# Comparison of missing data handling methods in building variant pathogenicity metapredictors

Mikko Särkkä[1,2], Sami Myöhänen[1], Inka Saarinen[1], Kaloyan Marinov[1], and Jussi Paananen[1,2]

[1]Blueprint Genetics
[2]University of Eastern Finland

## 1 Introduction

### 1.1 Variant pathogenicity prediction

- existing methods, how they deal with missingness

### 1.2 Missingness handling

*Missing data* is common in real datasets. Consider a matrix of $A$ that represents the unobserved, underlying values that would be obtained by data collection in the absence of any missing data generation mechanisms. The subset of values of $A$ that are observed in data collection is denoted $A_{obs}$, and the subset of missing values of $A$ is denoted $A_{mis}$. Of course, the values of $A_{mis}$ will not be known when analysing any real dataset. $M$ is the missingness indicator matrix whose values are 0 when the corresponding value of $A$ is observed, and 1 when the corresponding value of $A$ is missing.

- Define prediction accuracy

Missing data processes may be classified into *missing completely at random* (MCAR), *missing at random* (MAR) and *missing not at random* (MNAR).[**little-rubin**] In a missing data process with a MCAR mechanism[1], the probability of a value being missing does not depend on any observed or unobserved values. In a missing data process with a MAR mechanism, the probability of a value being missing may depend on observed values. In a missing data process with a MNAR mechanism, the probability of a value being missing may depend on both observed and unobserved values.

[1] Check whether one should only use mechanism, process, or both

### 1.2.1 Statistical inference[2] vs. prediction

**1.2.1.1 Statistical inference** Much of missing data literature is focused on treatment of missing data in the context of statistical inference[3].

- Validity of complete cases

In such cases, the methods are designed to ensure, in addition to unbiasedness, that the uncertainty introduced by missingness is correctly reflected in the standard errors. This is in contrast to designing methods simply for accurate prediction of the underlying value[4]. Indeed, naïvely imputing data with a single imputation method[5] is misleading as use of even highly accurate single imputation methods will cause underestimation of the standard errors[6].

- Explain problematicity of RMSE and (predictive) regression imputation

The uncertainty can be properly incorporated via two main approaches: *multiple imputation* (MI) and *maximum likelihood*[7] (ML) estimation. The latter naturally accounts for missing values by modeling both the data generating process and the missingness generating process at the same time[8]. A drawback of this method is the restriction to models that can be estimated via maximum likelihood[9]. The former instead is based on production of multiple complete datasets, on which separate models are fitted and whose estimates are then pooled.

**1.2.1.2 Prediction** A less commonly studied problem[10] is missingness handling in the prediction, as opposed to explanation. See [1] and [10] for discussion of the differences of predictive and explanatory statistics[11].

An important observation regarding the distinction is that the *theoretically correct model* may not be the *best model* with regards to prediction accuracy[10].

**1.2.1.3 Missing values at estimation time** Most studies introducing imputation methods focus on problems where missingness may occur in the model estimation phase, but where data is assumed to be complete at prediction time. In this context, it suffices to design methods which facilitate model estimation[12] in a way that maximizes prediction accuracy.

[2] Use word "explanation" like Shmueli, or "statistical inference"; or "estimation", like Sarle?

[3] Cite examples; kind of a difficult statement to properly evaluate. Stated in a different format [10] on page 296.

[4] Statement needs a lot more precision, and citations

[5] Define single imputation

[6] Find citation from Van Buuren. Use quote about machine learning imputation being the even more dangerous version of regression imputation

[7] Might want to switch up the order, or not.

[8] Add citations

[9] Elaborate, examples of models both MLE and not MLE

[10] can we to quantify this?

[11] Reread these and possibly elaborate

[12] Should I use "model estimation" as a term at all, and just refer to model training? There are two perspectives here, the machine learning and the statistics perspective. Machine learning might be considered a subset of predictive statistics.

