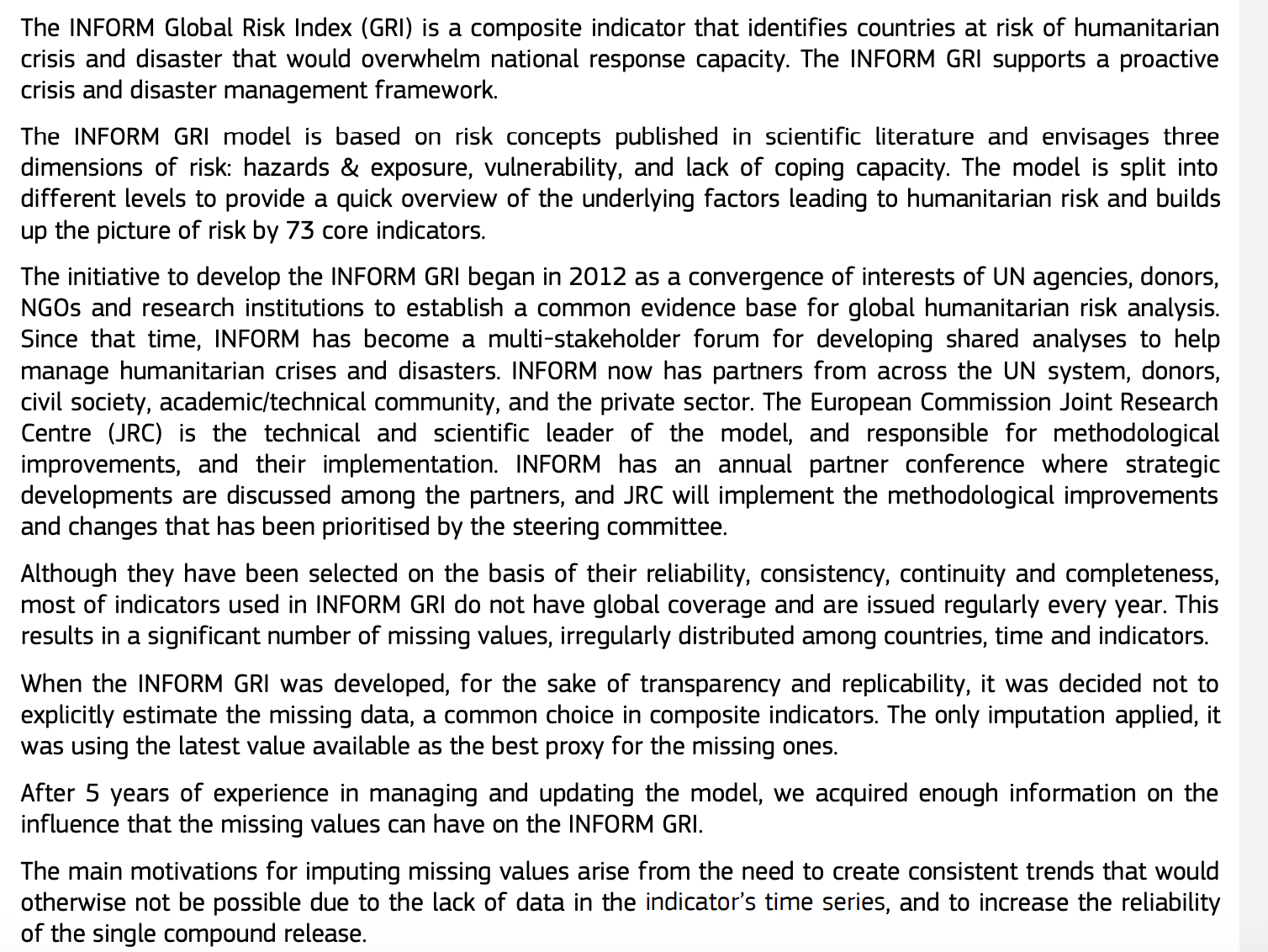
# 1. Introduction

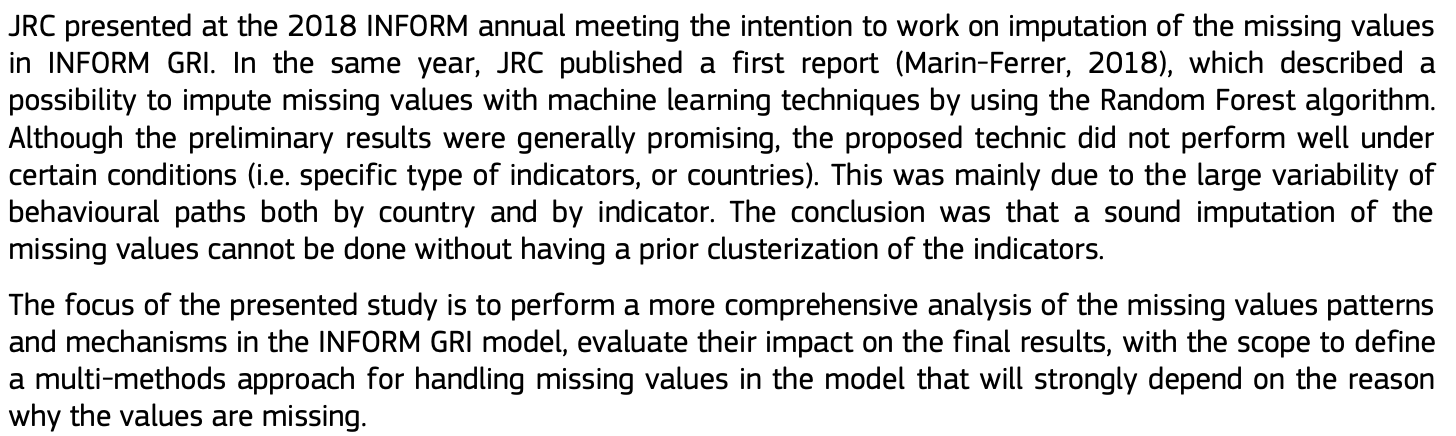
# 2. Background Information

Here I will explain the difference between national and modelled mmr

# 3. Related Work

<https://drmkc.jrc.ec.europa.eu/inform-index/Portals/0/InfoRM/Publications/EUR%2030037%20-%20Imputation%20of%20missing%20values%20in%20the%20INFORM%20Global%20Risk%20Index%20-%20print.pdf>





Unlike in my study, where modelled results were available

# 4. Materials and Methodology

All code was written using Python3 and run in Visual Code Studio or on the Gadi supercomputer at the National Computational Infrastructure.

**Figure 1**: Overview Flowchart of Experimental Workflow

## 4.1 Data Sources and Merging

### 4.11 Data Sources

Data was sourced from a variety of World Health Organisation (WHO) and World Bank Group (WB) data repositories. The final, merged dataset socioeconomic, health-related, and environmental indicators. Information about the specific datasets used in this study was summarised in Table 1, below. The specific variables gathered from each data source are listed in the Appendix, section 9.1. The specific features included in each dataset were originally gathered from a diverse range of sources. See the webpages for more information.

