen to replace MD as the most suitable. As shown in Tab. IV, all examined variables improved their quality after MD imputation. The replaced values are reliable, since a scenario of missing and pretended missing values was investigated. The pretended missing values yield similar and almost equal values so that suspect unreliable replacements is not considered as a possible error source (or not seriously). Only very low values (decimal numbers below value 1 of the pretended missing values was estimated as too low (even often under "–100"), but was still not considered serious. The MD imputed χ^2 values improved the data quality extremely. The χ^2 value improved about 61% for the variable CA-724 in a study period of 6 years and even 76% in a study period of one year.

The variable CRP never suffered from MD or data quality as observed in other variables. Although the χ^2 value tended to rise, it rose without falling under the 5% significance level. While the improvement after MD shows a positive effect, the study period itself is an important quality point, too, assuming not only a reduction in the study period, but also reducing possible missing patients in the data matrix. Independent of MD handling, just by reducing the study period to only two years of data collection, the overall χ^2 value improved for all variables on average by over 64% (and more when reducing still further to one year) independent of MD handling. So with both together, reducing the source of bias and years of costs, study costs can be saved doubtless after reliable statistical ensuring over 66% Euros or further counting only one year 83% maximally from 3.138.760 Euros (SD = 907,313.78 for 6 years) to maximally 1,055,560 Euros (SD = 308,102.39 for 2 years) or maximally 535,240 (SD = 158,004.86 for 1 year).

The distinction between fixed and variable costs depends on the studied amount and time. The more variables and decisions in the study considered, the more variable costs must be included in the calculation. Generally seen, the time horizon is a sensitive point. The shorter the study period, the better the data quality and the total study costs in the end. The cost function and detailed costs summed up per patient show how actions additionally influence the cost calculation and their dependency, but although the quality improves when the time horizon gets shorter, one should consider also that conversely certain knowledge can be found only with a longer time period (even when the data quality decreases).

Returning to the BL analysis, BL has proven to be a suitable approach to measure the quality and reliability of study results. Besides supporting decisions to reduce the study period and save costs, BL also helps to detect manipulation in the stage of data entry. The applicability is widely used and depends on specific issues: Investigating data quality (any kind of data) and is suitable for the sensitive purpose (e.g. the pharmaceutical registry), survival distribution research (Leemis et. al., 2000), biased conclusions in survey data (Judge and Schechter, 2009), management manipulation (Archambault and Archambault, 2011) or data fraud detection (Tam *et al.*, 2007).

Still, the disadvantages of BL are that investigation of categorical or dichotomous data is not possible. BL can further find signals that patients are missing in a data matrix, but cannot estimate the total number of missing patients (as with CARE techniques). Whereas an outlook could be to create artificial patients as phantom-patients replacing missing patients and their values for variables and analyse again by means of BL. Distance-based methods can also be chosen to expand the idea of phantom-patients by only given information, which is the location (geographical coordinates of existing observations).

REFERENCES

ARCHAMBAULT, J. and ARCHAMBAULT, M. E. 2011. Earnings Management among Firm during the Pre-SEC-Era: A Benford's Law Analysis. *The Accounting Historians Journal*, 38(2): 145–170.

BANKHOFER, U. and PRAXMEIER, S. 1998. Zur Behandlung fehlender Daten in der Marktforschungspraxis. *Marketing: Zeitschrift für Forschung und Praxis*, 20(2): 109–118.

BANSAL, G. et al. 2008. Tuning Data Mining Methods for Cost-Sensitive Regression: A Study in Loan Charge-Off Forecasting. *Journal of Management Information Systems*, 25(3): 315–336.

BENEISH, M. 1997. Detecting GAAP violation: Implications for assessing earnings management among firms with extreme financial performance. *Accounting Public Policy,* 16(3): 271–309.

BENEISH, M. 1999. The detection of earnings manipulation. *Financial Analysts J.,* 55(5): 24–36.

BENFORD, F. 1938. The Law of Anomalous Numbers. *Proceedings of the American Philosophical Society,* 78(4): 551–572.

BISHOP, Y. M. 2007. *Discrete Multivariate Analysis. Theory and Applications*. New York: Springer Science and Business Media.

BREDL, S., WINKER, P. and KÖTSCHAU, K. 2008. *A statistical approach to detect cheating interviewers*. ZEU Discussion Paper Nr. 39. Giessen: ZEU.

CAPPIELLO, et al. 2003. Time-Related Factors of Data Quality in Multichannel Information Systems. *Journal of Management Information Systems*, 20(3): 71–91.

CECCHINI, M. et al. 2010. Detecting Management Fraud in Public Companies. *Management Science,* 56(7): 1146–1160.

CHAO, A. et al. 2001. The applications of capture-recapture models to epidemiological data. *Stat Med,* 20: 3123–3157.