**1.2.1.4 Missing values both at estimation and prediction time** Sarle notes that "The usual characterizations of missing values as *missing at random* or *missing completely at random* are important for estimation but not prediction."[**sarle**][13] Ding & Simonoff [**ding-simonoff**] provide real-world evidence[14] in support of this statement in the use of classification trees[15]. Ina predictive context, the presence of *informative missingness* (as defined by Sarle[**sarle**]) in the data, i.e. missingness being dependent on the response variable conditional on $X_{obs}$, may actually lead to improved predictive accuracy compared to complete data.[**citeproperly**]

One additional challenge that arises in this context is that most imputation methods are not implemented in a way that easily allows reuse of learned parameters. That is, it is difficult to first estimate imputation method parameters on the training set, and then impute the test set using those same parameters. This leads to diminished prediction accuracy, as the data the distributions of imputed data differ in the training and test sets. Even worse, since the parameters for test set imputation are estimated from the test set, the content and size of the test set itself may affect prediction accuracy.

- Saar-Tsechansky & Provost

# 2 Materials and methods

## 2.1 Data

### 2.1.1 ClinGen dataset

The dataset consists of ClinGen[8] expert-reviewed single-nucleotide variants from ClinVar[3, 2, 4].[16]

Variants were annotated using the Ensembl Variant Effect Predictor (VEP)[7] version 96 and dbNSFP3.5[5, 6]. In addition, we incorporated annotations used by CADD[9], matching by VEP-annotated Ensembl transcripts. [17]

### 2.1.2 Features

The initial feature set was defined manually[18].

- Observed missingness patterns

[13] Check correct academic formatting for this sort of inline quote

[14] Can I use this expression?

[15] Check this

[16] include date obtained?

[17] add variant number

[18] add selection rationale

## 2.2 Preprocessing

For each variant, transcript specific values from dbNSFP were chosen to match the Ensembl canonical transcript annotated by VEP. Variants whose canonical VEP-annotated transcript ID did not match that from dbNSFP were discarded.

Missing values were replaced by default values for features where the missingness implied the default value *a priori* (e.g. a prediction of effect of amino acid substitution for a protein may be imputed with the neutral value (usually 0) when a variant is intronic) .[19]

Categorical variables were processed to dummy variables.

A binary outcome vector was formed by defining that variants classified as pathogenic or likely pathogenic as belonging to the positive class, and variants classified as benign or likely benign as belonging to the negative class. [20]

The final feature vectors of some sets of variants[21] may be equal (i.e. duplicated). In such cases, only one variant [22] was kept.

Use of categorical variables with high class imbalance within certain levels may obfuscate the performance of the imputation methods due to allowing the classifier to learn to classify all variants with that level into either the positive or the negative class, and therefore ignoring all other features upon which imputation may have been performed. VEP-predicted variant consequence is one such feature. Variants with consequences for which either class had less than 5% of overall variants of that consequence were removed[23][24].

To match an ordinary machine-learning process and to avoid issues with certain imputation methods[25], features with fewer than 1% unique values were removed.[26]

For feature pairs with high correlation, only one of the features was kept[27].

The final processed dataset contains $n$ [28] variants characterized by $m$[29] features.

- Note that features removal leads to "largest feature set that works with all methods"; more elaborate analysis might find larger sets for some of them and thus give an advantage that is now lost

[19] Add table here or in supp. materials

[20] How many in each?

[21] How many?

[22] With the lowest chromosomal position; write this in supplementary notes

[23] Removing how many?

[24] Explain that this would not be done in ordinary training practice.

[25] e.g. in PMM one often got failures due to singular matrices before application of these steps

[26] List these

[27] List removed features

[28] How many?

[29] How many?

### 2.2.1 Data split

The data was randomly split into training and test subsets, with 70% ($N = n^{30}$) of variants in the training set and 30%[31] ($N = m^{32}$) of variants in the test set.

## 2.3 Imputation methods

### 2.3.1 Univariate imputation

The simplest imputation methods impute every missing value within a feature with the same value, which may either be a constant or a statistic estimated from the observed values of the feature.

| Simple imputation methods | Value |
|---|---|
| Zero imputation | 0 |
| Maximum imputation | Maximum observed value within feature |
| Minimum imputation | Minimum observed value within feature |
| Median imputation | Median observed value within feature |
| Mean imputation | Mean observed value within feature |
| Unique-value[33] imputation | For observed values $F_{obs}$ of feature $F$,<br>$\lvert \max(F_{obs}) - \min(F_{obs}) \rvert \cdot 10$ |
| Missingness indicator | For each feature, perform zero imputation and create a binary feature[34] indicating original missing values |

### 2.3.2 Multiple imputation by chained equations

Multiple imputation by chained equations (MICE)[**mice**]

- PMM
- RF
- norm
- norm.predict

### 2.3.3 Other stuff, need name

- BPCA
- k-NN
- MissForest

[30] How many?