**Table 1:** Summary information about the datasets used in this study.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Type of Dataset** | **Number of Features** | **Date Range** | **Number of Areas Covered** | **Demographic Subsets Chosen from Dataset** | **Source** |
| MMR modelled and estimates | 2 | 1985-2023 | 242 | NA | World Bank Group’s Gender Data Portal |
| Overall health literacy, metrics, agency | 198 | 1960-2023 | 265 | NA | World Bank Group’s Gender Data Portal |
| Illness incidence and prevalence | 193 | 2000-2019 | 194 | Sex | WHO Health Inequality Data Repository & IHME |
| Empowerment | 9 | 1991-2023 | 120 | Economic status (quintiles 1, 5) | WHO Health inequality Data Repository after re-analysis by the WHO Collaborating Center for Health Equity Monitoring |
| Socioeconomic, education, environmental | 64 | 1970-2023 | 195 | Sex, economic status (quintiles 1, 5), residence (urban, rural) | WHO Health Inequality Data Repository, sourced from The World Bank Data Catalogue |
| Income level | 1 | 2024 |  | NA | WHO Health Inequality Data Repository, produced by the WHO’s Global Health Observatory |

The modelled and national estimates for the maternal mortality ratio (MMR) was sourced from the World Bank Group’s Gender Data Portal (1). This dataset contains MMR estimates collected yearly between 1985 and 2023 for 242 regions, countries, territories, and areas.

