



## Addressing data deficiencies in assistive technology by using statistical matching methodology: a case study from Malawi

Monica Jamali-Phiri, Juba Alyce Kafumba, Malcolm MacLachlan, Emma M. Smith, Ikenna D. Ebuenyi, Arne Henning Eide & Alister Munthali

To cite this article: Monica Jamali-Phiri, Juba Alyce Kafumba, Malcolm MacLachlan, Emma M. Smith, Ikenna D. Ebuenyi, Arne Henning Eide & Alister Munthali (2020): Addressing data deficiencies in assistive technology by using statistical matching methodology: a case study from Malawi, *Disability and Rehabilitation: Assistive Technology*, DOI: [10.1080/17483107.2020.1861118](https://doi.org/10.1080/17483107.2020.1861118)

To link to this article: <https://doi.org/10.1080/17483107.2020.1861118>



Published online: 28 Dec 2020.



Submit your article to this journal [↗](#)



Article views: 55



View related articles [↗](#)



View Crossmark data [↗](#)






Citing articles: 1 View citing articles [↗](#)

ORIGINAL RESEARCH



## Addressing data deficiencies in assistive technology by using statistical matching methodology: a case study from Malawi

Monica Jamali-Phiri<sup>a</sup>, Juba Alyce Kafumba<sup>a</sup>, Malcolm MacLachlan<sup>b,c</sup> , Emma M. Smith<sup>b</sup>, Ikenna D. Ebuenyi<sup>b</sup> , Arne Henning Eide<sup>d</sup> and Alister Munthali<sup>a</sup> 

<sup>a</sup>Centre for Social Research, Chancellor College, University of Malawi, Zomba, Malawi; <sup>b</sup>Assisting Living & Learning (ALL) Institute, Department of psychology, Maynooth University, Maynooth, Ireland; <sup>c</sup>Olomouc University Social Health Institute(OUSHI), Palacky University Olomouc, Czech Republic; <sup>d</sup>SINTEF, Oslo, Norway

### ABSTRACT

**Purpose:** To address the data gap on efforts to assess use of assistive technology among children with disability in sub-Saharan Africa. Contribute towards efforts examining access to assistive technologies in sub-Saharan Africa.

**Materials and methods:** The paper uses data from the 2017 survey on Living conditions among persons with disabilities in Malawi and the 2015-16 Malawi Demographic and Health survey to address the objective of the study. The two datasets were statistically matched through random hot deck technique, by integrating the two datasets using randomly selected units from a subset of all available data donors.

**Results:** Results indicate that statistical matching technique produces a composite dataset with an uncertainty value of 2.2%. An accuracy assessment test of the technique also indicates that the marginal distribution of use of assistive technology in the composite dataset is similar to that of the donor dataset with an Overlap index value of close to 1 (Overlap = 0.997).

**Conclusions:** The statistical matching procedure does enable generation of good data in data constrained contexts. In the current study, this approach enabled measurement of access to assistive products among children with disabilities, in situations where the variables of interest have not been jointly observed. Such a technique can be valuable in mining secondary data, the collection of which may have been funded from different sources and for different purposes. This is of significance for the efficient use of current and future data sets, allowing new questions to be asked and addressed by locally based researchers in poor settings.

### ARTICLE HISTORY

Received 7 July 2020  
Revised 3 December 2020  
Accepted 3 December 2020

### KEYWORDS

Assistive technology; children with disabilities; statistical matching; sub-Saharan Africa; distinct data-sources

### ► IMPLICATIONS FOR REHABILITATION

- In resource-poor settings, the technique of statistical matching can be used to examine factors that predict the use of assistive technology among persons with disabilities.
- The statistical matching technique is of significance for the efficient use of current and future datasets, allowing new questions to be asked and addressed by locally based researchers.

## Introduction

The need to provide safe and efficient assistive products for children with disabilities in Sub-Saharan Africa necessitates the availability of reliable and efficient information on the use of assistive products [1,2]. Unfortunately, the collection of information on the joint distribution of children with disabilities and their use of assistive products poses several challenges to national statistical agencies in the region. For example, budgetary constraints may make the designing of new nationally representative surveys that target children with disabilities' use of an assistive product, unfeasible [3,4]. Furthermore, collecting large amounts of data in a single survey may cause an undue burden to survey participants. A more realistic solution may be to add questions on the use of assistive products among children with disabilities to the existing survey efforts, such as Demographic and Health Surveys (DHS), Multiple Indicator Surveys (MICs), and Indicator Household

Surveys (HIS). If this is not possible, then statistical matching techniques could be a valid alternative.

