Health Informatics Meets eHealth
G. Schreier and D. Hayn (Eds.)

doi:10.3233/978-1-61499-858-7-116

Cleansing and Imputation of Body Mass Index Data and Its Impact on a Machine Learning Based Prediction Model

Stefanie JAUK[a,1], Diether KRAMER[b] and Werner LEODOLTER[b]
[a] *CBmed, Graz, Austria*
[b] *Steiermärkische Krankenanstaltengesellschaft m.b.H. (KAGes), Graz, Austria*

Abstract. *Background:* A challenge of using electronic health records for secondary analyses is data quality. Body mass index (BMI) is an important predictor for various diseases but often not documented properly. *Objectives:* The aim of our study is to perform data cleansing on BMI values and to find the best method for an imputation of missing values in order to increase data quality. Further, we want to assess the effect of changes in data quality on the performance of a prediction model based on machine learning. *Methods:* After data cleansing on BMI data, we compared machine learning methods and statistical methods in their accuracy of imputed values using the root mean square error. In a second step, we used three variations of BMI data as a training set for a model predicting the occurrence of delirium. *Results:* Neural network and linear regression models performed best for imputation. There were no changes in model performance for different BMI input data. *Conclusion:* Although data quality issues may lead to biases, it does not always affect performance of secondary analyses.

Keywords. Electronic health records, body mass index, machine learning, data imputation, data cleansing, predictive modelling.

1. Introduction

Within the last years, several prediction models based on machine learning (ML) methods have been published, many of them based on data from electronic health records (EHRs) [1,2]. The advantage of using EHRs as an input for prediction modelling is the big amount of collected longitudinal data, which can be used as training and validation sets [3].

Although the use of EHRs may be beneficial for prediction models, challenges for the secondary use of data need to be considered. Criticism of using EHRs in secondary analyses is often centred on data biases and quality issues. Not only incomplete but also incorrect data may affect the outcome of prediction models. Cruz and Wishart [4] stated that if data used for machine learning is of poor quality the results will be as well. The term "garbage in, garbage out" is commonly used to describe this scenario. Hence, increasing the quality of the data used as input for prediction models may improve model performance and lead to more accurate prediction.

One example of treating missing data in EHR is the work of Jerez et al. [5]. The authors compared different imputation methods in a data set consisting of 3,679 records for modelling early breast cancer relapse. Six artificial neural networks were trained

[1] Corresponding Author: Stefanie Jauk, CBmed GmbH – Center for Biomarker Research in Medicine, Stiftingtalstrasse 5, 8010 Graz, Austria; E-mail: stefanie.jauk@cbmed.at.

using different imputation methods and prognostic accuracies of all models were compared. However, as their aim was to improve the accuracy of the prediction model, they did not evaluate the quality of imputed values but only final accuracy.

While some data in EHRs are missing at random, other records include systematically missing values which may lead to biases. An example for the latter case is body mass index (BMI). BMI is more likely to be measured in patients suffering obesity or anorexia, and records are more likely to be missing in patients within the normal range of BMI [6]. Recording of BMI is also correlated with disease status, e.g. BMI data is more frequent for patients with diabetes mellitus than for those without [7]. Apart from recording bias, reporting biases may exist in BMI data if self-reported, e.g. more missing values for women who categorized themselves as overweight or underweight [8].

In healthcare, BMI data is an important parameter for risk assessment, such as for fractures [9], heart failure [10] or preeclampsia [11]. Besides, BMI data has been used as a modelling feature in several risk prediction models developed with ML. An example is the work of Kramer et al. [12] who developed a model predicting the occurrence of delirium in hospitalized patients. Further investigations of the developed model showed that BMI was one of the variables with highest importance. Furthermore, malnutrition has been identified as a risk factor for delirium [13] and we assume that a low BMI can be used as an indicator.

Due to its importance for risk prediction and keeping in mind the biases of BMI recording, BMI cleansing and imputation methods are essential though challenging. A common problem of clinical data cleansing is the observation of extreme values. Cleansing of clinical BMI data can be difficult, as distributions for weight and height may differ from healthy samples. Freedman et al. [14] illustrated that plausibility intervals for weight and height like the ones set by WHO often result in numerous outliers which are set to missing, even though the values are correct.

Facing the problem of missing data in EHRs, Kontopantelis et al. [6] developed a multiple imputation algorithm for longitudinal body mass index data. Their aim was to produce imputation values for variables with very low individual variability. However, their newly established algorithm did not always perform better than the reference algorithm and appeared to have very long computation times. Another limitation of their work is its applicability on EHRs of hospital information systems (HIS). Many patients do not have any BMI documented in their EHR. In these cases, the algorithm cannot be used as it ignores patients with less than two BMI values.