Some of the datasets used in this study contained disaggregated data. For example, features were sex or economic status specific. However, the MMR estimations were not disaggregated. Including disaggregations as a separate column would therefore produce a missing value in the MMR estimates columns when merging the datasets. To prevent this label variable from having missing values, I did not include the disaggregated demographic groups as a separate variable. Instead, the indicator was replicated for each included subgroup. For example, rather than a single feature, ‘Feature 1’, multiple versions of the feature – one for each subgroup – were included. See Table 2 for an illustrative example.

**Table 2**: Illustrative example of a feature being presented with demographic specific data. The bolded text represents the subgroup being represented.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Country | Date | Indicator 1 **Female** | Indicator 1 **Male** | Indicator 1**Rural** | Indicator 1 **Urban** | Indicator 1 **Quintile 1** | Indicator 1 **Quintile 5** |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

A dataset containing 198 health-specific indicators, including health literacy and health-related decision-making ability, as well as mortality and disease rates, was collected from the World Bank Group’s Gender Data Portal (2). Its data was collected annually between 1960 and 2023, and covers 265 regions, countries, territories, and areas. This dataset was not disaggregated. For brevity, this dataset will be referred to as the ‘overall health’ dataset from this point forward.

Similarly, a dataset containing 193 indicators detailing the incidence and prevalence of a greater variety of illnesses was downloaded from the WHO Health Inequality Data Repository (3). The estimates were produced by the Institute for Health Metrics and Evaluation (IHME) Global Burden of Diseases, Injuries and Risk Factors Study (GBE). The specific data was sourced from the Global Burden of Disease Study 2019 on the Global Health Data Exchange website. This dataset contained nation-specific information collected annually between 2000 and 2019 for 194 countries, territories or areas. I used features from this dataset that were disaggregated by sex. For ease of reference, this dataset will be referred to as the ‘illness dataset’ for the rest of this paper.

A dataset containing 9 indicators describing women’s social independence, decision-making agency, and attitude to violence was downloaded from the WHO Health inequality Data Repository after re-analysis by the WHO Collaborating Center for Health Equity Monitoring (4). The data was collected between 1991 and 2023, describing 120 countries, territories, and areas. I included this dataset’s economic status specific disaggregations (quintiles 1 (poorest) and 5 (richest)). This dataset will be referenced as the ‘empowerment dataset’ from this point onwards.

A dataset containing 64 features related to economic status, education level, and location was gathered from the WHO Health Inequality Data Repository (5) and sourced from The World Bank Data Catalogue. It contains nation-specific information recorded between 1970 and 2023 for 195 countries, territories or areas. I used this dataset’s sex (female, male), place of residence (urban, rural), and economic status (quintiles 1 (poorest) and 5 (richest)) disaggregations. This dataset will be referred to as the ‘socioeconomic dataset’ for the rest of this paper.

Finally, the World Bank’s 2024 categorisation of an area’s income level (1 to 4, with 4 being the highest income level) was sourced from the WHO Health Inequality Data Repository and was produced by the WHO’s Global Health Observatory (GHO) (6).