An assistive product is defined by the Global Cooperation on Assistive Technology (GATE) as any product (including devices, equipment, instruments, and software) either specially designed and produced or generally available, whose primary purpose is to maintain or improve an individual's functioning and independence and thereby promote their wellbeing [5]. Among children with disabilities, assistive products have been found to be fundamental in their educational and societal inclusion through increased levels of independence in daily living and greater access to learning opportunities [6,7]. However, there is enough evidence to indicate that there is a general lack of information regarding the availability of affordable, accessible, contextual, and relevant assistive products among children with disabilities in the sub-Saharan African region [6,7]. Such lack of information undermines

efforts that could assist in the inclusion of children with disabilities in societies.

In Malawi, information on the use of assistive products among children with disabilities is collected during Population Censuses and Demographic and Health Surveys. The collected information only includes the use of eyeglasses and hearing aids. The use of mobility devices, such as white canes and wheelchairs, and other technologies, such as computers, is not collected. To address this limitation, this paper describes a methodology to combine data from the 2017 survey on “Living Conditions among persons with disabilities in Malawi” (LCS) and the 2015-16 “Malawi Demographic and Health Survey” (MDHS) to produce a composite dataset for studying the use of assistive products among children with disabilities.

### Statistical matching

Statistical matching is a technique used by practitioners to combine information from distinct data sources referring to the same target population [4,8]. The technique often involves two data files, A and B, where A and B share a set of common variables ( $X$ ), with variables  $Y$  observed only in A and variables  $Z$  observed only in B. The objective of statistical matching is to estimate the correlation coefficient between  $Y$  and  $Z$  conditional on  $X$  variables at a macro level, or to create a synthetic data source in which all the variables  $X$ ,  $Y$  and  $Z$  are available – the micro case [4,9].

Statistical matching is used in situations where variables of interest are not readily available in one data source and when two or more data sources do not have unique identifiers for merging or linking the variables [4,9,10]. For example, Simonson et al. [11] in their study of life course and old age incomes of Germany baby boomers failed to obtain a dataset that contained information on life course and old age income. To obtain such a dataset, they statistically matched the German Ageing survey and the Active Pension accounts to estimate the effect of changes in life course on an individual’s financial situation. In addition to the absence of unique identifiers, statistical matching can also be used in situations where detailed information for a particular topic entails development of long questionnaires which tend to have a lower response quality and a higher frequency of missing responses [8]. In these situations, statistical matching is used to reduce high missing response rates and improve response quality.

The inherent challenge with statistical matching is its outcome measures, which contain some levels of uncertainty due to the inability of the statistical matching technique to create true  $Y$  data for File A or true  $Z$  data for file B [10]. To solve this challenge, a number of researchers, including Rubin, Marcello et al. and Zhang have devised techniques of file concatenation with adjusted weights [10], use of logical constraints [12], and use of proxy variables [13]. In file concatenation, a database with imputed values is created by treating the two databases (A and B) as probability samples from the same population. The imputed values reflect the uncertainty of the values from which they have been imputed [10]. Logical constraints, on the other hand, are rules that make some of the parameter vectors in the joint distribution illogical for the investigated phenomenon. Logical constraints are introduced in statistical matching to eliminate impossible worlds [8,12]. For example, in the matching of datasets by age and marital status, a rule can be introduced such that it is not possible for a unit in a population to be both ten years old and married.

The other critical challenge in statistical matching is the assumption that the distribution of  $Y$  given  $X$  is independent of

the distribution of  $Z$  given  $X$  (*Conditional Independence Assumption*) [8]. The problem with this assumption is that it rarely holds in practice and that it cannot be tested from the datasets [14]. In situations where the assumption does not hold and no additional information is available to exploit the distribution of  $Y$  and  $Z$ , it is assumed that the model used to estimate the association between  $Y$  and  $Z$  has identification problems and that the artificial dataset produced may lead to incorrect inferences. To overcome the conditional independence assumption problem, two solutions have been suggested; the first is the use of some auxiliary information in the form of a small subset containing all the variables ( $X, Y, Z$ ) or just ( $Y, Z$ ) to explore the joint distribution of  $Y$  and  $Z$  [14,15]. The second one is the use of proxy variables with high predictive power. The proxy variables help mediate the relationship between  $Y$  and  $Z$  and make the conditional independence assumption hold true [14].

This paper, therefore, applies the statistical matching technique of the two distinct surveys. The two datasets have been combined using a variable on the use of assistive devices in the 2017 LCS and on disability in the 2015-16 MDHS. This study is part of the preparatory work for the Assistive Product list Implementation Creating Enablement of inclusive SDGs (APPLICABLE) project which seeks to develop a framework for creating an effective national Assistive Technology (AT) policy and specify a system capable of implementing that policy in Malawi [16].

## Method and data for addressing data deficiencies

### Data sources

The data sources used here to conduct statistical matching are the 2017 “Living Conditions among persons with disabilities in Malawi” survey and the 2015-16 MDHS. The 2017 LCS survey is a nationally representative dataset that draws its understanding of disability from the International Classification of Functioning, Disability, and Health (ICF) framework [17,18]. The information in this dataset was collected for the purpose of mapping out the living conditions of persons with disabilities and comparing it with that of the non-disabled population. The information on the living conditions among persons with disabilities was collected from 19946 individuals with disabilities and 10631 individuals without disabilities. The 2017 LCS survey is used as a “donor” dataset in this paper.

