Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240780

Comparison of Imputation Methods for Categorical Real-World Prostate Cancer Data with Natural Order

Susanne SCHMITT [a,b,1] and Franz ROTHLAUF [a]
[a] *Johannes Gutenberg University, Mainz, Germany*
[b] *Cancer Registry of Rhineland-Palatinate in the Institute for Digital Health Data, Germany*
ORCiD ID: Susanne Schmitt https://orcid.org/0009-0007-3183-8366

Abstract. Missing values (NA) often occur in cancer research, which may be due to reasons such as data protection, data loss, or missing follow-up data. Such incomplete patient information can have an impact on prediction models and other data analyses. Imputation methods are a tool for dealing with NA. Cancer data is often presented in an ordered categorical form, such as tumour grading and staging, which requires special methods. This work compares mode imputation, k nearest neighbour (knn) imputation, and, in the context of Multiple Imputation by Chained Equations (MICE), logistic regression model with proportional odds (mice_polr) and random forest (mice_rf) on a real-world prostate cancer dataset provided by the Cancer Registry of Rhineland-Palatinate in Germany. Our dataset contains relevant information for the risk classification of patients and the time between date of diagnosis and date of death. For the imputation comparison, we use Rubin's (1974) Missing Completely At Random (MCAR) mechanism to remove 10%, 20%, 30%, and 50% observations. The results are evaluated and ranked based on the accuracy per patient. Mice_rf performs significantly best for each percentage of NA, followed by knn, and mice_polr performs significantly worst. Furthermore, our findings indicate that the accuracy of imputation methods increases with a lower number of categories, a relatively even proportion of patients in the categories, or a majority of patients in a particular category.

Keywords. Cancer registry, MCAR, missing values, ordinal categorical data, prostate cancer, single and multiple imputation

1. Introduction

Missing values (NA) are a large problem for clinical-epidemiological studies using real datasets [1]. In cancer research, incomplete or missing patient information can influence findings about tumour characteristics or treatment methods [2]. This may impact the improvement and further development of individual forms of therapy, prognoses, and prediction models. In addition to data protection reasons, NA occur for various reasons, such as missing examination results on patient characteristics, data loss, patient refusal, illegibility, technical problems, or missing follow-up data [1,3]. The correct handling of

[1] Corresponding Author: Susanne Schmitt, Jakob-Welder-Weg 9, 55128 Mainz, Germany; E-mail: susanne.schmitt@uni-mainz.de.

NA in clinical-epidemiological research and the associated trade-off between information gain and information loss is a controversial topic, where the question is whether all available data should be used or NA removed.

There are a variety of studies on imputation methods for numeric, mixed or binary data where NA are included in the dataset to avoid information loss, error, and bias [2,4,5,6]. However, cancer datasets often contain exclusively categorical variables such as tumour gradings and cancer staging where a natural order exists [2,5], which is why specific categorical imputation methods for NA are required. Although the analysis of such data is relevant for many areas of medical research, there are no general guidelines or specifications for dealing with NA in ordinal categorical cancer data.

This work studies different single and multiple imputation methods for dealing with different proportions of NA. The methods are evaluated and ranked based on the accuracy per patient and per variable to the original dataset without NA. We use a real-world dataset, which contains reportable information on the main characteristics of the primary tumour of patients with prostate cancer (ICD-10 code: C61) and the survival time after diagnosis. All methods benefit from a low number of categories, a roughly equal proportion of the categories of one variable, or if the majority of patients are assigned to one category. Our results can support the selection of methods to deal with NA in categorical cancer data with a natural order and thus improve the quality and reliability of medical data analysis.

2. Methods

2.1. Dataset

Our dataset contains information from outpatient and inpatient male patients with prostate cancer between January 1, 2016 and June 30, 2023. The data is collected by the clinical-epidemiological Cancer Registry of Rhineland-Palatinate in Germany with 2,026,105 male inhabitants in 2020 [7].

Table 1 shows the variables in our dataset. The characteristics of the primary tumour are classified using the international risk classification of prostate cancer [8,9] and the classification system of the World Health Organisation (WHO) [10]. The original dataset consists of 16,709 observations and six variables without NA. All variables have a natural order and are ordinal categorical variables. AGE, SURVIVAL, and PSA are interval-based.

2.2. Imputation Methods

For the comparison, we insert either 10%, 20%, 30%, or 50% NA into the dataset using one mechanism from Rubin (1976): Missing Completely At Random (MCAR) [1,11]. In contrast to the other mechanism of Rubin (1976), Missing At Random (MAR) and Missing Not At Random (MNAR), in which the NA are systematically missing, MCAR assumes that NA are not associated with other observations in the dataset but are selected completely at random. Since we do not know the exact dependencies between the structures and patterns of the variables from our dataset, we use the MCAR mechanism of Rubin (1976). In addition, some single value imputation methods (e.g. mode imputation) require this mechanism to be bias-free and return interpretable results [3,5].

