UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Filling the Void: Imputing Missing Educational Level Information in Cause of Death Data

Carol Ogira
Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :
Prof. dr. Christel FAES

SUPERVISOR :
Prof. dr. Brecht DEVLEESSCHAUWER

2024
2025

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Filling the Void: Imputing Missing Educational Level Information in Cause of Death Data

Carol Ogira

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :
Prof. dr. Christel FAES

SUPERVISOR :
Prof. dr. Brecht DEVLEESSCHAUWER

Acknowledgements

I would like to express my sincere gratitude to everyone who provided support in any kind; intellectually, emotionally, and even through the simplest of light moments that gave me the zeal to keep going.

To my supervisors, Prof. dr. Brecht Devleesschauwer, Prof. dr. Christel Faes, and Aline Scohy, I am deeply grateful for your guidance, support and for always being available throughout this research. Your feedback, insights and direction through your expertise have been instrumental and have truly contributed to making this journey enjoyable. Working under your supervision has been a privilege and a great learning experience.

In a special way, my gratitude goes to VLIR-UOS for the opportunity to study at UHasselt. I can't fail to thank the entire UHasselt fraternity, especially my professors, who have collectively built the foundation upon which this work stands.

To my family, your support and belief in me were always felt, even from miles away. I am forever grateful for every word of encouragement and your constant prayers. To my friends, far and near, you were my cheerleaders. Thank you for replacing my worries and stress with cheer and laughter.

Above all, all glory and honour to God Almighty for making all things possible.

Abstract

Background: Educational attainment is a key socioeconomic indicator widely used in mortality analyses to evaluate health inequalities in a population. However, missing education levels (ELs) in the data limit these analyses. This study analyses the patterns of missingness of this variable in the Belgian linked cause of death (COD) data, and implements a method to impute the missing ELs.

Methods: Linked COD data were provided by Statbel, and ill-defined deaths redistributed by Sciensano. Patterns of missingness were evaluated to determine key variables for the imputation process. To complement the descriptive analyses, a mixed-effects ordinal logistic regression model was also fitted to further inform the methodology. A probabilistic imputation procedure was then developed and implemented based on Bayes' rule, using COD, age groups, sex and region as the explanatory variables. The model-based approach was also used as a comparative method to the manual probabilistic procedure.

Results: The proportion of missing data in this study was approximately 10.4%, among which the ELs for the cohort below 15 years old were completely missing. The covariates indicated subtle systematic patterns in missingness, especially when examined in combination with one another. Age group was the most influential variable in the data, and inclusion of COD information better captured the underlying structure in the data. The imputation process generated datasets that preserve the underlying distribution in the data, even in relation to the covariates; an observation that was made from both the probabilistic and model-based approaches. The complete dataset indicated that the majority of deaths were more likely to be for individuals with a lower education level.

Conclusion: The imputation procedure filled the gaps in educational attainment, leading to complete datasets that would be used to examine health disparities by ELs. The methodology can be flexibly tailored to datasets with similar challenges, thus improving mortality-related analyses based on the complete datasets.

Key Words: Education level, Missing data, Probabilistic imputation, Cause of death, Burden of disease, Health inequalities, Belgium

Contents

List of Abbreviations

COD	Cause of Death
EL	Education Level
GBD	Global Burden of Disease
IDD	Ill-defined Deaths
ISCED	International Standard Classification of Education

List of Tables

List of Figures

1 Introduction

1.1 Background

Effective assessment of health interventions and formulation of public health policies require accurate and comprehensive vital statistics. Among these, statistics on causes of death (COD) are essential for comprehending mortality patterns and guiding public health initiatives. They inform the allocation of resources in the health sector, and provide critical insights for epidemiological research [1]. Historically, the procedures and systems for designating the underlying cause of death have evolved from as basic as mortality announcements in the 16th century [2], to civil registration systems in the 18th century, advancing and laying the groundwork for the International Classification of Diseases (ICD), which is currently used for uniform COD reporting worldwide [3]. This has led to high-quality COD data for monitoring the health of populations, thus transforming the ability of governments, institutions and researchers to analyse mortality and inform targeted interventions.

In Belgium, COD data is acquired in a series of steps. The process begins when the standardised death certificate (Model IIIC or IIID) is filled in by a certifying doctor upon the demise of an individual. The forms are subsequently finalised by the municipal authorities, and then sent to the regions, which review, code, and add the information to their systems. Thereafter, the information is sent to Statbel, the Belgian statistical office, where the databases are consolidated [4]. As part of the process, Statbel links the death certificate data to the National Register of Natural Persons (RNPP), ensuring completeness of the information and enabling alignment of deaths of Belgian residents and non-residents. This linkage allows the addition of more demographic and socioeconomic variables to enrich the COD data. Among these variables is educational attainment, which originates from the population census based on administrative data.

Educational attainment is a key socioeconomic determinant of health, as it has been demonstrated to be strongly linked to various health indicators such as access to healthcare services, mortality rates and morbidity prevalence [5, 6]. Highly educated individuals tend to live healthier lifestyles and are likely to be associated with pronounced health-seeking behaviour, leading to better health outcomes and increased longevity. On the other hand, previous studies and reports on the health of the Belgian population have demonstrated a strong association between lower education levels (ELs) and adverse health-related behaviours. The 2021 health status report [7] outlined that individuals with low EL are more likely to smoke daily, consume sugary drinks more frequently, and have higher obesity rates (BMI $\geq$ 30). Additionally, a study on the evolution of life expectancies between 2001 and 2011 by Renard et al. [8] indicated that life expectancies increased across all ELs, but the increase was more substantial among highly educated individuals.

In addition to the relationship between educational attainment and various health-related behaviours and outcomes, it is important to evaluate how educational inequalities manifest in mortality patterns. Incorporating EL information into analyses focused on COD allows

for an examination of which CODs disproportionately affect the various levels of education, thus offering essential insights for targeted interventions and public health policies. However, the linked COD databases have missing EL information for some records, hence limiting the ability to fully investigate the health disparities. Earlier studies conducted by Renard et al. [9, 10] on educational disparities in premature mortality have also identified this missingness as a significant limitation in their work. In these studies, the missing EL information was either treated as a separate category or excluded from analyses, potentially introducing bias in the results due to underestimation of the inequalities.

1.2 Sources of Missingness

In earlier censuses, some of the sources of missingness for EL data were as a result of non-response in census forms, where individuals with low EL were less likely to declare, and very sick people were unlikely to complete the forms. Due to changes in data collection methods in the 2011 census, where questionnaires were no longer used, missing EL data was then a result of new migrants whose information was not in the existing databases [8]. These sources have evolved such that in the 2021 census data, missingness was mainly linked to international migration and the increasing diversity of educational backgrounds. Some of the reasons for missingness include migrants, as well as Belgians whose highest EL was obtained abroad, and the diplomas do not have direct equivalence in Belgium. This issue also extends to European schools in Belgium, whose credentials are not included in the databases [11].

1.3 Research Question and Study Rationale

The main research question considered in this study is: 'To what extent does COD information provide added value in imputing missing EL?' Addressing the missingness of EL information in mortality datasets is crucial because it poses a challenge in reliably examining health inequalities by educational attainment. This study seeks to address the issue by evaluating the contribution of COD information to the imputation process, which would then guide the steps followed in the multiple imputation process. Specifically, the study implements a two-dimensional probabilistic redistribution approach to fill in the missing information. This two-dimensional approach refers to a sequential methodology where, in the first dimension, deaths for which the underlying cause is not clearly specified or not well defined (commonly referred to as ill-defined deaths (IDDs)) are first redistributed, then the information is used to impute missing ELs in the second dimension.

The first dimension has been previously developed and implemented by Devleesschauwer et al. [12]. The process entails the redistribution of IDDs to specific causes following a four-step probabilistic procedure. This procedure is based on a stratification approach, where probabilities of specific ICD-10 codes are calculated within each stratum of age group and sex, and the target code is randomly sampled from the causes in the stratum. In the first step, the redistribution is done for IDDs with clearly explained codes; in cases where a stratum has a small number of deaths, the target distribution is obtained from sex only. In the second step, intrinsically uninformative causes such as 'unspecified heart failure' are

redistributed based on packages that consist of ICD-10 codes that are related and relevant to the CODs. The third step involves an internal redistribution, where uninformative codes are randomly assigned to cause mentioned in the death certificate. In the final step, all remaining IDDs are proportionally redistributed over specific causes within the last 5 years.

This study builds upon the work of Devleesschauwer et al. [12], developing a probabilistic technique for the imputation process that uses the redistributed CODs together with other demographic characteristics in the linked dataset, as a vital step towards achieving complete datasets for producing valid and reliable indicators of health inequalities.

1.4 Study Objectives

The primary objective of this study is to develop and evaluate a reliable probabilistic technique to impute missing ELs in the Belgian linked COD dataset. Specifically, the study aims to:

I. Analyse the patterns of missingness for education level in the linked COD data.

II. Examine the extent to which COD data can be used to impute missing education level

III. Impute missing education level via a two-dimensional redistribution process using Monte Carlo simulations.

IV. Assess the validity of the imputed results.

2 Data Description

The dataset used in this study is a subset of the data obtained from the work of Devleesschauwer et al. [12] for the year 2022, where ill-defined deaths have been probabilistically redistributed in 100 iterations, thus 100 datasets with completely imputed COD. Each dataset includes the demographic characteristics of the individual, the calculated years of life lost (YLL) and information on the underlying COD. Only CODs vary from dataset to dataset as a result of the imputation procedure, while the rest of the variables are identical in all the datasets. For purposes of this analysis, only one of the 100 datasets is used, such that the methods developed can be applied to the other datasets. Table 1 provides a summary of the selected variables in the dataset that are relevant to the study.

Table 1: Overview of the variables in the dataset

Variable	**Description**	**Type**	**No. of categories**
Age group	Age group at time of death	Categorical	6
Sex	Sex	Categorical	2
Region	Region of death	Categorical	3
Province	Province of death	Categorical	11
Education Level	Highest education level attained	Categorical	9
YLL	Years of life lost due to premature mortality	Numerical	-
ICD	Redistributed ICD-10 code for the underlying COD	Categorical	1663
GBD3	Level 3 GBD clusters	Categorical	121
GBD2	Level 2 GBD clusters	Categorical	20
GBD1	Level 1 GBD clusters	Categorical	3

The age groups are defined within the following categories: [0-5), [5-15), [15-45), [45-65), [65-85) and 85+. Years of life lost (YLL) is a pre-calculated metric that indicates the number of years lost due to premature mortality. It is calculated by multiplying the number of age-specific deaths by the standard expected residual life expectancy at age of death, from the GBD 2019 reference life expectancy table [13]. ICD refer to the ICD-10 code indicating the underlying COD, which is further grouped into hierarchical nested categories called "Levels", with Level 1 (GBD1) as the highest and Level 3 (GBD3) as the lowest in the dataset. For example, Tuberculosis, a level 3 cause, is nested within HIV/AIDS and tuberculosis (level 2), which is nested within Communicable, maternal, neonatal, and nutritional diseases (level 1). GBD level 3 was used as the covariate for COD, consistent with the majority of Sciensano's reports, especially the website displaying inequality estimates related to the Belgian Burden of Disease [14]. Overall, in the Belgian National Burden of Disease Study (BeBOD), there are 130 unique GBD3 categories. However, the dataset used in this study had only 121 categories since some CODs were not represented in the redistributed iteration.