The 2015-16 Malawi Demographic and Health Survey, on the other hand, is used as a “recipient” dataset. The 2015-16 MDHS is a nationally representative survey, which was conducted with the purpose of providing current estimates of basic demographic and health indicators of the Malawian population. The survey collected information from 24,562 women aged 15 to 49 and 7478 men aged 15 to 54. The assumption in statistically matching the two datasets is that they were drawn from the same population as such the demographic characteristics (i.e., age, sex and place of residence) of the sampled population in the 2017 “Living Conditions among persons with disabilities in Malawi” sample are similar to the characteristics of the 2015-16 Malawi Demographic and Health Survey. For example, the mean age for children aged 2 to 17 is 9 for both datasets and the standard deviation is almost similar ( $SD = 4.31$  in the MDHS and  $SD = 4.28$  in the LCS).

### Study variables

The variables of interest in this analysis are “disability” and “use of assistive device”. Disability is a variable of interest because the desired objective is to match the number of children with

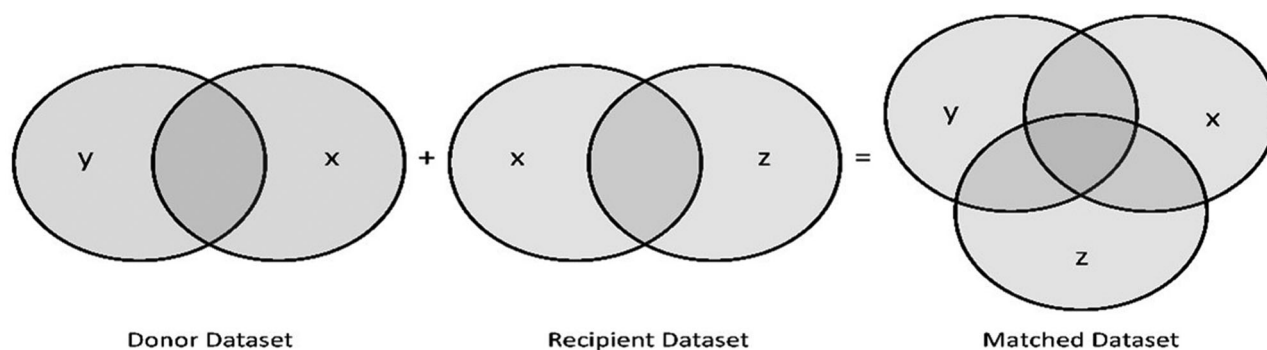


Figure 1. Graphical representation of the statistical matching technique.

disabilities in the 2015-16 Malawi Demographic and Health survey with those of 2017 Living conditions survey who were using assistive devices. The study focuses on children with disabilities because the MDHS only collected disability information for children aged between 2 to 17 years. Disability ( $Z$ ) is a bivariate variable created from self-reported responses about difficulties in seeing, hearing, communicating, walking, remembering things and washing oneself [19]. To create the disability variable, responses to the functioning domains were summed up to create a composite disability score ranging from zero (absence of disability) to 72 (extensive disability). Respondents with a total score of more than zero were then grouped together to create a binary variable with values “0” no disability (total score of 0) and 1 “having a disability” (total score from 1 to 72). The variable use of assistive device ( $Y$ ) has been chosen for statistical matching because research indicates that children who use assistive devices are able to achieve greater independence. Have reduced need for formal support services [20,21], as well as reduced time and physical burden to care givers [21,22]. Use of assistive device ( $Y$ ) in this study is a categorical variable with categories “does not use any assistive device” and “currently use an assistive device”. The two categories are the categories that were used when collecting the information on use of assistive devices during the survey.

The common variables denoted by  $X$  are place of residence, the age of participant, level of education and sex of respondent. These variables are available in both the 2017 LCS and the 2015-16 MDHS datasets. The statistical matching procedure thus involves the integration of the 2017 LCS with the 2015-16 MDHS, as illustrated in Figure 1.

## Analysis of data using statistical matching

### Harmonization of sources

Before starting the process of statistical matching, common variables in the 2017 “Living Conditions among persons with disabilities in Malawi” survey and the 2015-16 MDHS were harmonized to facilitate the coherence of the datasets. The name of the participant-id variable in the 2017 “Living Conditions among persons with disabilities in Malawi” survey, was renamed to match that of the 2015-16 MDHS dataset. The categories for age, sex, place of residence, and level of education of both datasets were also re-categorized into same value labels. The value labels of a place of residence in the 2017 “Living Conditions among persons with disabilities in Malawi” survey were also reclassified into a dummy variable 0=rural and 1= urban to match 2015-16 MDHS variable labels.