1. <https://genderdata.worldbank.org/en/indicator/sh-sta-mmrt>
2. <https://genderdata.worldbank.org/en/topics/health#idAllIndicators>
3. Illness prevalence
4. empowerment
5. <https://www.who.int/data/inequality-monitor/data> check source
6. income level

### 4.12 Merging Data

All datasets used in this report contained a column for the country/region and a column with the associated ISO3 country codes, as described in the ISO 3166 international standard. Given that different versions of the same country’s name could be used in different datasets (e.g. United States versus United States of America), I joined the datasets on ISO3 code and year. Thus, the final merged dataset contained a row per area, year unique ID. The rest of the columns were features extracted from the original dataset.

The final, merged dataset contained 16,948 samples and 733 features.

### 4.13 Missing Data Exploration

#### 4.131 MMR Estimates

As described in Table 1, Section 4.11 and reinforced in Figure 2, national and modelled MMR estimates are only available from 1985. Therefore, I will exclude all data collected before 1985 from my analysis.

A graph with orange dots

AI-generated content may be incorrect.

**Figure 2:** Proportion of missing national and modelled MMR estimates per year. The national estimate is shown in orange, and the modelled estimate is depicted in blue.

Figure 2 demonstrates the difference between the national and modelled estimates, with the latter having a much lower proportion of missing data per year (8-10% versus 60-100%). As explained in the background information, this is likely due to the challenges involved in collecting national MMR estimates, especially for lower income countries without vital registration systems. This is highlighted by the difference between the median modelled MMR estimate (86) and median national MMR estimate (19), as the lower median national estimate reflects how the majority of data informing the national estimate is obtained in higher income countries, which tend to have lower MMR estimates. Merging these estimates will be discussed in more detail in section 4.24, below.

What is the justification for combining the metrics?

#### 4.132 Feature Data Availability

As described in the background information, choosing the correct imputation method relies on understanding the pattern of missing data within the dataset. Therefore, I first plotted the proportion of non-missing data per year across all features and countries for all years post-1984 where data was recorded (see Figure 3).

A graph with blue dots

AI-generated content may be incorrect.

**Figure 3:** Proportion of missing data across all countries and features per year after 1984.

Before 2000, the dataset had close to or greater than 90% missing data. In contrast, between 2000 and 2020, the dataset had been 40 and 80% missing data. After 2020, the proportion of missing data again increases to above 90%, likely due to the COVID-19 pandemic increasing the difficulty of data collection. *Therefore, I only used data collected between 1984 and 2020 in my analysis.* This reduced the number of rows in the dataset to 9,550.

There are 5 years where the proportion of missing data is less than 50%. The specific figures are shown below in Table 3.

**Table 3:** Years with the lowest proportion of missing data across all countries and features, rounded to two decimal places.

|  |  |
| --- | --- |
| **Date** | **Proportion of Non-Missing Data** |
| 2000 | 0.44 |
| 2005 | 0.49 |
| 2010 | 0.42 |
| 2015 | 0.43 |
| 2019 | 0.43 |

This pattern is likely due to many of the indicators being reported in multi-year increments.

#### 4.133 Missing Data Pattern

As discussed in the background information, past studies have stated that it is nearly impossible to definitively state whether a dataset is missing at random (MAR), missing completely at random (MCAR), or missing not at random (MNAR) [1]. As a result, researchers have found that treating all data as MAR is a reasonable approach between it lies in the middle of the MCAR to MNAR spectrum [1]. Thus, it is reasonable to treat this dataset as MAR for future data processing and imputation.

Moreover, the pattern of missingness may be MAR because the data is more likely to be missing if it is collected in a year other than 2000, 2005, 2010, 2015, 2019. Thus, the probability of a datapoint being missing is related to observed data (e.g. year). While, there is the possibility that the data is MNAR, as potentially a country may have reduced willingness data that reflects negatively, this may be unlikely due to international reporting obligations [citation]

Nevertheless, in this paper I assume that the data is missing at random and employ imputation methods that fit this case. These methods are discussed in more depth in section 4.26, below.

1. <https://link.springer.com/article/10.1186/s40537-021-00516-9>

## 4.2 Data Cleaning and Pre-Processing

### 4.21 Initial Data Removal

#### 4.211 Feature Data Removal

Should this be part of the iterative thresholding process, or should I start with this?