Educational Attainment

Classification of the education levels follow the 9-point scale according to the International Standard Classification of Education (ISCED) [15], adopted by Statbel. Level 0 refers to individuals whose highest attainment was below primary education; this level includes the early childhood development and pre-primary education programs. Level 1 represents primary education, Level 2 refers to lower secondary education, and Levels 3 and 4 represent upper secondary and post-secondary non-tertiary education, respectively. Level 5 is for short-cycle tertiary education, which encompasses practical and job-oriented programmes that prepare students for the job market or for other tertiary education programmes [16]. Level 6 indicates completion of a bachelor's degree or equivalent, Level 7 is completion of a master's degree, and Level 8 represents the highest level of attainment, i.e., doctorate or equivalent.

For analytical purposes, a new variable was created that consolidates the 9 levels into three categories, following the ISCED standard. Levels 0 to 2 were categorised as "Low", Levels 3 and 4 as "Medium", and Levels 5-8 categorised as "High". Additionally, a binary variable was created to indicate whether the education level was observed or missing for each individual:

$$R_i = \begin{cases} 1, & \text{if EL is observed} \\ 0, & \text{if EL is missing} \end{cases}$$

All the variables in the dataset were fully observed with no missing values, except education level, which is the target variable for imputation.

3 Methods and Materials

3.1 Patterns of Missingness

Exploratory Analysis

To address this objective, an exploration of the patterns and mechanisms of missingness in the education level variable in the linked dataset was conducted. The variables considered were COD, sex, age group and region. These variables were selected based on their established relationship with educational attainment and their relevance in Belgian health research. Besides COD, which is the primary variable of interest, sex and age group were considered because they are fundamental demographic determinants of educational attainment. In addition to these, region was considered important due to the documented differences in educational levels, as well as health inequalities among the three regions. For instance, Statbel reports that 77% of the 26-64 year-olds in Flanders have at least upper secondary EL. In Wallonia, the percentage of this population is 69%, and even lower in Brussels, where the proportion is 66%.

This exploration was conducted in three parts. First, the distributions of observed ELs across the variables, as well as in strata created by combining the variables, were explored as a basis for understanding the composition of the data and any existing associations. For example, if from the data it would be that individuals in one age cohort are more likely to have a certain educational attainment compared to another cohort, the relationship would be an indication of the importance of age in explaining educational attainment. Second, patterns of missingness were also examined, where for each stratum based on a single or combination of variables, the proportion of missing values relative to all records in that stratum was evaluated. This enabled the identification of any systematic patterns of missingness in the covariates to provide guidance on the variables that would be essential for imputing the missing information. Additionally, these results would provide preliminary evidence of the potential underlying missingness mechanism. Specifically, systematic associations between the missing education information and the observed variables would suggest that the data are likely to be missing at random (MAR). Otherwise, it would be that the data are likely to be missing completely at random (MCAR). Third, comparisons of the characteristics of the fully observed and the missing subgroups were done to determine whether the demographic characteristics differed in these subgroups. Notable differences from this comparison would mean that complete-case analyses on the data would lead to biased estimates and conclusions, hence the relevance of imputation.

Little's MCAR Test

A test for formally assessing whether data are MCAR was developed by Little, known as Little's MCAR test [17]. The null hypothesis in this test is that the data are MCAR, against the alternative that the data are not MCAR. The test evaluates mean differences across various subgroups in the data for cases that share the same pattern of missingness [18], with the distribution of the test statistic being asymptotically chi-squared. A key assumption

of this test is multivariate normality, and departures from this assumption would lead to unreliable results. Considering the categorical nature of the EL covariate, this test was deemed inappropriate for this study. Additionally, among the various limitations of this test highlighted by Enders (2010) [18] is that when evaluating the mean differences, the test assumes that the missing data patterns have a common variance-covariance matrix. Hence, deviation from MCAR due to covariance would not be detected. Another important limitation of this test-based approach is that a statistically significant p-value from the test would only reject the hypothesis that the data are MCAR, implying that they are MAR or MNAR, but does not distinguish between the two. The test is also not conclusive for the MCAR mechanism since a non-significant result does not necessarily prove MCAR. This study, therefore, relied on the insights from the exploratory analyses to determine whether or not the data are MAR.

3.2 Model-based Assessment

To complement the exploratory analyses, a model-based approach was used to further examine the extent to which COD, as well as the other variables, explain educational attainment based on the observed data. Given the nature of the variable of interest, which is categorical and has a natural ordering, an ordinal logistic regression model was fitted. A series of models were fitted, considering each covariate separately and all possible combinations of the covariates. For models with COD, mixed effects models were fitted where COD was treated as a random effect.

The rationale for the mixed effects model was based on the large number of COD categories (121), resulting in 120 parameters for a single covariate if considered as a fixed effect, hence increasing the risk of overfitting and affecting the parsimony of the model. Moreover, the model efficiency would also be affected if there are CODs with few observations, leading to unreliable estimates. The random effects, on the other hand, introduce the advantage of "borrowing strength", where estimates for the sparse CODs are shrunk towards the overall mean, leading to more reliable estimates.

Ordinal Logistic Regression Model

The model considered in this analysis is the proportional odds model described by McCullagh [19]. In this model, the three-level categorisation of EL (low, medium and high) was considered as the outcome instead of the nine ISCED levels for interpretability and to ensure adequate sample sizes in each level, hence reducing sparsity issues. For the fixed effects models, i.e., models without COD, the general equation of the model is of the form:

$$\text{logit}[P(Y \leq j|\mathbf{X})] = \log\left[\frac{P(Y \leq j)}{P(Y > j)}\right] = \alpha_j - \boldsymbol{\beta}\cdot\mathbf{X} \tag{1}$$

Where:

- Y is the outcome (education level)
- $j = 1, 2$ refers to the two thresholds (or cut-points) such that:

  - α_1 is the log odds of being in low vs medium or high ELs, and
  - α_2 is the log odds of being in low or medium vs high EL
- $\mathbf{X}$ is the vector of the covariates in the model, and
- $\boldsymbol{\beta}$ is the vector of their regression coefficients

On the other hand, the general form of the equation for mixed effects models is:

$$\text{logit}[P(Y \leq j|\mathbf{X}, \mathbf{b})] = \alpha_j - (\boldsymbol{\beta}\cdot\mathbf{X} + \mathbf{b}) \tag{2}$$

Where $\mathbf{b}$ is a vector of the random intercepts for the 121 CODs, and $b_i \sim N(0, \sigma^2)$. For instance, the equation for the model with sex, region and COD as the covariates becomes:

$$\text{logit}[P(Y \leq j)] = \alpha_j - (\beta_1 \cdot \text{Sex} + \beta_2 \cdot \text{Flanders} + \beta_3 \cdot \text{Wallonia} + b) \tag{3}$$

,with the reference category for sex being male, and for region is Brussels.

To check the proportional odds (PO) assumption, a numerical approach was used, where two ordinary binary logistic regression models were fitted separately and their empirical odds ratios (OR) compared to the OR from the ordinal model. Below is the approach used in creating the variables:

$$\text{bin1} : \begin{cases} 0, & \text{for low EL} \\ 1, & \text{for medium or high EL} \end{cases} \qquad \text{bin2} : \begin{cases} 0, & \text{for low or medium EL} \\ 1, & \text{for high EL} \end{cases}$$

The PO assumption implies that for any covariate in the model, the binary ORs should be similar to the OR from the ordinal model, such that if from the ordinal model, a variable increases the odds of being in a higher category, both binary ORs would reflect this phenomenon. In instances where this is not the case, e.g., if some estimates vary greatly, to the extent of changing directions, then the partial proportional odds models were to be considered since the PO assumption is violated, and the affected variable would be allowed to have different slopes for each EL threshold [20]. The choice of this approach was due to the presence of random effects, which makes it difficult to use the standard tests to check the PO assumption.

The ordinal logistic models were implemented using appropriate functions from the *ordinal* package in R [21], which has functions that allow inclusion of random effects and is flexible to allow fitting a partial proportional odds model in case of violation of the PO assumption. For binary models used to check the PO assumption, the *glm()* and *glmer()* functions were used accordingly.

The Akaike Information Criterion (AIC) values for each model were thereafter evaluated to determine the model that provides a better fit to the data. Importantly, since one of the objectives of the study was to examine the extent to which COD information can be used to impute missing ELs, comparison of AIC for the models with and without COD would provide insights to assess the contribution of COD relative to the other covariates.

3.3 Imputation of Missing Education Levels

3.3.1 Probabilistic Approach

In this second dimension of the redistribution process, missing ELs were imputed in two steps. In the first step, the imputation was done for individuals under 15 years old, i.e., age groups [0,5) and [5,15). The regulation implemented in the 2021 Census [22] stated that the education level for individuals under 15 years should be stated as Not applicable (NAP). Therefore, missing ELs for these individuals were assigned the code "NAP". In the second step, the probabilistic approach was used to impute data for the remaining age groups, using the 9-level ISCED classification to retain the granularity of the ISCED scale. Here, the probability of an EL was calculated conditional on COD and the demographic variables, making use of Bayes' rule [23, 24]:

$$P(E|\text{COD}, \mathbf{X}) = \frac{P(\text{COD}|E, \mathbf{X}) \cdot P(E|\mathbf{X})}{P(\text{COD}|\mathbf{X})} \tag{4}$$

Where E is education level and $\mathbf{X}$ is a vector of the demographic variables; age group, sex and region. On the left-hand side (LHS) of the equation is the set of probabilities of each possible EL of the individual, given the specific values of their characteristics (i.e., COD, age group, sex and region). The components on the RHS, used to estimate these probabilities, are calculated specific to the values of $\mathbf{X}$ and COD for the missing record. As an illustration, suppose for a record with missing EL, the characteristics were as follows: COD - Ischemic heart disease, sex - Male, age group - [45-65) and region - Wallonia. Then;

- $P(\text{COD}|E, \mathbf{X})$ is the probability of COD being Ischemic heart disease, given the demographic characteristics above and each educational attainment. These values can only be calculated from the subset with ELs observed for this disease.
- $P(E|\mathbf{X})$ is the probability of an EL given the age group is [45-65), sex is male, and region is Wallonia. This is calculated from the broader population, for each EL.
- $P(\text{COD}|\mathbf{X})$ is the aggregate probability of dying of Ischemic heart disease, given that sex is male, age group is [45-65), and region is Wallonia, across all ELs.

This results in a set of estimated probabilities for each possible EL, representing how likely the individual is to have attained each level based on their observed characteristics.

These calculations were done for every combination of the covariates, with the ELs and their probabilities stored in a list of lists. Imputation was then done by randomly sampling an EL from the list, depending on the characteristics of the individual whose EL information was missing. For instance, if from a given combination of values of the variables in the observed subset the following probabilities were calculated for each EL: (ISCED 0: 0.61, ISCED 2: 0.10, ISCED 7: 0.29), the imputation value would be obtained by randomly sampling between ELs 0, 2 and 7, using these probabilities. In case of combinations of COD and $\mathbf{X}$ (sex, age group, region) in the missing subset that were not observed in the complete subset, and cases where COD is sparse in the data (i.e., less than 10 records) a

fallback strategy was adopted, where the probabilities were calculated based on age group only. This method was preferred to a stratification approach (i.e., creating strata defined by combinations of sex, age group, region and COD, then calculating probabilities empirically), since the stratification technique would be easily affected by sparse cells.