In our study, we compare different imputation methods: *Mode* imputation, *k nearest neighbour* (knn) imputation, and within *Multiple Imputation by Chained Equations* (MICE) the logistic regression model with proportional odds (mice_polr) and the random forest model (mice_rf). Mode imputation replaces each NA with the mode of each category [1]. Knn imputation is part of multiple imputation and calculates the average of the knn of the remaining data with the most similar categories determined by Hamming distance [5,6]. For our experiments, we set k=5 after systematic variation, as this leads to the highest accuracy. MICE uses either mice_polr, which is particularly suitable for ordinal variables, or mice_rf which is suitable for all categorical variables [12]. In our experiments, we used m=5 iterations to estimate the NA for each variable.

We measure the performance of the imputation methods by the number (in percent) of patients where all NA are correctly imputed (ACC) and by the number (in percent) of correctly imputed values of a variable (ACC_variable). For one particular patient, ACC becomes 100% if all NA of the patient are correctly imputed. Of course, ACC_variable is always equal or larger than ACC. For each method, we perform 30 runs and present the mean performance value. The calculations are carried out in R.

Table 1. Selection of variables, their categorisation and meaning in our dataset.

Variable	Categorization	Meaning
AGE	<60, 60-65, ..., ≥ 80	Age of patients in years at the time of observation
SURVIVAL	<1, 1-2, …, ≥ 6, x	Time between the date of diagnosis and the date of death in years. x indicates the survival at the time of observation.
GRADING	l, h	Tumour histological classification system of the WHO l(ow) = well-differentiated (1-2) and mucinous (m) carcinomas h(igh)=moderately poorly-(3, h) and highly-differentiated carcinomas (4)
GLEASON	6, 7, 7a, 7b, 8, 9, 10	Gleason Score
PSA	<10, 10-20, ≥20	Prostate-specific antigen value (measured in ng/m)
TNM_T	1, 2, 2a, 2b, 2c, 3, 4	The Tumour-component of the TNM classification system

3. Results

Table 2 shows the results for each imputation method for 10%, 20%, 30%, and 50% NA. The last row lists the rank of the methods according to their ACC. The rank as well as the relative differences between the different methods do not change with increasing number of NA. Mice_rf achieves the highest performance for all proportions of NA and is ranked first. An exception at ACC_variable level is the variable SURVIVAL, where mode performs best and knn performs better than mice_rf at 10% and 50%. For SURVIVAL, most observations of the original dataset fall into category x with 93% of the patients. Thus, mode returns the best results. Mice_polr achieves the lowest accuracy for all proportions of NA.

Overall, performance decreases with larger number of NA. ACC decreases with increasing number of NA for all methods and is more than five times smaller at 50% for mice_polr and more than two times smaller at mice_rf than at 10%. In all methods, the imputation of NA in the SURVIVAL variable is most accurate, as most observations fall into the x category. For both, GRADING and PSA, we observe good results. Results for GRADING are good as the variable has only two categories (low and high) with about equal share (53% and 47%). Similarly for PSA, most patients have a PSA value of <10 (58%) and 10-20 (21%) in the original dataset resulting in good performance of the

imputation methods. Performance is always lowest for AGE and GLEASON. AGE consists of six categories, where each category has only a share between 10%-20% patients. The same applies to GLEASON with a total of seven categories, with GLEASON=10 occurring in only 4% of patients and GLEASON=7 in only 2% of patients.

We use pairwise Mann-Whitney U tests to test the hypothesis that the ACC values are generated by the same distribution. We find that all pairwise comparisons result in p-values < 0.01, which indicates that all ACC values are significantly different.

Table 2. Performance (accuracy) of imputation methods per patient (ACC) in percent and per variable (ACC_variable) in percent for 10%, 20%, 30%, and 50% NA.