DE VOCHT, F. and KROMHOUT, H. 2012. The Use of Benford's Law for Evaluation of Quality of Occupational Hygiene Data. *Ann. Occup. Hyg.,* 57(3): 296–304.

ENDERS, C., K. 2010. *Applied Missing Data Analysis (Methodology in the Social Sciences)*. 1st Edition. New York, London: Guilford Press.

ESTEVA, M. et al. 2013. Data Mining for "Big Archives" Analysis: a Case Study. *Proceedings of the American Society for Information Science and Technology,* 50(1): 1–10.

FEWSTER, R. M. 2009. Teachers Corner. A Simple Explanation of Benford's Law. *The American Statistician,* 63(1): 26–32.

GLICKLICH, R. E. et al. 2014. *Registries for Evaluating Patient Outcomes: A User's Guide*. 3rd Edition. Rockville (MD) Agency for Healthcare Research and Quality (US).

HILL, T. 1995a. Base-Invariance Implies Benford's Law. *Proc. Amer. Math. Soc*., 123(3): 887–895.

HILL, T. 1995b. A Statistical Derivation of the Significant-Digit Law. *Statis. Sci*., 10(4): 354–363.

JUDGE, G. and SCHLECHTER, L. 2009. Detecting Problems in Survey Data Using Benford's Law. *The Journal of Human Resources,* 44(1): 1–24.

LEE, Y. W. 2003. Crafting Rules: Context-Reflective Data Quality Problem Solving. *Journal of Management Information Systems*, 20(3): 93–113.

LEE, Y. W. and STRONG, D. M. 2003. Knowing-Why about Data Processes and Data Quality. *Journal of Management Information Systems*, 20(3): 13–39.

LEEMIS, L. et al. 2000. Survival Distributions Satisfying Benford's Law. *The American Statistician,* 54(4): 236–241.

LEWIS, W. R. et al. 2008. An organized approach to improvement in guideline adherence for acute myocardial infarction: results with the Get with The Guidelines quality improvement program. *Arch Intern Med,* 168(16): 1813–1819.

MA, M., CHEN, S., ZHU, B., Y., ZHAO, B. W., WANG, H. S., XIANG, J., WU, X. B., LIN, Y. J., ZHOU, Z. W., PENG, J. S. and CHEN, Y. B. 2015. The clinical significance and risk factors of solitary lymph node metastasis in gastric cancer. *PLoS One*, 10(1): e0114939.

MARITZ, S. G. 2003. Data management: managing data as an organisational resource. *Acta Comercii*, 3(1): 75–85.

NAGEL, G. et al. 2012. Potential of register-based studies to investigate rare diseases. Example of the first German population based amyotrophic lateral sclerosis registry. *Akt Neurol,* 39(1):12–17.

NAGEL, G. et al. 2013. Implementation of a population based epidemiological rare disease registry: study protocol of the amyotrophic lateral sclerosis (ALS) registry Swabia. *BMC Neurology,* 13: 22.

ROTHENBACHER, D. et al. 2015. New opportunities of real world data from clinical routine settings in life cycle-management of drugs: example of an integrative approach in multiple sclerosis. *Curr Med Res Opin,* 11: 1–39.

SILVIA, P. J. 2014. Planned Missing Data Designs in Experience Sampling Research: Monte Carlo Simulations of Efficient Designs for Assessing Within-Person Constructs. *NIH,* 46(1): 41–54.

SPENCER, B. D. 1985. Optimal Data Quality. *Journal of the American Statistical Association,* 80 (391): 564–573.

TAM, C. et al. 2007. Breaking the (Benford) Law: Statistical Fraud Detection in Campaign Finance. *The American Statistician,* 61(3): 218–223.

UENAL, H. et al. 2014a. Incidence and geographical variation of amyotrophic lateral sclerosis (ALS) in Southern Germany The ALS registry Swabia. *PlosOne,* 9(4): e93932.
UENAL, H. et al. 2014b. Choosing Appropriate Methods for Missing Data in Medical Research: A Decision Algorithm on Methods for Missing Data. *JAQM,* 9(4): 10–21.
WANG, R. Y. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers, *Journal of Management Information Systems,* 12(4): 5–33.
WARD, D., H. 1968. "Counting the Cost" – Statistical Methods and Profitability. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* 17(3): 274–276.
WILLIAMS, B. 2002. *Analysis and Management of Animal Populations*. Academic Press.
YEWDALL, G. A. 1969. Cost-Effective Operational Research, Sessions 1 and 2, *Operational Research Society,* 20: 23–24.
ZHANG, S. et al. 2007. "Missing is Useful": Missing Values in Cost-sensitive Decision Trees. In: ZHANG, Z. and SIEKMANN, J. (Ed). *Knowledge Science, Engineering and Management*. Melbourne.

Contact information

David Hampel: qqhampel@mendelu.cz