Since the datasets from the first dimension (IDD redistribution) were probabilistically redistributed using 100 iterations, the procedure from this second dimension was only performed once for each dataset. This way, at the end of the two-dimensional redistribution process, there would be 100 datasets with complete datasets, i.e., all IDDs are redistributed and missing EL data imputed, hence accounting for the uncertainty of the imputed values for these two variables.

3.3.2 Model-based Imputation

While the focus of the study was to use a probabilistic framework to impute the missing EL information, a model-based alternative was also considered to evaluate the robustness and possible differences between the two methods. The ordinal logistic regression model described in Section 3.2 was fitted to the fully observed subset, and the predicted probabilities were obtained based on the estimates of the coefficients in the model. These predictions were then used to impute the missing information. Similar to the probabilistic approach, this technique was applied to the 100 datasets, hence 100 fully-imputed datasets.

One drawback of this method is that it uses and imputes the three aggregated levels of education (low, medium and high) instead of the ELs imputed in the probabilistic approach (ISCED 0-8). Therefore, for comparisons, the imputed values from the probabilistic method were regrouped into the three categories.

This comparison helped to evaluate whether the simpler, probabilistic method sufficiently captures the underlying structure in the data without the need for a model that would require additional complexities, such as validity of the assumptions on which the model is based, or whether the model-based approach leads to more plausible results.

Evaluation of the Imputed Results

To evaluate the plausibility of the imputed values, the distribution of the imputed ELs in the final dataset (fully imputed) was compared to the complete subset (fully observed data, pre-imputation) across the key variables. Through this comparison, it was possible to assess whether the imputation preserved significant patterns in the data and maintained consistency with the population's known structure. Using insights from the comparison of the differences in demographic characteristics initially done in the exploratory step as the basis, outstanding variations between the fully imputed dataset from the original would suggest potential issues with the imputation process.

These checks, however, were considered as informal evaluations of internal consistency and plausibility of the results and do not confirm the correctness of the imputed values since

the true education levels for these missing cases are unknown.

3.4 Software

All analyses were conducted using the R software version 4.4.2 [21]. Statistical tests were performed at a 5% level of significance.

4 Results

4.1 Distribution of Education Levels

Overall distribution

The subset of linked COD data considered for this analysis had a total of 116,378 records, with only 12,113 (10.41%) missing. Educational attainment for all individuals under 15 years, i.e., age groups [0,5) and [5, 15) was missing, the total count being 522 (0.45% of the dataset). ISCED levels 1, 2 and 3 had the highest proportions in the dataset, with ISCED level 5 having the least representation(approximately 0.3%) as shown in Figure 1.

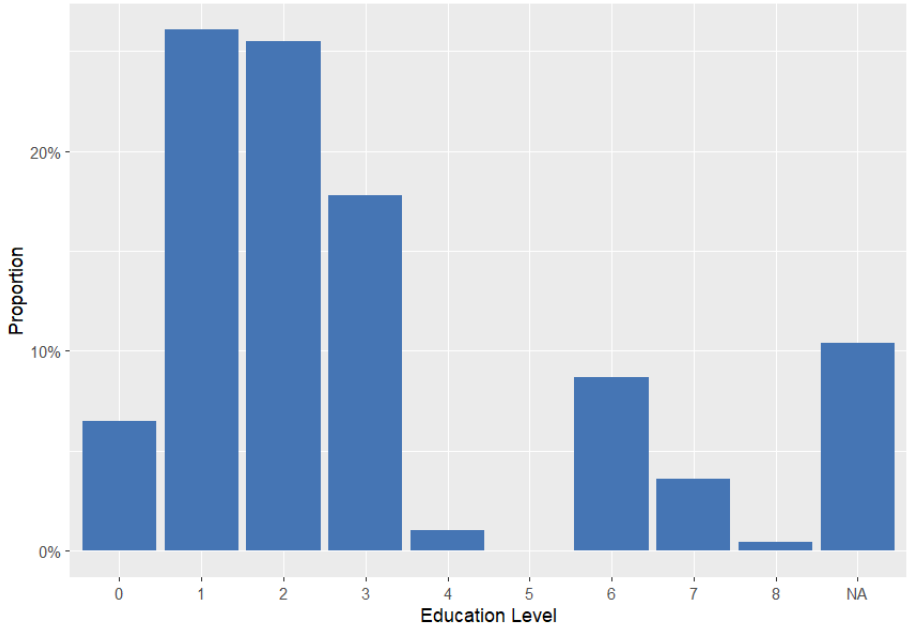


Figure 1: Overall distribution of education levels in the dataset

Due to the complete missingness of EL information in the two age groups below 15 years, analyses on the fully observed subset had only four age groups.

Educational Attainment by COD

A heatmap displaying the proportions of ISCED levels across various GBD level 3 causes of death was generated to explore whether educational attainment differs systematically between individuals who died from more common versus rare diseases. Each line in Figure 2 represents a specific COD, ordered from rare CODs (top) to the most frequent ones. While there was no clear systematic pattern between the rare and common diseases across the ELs, the plot generally indicated dominance in the lower ELs, i.e., ISCED levels 1, 2, and 3, which was reflected in the common and rare cases alike. For instance, educational attainment for individuals who died of rare cases such as COD with very few to only one count in the whole dataset indicated proportions close to 100% (represented by the bright yellow shades), and were more dominant in ISCED levels 1,2,3 and in a few cases, level 4. However, this appeared to be a reflection of the general distribution of the data (shown

in Figure 1). Due to the high cardinality of the COD variable (121 categories), it was relatively difficult to observe or characterise distinct patterns across the ELs; hence, the need for model-based assessment conducted in the sections that follow.

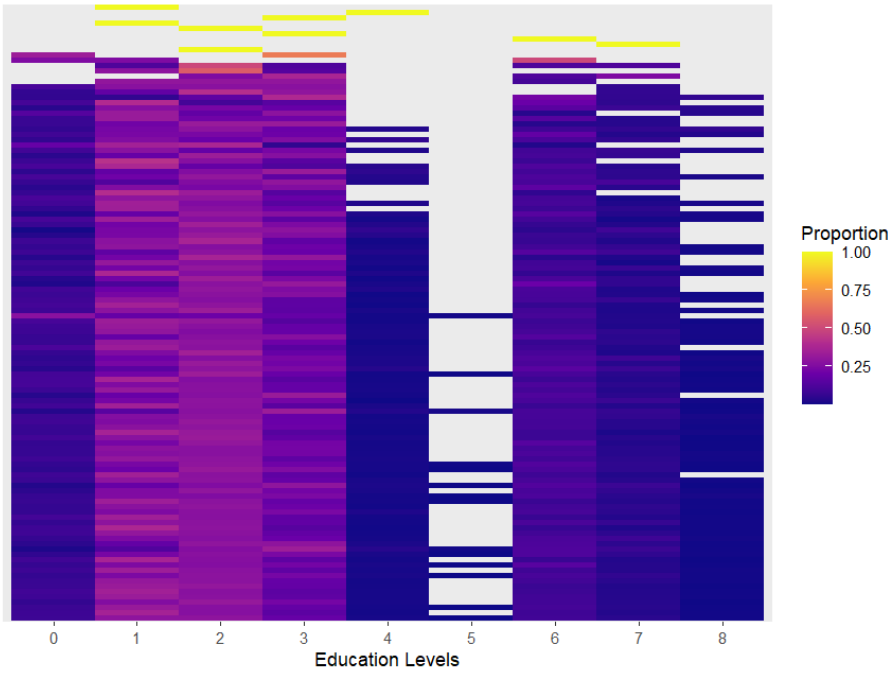


Figure 2: Distribution of Educational attainment across GBD Level 3, with CODs arranged from rare CODs (top) to the most common ones in the dataset

Educational Attainment by Age groups and Sex

The overall proportions of observations among males and females were fairly close, with 50.8% of the individuals being females and 49.2% males. Sex, by itself, showed similar proportions among males and females across most of the ELs, except for ISCED level 7 (master's or equivalent), where males had a notably higher proportion compared to females (**Appendix A.1**). For age groups, the older population (85+) mostly had up to ISCED level 1 (primary education), while for most 15-65 year olds the educational attainment was generally above primary education (EL beyond level 1). Upon combining age groups and sex, both males and females indicated similar trends in how proportions of ELs shift across age groups as shown in Figure 3. For instance, among both sexes, proportions for the lowest EL (ISCED level 0) tend to increase in the older age groups. However, subtle differences were observed, especially in higher ELs (ISCED level 7 and 8, which represent master's level and doctorate or equivalent, respectively). For females, proportions for these ELs indicated a clear decline as the age groups progressed, while in the male population, the proportions tend to be relatively the same across the age groups, which suggests historical gender disparities in regard to educational attainment. A similar observation was made for ISCED level 6 (bachelor's degree or equivalent), where the proportions fluctuated only slightly among males, with 45-85 year olds having higher proportions. Females, on the

other hand, had substantial proportions in this EL, though with a notable decline in the older cohorts. This indicates that for the population in the dataset, age could be considered a good indicator of educational attainment. Sex, by itself, had a less pronounced role but indicated some subtle differences when considered in combination with age group.

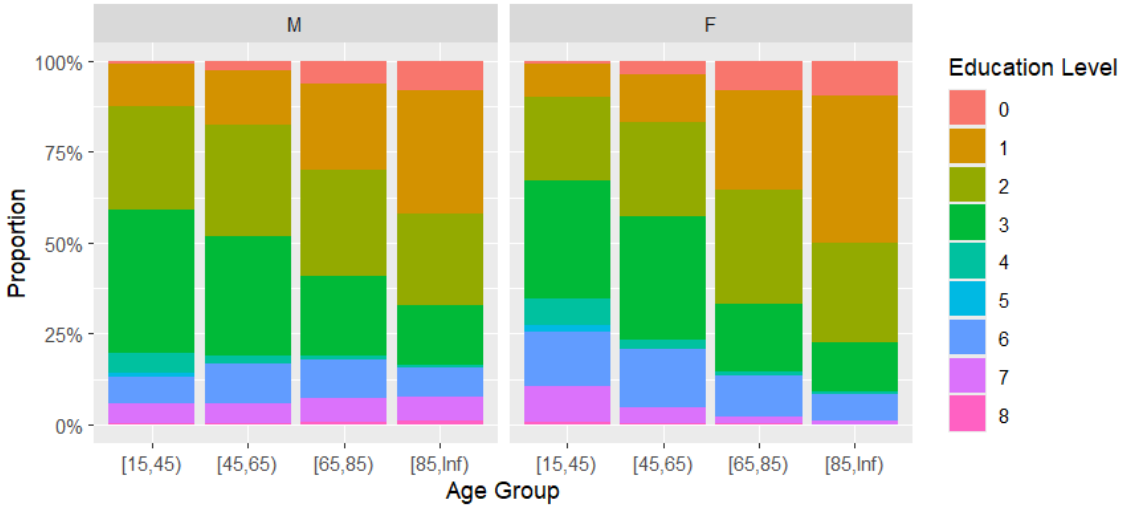


Figure 3: Distribution of ELs across age groups and sex

Distribution by Region

Overall, Flanders had the highest population in the dataset, with a proportion of approximately 58.6%; while Wallonia and Brussels had 34.2% and 7.2% respectively. The summary presented in Table 2 highlights the differences in EL composition, especially in the order of ELs with the highest proportions. Although with minimal differences in the percentages, Brussels indicated ISCED level 2 as the most dominant, followed by level 3, then level 1. Different orders for the highest proportions were observed in Flanders (levels 1, 2, then 3) as well as in Wallonia (2, 1, 3).