Apart from variable harmonization, the two datasets were also adjusted for missing values. This was achieved by removing

irrelevant values. For example, in the 2017 “Living Conditions among persons with disabilities in Malawi” survey, all-male participants and female participants aged 18 and above were removed from the dataset. This is because disability information for the 2015-16 MDHS was only collected from children aged 2 to 17. Participants aged below 2 in the 2017 “Living Conditions among persons with disabilities in Malawi” survey were thus removed to match those of the 2015-16 Malawi Demographic and Health survey.

In addition to adjusting for missing values and sample populations, frequency analysis of the common variables was conducted to examine proportional distributions. Examination of the proportional distributions was important because it is the proportional distribution of the matching variables that determine the marginal distribution of the imputed values.

### Selection of matching variables

In statistical matching applications, datasets A and B may share many common variables, but it is only the most *relevant variables* (variables that significantly explain the variation in the target variables, in this case, disability and use of assistive devices) that are used in the matching process [15,23]. The selection of these variables is performed using descriptive or inferential methodologies. In this study, the selection of matching variables involved the use of Chi-square and uncertainty measures of association.

### Chi-square test

The Chi-square test is a measure of association that is used to determine the association between two categorical variables. Disability and use of assistive devices are categorical variables, hence it was appropriate to use the Chi-square test of association to measure the relationship between these variables and the common variables. The Chi-square test produces a number of test statistics, but this paper concentrates on the post-estimation outputs, because the focus is on assessing the power of common variables in predicting the variation in disability and use of assistive devices. Therefore, only Cramer’s  $V$ , Goodman–Kruskal lambda ( $\lambda$ ) and Goodman–Kruskal tau ( $\tau$ ) were used as measures of association.

Cramer’s  $V$  is a Chi-square measure of association that is used to determine the strength of association between two categorical variables [24]. Its values range from 0 (no relationship) to 1 (a strong relationship between the two variables).

Goodman–Kruskal lambda ( $\lambda$ ) is another measure of Chi-square based association. It measures the proportional reduction in error that is achieved when membership of a category of one variable is used to predict category membership of the other variable [25].

Its values range from 0 (one variable does not predict the other) to 1 (one variable perfectly predicts the other).

In addition to Goodman–Kruskal lambda ( $\lambda$ ), Goodman–Kruskal tau ( $T$ ) was also used to select common variables. Goodman–Kruskal tau ( $T$ ) is the same as Goodman–Kruskal lambda ( $\lambda$ ), except that it measures the proportional reduction in error that is achieved by assigning probabilities specified by marginal or conditional proportions [26]. Goodman–Kruskal tau ( $T$ ) has values 0 (no association) and 1 (complete or perfect association).

### Uncertainty test

In addition to pairwise association, a test for uncertainty reduction was also conducted to assist in the selection of matching variables. This is done by selecting just those common variables with the highest contribution to the reduction of *uncertainty* i.e., the impact of the absence of joint information on use of assistive device ( $Y$ ) and disability on the estimates of the joint ( $Y, Z$ ) parameters [15,23,27]. The reduction of uncertainty technique allows exploration of uncertainty when all the variables ( $X, Y$ , and  $Z$ ) are categorical. It estimates the likely interval values for the probabilities in the contingency table  $Y \times Z$  as given by the Fréchet bound:

$$\max\{0, P(Y) + P(Z) - 1\} \leq P(Y \cap Z) \leq \min\{P(Y), P(Z)\}$$

where  $P(Y)$  is the probability of event  $Y$  happening and  $P(Z)$  is the probability of event  $Z$  happening independently.

Assuming that  $X_D$  relates to the complete crossing of the matching variables  $X_M$ , it can be shown that

$$P_{j,k}^{(low)} \leq P_{Y=j, Z=k} \leq P_{j,k}^{(up)}$$

where

$$P_{j,k}^{(low)} = \sum_i P_{X_D=i} \times \max\{0; P_{Y=j|X_D=i} + P_{Z=k|X_D=i} - 1\}$$

$$P_{j,k}^{(up)} = \sum_i P_{X_D=i} \times \min\{P_{Y=j|X_D=i}; P_{Z=k|X_D=i}\}$$

For  $j = 1, \dots, J$  and  $k = 1, \dots, K$  where  $J$  and  $K$  are categories of  $Y$  and  $Z$ , respectively [15].