In order to improve the quality of the secondary use of EHRs, we wanted to investigate and improve the shortcomings of BMI data on routine data of a HIS. The data for analyses belongs to Steiermärkische Krankenanstaltengesellschaft m.b.H (KAGes), the regional health care provider in Styria (Austria). The HIS of KAGes hosts longitudinal health records of around 90 % of all Styrian inhabitants, including hospital stays and outpatient visits over 15 years [12].

Hence, the aim of this study is to (1) establish an automatic method for data cleansing and imputation of missing values for BMI data of a hospital information system, and (2) assess the impact of cleansing and imputation on the performance of an already established prediction model.

In a first step, we evaluate common sources of incorrect values for weight and height due to data entry errors and predefine intervals for plausible values. After cleansing of existing BMI values, missing values will be addressed. In contrast to Kontopantelis et al. [6] we believe that even though height will be quite stable over time, weight

recordings may undergo changes, depending on a patient's health status. In addition, there may be good indicators for extreme values, e.g. correlations between weight and diseases like anorexia and obesity. Imputation approaches may reflect those correlations and result in accurate imputed values. In contrast to the work of Jerez et al. [5] our aim is to obtain imputations closest to real data, not imputation methods that perform best in prediction modelling. If quality of clinical data can be influenced via a data cleansing and imputation step, we expect a change in the accuracy of risk prediction. Therefore, we will assess whether such a step improves the performance of a prediction model.

2. Methods

2.1. Information Extraction

All information was extracted from the KAGes hospital information system *openMEDOCS,* based on IS-H/i.s.h.med information systems and implemented on SAP platforms. In openMEDOCS, height and weight values are recorded via nursing assessment. Patients self-report their current values and nursery personnel enter the numbers in a free-text entry field. BMI values are automatically calculated once height and weight values are available. As long as height and weight are self-reported and not measured, we assume reporting biases in the data.

Between 2011 and 2017 753,701 patients have had at least one admission in a KAGes hospital. Out of those, 39,015 (5.2 %) patients had only height data and 5,300 (0.7 %) only weight. For 346,010 (45.9 %) patients both values were available.

Before extracting lab data and diagnoses for all patients, we defined inclusion criteria: Patients must have at least one admission between 2011 and 2017, height and weight recordings must be available, and age must be over 18 years. The sample resulted in 328,283 patients, with a total of 709,003 admissions. The median BMI for all patients was 25.88 kg/m², with a lower quartile of 22.89 kg/m² and an upper quartile of 29.30 kg/m².

2.2. Data Cleansing

As longitudinal records were available for some patients, we cleansed the data for each admission and then extracted the latest plausible information for each patient.

First, we defined plausible ranges for height and weight data in hospitalized patients. Considering the research of Freedman et al. [14] we chose ample intervals to include correct outliers. For weight and height values we allowed the intervals [20 kg, 250 kg] and [120 cm, 250 cm], respectively.

Second, we analysed the extracted data and discovered several systematic errors that most probably occurred during data entry. We adjusted the implausible values due to these errors, before applying the plausibility intervals for cleansing:

- A small number of EHRs had height values within plausible ranges, but values for weight over 400. We assumed this to be due to a missing decimal point and multiplied weight values by 0.1.
- Some EHRs showed weight within plausible ranges, but height values between 12 and 20. The most probable reason for such values is a missing last digit and multiplied height by 10.

- A larger group of EHRs showed height values below 130 and weight above 130 at one admission. After a visual examination we considered those records to be inverted and simply swapped them.
- In some EHRs we found plausible weight values, but values for height ranged between 50 and 99. Again, we assumed this to result because of a missing digit and corrected this group of records by adding 100 centimetres.

KAGes has already addressed some of those in a data entry control function. For implausible BMI values an alert is triggered during the documentation process. Nevertheless, errors are still possible and previous data needs to be corrected in order to use it for the training of prediction models.

2.3. Imputation of Missing Height and Weight Values

One aim of our study was to compare different imputation methods and to choose the best for further analyses on prediction modelling. We extracted a random sample of 40,000 hospital admissions corresponding to 30,104 patients for whom we compared several statistical methods and machine learning methods. Among possible statistical imputation methods we chose the median imputation, a linear regression model and a linear regression model based on multiple imputation. Machine learning uses the k-nearest neighbor method (KNN), a random forest (RF), a neural network (NN) and a support vector machine (SVM).