Table 2: Proportions (in %) of observed ISCED levels in each region

ISCED Level	Brussels	Flanders	Wallonia
0	8.450	5.344	8.018
1	16.968	29.487	22.704
2	20.803	25.042	27.611
3	18.224	18.193	17.301
4	1.358	1.045	0.868
5	0.023	0.048	0.003
6	10.996	8.090	9.306
7	7.262	3.231	3.426
8	1.052	0.382	0.375

4.2 Patterns of Missingness

Univariate Analyses

Sex, by itself, showed relatively similar probabilities of missing data, with a 9.32% proportion among males and 10.7% among females. Among the four age groups, the 85+ year-olds were more likely to have missingness compared to the younger population, and , among the three regions, Brussels was the most likely to have missingness (**Appendix A.1**). Figure 4 shows the pattern of missingness across COD categories. The proportions appeared fairly similar across the levels, and the higher proportions seen in the least frequent can be attributed to the sparsity of these CODs. However, due to the high number of categories and the nominal nature of this variable, the patterns could not be easily deciphered without the risk of overinterpreting the plot. Therefore, a higher level of the COD (GBD1), with only three categories, was examined to determine any substantial differences. Contrary to the overall distribution of data among the GBD1 categories, where non-communicable diseases had the highest proportion in the data, the category of communicable, maternal, neonatal and nutritional diseases was more likely to have missingness of EL (11.2%), compared to non-communicable diseases (9.9%) and injuries (8.7%).

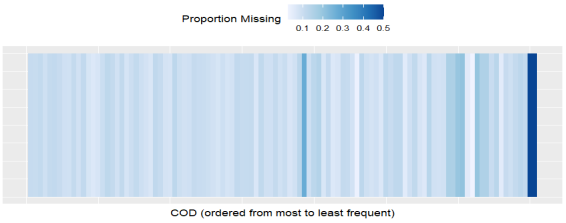


Figure 4: Patterns of missingness for the 121 COD categories ordered by most frequent (left) to least frequent (right)

Patterns of Missingness by Age groups, Sex and Region

Figure 5 displays a notable variation of missingness across the strata defined by age groups and sex. 45-65 year-olds had lower proportions of missing data for both sexes. In addition to that, younger ([15, 45)) and older (85+) cohorts had higher proportions of missingness, with the female population being slightly more prone to having missing data. The middle age group ([45,65)) indicated a slightly different pattern compared to the others, where males had a slightly higher proportion of missingness. Given these subtle patterns, it is probable that these demographic characteristics are likely to influence the probability of missingness.

Adding regions to the strata showed even more conspicuous patterns, where Brussels generally had higher proportions of missing data compared to the two other regions. In Brussels, the proportions for males decreased across the age groups; an occurrence that is observed in this region only. Additionally, compared to other regions where females exhibited a tendency to have similar or higher proportions of missingness compared to males in the majority of the age groups, Brussels had a reverse pattern, where in all age groups except in the oldest cohort (85+ year-olds), males tend to have higher proportions of missingness. These observations, though relatively nuanced, suggest that the missingness mechanism

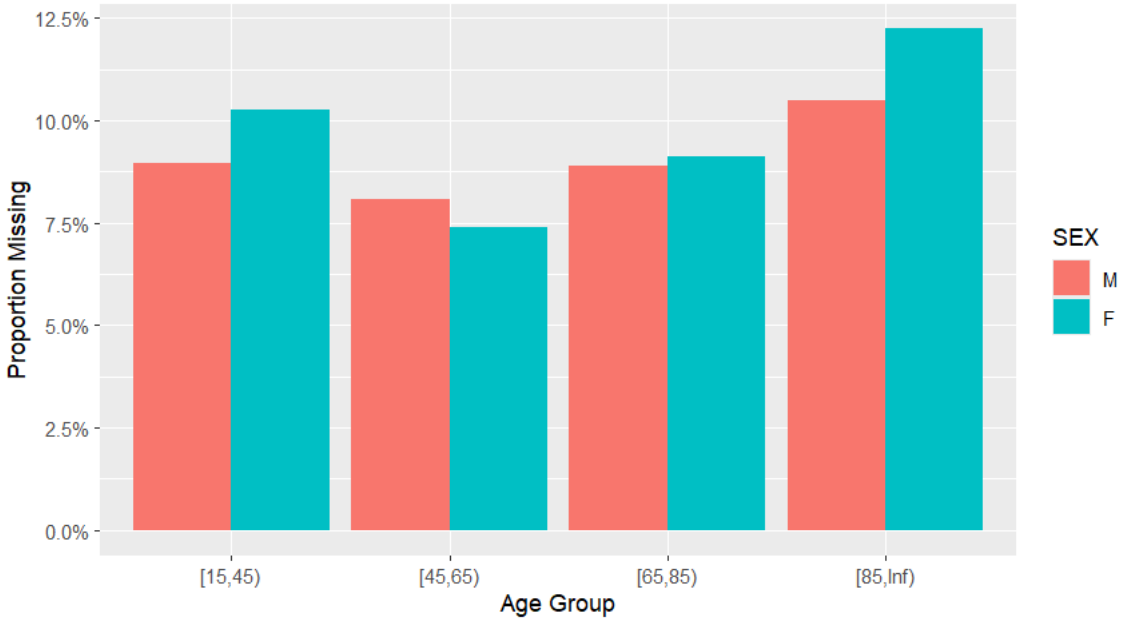


Figure 5: Patterns of missingness by age groups and sex

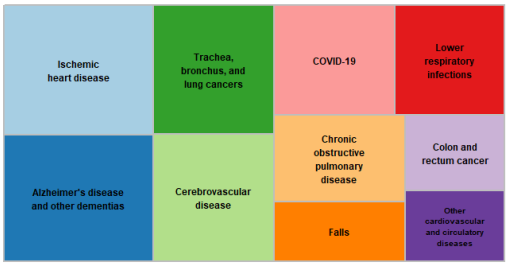
Figure 6: Patterns of missingness by age groups, sex and regions

in this dataset is more likely not to be MCAR. These findings highlight the necessity to address the issue of missingness through imputation.

Characteristics of the demographic variables in the complete and missing subsets

Considering only the ten most frequent CODs for illustration, Ischemic heart disease and Alzheimer's disease were the top two; the remaining causes differed only by order, but were the same in both subsets.

(a) Fully Observed



(b) Missing

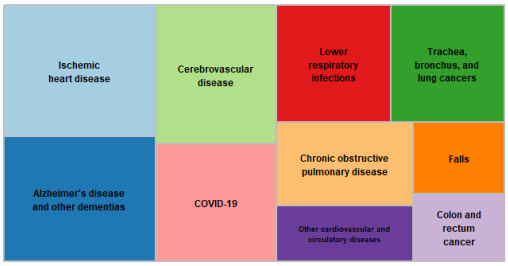
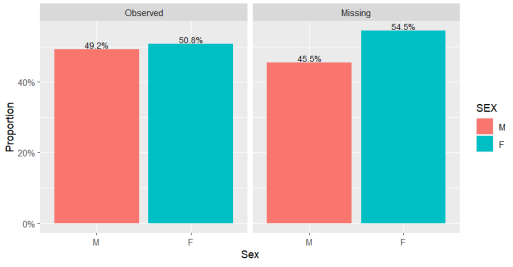


Figure 7: Top 10 CODs in complete and missing subsets

In the complete subset, females had slightly higher proportions compared to males, although the values were very close, as displayed in Figure 8a. A similar observation was seen in the missing subset; however, the difference was a bit more pronounced. For age groups, the two older cohorts had almost the same proportions in the complete subset, but in the missing subset, a consistently increasing trend was observed across the age groups. Combination of age groups and sex reflected the differences already observed in the individual plots (**Appendix A.1**). For regions only, both subsets had similar observations, where Flanders had the highest proportion, followed by Wallonia, then Brussels. Similar to the sex-age group strata, combining regions, age groups and sex indicated similar compositions in the two subsets, with higher proportions in the older populations as shown in Figure 9.

(a) Differences by sex only



(b) Differences by age group only

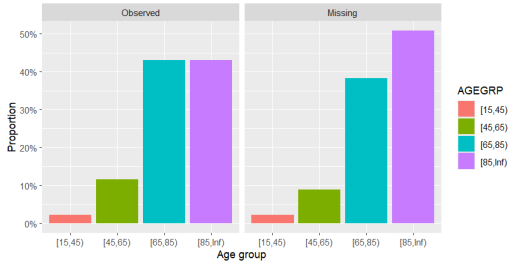


Figure 8: Composition within sex and age groups

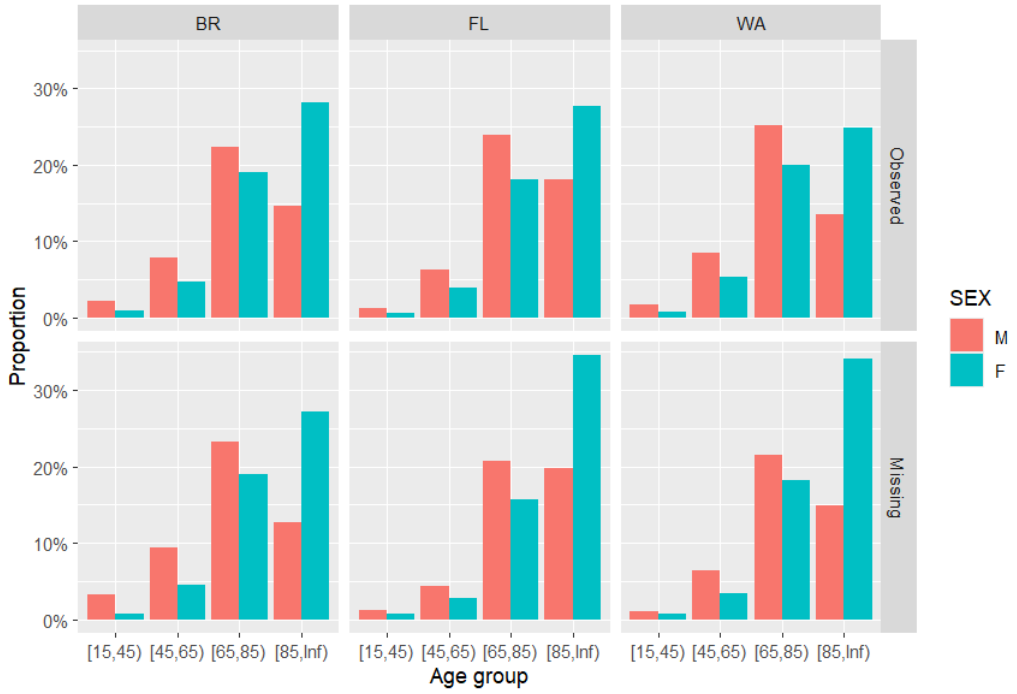


Figure 9: Differences by age groups, sex and regions

4.3 Ordinal Logistic Regression Model

For each combination of the variables, the ordinal logistic regression model was fitted, and the PO assumption evaluated. All models with age group as a covariate violated the PO assumption, while for the other models, the assumption was not violated. In the cases with violations, the parameter estimates of the binary variables (specifically, bin2, which represents low or medium EL vs high) varied from the estimate obtained from the ordinal model for some age groups. An illustration of this finding is displayed in Table 3, which has a summary of estimates for the fixed effects only for the model with age group and GBD3 as the covariates. The estimates from the binary variables vary notably to the extent of changing signs, e.g., for the 45-65 year-olds.