Figure 3 shows a large amount of missing data. Many studies published in the public health domain use a threshold of 60 to 90%, where variables with more than 90% missing values are excluded from the dataset, as simulation studies have shown that principled imputation methods can produced unbiased results for up to 90% missingness assuming the data is missing at random. [1] 541 of the 733 columns have a missing data proportion of less than 90% (with the MMR estimates included in the count).

1 . <https://www.jclinepi.com/article/S0895-4356(18)30871-0/pdf>

A graph of a number of years

AI-generated content may be incorrect.

**Figure 4:** Proportion of missing data across all countries and features per year between 1985 and 2020 after excluding variables with more than 90% missing data.

Excluding variables with more than 90% missing data resulted in a small shift downward in the amount of missing data per year, such as from close to 90% to close to 80% between 2000 and 2020 (Figure 4). This decrease can be seen clearly in the years with the lowest quantity of missing data, where the proportion of missing data decreased by roughly 10% (Table 4).

**Table 4:** Years with the lowest proportion of missing data across all countries and features, rounded to two decimal places, after variables with more than 90% missing data were excluded.

|  |  |
| --- | --- |
| **Date** | **Proportion of Non-Missing Data** |
| 2000 | 0.33 |
| 2005 | 0.32 |
| 2010 | 0.31 |
| 2015 | 0.31 |
| 2019 | 0.33 |

#### 4.212 MMR Estimate Data Removal

After merging the estimates, any year with a national estimate reported without a modelled estimate could use the national estimate, and vice versa. Therefore, I only removed the area/year combinations with both the national and modelled MMR estimates missing. This removal prevents label variables from needing to be imputed.

As a result, the number of rows in the dataset decreased from 9,550 to 8,712. These rows belonged to countries for whom MMR ratios had not been calculated, and thus model predictions could not be verified.

#### 4.213 Data Removal Summary

As a result of removing variables with more than 90% missing data and samples that were missing both a national and modelled MMR estimate, the final, merged dataset shrank to *8,712 samples and 541 columns*.

### 4.22 Iterative Thresholding

sssssss

### 4.23 Optional Correlation-Based Imputation

The original dataset contained 358 cases where pairs of numeric variables had a correlation coefficient of greater than 0.9 or less than -0.9. For example, the variables ‘Mortality rate, infant (per 1,000 live births)’ and ‘Survival to age 65, female (% of cohort)’, had a correlation coefficient of -0.909. I took advantage of these inter-variable relationships to impute missing values, as the high correlation implies that the columns follow a similar relationship [citation].

Each feature variable that was missing more than 75% of its values was compared to each feature value that had a missing data proportion of less than 50%. The 75% threshold was chosen to ensure that the imputed features had enough missing data that it would be difficult to impute the non-missing data solely from their own values [citation?]. The 50% threshold was chosen to ensure that the paired column used for imputation had sufficient values to be useful. The Pearson’s correlation coefficient was calculated between each pair of columns. After comparing the missing data column with all possible more complete columns, the pairing with the highest correlation was isolated. If the correlation coefficient had an absolute value of greater than 0.9, the more complete column was used to impute the missing column’s values. More specifically, a first order regression model was fit to the more complete column and used to impute the values of the missing data column.

This correlation-based imputation was applied to the original and thresholded datasets, producing four datasets that could later be split into train/test datasets. These datasets were:

1. Original dataset with correlation-based imputation
2. Original dataset without correlation-based imputation
3. Thresholded dataset with correlation-based imputation
4. Thresholded dataset without correlation-based imputation

### 4.24 Splitting into Train/Test Sets

Each of the four datasets listed above were split into separate train and test datasets using the following methodology in a 90:10 ratio.