All imputation analyses were carried out in R. For the linear regression model with multiple imputation we used the mice package [15]. KNN was modelled with the VIM package and the SVM with the e1071 package. For RF and NN we used the caret package [16] with a 10-fold cross validation as recommend by Kuhn and Johnson [17]. The NN was modelled inside the caret package with nnet, using a feed-forward neural network with a single hidden layer.

We used 75 % of the data for training and tested the results on the remaining 25 % of the data. The training set consisted of 114 features for prediction, including gender, diagnoses and lab data. The features were selected during pre-processing using bivariate analytics. After modelling, we compared the predicted values for the test data set with the real values obtained from the EHR. In order to determine the precision of predictions, we computed the root mean square error (RMSE) for the predictors of BMI between all methods. The RMSE represents the square root of the average of the differences between the real values and the estimated ones.

2.4. Effects of Imputation on a Prediction Model

In order to examine the effect of cleansing and imputation of BMI values on a prediction model we used an already defined model predicting an occurrence of delirium in hospitalized patients. The sample for modelling resulted in 29,568 patients with or without an occurrence of delirium. Out of those, 13,325 patients (45.1 %) had no BMI values. A random forest with down-sampling was used as a classification method and modelled with the caret package in R [16]. The feature set was based on 321 features, including demographic data, ICD-10 diagnoses codes, lab data and BMI. More details of modelling and feature selection can be found in [12].

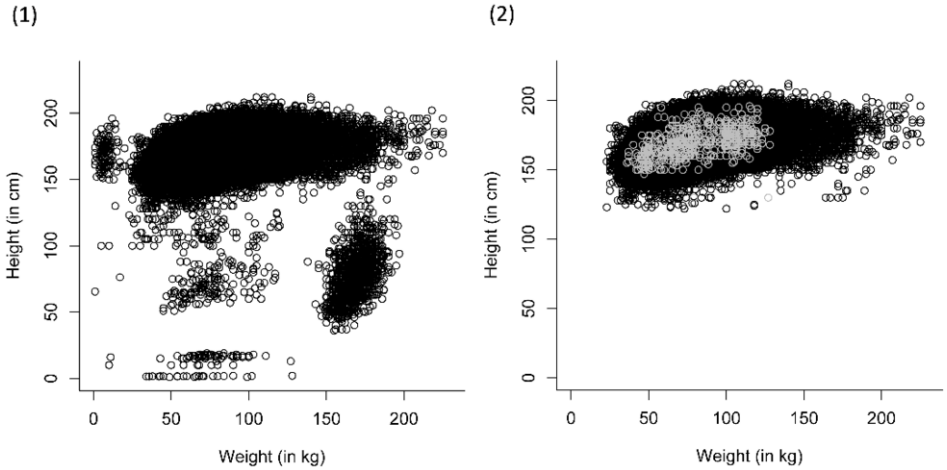

Figure 1. Height and weight records for 709,003 EHRs of adults before (1) and after (2) data cleansing. In (2), values that were affected by data cleansing are highlighted in grey.

In a first scenario, the prediction model was trained and evaluated without including BMI data. For a second simulation, we used the BMI records obtained from the HIS (including wrong or missing values) as a feature. In a third modelling process, the cleansed and imputed BMI values were used as BMI feature. Performance of the prediction model was evaluated using accuracy, sensitivity and specificity.

3. Results

3.1. Data Cleansing of Height and Weight Values

Results of performing the cleansing method on the total sample are shown in Figure 1, comparing original height and weight records and cleansed data.

The group of EHRs with weight values above 130 kg, and height below 130 cm needs to be highlighted. These values build a cluster which is due to the exchange of height and weight entry. The cleansing affected 2,077 records in total. Apparently incorrect values due to data entry were successfully cleansed, as shown in Figure 1.2 with grey data points.

3.2. Comparison of Imputation Methods

To assess the precision of the different imputation methods, we compared the predicted values of all methods with the values obtained from the EHR data for the test data set (Table 1). As expected, computation times were much higher for machine learning methods than for statistical methods. The lowest RMSE was achieved by the NN method with 4.34. However, computation time for NN was the highest with more than ten hours. Results for the linear regression model with simple imputation and the one with multiple imputation were comparable to the NN results, but with computation times below 30 seconds We compared scatter plots of all methods (excluding median imputation) for original and fitted values (Figure 2). For further analyses, we used the linear model with simple imputation as model for imputation.

Table 1. Root mean square errors (RMSE) for predictors of body mass index and computation times using different methods for imputation of missing values.