Table 3: Parameter estimates to illustrate a model that violates the PO assumption (with calculated ORs in brackets)

	Ordinal		**Low vs Medium or High**		**Low or Medium vs High**	
Covariate: Age group	**Estimate**	**95% CI**	**Estimate**	**95% CI**	**Estimate**	**95% CI**
[45,65)	-0.1142 (0.8921)	[-0.1977; -0.0306]	-0.2385 (0.7878)	[-0.3338; -0.1432]	0.0760 (1.0790)	[-0.0423; 0.1943]
[65,85)	-0.5916 (0.5534)	[-0.6733; -0.5100]	-0.8560 (0.4249)	[-0.9487; -0.7633]	-0.0594 (0.9423)	[-0.1746; 0.0558]
[85,Inf)	-1.0720 (0.3423)	[-1.1553; -0.9887]	-1.3405 (0.2617)	[-1.4346; -1.2464]	-0.4604 (0.6310)	[-0.5780; -0.3428]

The estimates in the table represent the log odds for the covariates. Given the negative sign in the regression equation 2, the odds of $Y \leq j$, comparing individuals at any level k (the levels of the categorical variable) with those at the reference point is given by $e^{-\beta_k}$. This implies that the OR presented in the table, i.e., e^{β_k} is therefore the odds of $Y > j$ in this comparison. Based on the ordinal OR, individuals aged [45–65), for instance, have about 11% lower odds of having a higher educational attainment compared to those in the reference group ([15–45) year-olds), and the effect is assumed to be the same across all thresholds for EL. The binary ORs, on the other hand, show that for this age group, the odds are 23% lower for the low vs medium or high threshold, but 8% higher for the low or medium vs high threshold. This means that individuals in this age group had a higher chance of having the "high" EL compared to low or medium, compared to individuals in the 15-45 years cohort; conditional on the cause of death. Consequently, partial proportional odds (PPO) models were fitted for those that violated the assumption, allowing age groups to have non-proportional odds (i.e., different effects for each threshold), while the remaining covariates had proportional odds.

Table 4 displays the respective AIC values for the models fitted. The AIC values for models with age group as a covariate are from the PPO models. Among the models without GBD3, all models with age group as a covariate indicated notable reduction in AIC values, implying that age group improves the fit. GBD3 by itself, i.e., a random effects only model, had worse performance compared to most of the models in which it was not included, especially those with age group as one of the covariates. However, when GBD3 was added as a random effect to the models with the other covariates (hence a mixed effect model), a substantial reduction in AIC was observed for all of them. Additionally, the reduction in AIC when age group is present was still consistent in these models. This improvement in fit shows that with GBD3 as one of the explanatory variables, the underlying structure in the data is better captured compared to when it is excluded.

Table 4: Table with AIC values from the models fitted

	Without GBD3	**With GBD3**
Variable(s) in the model	**AIC**	**AIC**
GBD3 only	-	183101.28
Sex	183596.50	182099.34
Agegroup	181330.00	179912.32
Region	184337.33	182612.91
Sex + Agegroup	179905.01	179326.68
Sex + Region	183075.08	181582.84
Agegroup + Region	180010.24	179452.63
Sex + Agegroup + Region	179427.20	178848.05

The parameter estimates for the model with all the covariates (sex, age group, region and GBD3), which had a better performance among all the models considered, are summarised

in Table 5. The equation of this partial proportional odds (PPO) model is:

$$\log\left[\frac{P(Y \leq j)}{P(Y > j)}\right] = \alpha_j - \left(\sum_{k=1}^{3} \beta_{k,j} \cdot \text{Agegrp}_k + \beta_4 \cdot \text{Sex} + \beta_5 \cdot \text{Flanders} + \beta_6 \cdot \text{Wallonia} + b\right) \tag{5}$$

In this equation, α_j represents the baseline log-odds for each threshold, i.e., threshold-specific intercept. $\beta_{k,j}$ represents varying coefficients for age groups, which depend on the threshold j due to the relaxed PO assumption. β_4, β_5 and β_6 are coefficients for sex, Flanders and Wallonia regions, respectively, and these are constant across all the thresholds. The random intercepts for GBD3 are represented by b, assumed to be normally distributed with a mean of 0 and variance σ^2.

Table 5: Parameter estimates for the model with all covariates

Covariate	**Parameter**	**Estimate**	**Odds Ratio**	**SE**	**p-value**
AGEGRP : [45,65)					
Low|Medium	$\beta_{1,1}$	-0.2095	0.8110	0.0489	< 0.001
Medium|High	$\beta_{1,2}$	0.1058	1.1116	0.0605	0.6189
AGEGRP : [65,85)					
Low|Medium	$\beta_{2,1}$	-0.8168	0.4418	0.0474	< 0.001
Medium|High	$\beta_{2,2}$	-0.0034	0.9966	0.0582	0.0169
AGEGRP : [85,Inf)					
Low|Medium	$\beta_{3,1}$	-1.2537	0.2854	0.0483	< 0.001
Medium|High	$\beta_{3,2}$	-0.3438	0.7091	0.0592	< 0.001
SEX : Female	β_4	-0.3403	0.7116	0.0138	< 0.001
REGION : FL	β_5	-0.5331	0.5868	0.0243	< 0.001
REGION : WA	β_6	-0.5293	0.5890	0.0254	< 0.001
Threshold	**Parameter**	**Estimate**	**Inverse logit**	**SE**	
Low|Medium	α_1	-0.9587	0.2771	0.0555	
Medium|High	α_2	1.0231	0.7356	0.0645	
Random Effects	**Parameter**	**Variance**			
GBD3	b	0.0420			

The thresholds are labeled Low|Medium and Medium|High, illustrating that the log odds correspond to P(EL $\leq$ Low)/P(EL $>$ Low) and can be alternatively presented as P(EL $\leq$ Low)/P(EL $\geq$ Medium) for the first one, with a similar derivation for the second threshold. The results from the model imply that if the linear combination of the covariates and the random effect is less than -0.9587, the individual would be more likely to have low educational attainment; between -0.9587 and 1.0231, they would be likely to have medium, and greater than 1.0231, they would be likely to have high educational attainment. From the calculated inverse logits [1], if after computing all the values on the RHS the value of the

[1] Formula: $\frac{e^{\alpha_j}}{1+e^{\alpha_j}}$

inverse logit (the probability) would be less than 0.2771, the individual would be likely to have "low" EL; "medium" if the probability is between 0.2771 and 0.7356, and "high" if the probability is greater than 0.7356.

The OR for age group [45, 65) for the comparison of low vs medium or high was 0.8110, meaning this cohort has approximately 19% lower odds of being in medium or high level, compared to the [15,45) cohort when all the other variables are controlled for, and conditional on the COD. This effect was statistically significant (p-value < 0.001). For the comparison of being in high vs low or medium, this age group had approximately 11% higher odds compared to the reference group. However, this effect was not statistically significant (p-value = 0.6189). In the other age groups, the odds, compared to the reference group, were significantly lower for both thresholds.

The OR for females was 0.7116 (p-value < 0.001), implying that while controlling for the other covariates, the odds of females having higher categories of educational attainment was significantly lower by about 29% compared to the males. The odds for the two regions were similar (41% lower) compared to the reference group, which is Brussels, conditional on the COD and controlling for the other covariates. These results confirm the observations from the exploratory analyses for variables like age group and sex, and uncover some insights that were otherwise difficult to observe from the visualisations.

The intra-cluster correlation for the random effects was approximately 1.3%, which represents the variability in educational attainment explained by GBD3. This value is calculated as;

$$ICC = \frac{\sigma^2}{\sigma^2 + \pi^2/3} = \frac{0.042}{0.042 + \pi^2/3}$$

4.4 Imputation

Probabilistic Imputation

Prior to the imputation process, the fully observed subset and the subset with missing data were examined for any differences, especially in the composition of the covariates necessary for the procedure. This was done to identify whether there were CODs with missing values that had no presence whatsoever in the observed subset, and if there were some COD, age group, region and sex combinations in the missing subset but not previously observed. From this check, all CODs in the missing subset had at least some records in the fully observed subset. Of the 11,591 records with missingness (excluding the 522 records for under 15-year-olds), approximately 99.6% had combinations of all four covariates that were already present in the observed data, with only 48 (approximately 0.4%) having combinations of COD with only two or one of the other covariates. These 48 observations were imputed using the fallback strategy, i.e., probabilities calculated from age groups only from the complete subset.

Following the imputation procedure, 0.45% of the data, representing age groups [0,5) and [5,15), were imputed using the "NAP" code. For the remaining age groups, the resulting distribution of education levels in the fully imputed dataset indicated very minimal deviations from the proportions previously observed from the complete subset. This small difference could be a result of the proportion of missingness being quite small in the dataset of 116,378 records, hence the possibility of the results being masked by proportions in the already observed records. For this reason, a subset of only the records whose ELs were imputed was also examined, and the proportions displayed in Table 6, based on the higher categorisation of educational attainment (low, medium and high). These summaries are based on only one of the datasets out of the 100, therefore, these proportions would differ because of the difference in the composition of CODs in each dataset achieved in the process of IDD redistribution [12].

Table 6: Summary of proportions (in %) in the complete subset, subset of imputed and the fully imputed dataset from the probabilistic approach

EL	**Observed**	**Imputed subset**	**Fully imputed**
Low	64.78	63.78	64.68
Medium	21.00	18.75	20.76
High	14.22	13.16	14.11
NAP	-	4.31	0.45

The proportions in the fully imputed dataset were very similar to the observed proportions. This similarity was also reflected in the relationships between educational attainment and other variables. For the subset with only imputed records, slight differences were observed for these proportions, indicating some difference in proportionality as a result of the imputation method used. However, based on the similarities initially observed from the comparison of the demographic characteristics between the complete and missing subsets, the imputation procedure was less likely to result in very pronounced differences. Figure 10 juxtaposes the distribution of ELs across age groups for the fully observed subset against the imputed subset.

Generally, the imputation process preserved the overall pattern of educational attainment across age groups, with older cohorts being more likely to have lower ELs. Additionally, ISCED level 5 continued to exhibit the lowest frequency as in the complete subset. However, a closer examination of the patterns revealed some subtle differences, especially at lower ELs. In ISCED level 2, for instance, the age groups with the highest proportions switched, such that in the observed, the [65,85) cohort had the highest, followed by [85, Inf), while in the imputed subset, this order was reversed. Furthermore, the absolute differences in proportions for lower ELs were more pronounced in the imputed dataset. These differences, though subtle, indicate that the imputation procedure did not provide a perfect replicate of the relationships in the complete subset, especially with the key demographic variables.