#### 4.241 Train/Test Split

Each dataset listed above was split into four mini-datasets, each of which contained all the data involving areas categorised into a specific World Bank defined income level (high income, upper middle income, lower middle income, low income). Each of the mini-datasets were further divided into train/test datasets using a 90:10 split. To ensure the train and test sets remained independent, all data pertaining to a specific area (e.g. The United Kingdom) was placed in either the train or test set, with the list of countries shuffled before being split to prevent alphabetical ordering from affecting whether a country went into the test or train dataset. This independence preserving measure meant that the train/test split was not exactly 90:10, but instead within one or two percentage points, as the country data did not always allow for this proportional split. Each income level’s train set was merged, and each income level’s test set was merged to form one overall train/test pair.

This method was repeated for each of the four datasets listed above. Citation for 90:10

#### 4.242 Cross-Validation

Each of the four train datasets were further split into train/validation pairs in an 80:20 ratio. As with the initial train/test split, a specific area could only be in the train or validation set, again resulting in slight deviations from the 80:20 ratio. This split was performed five times to create a 5-fold cross-validation set, where there were 5 permutations of the train/validation 80:20 split. By performing cross-validation, I assessed whether slight changes in the training dataset significantly affected performance, thus commenting on my models’ generalisability. Citation and motivation for this approach

Cross-validation splitting was performed by scikit-learn’s GroupKFold method.

#### 4.243 Encoding Categorical Variables

Many regression-based machine learning models cannot work with categorical data. Thus, I used scikit-learn’s OrdinalEncoder to convert all categorical variables to numeric. For example, this method replaced a country’s name with a number. This encoding was performed on the train and test sets separately to prevent data leakage.

### 4.25 Treatment of Labels: Modelled versus National MMR Estimates

Need citation for the following method

Rather than disregarding the national estimates, I combine the two estimates to produce a more robust overall estimate that benefits from the strengths of both estimation methods. Given that the estimates have different medians and standard deviations (~396 for the modelled estimate and ~165 for the national estimate), they could not be directly averaged. Instead, I first standardised the modelled and national estimates separately using Equation 1a to produce estimates with zero mean and unit standard deviation. Next, the standardised estimates were averaged and combined into a single metric, before being rescaled to the number system of the modelled estimate using Equation 1b (which is simply a re-arranged version of Eq. 1a). This rescaling puts the combined estimate back into a practical scale, with the units of the original ratio.

**Equation 1a:** Standardisation formula

**Equation 1b:** Rescaling formula

### 4.26 Imputation Methods

I applied

### 4.27 Normalisation

S

### 4.28 Bootstrapping

S

## 4.3 Machine Learning Models

### 4.31 Choice of Model

### 4.32 Hyperparameter Tuning

### 4.33 Model Evalation

Imputing\_target.ipynb has the correlations between target and other feature columns

When imputing the test data

* Using 95% thresholded to prevent empty columns
* Always has some imputation (did not use no imputation nor solely std)
* Later dropped rows with missing data to prevent any missing data from preventing use