[31] Check exact final percentages and add them

[32] How many?

[33] or "outlier"

[34] Make sure to note that duplicated indicator vectors added here are removed

### 2.4 Classifiers

- Logistic regression
- Random Forest

### 2.5 Simulation

- Addition of missingness to already incomplete dataset
- Added missingness patterns
- Note that different simulations may lead to different feature sets due to preprocessing

### 2.6 Evaluation metrics

- MCC
- AUC-ROC
- RMSE for simulations
- Ranging over multiple imputation datasets
- (Non-)reuse of parameters from training in test process

## 3 Results

## 4 Discussion

discussion[35]

## References

[1] Leo Breiman. "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)". In: *Statistical Science* 16 (Aug. 2001), pp. 199–231. DOI: 10.1214/ss/1009213726.

[2] Melissa J. Landrum et al. "ClinVar: public archive of interpretations of clinically relevant variants". In: *Nucleic Acids Research* 44.D1 (Nov. 2015), pp. D862–D868. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1222. eprint: http://oup.prod.sis.lan/nar/article-pdf/44/D1/D862/9483060/gkv1222.pdf. URL: https://doi.org/10.1093/nar/gkv1222.

[3] Melissa J. Landrum et al. "ClinVar: public archive of relationships among sequence variation and human phenotype". In: *Nucleic Acids Research* 42.D1 (Nov. 2013), pp. D980–D985. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1113. eprint: http://oup.prod.sis.lan/nar/article-pdf/42/D1/D980/3584314/gkt1113.pdf. URL: https://doi.org/10.1093/nar/gkt1113.

[35] Commit to either "missing data handling" or "missingness handling" over the whole thing

[4] Melissa J Landrum et al. “ClinVar: improving access to variant interpretations and supporting evidence”. In: *Nucleic Acids Research* 46.D1 (Nov. 2017), pp. D1062–D1067. ISSN: 0305-1048. DOI: 10.1093/nar/gkx1153. eprint: http://oup.prod.sis.lan/nar/article-pdf/46/D1/D1062/23162472/gkx1153.pdf. URL: https://doi.org/10.1093/nar/gkx1153.

[5] Xiaoming Liu, Xueqiu Jian, and Eric Boerwinkle. “dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions”. In: *Human Mutation* 32.8 (2011), pp. 894–899. DOI: 10.1002/humu.21517. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.21517. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/humu.21517.

[6] Xiaoming Liu et al. “dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs”. In: *Human Mutation* 37.3 (2016), pp. 235–241. DOI: 10.1002/humu.22932. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.22932. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/humu.22932.

[7] William McLaren et al. “The Ensembl Variant Effect Predictor”. In: *Genome Biology* 17.122 (2016). DOI: 10.1186/s13059-016-0974-4.

[8] Heidi L. Rehm et al. “ClinGen — The Clinical Genome Resource”. In: *New England Journal of Medicine* 372.23 (2015). PMID: 26014595, pp. 2235–2242. DOI: 10.1056/NEJMsr1406261. eprint: https://doi.org/10.1056/NEJMsr1406261. URL: https://doi.org/10.1056/NEJMsr1406261.

[9] Philipp Rentzsch et al. “CADD: predicting the deleteriousness of variants throughout the human genome”. In: *Nucleic Acids Research* 47.D1 (Oct. 2018), pp. D886–D894. ISSN: 0305-1048. DOI: 10.1093/nar/gky1016. eprint: http://oup.prod.sis.lan/nar/article-pdf/47/D1/D886/27436395/gky1016.pdf. URL: https://doi.org/10.1093/nar/gky1016.

[10] Galit Shmueli. “To Explain or to Predict?” In: *Statist. Sci.* 25.3 (Aug. 2010), pp. 289–310. DOI: 10.1214/10-STS330. URL: https://doi.org/10.1214/10-STS330.