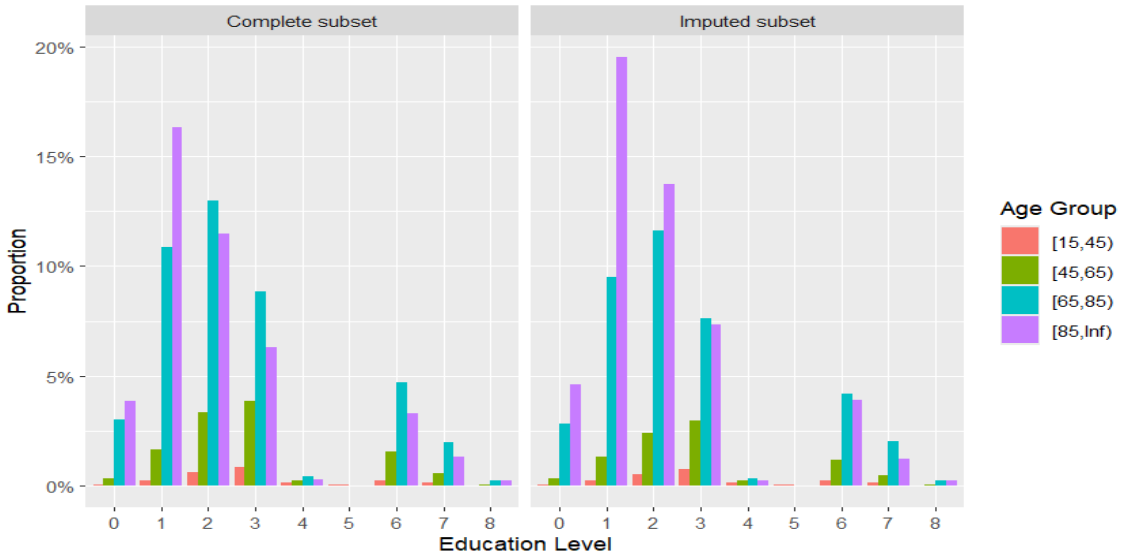


Figure 10: Side by side comparison of the distribution of ELs in the complete subset and the imputed subset

Model-based Imputation

The model whose estimates have been presented and interpreted in Section 4.3 was fitted to the data. Since all CODs in the subset with missing ELs were also observed in the complete cases, the model fitted on the observed data provided all the required random effects estimates for this procedure. For each missing record, the systematic component of Equation 5 was obtained by using the specific estimates depending on the value of their characteristics (COD, sex, age group and region). This generated predicted probabilities for each EL (low, medium and high) for every record with missing information.

Table 7 summarises the proportions obtained after imputation. Similar to the observations made from the probabilistic approach, the fully imputed dataset had similar proportions to the observed ones, but the subset of only imputed records had small differences.

Table 7: Summary of proportions (in %) in the complete subset, subset of imputed and the fully imputed dataset from the model-based approach

EL	Observed	Imputed subset	Fully imputed
Low	64.78	63.11	64.61
Medium	21.00	19.66	20.86
High	14.22	12.92	14.08
NAP	-	4.31	0.45

5 Discussion

The challenge of missing information on educational attainment is a fundamental methodological issue in studies focusing on health inequalities using this socioeconomic indicator. In Belgium, as in other countries, misclassified or, in some cases, missing educational levels in linked death datasets can lead to biased estimates of educational inequalities in mortality, making cross-country comparisons (e.g., for OECD countries) difficult [25]. International studies have also reflected this challenge as one that may lead to bias in mortality-related estimates. For instance, in Switzerland, a lack of high-quality linked data, caused by missingness and misclassification, resulted in the underestimation of socioeconomic inequalities in death rates [26].

In light of this, the study sought to address the challenge of missingness in the linked COD data using a two-step probabilistic procedure. In the first step, individuals under 15 years of age were assigned a 'Not Applicable' (NAP) code. For the remaining age groups, a probabilistic imputation approach based on Bayes' rule was used, where the choice of variables considered was driven by the patterns observed in the data, as well as known relevance based on prior studies. This analysis was therefore grounded in a thorough analysis of the structure of the dataset used, as a foundation for the process. By leveraging the probabilistic, as well as a model-based approach, the study aimed at generating plausible imputation values, thus minimizing bias in subsequent studies that rely on EL information.

Summary of findings

In the linked COD data analysed in this study, approximately 10.41% of the records had missing EL information; this included all individuals under the age of 15 years for whom EL data were completely missing. For this reason, records for those below 15 years old were excluded from the exploratory analyses. Overall, the representation of males and females was fairly equal in the dataset. More records were observed among the older population, which is in line with the Belgian life expectancy [7]. The relationship between educational attainment and the covariates revealed slight differences across various strata. In the lower ELs, the proportions increased across age for both sexes, indicating that the older population were less likely to have high educational attainment. On the other hand, in higher ELs (ISCED 6-8) the proportions among females consistently declined across age groups, compared to the proportions of males for the same levels, where the values did not fluctuate much across age. This is a reflection of the historical educational inequality among males and females, as reported by Ronsijn (2014) [27], that in the earlier years, the impact of educational expansion was more pronounced among males than females. COD, by itself, did not show notable differences across groups due to the high cardinality of the variable.

The covariates indicated subtle patterns of missingness, such that the oldest cohort (85+) and the cohort of [45,65) were more likely to have missing ELs, with females having a slightly higher probability compared to males. For COD, patterns could not be easily deduced from the visual inspections due to the high number of categories. Generally, individuals from

Brussels were more likely to have missing data compared to the other regions. In addition to that, contrary to Flanders and Wallonia where males and females had comparable probabilities across age, in Brussels, males were more likely to have missingness across age. These differences, though subtle, highlighted the importance of considering these variables when imputing EL. A comparison of the demographic characteristics between the fully observed and missing subset revealed that the two populations were similar in this regard.

From the ordinal logistic regression models, improvement in fit was observed when COD was included in the model. This underscores the additional value of GBD3 in capturing the underlying structure in the data. Specifically, the model with all the covariates provided a better fit to the data compared to the other models. Inferences from this model mostly reflected the insights from the visualisations, e.g., the age group [45,65) having higher odds of attaining the highest ELs compared to the [15, 45) cohort. The model, however, uncovered an important association between COD and educational attainment, a relationship that was otherwise difficult to observe from the visualisations.

The distribution of educational levels in the fully imputed dataset was considerably similar to that observed in the complete subset; both overall and in relationship to the key variables. This indicates that the underlying structure of the data was preserved during the imputation process; an observation that was not surprising because both the fully observed and missing subsets had similar demographic characteristics. Moreover, in line with the demographic reality that deaths are more common among the older population, who, historically, had lower educational attainment, it is noteworthy that a larger share of imputed values fell into the lower ELs. Although the imputation process did not lead to values that are drastically different from the complete cases, it is still advantageous to impute missing values rather than completely ignore them for robust estimates.

Comparison with other studies

Similar to the insights from this study, both from the complete cases and fully imputed data, various studies have consistently shown an association between higher mortality risks and lower educational attainment. This relationship can be attributed to the older population, who, overall, are less likely to have attained the higher ELs, as well as the fact that certain causes of death, especially those related to health-seeking behaviours, are more likely among less educated individuals [7, 28].

In a study of mortality in the US, Lourés et. al [24] also implemented a similar probabilistic approach to impute missing educational attainment. While they did not directly compare the distributions in the complete cases to those in the imputed datasets, they leveraged the fact that this approach makes optimal use of the available information in the dataset through the components used to calculate the probabilities (Equation 4) [23]. As a result, the complete datasets enabled them to conduct comprehensive analyses of mortality using educational attainment as a stratifier.

Strengths and Limitations

A key strength of the probabilistic process used in this study is the generation of complete datasets, and more importantly, imputation of ELs at a granular level, in line with ISCED levels (0-8), which can be useful in other analyses. Additionally, the process followed in creating the linked COD dataset used ensures the data covers the entire Belgian population, hence the data can be reliably used in mortality analyses [8]. This process is also flexible and can be adapted for similar use cases, with variables tailored to the patterns observed in those datasets.

Nonetheless, the study has several limitations. The variable selection process is data-driven, since imputation was done using variables identified in the data as potentially associated with educational attainment. For this reason, a different dataset with varying population characteristics may require other variables that better explain the structure in the data. Additionally, all the individuals under 15 years of age had missing EL, thus limiting the ability to assess educational attainment for this cohort, given there is a likelihood that some may have completed certain educational levels prior to their death, especially the lower educational levels. Furthermore, this process was computationally intensive, such that the probabilistic procedure required a long processing time even for one dataset.

It is worth noting that since this procedure uses COD information to impute ELs, and the imputed EL would be later used to analyse health inequalities, careful assessment is required, especially in cases where the proportion of missingness is high, such that the associations observed are not artificially introduced in the imputation process.

Future Work

Future studies could leverage the imputed datasets to conduct comprehensive analyses of the linked COD dataset, such as exploring health disparities using educational attainment to yield insights that can be used to inform public health policies. Moreover, if available, additional socioeconomic variables such as information on employment would further refine probable education levels, taking into account that certain types of jobs require individuals to have attained specific levels of education.

6 Ethical Thinking, Societal Relevance, and Stakeholder Awareness

6.1 Ethical Standards Relevant to the Study

The data used in this study are for deceased persons, whose personal information is exempt from data protection laws according to the GDPR. Nonetheless, there were no variables in the dataset that could be used to re-identify the individuals. Moreover, to ensure privacy and integrity of the data provided by the company, all data processing and analyses were done on Sciensano's virtual machine.

6.2 Societal Relevance

This study contributes to the accurate analysis of health disparities within the Belgian population by ensuring the completeness of the linked COD data. Results generated from the downstream analyses that use this data would provide helpful insights that can be used by public health officials to make informed, evidence-based decisions on public health policies, as well as implement targeted interventions. This would consequently enhance health equity and lead to a healthier nation by meeting the needs of the population-at-risk.

6.3 Stakeholder awareness

This study is directly relevant to Sciensano Service Health Information, the government and other decision-making bodies, as well as the general population. Sciensano Service Health Information is particularly interested in complete and high-quality datasets that can be used in generating vital statistics. This would ultimately lead to the calculation of accurate estimates of the burden of disease and other key health indicators that are relevant to the general population.

7 Conclusion

This study developed and implemented a probabilistic procedure to impute missing EL information. Building upon the IDD redistribution framework [12], the study leveraged the COD information, in combination with key demographic variables, to "fill the gaps" in EL information. The procedure resulted in a fully imputed dataset that conserves the overall distribution of data in the population, with the majority of the deceased persons having lower educational levels, which aligns with previous studies. The results from the procedure developed would be useful in future applications, including complementing the data used in tracking the Belgian National Burden of Disease (BeBOD).

References

[1] L. T. Ruzicka and A. D. Lopez, "The use of cause-of-death statistics for health situation assessment: national and international experiences," *World Health Stat Q*, vol. 43, no. 4, pp. 249–58, 1990.

[2] The Lancet, "Icd-11," *The Lancet*, vol. 393, no. 10188, p. 2275, 2019.