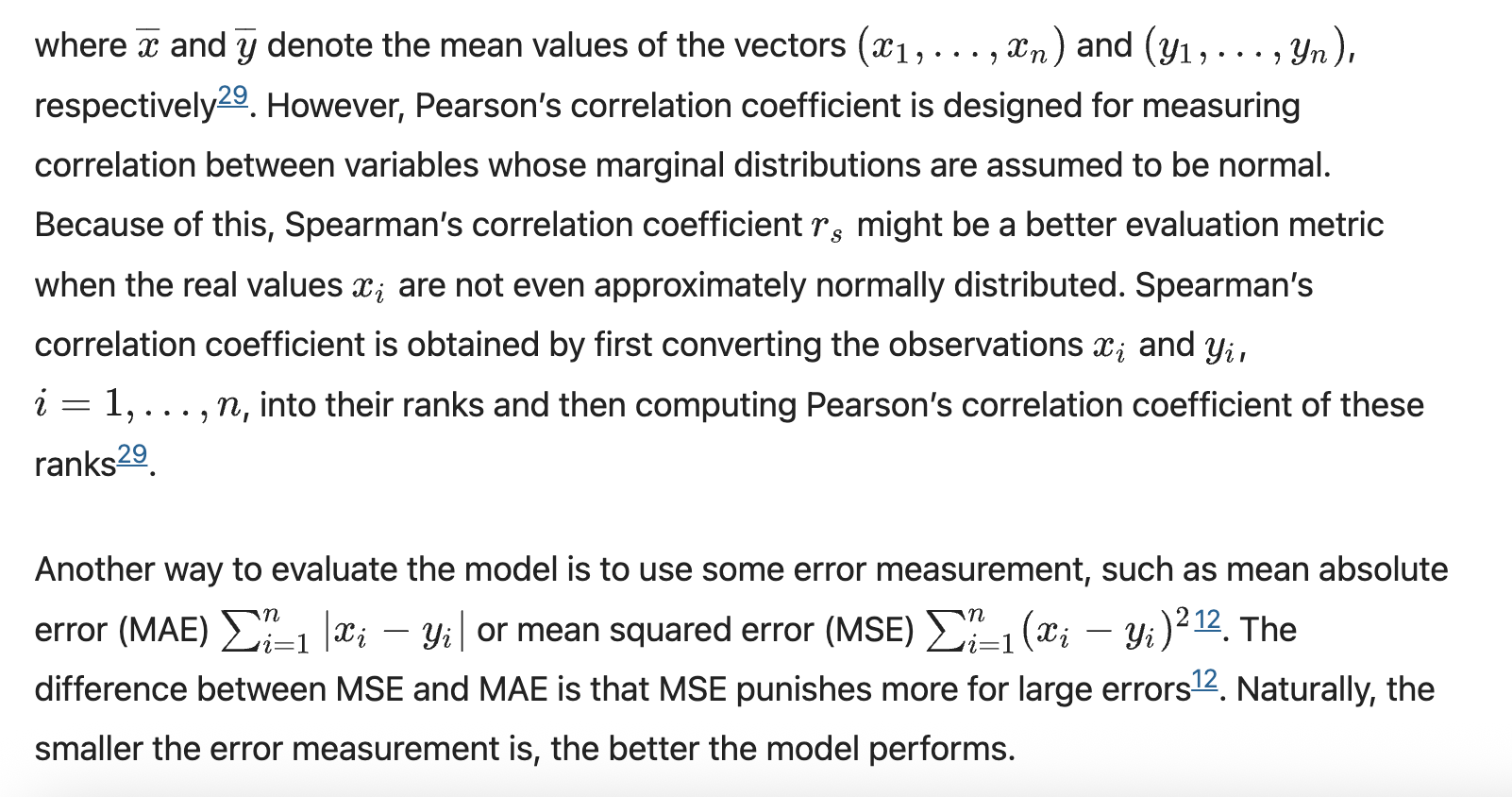
Combined data from all imputation methods to prevent a model trained on one to have a significant advantage

* Could maybe reflect natural variation across time and demographics

Consider:

* The training and validation values are split randomly assuming that data samples are split iid
  + No, because they could be different years

Loss vs performance metrics

* <https://link.springer.com/article/10.1007/s10462-025-11198-7>
* <https://www.nature.com/articles/s41598-024-56706-x>
  + 

Standardising y-values for each separate fold to prevent data leakage

Use anova to see differences in model/imputation/dataset methods

**MICE method:**

* call method to iteratively update missing values multiple times (say 10) based on all others
* take one of the datasets imputed using this method for training
* then use full mice method to see confidence intervals of variables

usually have knn, mice, regression, simple, em, ml, missforest, svm

idea for method: first test all of the imputation methods single imputation, then use multiple imputation for best to get the confidence intervals on the imputation

<https://pmc.ncbi.nlm.nih.gov/articles/PMC6902303/#sec1-7>

A white paper with black text

AI-generated content may be incorrect.

<https://www.nature.com/articles/s41598-024-56706-x#Sec10>

A screenshot of a test

AI-generated content may be incorrect.

# 5. Results

# 6. Discussion

# 7. Conclusion

# 8. References

# 9. Appendix

### 9.1 Features From Each Data Source