And

$P_{X_D}$ : probability of the complete crossing of the matching variables

$P_{Y=j|X_D=i}$ : probability of disability given a complete crossing of matching variables

$P_{Z=k|X_D=i}$ : probability of parity given a complete crossing of matching variables

Therefore, for each cell in the contingency table  $Y \times Z$  for all possible combinations of the input  $X$  variables, the reduction of uncertainty is measured by the average widths of the interval:

$$\bar{d} = \frac{1}{J \times K} \sum_{j,k} (\hat{p}_{j,k}^{(up)} - \hat{p}_{j,k}^{(low)})$$

The reduction of uncertainty output reports the possible combination of  $X$  variables that can be used for matching. It also reports the number of cells in each of the input tables and the corresponding number of cells with a frequency equal to 0. The analysis also provides the average width of the uncertainty intervals  $[0, 1]$  and its relative value  $[0, 1]$  when compared with the average widths of the uncertainty intervals when no  $X$  variables are considered [9,15,28]. For our purposes, common variables that were not strongly associated with use of assistive devices and

disability were regarded as redundant predictors and were removed from the matching set.

### Statistical matching of data sources

The statistical matching technique used in this study is the random hot deck. This non-parametric technique method is often used under the Conditional Independence assumption (CIA). This technique integrates the two datasets by randomly selecting each of the donors from a subset of all the available donors [14,15]. This subset is formed by considering all the donors that share characteristics that are similar to that of the recipients [15]. The subset can be defined according to some  $X_M$  variables such as place of residence and age of the respondent. This process ensures the preservation of the marginal distribution of the imputed variables in the synthetic dataset [8,14]. The main concern with this technique is that each record in the donor file can be used more than once. This choice of multiple donors then reduces the effectiveness of the sample size and the empirical distribution of the imputed  $Z$  variable in the statistical matching file [14]. To address the concern of having multiple donors for each recipient file, a penalty weight is introduced to the donors already used and an algorithm is established that limits the factor of dependence which is introduced by the used donor units [14,15].

### Assessment of the accuracy of the statistical matching results

Following the statistical matching procedure, it was necessary to evaluate the accuracy of the matching results, even though research has proven that it is difficult to do so [14,28]. Accuracy assessment of statistical matching results is difficult because in statistical matching the relationship of phenomena not jointly observed is studied [29]. The statistical matching process may also provide different outputs, like a synthetic data set in the micro case or estimates of parameters (e.g., correlation coefficient) in the macro case. The available data sources may also have different quality levels (sampling design, sample size and data processing steps).

The aim of conducting statistical matching in this study was to produce a synthetic dataset that will be used for statistical inference. Therefore, it was necessary to evaluate the accuracy of the synthetic dataset. This was achieved by first examining how the synthetic dataset preserved the marginal distribution of the imputed variable use of assistive devices, by comparing it with the marginal distribution of use of assistive device estimated from the donor dataset (2017 LCS dataset). The second step was to examine how the synthetic data set preserved the joint distribution of the imputed variable with the matching variables, with the reference as the joint distribution of the estimates from the donor data set (2017 LCS dataset) [29]. The comparison of the marginal distribution of use of assistive device between the synthetic dataset and the donor dataset was accomplished by means of similarity or dissimilarity measures (i.e., total variation distance, overlap, Hellinger's Distance and Bhattacharyya coefficient).

A descriptive analysis of the imputed variables including use of assistive devices, information, communication, personal mobility, household items, personal care and protection and computer technology was also conducted to compare the proportional distribution of the imputed variables from the donor dataset (2017 LCS).



**Table 1.** Proportional distribution of common variables.

Variable	2017 Living Conditions study		2015-16 MDHS	
	%	n	%	N
Age group				
2–4	17.2	253	19.1	11,049
5–9	35.5	524	34.5	20,019
10–14	33.3	491	33.2	19,252
15–17	14.0	207	13.2	7671
Sex				
Male	53.5	789	50.3	29,156
Female	46.5	686	49.7	28,835
Place of residence				
Urban	7.0	103	15.9	9195
Rural	93.0	1372	84.1	48,796
Level of education				
No Education	43.3	638	26.3	15,234
Primary	55.0	811	70.3	40,758
Secondary	1.8	26	3.4	1980
Tertiary	0.0	0	0.0	19
Total	100.0	1475	100.0	57,991

## Results

### Harmonization of data sources

In statistically matching the two distinct datasets, that is the 2017 LCS survey and the 2015-16 MDHS, variables that were commonly found in the two datasets were used. The common variables include the age of the participant, sex, place of residence, and level of education. Table 1 presents the proportional distribution of these common variables. The tables indicate that both datasets had a high proportion of children aged 5 to 9 and 10 to 14. For instance, in the 2017 LCS, more than 30% of the sample were children aged 10 to 14, the same applied to the 2015-16 MDHS.

With regards to sex, Table 1 indicates that the 2017 LCS survey had a high proportion of males (53.5%) compared to females (46.5%), whilst the 2015-16 MDHS survey had a slightly higher proportion of males (50.3) compared to females (49.7). In terms of residence, Table 1 indicates that the proportional distribution of children in the living conditions survey was not representative of the Country's population distribution. According to the 2018 Population and housing census, 16% of the country's population lives in urban areas whilst 84% live in rural areas [30]. The 2017 LCS on the other hand, indicates that 7% of the sampled children were from the urban areas whilst 93% were from the rural areas. Concerning education, 43.3% of the children in the 2017 LCS had no education, 55% had primary education and only 1.8% had secondary education. On the other hand, the 2015-16 MDHS data indicates that 26.3% of the sampled children had no education whilst 70.3 and 3.4% had primary and secondary education respectively.