[3] I. M. Moriyama, R. M. Loy, A. H. T. Robb-Smith, H. M. Rosenberg, and D. L. Hoyert, "History of the statistical classification of diseases and causes of death," 2011.

[4] Statbel, "Causes of death," 2024. [Online]. Available: https://statbel.fgov.be/en/themes/population/mortality-life-expectancy-and-causes-death/causes-death.

[5] J. P. Mackenbach, J. R. Valverde, B. Artnik, M. Bopp, H. Brønnum-Hansen, P. Deboosere, R. Kalediene, K. Kovács, M. Leinsalu, P. Martikainen, *et al.*, "Trends in health inequalities in 27 European countries," *Proceedings of the National Academy of Sciences*, vol. 115, no. 25, pp. 6440–6445, 2018.

[6] S. Stringhini, C. Carmeli, M. Jokela, M. Avendaño, P. Muennig, F. Guida, F. Ricceri, A. d'Errico, H. Barros, M. Bochud, *et al.*, "Socioeconomic status and the 25× 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1· 7 million men and women," *The Lancet*, vol. 389, no. 10075, pp. 1229–1237, 2017.

[7] D. P. R. J. J. D. B. Renard F., Scohy A., "Health status report 2021 – The state of health in Belgium," 2021. Deposit number: D/2022/14.440/06 Available on https://www.healthybelgium.be/en/health-status.

[8] F. Renard, B. Devleesschauwer, H. Van Oyen, S. Gadeyne, and P. Deboosere, "Evolution of educational inequalities in life and health expectancies at 25 years in Belgium between 2001 and 2011: a census-based study," *Archives of Public Health*, vol. 77, pp. 1–10, 2019.

[9] F. Renard, S. Gadeyne, B. Devleesschauwer, J. Tafforeau, and P. Deboosere, "Trends in educational inequalities in premature mortality in Belgium between the 1990s and the 2000s: the contribution of specific causes of deaths," *J Epidemiol Community Health*, vol. 71, no. 4, pp. 371–380, 2017.

[10] F. Renard, B. Devleesschauwer, S. Gadeyne, J. Tafforeau, and P. Deboosere, "Educational inequalities in premature mortality by region in the Belgian population in the 2000s," *Archives of Public Health*, vol. 75, pp. 1–16, 2017.

[11] Statbel, "Datalab: Census education," 2024. [Online]. Available: https://statbel.fgov.be/en/themes/datalab/datalab-census-education.

[12] B. Devleesschauwer, A. Scohy, R. De Pauw, V. Gorasso, A. Kongs, E. Neirynck, P. Verduyckt, G. M. Wyper, and L. Van den Borre, "Investigating years of life lost in Belgium,

2004–2019: A comprehensive analysis using a probabilistic redistribution approach," *Archives of Public Health*, vol. 81, no. 1, p. 160, 2023.

[13] R. De Pauw, V. Gorasso, A. Scohy, L. Van den Borre, and B. Devleesschauwer, "Belgian National Burden of Disease Study. Guidelines for the Calculation of Disability-Adjusted Life Years in Belgium," September 2023. Deposit number: D/2023.14.440/67.

[14] Sciensano, "Belgian National Burden of Disease Study (BeBOD)," 2025. [Online]. Available: https://burden.sciensano.be/shiny/ineq2022/.

[15] Unesco, *International standard classification of education (ISCED) 2011*. UNESCO, 2012.

[16] Cedefop, *Terminology of European education and training policy: a selection of 430 terms: third edition*. Cedefop reference series, Luxembourg: Publications Office, 2024.

[17] R. J. Little, "A test of missing completely at random for multivariate data with missing values," *Journal of the American statistical Association*, vol. 83, no. 404, pp. 1198–1202, 1988.

[18] C. K. Enders, "Applied missing data analysis.," 2010.

[19] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 42, no. 2, pp. 109–127, 1980.

[20] A. Agresti, *Analysis of ordinal categorical data*. John Wiley & Sons, 2010.

[21] R Core Team, "R: A language and environment for statistical computing," 2024.

[22] "Commission Implementing Regulation (EU) 2017/543." https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32017R0543.

[23] I. Sasson, "Trends in life expectancy and lifespan variation by educational attainment: United states, 1990–2010," *Demography*, vol. 53, no. 2, pp. 269–293, 2016.

[24] C. R. Lourés and A. J. Cairns, "Mortality in the US by education level," *Annals of Actuarial Science*, vol. 14, no. 2, pp. 384–419, 2020.

[25] J. Mackenbach, G. Menvielle, D. Jasilionis, and R. de Gelder, "Measuring educational inequalities in mortality statistics," 2015.

[26] M. Lerch, A. Spoerri, D. Jasilionis, and F. Viciana Fernandez, "On the plausibility of socioeconomic mortality estimates derived from linked data: a demographic approach," *Population health metrics*, vol. 15, pp. 1–15, 2017.

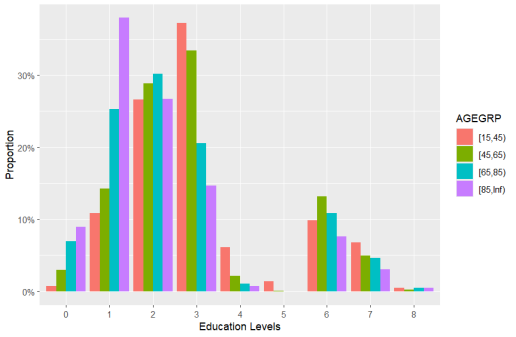
[27] W. Ronsijn, " Educational Expansion and Gender Inequality in Belgium in the Twentieth Century," *Histoire & mesure*, no. 1, pp. 195–218, 2014.

[28] R. G. Rogers, B. G. Everett, A. Zajacova, and R. A. Hummer, "Educational degrees and adult mortality risk in the United States," *Biodemography and social biology*, vol. 56, no. 1, pp. 80–99, 2010.

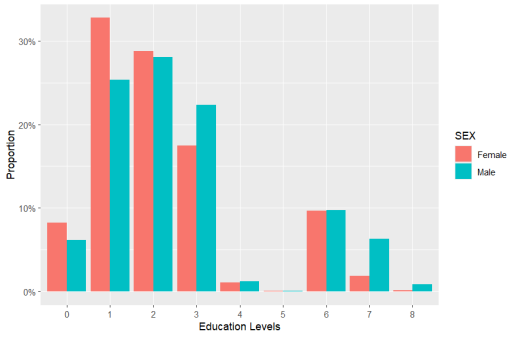
Appendix A: Supplementary Results

A.1 Results from the Exploratory Analyses

Univariate Plots for Distribution of ELs

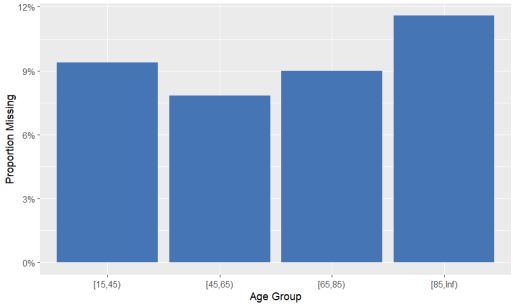


(a) Distribution of ELs by age

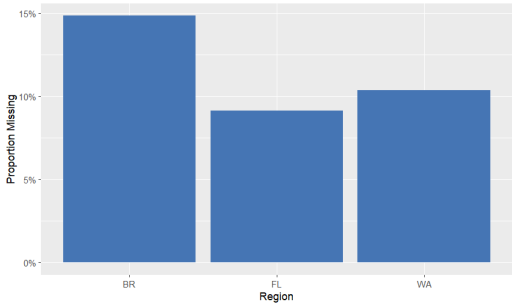


(b) Distribution of ELs by sex

Additional Plots for Patterns of Missingness

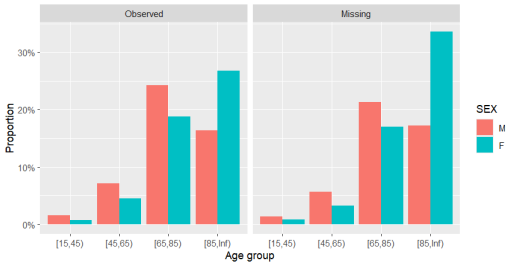


(a) Patterns of missingness by age

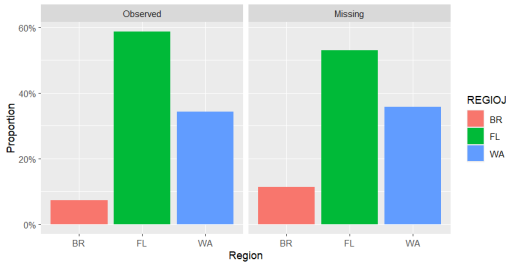


(b) Patterns of missingness by region

Additional plots for differences in characteristics between complete and missing subsets



(a) Differences by sex and age groups



(b) Differences by region

A.2 Estimates for the PO Assumption Check for the Selected Model

	Ordinal		**Bin1**		**Bin2**	
Covariate	**Estimate** (SE)	**p-value**	**Estimate** (SE)	**p-value**	**Estimate** (SE)	**p-value**
Sex: F	-0.3243 (0.0132)	< 0.001	-0.3238 (0.0135)	< 0.001	-0.3579 (0.0184)	< 0.001
Age group: [45,65)	-0.1542 (0.04032)	0.0001	-0.2921 (0.0466)	< 0.001	0.0345 (0.0586)	0.5559
Age group: [65,85)	-0.6602 (0.03796)	< 0.001	-0.9446 (0.0439)	< 0.001	-0.1058 (0.0553)	0.0554
Age group: [85,Inf)	-1.1160 (0.03845)	< 0.001	-1.4094 (0.0443)	< 0.001	-0.4609 (0.0560)	< 0.001
Region: Flanders	-0.5344 (0.02421)	< 0.001	-0.4739 (0.0253)	< 0.001	-0.6819 (0.0303)	< 0.001
Region: Wallonia	-0.5377 (0.02525)	< 0.001	-0.5026 (0.0264)	< 0.001	-0.5686 (0.0316)	< 0.001

Table 8: Parameter Estimates for the ordinal and two binary models. Standard errors are provided in brackets for each estimate