Imputation method	RMSE	Computation time
Median	5.43	-
Linear model	4.38	0.5 sec
Linear model with multiple imputation (MICE)	4.38	29 sec
K-nearest neighbour (VIM)	5.11	16 min
Random forest (CARET – RF)	4.40	4 h, 40 min
Neural network (CARET – NNET)	4.34	10 h, 12 min
Support vector machine (e1071)	4.49	13 min

3.3. Impact of Imputation on a Prediction Model

Finally, we compared the performance of a prediction model for the occurrence of delirium with three variations of data quality in BMI data. The model performance did not depend on the changes in BMI data and resulted in the same results for all three modelling scenarios. The accuracy for predicting the outcome of a delirium was 0.74, specificity 0.72 and sensitivity 0.88. Furthermore, the prediction outcome for those patients without a valid BMI did not differ between the scenarios with and without imputation.

Although there were no significant changes in the prediction performance, we observed a change in the variable importance when varying the BMI data quality. In the first modelling process the BMI data was excluded and therefore not presented as variable. In the second analysis, the variable for BMI was ranked as the 14th most important variable for prediction, whereas in the third scenario with cleansed and imputed data BMI was the second most important variable.

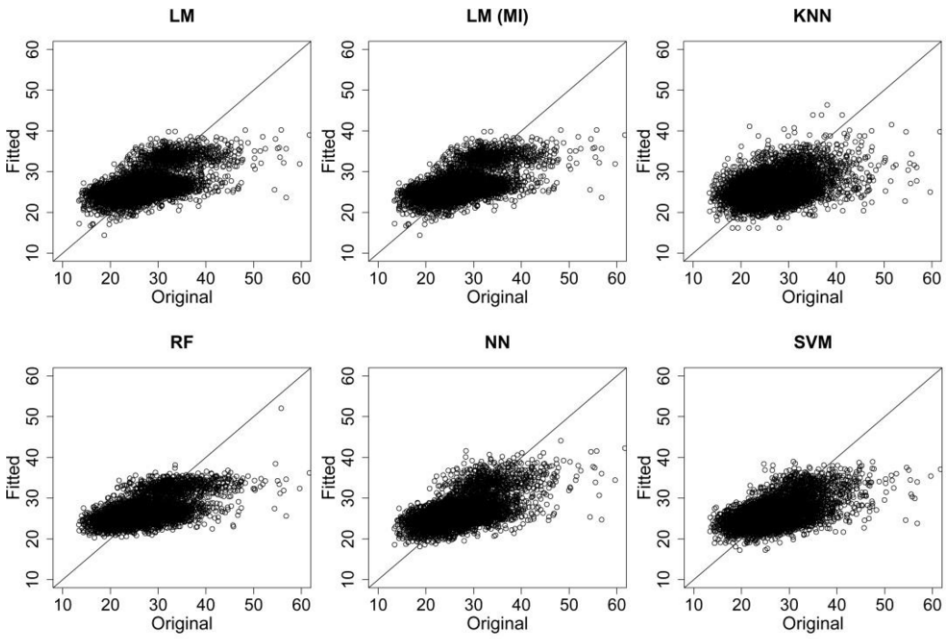

Figure 2. Scatterplots for original BMI values and fitted values for six imputation methods. LM – Linear model; LM (MI) – Linear model with multiple imputation; KNN – k-nearest neigbor; RF – Random forest; NN – Neural network; SVM – Support vector machine.

4. Discussion

In our study, we evaluated different imputation methods for BMI data and assessed their effect of cleansed and imputed data on a prediction model.

We cleansed the BMI data of 328,283 patients obtained from the EHR data of a Styrian HIS and compared imputation methods based on statistical methods and machine learning methods. RMSE was lowest for a NN predicting BMI in our test data set, but with the longest computation time. The fastest predictions with similar results were achieved by the linear regression models with simple and multiple imputation. Imputations based on ML methods varied in their quality of prediction: While RF and SVM resulted in RMSEs comparable to NN, the KNN method showed higher errors.

The use of prediction models in clinical practice requires high precision. Hence, there is a need for best performing prediction models. Despite the commonly used statement "garbage in, garbage out", the model performance in our simulation did not change with the variation of data quality. Even though BMI data was one of the most important predictors in our model and high in missing and incorrect values, the performance of the model was not affected. These results indicate that criticism of bad data quality should not be generalized for all scenarios of secondary use of data.

Although we did not observe a change in performance, the variable for BMI data gained in importance when cleansing and imputing methods were carried out. We assume this to be due to the information that was represented in imputed values: The imputation models included predictors that were afterwards used for risk prediction, e.g. diagnoses and lab data, resulting in a predictive model within a predictive model.