NA			**10%**				**20%**	
Method	**mode**	**knn**	**mice_polr**	**mice_rf**	**mode**	**knn**	**mice_polr**	**mice_rf**
ACC	73.47	78.98	69.17	80.55	53.13	61.74	46.57	64.37
AGE	92.14	94.18	91.37	94.8	84.35	88.21	83.56	89.3
SURVIVAL	99.35	99.32	98.63	99.31	98.67	98.57	97.29	98.59
GRADING	95.67	96.67	95.04	96.87	91.34	93.2	90.05	93.52
GLEASON	92.43	94.34	91.83	94.84	84.86	88.32	83.69	89.55
PSA	96.01	96.25	94.33	96.47	91.98	92.25	88.66	92.8
TNM_T	94.26	95.27	92.86	95.78	88.39	90.15	85.7	91.07
RANK	3	2	4	2	3	2	4	1
NA			**30%**				**50%**	
Method	**mode**	**knn**	**mice_polr**	**mice_rf**	**mode**	**knn**	**mice_polr**	**mice_rf**
ACC	37.92	47.24	30.33	50.56	18.07	24.48	11.36	28.23
AGE	76.54	81.69	75.37	83.45	60.9	66.92	58,73	69.82
SURVIVAL	97.95	97.75	95.89	97.79	96.56	96.06	93.19	95.94
GRADING	87.08	89.35	85.15	90.07	78.49	80.74	75.39	81.88
GLEASON	77.22	81.92	75.56	83.73	62.06	67.55	59.34	70.72
PSA	87.97	87.98	82.95	88.59	79.92	78.46	71.59	79.44
TNM_T	82.7	84.61	78.5	86.13	71.13	72.47	64.23	74.92
RANK	3	2	4	1	3	2	4	1

4. Discussion and Conclusions

When imputing NA for ordinal categorical real-world prostate cancer data, the imputation approach Multiple Imputation by Chained Equations (MICE) using a random forest model performs significantly best; in contrast MICE using a logistic regression model performs significantly worst. A drawback of MICE in comparison to fast models like mode imputation is its higher complexity and computational effort [5,12].

All methods achieve high accuracy if the number of categories is limited, the patients are relatively evenly distributed across the categories, or the majority of patients fall into a particular category. In contrast, all methods perform worse with a larger number of categories that have unequal proportions. Unfortunately, cancer data is often unevenly distributed across categories such as AGE or GLEASON. Nevertheless, we do not recommend merging categories, as underrepresented and marginalized patient groups would not be adequately represented and lead to a loss of information.

For one particular patient, the values of the variables GLEASON, PSA, and TNM_T are not necessarily independent of each other, but may be related depending on the corresponding risk classification. Thus, the approach Missing Completely At Random (MCAR), which we used to generate synthetic datasets with a given number of NA, may

be inappropriate for the given dataset although it can handle single and multiple imputation methods well. For cases, where NA depend on the structures and patterns in the already observed data, the approach Missing At Random (MAR) may be more appropriate [1,11]. Similarly, if there are systematic reasons for NA such as the exclusion of older patients or a specific risk group, approaches like Missing Not At Random (MNAR) may be more appropriate. Although MAR and MNAR lead to a bias when imputing single values, we want to study their influence on imputation methods in future research.

When selecting an imputation method, it is important to analyse the impact of the selection on subsequent studies (e.g. prediction models and the effectiveness of other machine learning methods) [1,2,5]. In addition to imputation methods, there are alternative options for dealing with NA to avoid loss of information. For example, keeping NA as a separate category can contain important information about treatment methods for prostate cancer such as Active Surveillance or Watchful Waiting [8,9]. We leave such analysis to future work.

References

[1] Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, et al. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol. 2017;9:157-66.
[2] Eisemann N, Waldmann A, Katalinic A. Imputation of missing values of tumour stage in population-based cancer registration. BMC Med Res Methodol. 2011;11:1-13.
[3] Heymans MW, Twisk JW. Handling missing data in clinical research. J Clin Epidemiol. 2022;151:185-8.
[4] Ibrahim JG, Chu H, Chen MH. Missing data in clinical studies: issues and methods. J Clin Oncol. 2012;30(26):3297.
[5] Memon SM, Wamala R, Kabano IH. A comparison of imputation methods for categorical data. Inform Med Unlocked. 2023;42:101382.
[6] Waljee AK, Mukherjee A, Singal AG, Zhang Y, Warren J, Balis U, et al. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open. 2013;3(8):e002847.
[7] Statistikamt Rheinland-Pfalz. Bevölkerung und Gebiet - Basisdaten Land: Tabelle 4 [Internet]. Available from: https://www.statistik.rlp.de/de/gesellschaft-staat/bevoelkerung-und-gebiet/basisdaten-land/tabelle-4/.
[8] Heidenreich A, Aus G, Bolla M, Joniau S, Matveev VB, Schmid HP, et al. EAU guidelines on prostate cancer. Eur Urol. 2008;53(1):68-80.
[9] Lancee M, Tikkinen KA, de Reijke TM, Kataja VV, Aben KK, Vernooij RW. Guideline of guidelines: primary monotherapies for localised or locally advanced prostate cancer. BJU Int. 2018;122(4):535-48.
[10] Humphrey PA, Moch H, Cubilla AL, Ulbright TM, Reuter VE. The 2016 WHO classification of tumours of the urinary system and male genital organs—part B: prostate and bladder tumours. Eur Urol. 2016;70(1):106-19.
[11] Rubin D. Inference and missing data. Biometrika. 1976;63(3):581-92.
[12] Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. J Stat Softw. 2011;45:1-67.