### Selection of matching variables

#### Chi-square test

Table 2 presents the Chi-Square test of association between the common variables and the use of assistive devices and childhood disability. Only Cramer's V results have been presented in the table because Goodman-Kruskal lambda ( $\lambda$ ) and Goodman-Kruskal tau (T) results indicated that not all the common variables predicted the category membership and the proportional reduction in error of predicting use of assistive devices and childhood disability. In-terms of Cramers' V, the results in Table 2 indicate that there is a weak association between the common variables, age, sex and place of residence, and use of assistive devices, and childhood disability.

**Table 2.** Chi-Square test of association.

Variable	Use of an assistive device			Disability in children		
	Cramer's V	df	p Value	Cramer's V	df	p Value
Age of respondent	0.05	15	0.16	0.20	15	0.00
Sex	0.04	1	0.13	0.01	1	0.12
Place of resident	0.07	1	0.00	0.03	1	0.00

**Table 3.** Table presenting the levels of uncertainty obtained from combining the common variables.

Variable combination	Cells with zero XY frequencies	Cells with zero XZ frequencies	Average width
Age	1	0	0.01895
Age $\times$ Sex	2	0	0.01904
Sex	1	0	0.01933
Residence $\times$ Age $\times$ Sex	3	3	0.02201
Residence	1	0	0.02273

#### Uncertainty test

Further to conducting the pairwise associations, an uncertainty test was also conducted to determine the combination of common variables with the highest contribution to the reduction of uncertainty. Looking at the average width of the cell bounds in Table 3 below, it appears that all the common variables (X) being considered should be used as matching variables. Unfortunately, the columns with zero frequencies indicate that a combination of all common variables produces a certain number of cells with zeros. Thus, a combination of place of residence, age and sex, produces 3 cells with zero frequencies in both their combination with use of assistive device (XY) and childhood disability (XZ). Regarding the impact of the absence of joint information on use of assistive devices (Y) and disability (Z) on the estimates of the joint (Y,Z) parameters, the results in Table 3, indicates that combining all the three common variables produces an uncertainty of 2.2%. This uncertainty value is not significantly higher than that of combining age and sex (1.9%). Therefore, it was ideal to use all three common variables as matching variables. However, there were not enough units on sex of the respondent in the donor file due to missing values that could be matched with the recipient file. Thus, only age and place of residence were used as common variables for the matching process.

#### Assessment of the accuracy of statistical matching

Upon completion of the statistical matching, it was necessary to assess the accuracy of the created synthetic dataset. The assessment was accomplished by comparing the marginal distribution of the imputed use of assistive devices with the original variable in the donor dataset through use of similarity and dissimilarity measures. The joint distribution of use of assistive devices with the matching variables (age and place of residence) in the synthetic dataset was also compared with the donor dataset.

The *similarity/dissimilarity* measurement results indicate that the marginal distribution of use of assistive devices in the synthetic dataset was similar to that of the donor dataset with an *Overlap* index value of close to 1 ( $Overlap = 0.997$ ) and associated *Bhattacharya coefficient* of close to 1 ( $Bhatt = 0.999$ ). With regards to the joint distribution of use of assistive devices with the matching variables in the synthetic dataset, in comparison to that of the donor dataset, also indicated that the joint marginal

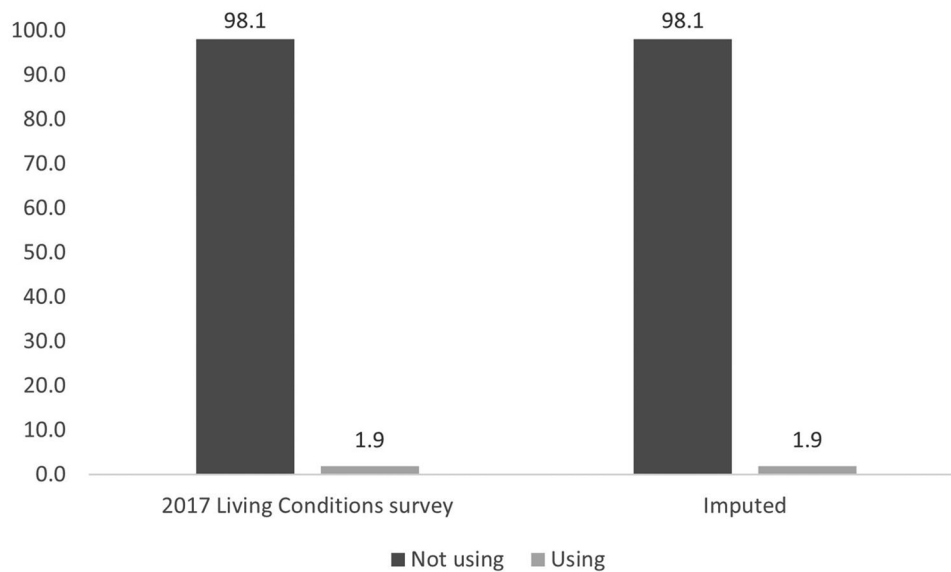


Figure 2. Graphical comparison of the proportional distribution of use of assistive products between the Imputed dataset ( $N=13,121$ ) and the 2017 Living Conditions survey ( $N=1475$ ).