Several authors reported biases in BMI records. Even though it was not our objective to determine biases, further analysis of our data is needed to assess the importance of these and its impact on imputation methods.

A limitation of our study is the correctness of the BMI values that were used for imputation modelling. Although the data was checked for common errors and plausibility ranges were applied, some values might still be incorrect. We do not know how this fact influences the imputation modelling.

Finally, the imputed values need to be used with reservation. All compared imputation models had standard deviations of more than a BMI value of 4 kg/m^2 for predicted values. Considering descriptive statistics for BMI values, patients are likely to change quartiles due to imputation.

To sum up, our results showed that accuracy of imputation did not differ significantly between ML methods and statistical methods in general. If computation times need to be kept down, linear regression models may be preferable. Even though there was an effect of a variation in data quality on a prediction model considering the changes of variable importance, model performance did not change.

Acknowledgements

This work has been carried out with the K1 COMET Competence Centre CBmed, which is funded by the Federal Ministry of Transport, Innovation and Technology (BMVIT); the Federal Ministry of Science, Research and Economy (BMWFW); Land Steiermark (Department 12, Business and Innovation); the Styrian Business Promotion Agency (SFG); and the Vienna Business Agency. The COMET program is executed by the FFG.

KAGes and SAP provided significant resources, manpower and data as basis for research and innovation.

References

[1] Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLOS ONE, 12(4), e0174944.
[2] Hao, S., Wang, Y., Jin, B., Shin, A. Y., Zhu, C., Huang, M., … Ling, X. B. (2015). Development, Validation and Deployment of a Real Time 30 Day Hospital Readmission Risk Assessment Tool in the Maine Healthcare Information Exchange. PLOS ONE, 10(10), e0140271.
[3] Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. A. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. Journal of the American Medical Informatics Association, 24(1), 198–208.
[4] Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. Cancer Informatics, 2, 59.
[5] Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial Intelligence in Medicine, 50(2), 105–115.
[6] Kontopantelis, E., Parisi, R., Springate, D. A., & Reeves, D. (2017). Longitudinal multiple imputation approaches for body mass index or other variables with very low individual-level variability: the mibmi command in Stata. BMC Research Notes, 10(1).
[7] Rea, S., Bailey, K. R., Pathak, J., & Haug, P. J. (2013). Bias in Recording of Body Mass Index Data in the Electronic Health Record. AMIA Summits on Translational Science Proceedings, 2013, 214–218.
[8] Mishra, G. D., & Dobson, A. J. (2004). Multiple imputation for body mass index: lessons from the Australian Longitudinal Study on Women's Health. Statistics in Medicine, 23(19), 3077–3087.
[9] De Laet, C., Kanis, J. A., Odén, A., Johanson, H., Johnell, O., Delmas, P., … Tenenhouse, A. (2005). Body mass index as a predictor of fracture risk: A meta-analysis. Osteoporosis International, 16(11), 1330–1338.
[10] Yang, H., Negishi, K., Otahal, P., & Marwick, T. H. (2015). Clinical prediction of incident heart failure risk: a systematic review and meta-analysis. Open Heart, 2(1), e000222.
[11] O'brien, T. E., Ray, J. G., & Chan, W.-S. (2003). Maternal body mass index and the risk of preeclampsia: a systematic overview. Epidemiology, 14(3), 368–374.
[12] Kramer, D., Veeranki, S., Hayn, D., Quehenberger, F., Leodolter, W., Jagsch, C., & Schreier, G. (2017). Development and Validation of a Multivariable Prediction Model for the Occurrence of Delirium in Hospitalized Gerontopsychiatry and Internal Medicine Patients. In Health Informatics Meets EHealth: Digital Insight–Information-Driven Health & Care. Proceedings of the 11th EHealth2017 Conference (Vol. 236, p. 32). IOS Press.
[13] Inouye, S. K. (2006). Delirium in Older Persons. New England Journal of Medicine, 354(11), 1157–1165.
[14] Freedman, D. S., Lawman, H. G., Skinner, A. C., McGuire, L. C., Allison, D. B., & Ogden, C. L. (2015). Validity of the WHO cutoffs for biologically implausible values of weight, height, and BMI in children and adolescents in NHANES from 1999 through 2012,. American Journal of Clinical Nutrition, 102(5), 1000–1006.
[15] Buuren, S. van, & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 1–68.
[16] Kuhn, M. (2012). caret: Classification and Regression Training.
[17] Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. New York: Springer-Verlag.