Appendix B: R Codes

The R Codes presented here are for the selected regression model, its PO assumption check, and the imputation methodologies implemented. The building blocks, including data processing, exploratory analyses and all the other models fitted, can be found on Github https://github.com/Carol0128/Filling-the-Void.

```
#libraries
library(ggplot2)
library(dplyr)
library(stringr)
library(tidyr)
library(ordinal)
library(lme4)

##::::::::::::Ordinal Logistic Regression Model ::::::::::::::

# ordinal model
allvars_olr_model <- clmm(CD_ISCED_CENSUS_2 ~ SEX + AGEGRP + REGIOJ +
                          (1|GBD3_RED4), data = df_observed)
summary(allvars_olr_model)

##:::Check for PO assumption
# creating bin1 and bin2 variables
df_observed <- df_observed %>% mutate(
  bin1 = ifelse(CD_ISCED_CENSUS_2 == "Low", 0, 1),
  bin2 = ifelse(CD_ISCED_CENSUS_2 == "High", 1, 0))

# binary model for bin1
allvars_glm_model1 <- glmer(bin1 ~ SEX + AGEGRP + REGIOJ + (1|GBD3_RED4),
                            family = binomial, data = df_observed)
summary(allvars_glm_model1)

# binary model for bin 2
allvars_glm_model2 <- glmer(bin2 ~ SEX + AGEGRP + REGIOJ + (1|GBD3_RED4),
                            family = binomial, data = df_observed)
summary(allvars_glm_model2)

##:::Partial Proportional Odds Model
all_vars_ppo_model <- clmm2(CD_ISCED_CENSUS_2 ~ REGIOJ + SEX, nominal = ~ AGEGRP,
                            random = GBD3_RED4, data = df_observed, Hess = TRUE)
summary(all_vars_ppo_model)
```

```r
##::::::::::::IMPUTATION ::::::::::::::

##::: Probabilistic Imputation
prob_impute <- function(data){
  #STEP1: NAP for under 15 years
  data <- data %>% mutate(CD_ISCED_CENSUS_IMP1 = ifelse(is.na(CD_ISCED_CENSUS) &
                          AGEGRP %in% c("[0,5)", "[5,15)"), "NAP",
                          as.character(CD_ISCED_CENSUS)))

  # column to hold imputed values in Step 2
  data$CD_ISCED_CENSUS_IMP2 <- as.character(data$CD_ISCED_CENSUS_IMP1)

  #records for which Ed level is not missing after step 1
  df_observed_imp1 <- data %>% filter(!is.na(CD_ISCED_CENSUS_IMP1))
  #records for which Ed level is missing
  df_missing_imp1 <- data %>% filter(is.na(CD_ISCED_CENSUS_IMP1))

  # Calculating the probabilities
  ## 1. p(C | e, X)
  p_cod_given_EX <- df_observed_imp1 %>%
    group_by(GBD3_RED4, CD_ISCED_CENSUS_IMP1, AGEGRP, REGIOJ, SEX) %>%
    summarise(n_cex = n(), .groups = "drop") %>%
    group_by(CD_ISCED_CENSUS_IMP1, AGEGRP, REGIOJ, SEX) %>%
    mutate(m_cex = sum(n_cex), p_cod_eX = n_cex / sum(n_cex))

  ## 2. p(e | X)
  p_E_given_X <- df_observed_imp1 %>%
    group_by(CD_ISCED_CENSUS_IMP1, AGEGRP, REGIOJ, SEX) %>%
    summarise(n_ex = n(), .groups = "drop") %>%
    group_by(AGEGRP, REGIOJ, SEX) %>%
    mutate(m_ex = sum(n_ex), p_e_X = n_ex / sum(n_ex))

  ## 3. p(C | X)
  p_cod_given_X <- df_observed_imp1 %>%
    group_by(GBD3_RED4, AGEGRP, REGIOJ, SEX) %>%
    summarise(n_cx = n(), .groups = "drop") %>%
    group_by(AGEGRP, REGIOJ, SEX) %>%
    mutate(m_cx = sum(n_cx), p_cod_X = n_cx / sum(n_cx))

  ## I Calculating the posterior p(E|C, X) from the 3 components
  posterior_df <- p_cod_given_EX %>%
    left_join(p_E_given_X, by = c("REGIOJ", "AGEGRP", "SEX",
                                  "CD_ISCED_CENSUS_IMP1")) %>%
```

```
  left_join(p_cod_given_X, by = c("REGIOJ", "AGEGRP", "SEX", "GBD3_RED4")) %>%
  mutate(p_E_given_all = (p_cod_eX * p_e_X) / p_cod_X)

# creating a list of lists with each EL and it prob for every COD,X combination
posterior_summary <- posterior_df %>% group_by(GBD3_RED4, AGEGRP, SEX,
                                                REGIOJ) %>%
  summarise(group_size = mean(n_cx), ed_prob_list =
              list(purrr::map2(as.character(CD_ISCED_CENSUS_IMP1),
              p_E_given_all, ~ list(ED = .x, p = .y))), .groups = "drop")

## II. Fallback probs when there are missing sets or sparse groups
p_E_given_age <- df_observed_imp1 %>%
  group_by(CD_ISCED_CENSUS_IMP1, AGEGRP) %>% summarise(n = n(), .groups = "drop") %>%
  group_by(AGEGRP) %>% mutate(m = sum(n), p_E_age = n/sum(n)) %>%
  summarise(ed_prob_list = list(purrr::map2(as.character(CD_ISCED_CENSUS_IMP1),
                          p_E_age, ~ list(ED = .x, p = .y))), .groups = "drop")

#Imputation begins here
for (i in 1:nrow(df_missing_imp1)){
  row <- df_missing_imp1[i,]

  #find matching group from the posterior summary table
  match_prob <- posterior_summary %>% filter(GBD3_RED4 == row$GBD3_RED4,
                  REGIOJ == row$REGIOJ, AGEGRP == row$AGEGRP, SEX == row$SEX)
  stratum_size <- match_prob$group_size

  cat("Imputing row", i, "\n")
  # If a match is found
  if (nrow(match_prob) > 0 && stratum_size > 9) {
    ed_prob_list <- match_prob$ed_prob_list[[1]]

    # Extracting Ed level and probs to simplify sampling step
    ed_levels <- sapply(ed_prob_list, function(x) x$ED)
    probs <- sapply(ed_prob_list, function(x) x$p)

    # Randomly sample Education level
    selected_ED <- sample(ed_levels, 1, prob = as.numeric(probs))
  }

  # If there is no match:
  else{
    match_fallback <- p_E_given_age %>% filter(AGEGRP == row$AGEGRP)
```

```
      ed_prob_list2 <- match_fallback$ed_prob_list[[1]]
      ed_levels2 <- sapply(ed_prob_list2, function(x) x$ED)
      probs2 <- sapply(ed_prob_list2, function(x) x$p)

      selected_ED <- sample(ed_levels2, 1, prob = probs2)
    }

    # impute the selected one in the initial dataset
    df_missing_imp1[i, "CD_ISCED_CENSUS_IMP2"] <- selected_ED

  }

  df_full <- rbind(df_observed_imp1, df_missing_imp1)

  #categorization for low, medium, high (and the new category - NAP)
  df_full <- df_full %>% mutate(CD_ISCED_CENSUS2_IMP2 =
              ifelse(CD_ISCED_CENSUS_IMP2 == "NAP", "NAP",
              ifelse(CD_ISCED_CENSUS_IMP2 %in% c(0,1,2), "Low",
              ifelse(CD_ISCED_CENSUS_IMP2 %in% c(3,4), "Medium", "High"))))%>%
    mutate(CD_ISCED_CENSUS2_IMP2 = factor(CD_ISCED_CENSUS2_IMP2,
                        levels = c("Low", "Medium", "High", "NAP"), ordered = T))

  return(df_full)
}

# implementing the function on a (preprocessed) dataset
prob_imputed_df <- prob_impute(df)

##::: Model-Based Imputation
# Here, the predicted probabilities are calculated using the inverse logit formula
# since the functions do not directly provide these values

model_impute <- function(data){
  #STEP1: NAP for under 15 years
  data <- data %>%
    mutate(CD_ISCED_CENSUS2_IMP1 = ifelse(is.na(CD_ISCED_CENSUS_2) &
                                            AGEGRP %in% c("[0,5)", "[5,15)"),
                                          "NAP", as.character(CD_ISCED_CENSUS_2)))

  # column for imputed values in Step 2
  data$CD_ISCED_CENSUS2_IMP2 <- as.character(data$CD_ISCED_CENSUS2_IMP1)

  #split observed and missing subsets
```

```r
df_obs2 <- data %>% filter(!is.na(CD_ISCED_CENSUS2_IMP1))
#records for which Ed level is missing
df_miss2 <- data %>% filter(is.na(CD_ISCED_CENSUS2_IMP1))

# dummies for the predictor variables; to make calculation of eta easy
df_missing_copy <- df_miss2 %>% transmute(GBD3_RED4,
                      SEXF = ifelse(SEX == "F",1,0),
                      REGIOJFL = ifelse(REGIOJ == "FL", 1, 0),
                      REGIOJWA = ifelse(REGIOJ == "WA", 1, 0),
                      `AGEGRP[45,65)` = ifelse(AGEGRP == "[45,65)",1,0),
                      `AGEGRP[65,85)` = ifelse(AGEGRP == "[65,85)",1,0),
                      `AGEGRP[85,Inf)` = ifelse(AGEGRP == "[85,Inf)",1,0))

model.fit <- clmm2(CD_ISCED_CENSUS_2 ~ REGIOJ + SEX,
                   nominal = ~ AGEGRP, random = GBD3_RED4,
                   data = df_obs2, Hess = TRUE)

m2 <- clmm(CD_ISCED_CENSUS_2 ~ (1 | GBD3_RED4), data = df_obs2)

# Fixed effects for proportional variables
po_betas <- model.fit$beta

X_po_cols <- names(po_betas)
X_po_fixed <- as.matrix(df_missing_copy[, X_po_cols])

eta_po_fixed <- X_po_fixed %*% po_betas

#Matrix with alphas and beta-coefficients for the uncommon slopes
theta_matrix <- model.fit$Theta

#-- extract cumulative thresholds (alpha_j) as an array
alpha <- matrix(theta_matrix["(Intercept)", ], nrow = 1)

# effects for non proportional variables
beta_age_low <- -1*as.vector(t(theta_matrix[-1, 1]))
beta_age_medium <- -1*as.vector(t(theta_matrix[-1, 2]))

X_npo_fixed <- as.matrix(df_missing_copy[, 5:7])

eta_age_low <- X_npo_fixed %*% beta_age_low
eta_age_medium <- X_npo_fixed %*% beta_age_medium

# Random effects
```

```
  re <- model.fit$ranef
  re_df <- data.frame(GBD3_RED4 = rownames(ranef(m2)$GBD3_RED4), RE_Int = re)

  # left join re_df with the df_missing_copy
  df_missing_copy <- df_missing_copy %>%left_join(re_df, by = "GBD3_RED4")
  b_i <- df_missing_copy$RE_Int # random effects for COD in missing df

  # RHS linear predictors
  eta_low <- eta_po_fixed + eta_age_low + b_i
  eta_medium <- eta_po_fixed + eta_age_medium + b_i

  # Cumulative probabilities
  cumprobs_low <- sapply(alpha[1], function(x) plogis(x - eta_low))
  cumprobs_medium <- sapply(alpha[2], function(x) plogis(x - eta_medium))

  # 'predicted' probabilities for each category
  probs <- matrix(NA, nrow = length(eta_low), ncol = 3)
  colnames(probs) <- c("Low", "Medium", "High")
  probs[, 1] <- cumprobs_low
  probs[, 2] <- cumprobs_medium - cumprobs_low
  probs[, 3] <- 1 - cumprobs_medium

  return(list(missing = df_miss2, observed = df_obs2,
              probs = probs))
}

model_output <- model_impute(df)
model_probs <- model_output$probs
model_df_miss <- model_output$missing # df that will be imputed
model_df_obs <- model_output$observed # complete-case df, to be merged with imputed

#sampling 1 category per row using the probabilities
el_sampled <- apply(model_probs, 1, function(p) {
    sample(categories, size = 1, prob = p) })

model_df_miss$CD_ISCED_CENSUS2_IMP2 <- el_sampled #impute
model_imputed_df <- rbind(model_df_obs, model_df_miss)
```