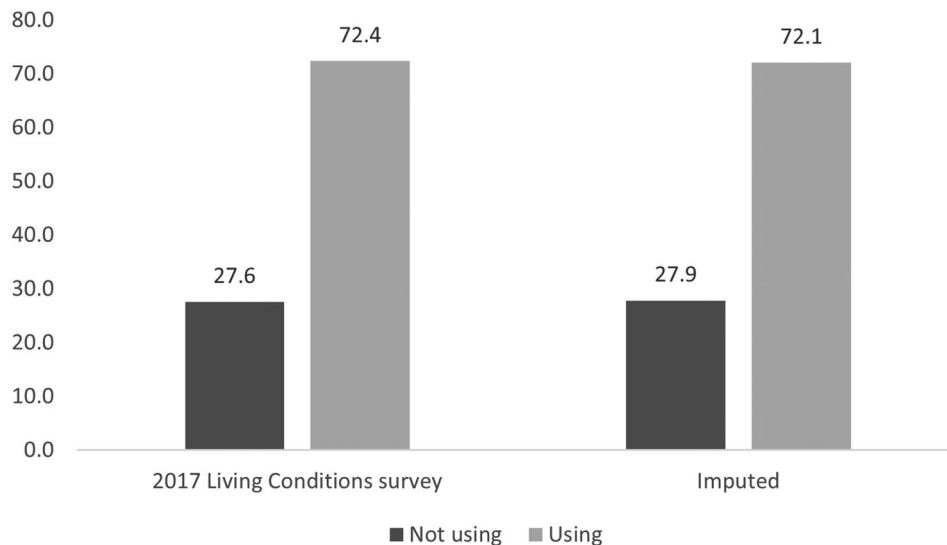


Figure 3. Graphical comparison of the proportional distribution of use of assistive devices for personal mobility between the imputed dataset ( $N=252$ ) and the 2017 Living conditions survey ( $N=28$ ).

distribution of the variables in the two datasets were the same with an *Overlap index* of 0.909 and an associated *Bhattacharyya coefficient* of 0.988.

Further to the use of the similarity/dissimilarity index to assess the accuracy of the statistical matching procedure, a descriptive analysis of the imputed variables “use of assistive devices” and use of assistive devices for personal mobility was also performed to graphically compare with the variables in the donor dataset. The results presented in Figures 2 and 3 below demonstrates the proportional distribution of the general question on use of assistive devices and the question on use of assistive devices for personal mobility. Figure 2 demonstrates that the proportional distribution of use of assistive devices in the synthetic dataset is similar to that of the donor dataset. The figure indicates that only 1.9% children with disabilities were using assistive devices. Figure 3 further demonstrates that among children who use assistive products the proportional distribution of children with disabilities

using assistive products for personal mobility in the synthetic dataset was similar to that of children with disabilities in the donor dataset or the 2017 LCS survey (72.1 and 72.4 respectively).

## Discussion

This paper has discussed the application of statistical matching to produce joint information on the use of assistive products and disability variables not jointly observed. The statistical matching procedure consisted of data harmonization, selection, and calibration of the matching variables, imputation of variables of interest through random hot deck method, and assessment of the accuracy of the outcome data. The statistical matching procedure has further shown that only 1.9% of children with disabilities in the country are using assistive products. Among those using assistive products, 27.9% are using assistive products for personal mobility.

Harmonization of the datasets through reclassification of the matching variables, assessment of missing values, and examination of the distribution pattern of matching variables through dissimilarity or similarity measures, assisted in the computation of representative imputed values. This process of data harmonization has not only been recommended as the first stage in statistical matching but has also been found to play a critical role in situations where there is a lack of consistency in the wording of similar questions in social surveys [3,8,14]. For example, in the statistical matching of the European Statistics on Income and Living conditions (EU-SILC) and the Household Budget Survey (HBS), Donatiello et al. [3] found harmonizing of the common variables in the two datasets improved the final estimations of household income, consumption, and wealth.

The pairwise association measures used for selecting matching variables have illustrated that there is a weak association between the age of the respondent and the use of an assistive device and disability in children. Age could not comprehensively explain the variation in disability in this study because of the complex relationship between age and self-reported disability. According to Jylhä et al. [31,32], age weakens the relationship between functional limitations and self-health assessment. As the age of participant increases the probability of reporting a functional limitation (disability) may stay the same.

The statistical matching of the 2017 LCS and the 2015-16 MDHS using the random hot deck method, demonstrates that the procedure preserves the marginal distribution of the variables after imputation, as shown by the dissimilarity and similarity indexes that were computed after the statistical matching procedure. These results correspond to the Leulescue et al. study, where the hot-deck method preserved the marginal distribution of life satisfaction, trust in institutions, and social exclusion variables, before and after imputation. Donatello et al. [3], also found the use of the hot-deck method to produce satisfactory results even though they are associated with high levels of uncertainty (more than 5%).

Concerning use of assistive technology, this paper has demonstrated that there is low usage of assistive technology among children with disabilities in the country. This low usage of assistive technology has also been observed in other countries in the sub-Saharan African region [7,33]. For example, in Ghana, Osam et al. observed that there was low usage of AT among children with disabilities. The low usage of AT was due to lack funds for purchasing the technologies and the high cost of ATs and rehabilitation services. In Tanzania, Mwajjande found children with physical disabilities to have no access to AT due material deprivation, low human development, lack of voice, and acute vulnerability to economic, social and health risks. With regards to Malawi, poverty could be the contributing factors to the low usage of AT. Nonetheless, there is need to further investigate factors contributing to the low use of assistive technology among children with disabilities in the country.

The main limitation of this study is the sample size of the 2017 LCS which was used as a donor dataset. The sample size of this donor dataset was smaller (1475) compared to that of the recipient dataset 2015-16 MDHS. The smaller sample size meant that other statistical matching techniques such as the nearest neighbour hot deck method could not be used to statically match the two datasets because it requires the donor file to be larger than the recipient file. Nonetheless, the 2017 LCS survey is amongst the most reliable nationally representative surveys that have collected data on use of assistive devices. Thus, it was the best representative dataset to use to address the issue of data deficiencies on use of assistive devices in Malawi.

## Conclusion

It can be concluded from this statistical matching procedure, that the matching procedure provides good data for measuring the use of assistive products among persons with disabilities in situations where the variables of interest have not been jointly observed. The data obtained from the matching procedure are also valid and reliable as shown by the similarity of the marginal distribution of the imputed variables and the donor dataset (2017 LCS). Nonetheless, there is a need for harmonization of the common variables in population surveys to improve the accuracy and consistency of the integrated datasets, since they play a critical role in the matching procedure. The data obtained from this statistical matching procedure can then be used to examine factors that predict the use of assistive products among persons with disabilities and so make an important contribution to systems strengthening in this area [34]. The need to do this is apparent both in the call for greater access to assistive products as a means to achieve the SDGs on a more equitable basis [35] and in the low rates of assistive produce use reported here. We have demonstrated how statistical matching can be used to combine distinct datasets that nonetheless have some relevant commonalities. Such a technique can be valuable in mining secondary data, the collection of which may have been funded from different sources and for different purposes. This is of significance for the efficient use of current and future datasets, allowing new questions to be asked and addressed by locally based researchers, including in more poorly resourced settings. It may also provide a scientific method that can contribute to addressing the political economy [36] of dominant donor agencies setting the research agenda in lower-income settings.

## Acknowledgements

The authors acknowledge the Malawi National Statistics Office and SINTEF for permitting use of the 2015-16 MDHS and 2017 Living Conditions survey; and the APPLICABLE project Action Research Group.

## Disclosure statement

The authors declare no conflict of interest.

## Funding

This work was supported by funding from the Irish Research Council (IRC) [grant number COALESCE/2019/114].

## ORCID

Malcolm MacLachlan  <http://orcid.org/0000-0001-6672-9206>

Ikenna D. Ebuenyi  <http://orcid.org/0000-0002-3329-6296>

Alister Munthali  <http://orcid.org/0000-0002-3495-3446>

## References

- [1] Borg J, Lindström A, Larsson S. Assistive technology in developing countries: national and international responsibilities to implement the Convention on the Rights of Persons with Disabilities. *Lancet*. 2009;374:1863–1865.



- [2] World Health Organization. Assistive technology for children with disabilities: creating opportunities for education, inclusion and participation. A discussion paper. 2015.
- [3] Donatiello G, et al. Statistical matching of income and consumption expenditures. *Int J Econ Sci.* 2014;3:50.
- [4] Moriarity C, Scheuren F. Statistical matching: a paradigm for assessing the uncertainty in the procedure. *J Off Stat.* 2001;17:407.
- [5] Khasnabis C, Mirza Z, MacLachlan M. Opening the GATE to inclusion for people with disabilities. *Lancet.* 2015;386:2229–2230.
- [6] Mji G, Edusei A. An introduction to a special issue on the role of assistive technology in social inclusion of persons with disabilities in Africa: Outcome of the fifth African Network for Evidence-to-Action in Disability conference. *Afr J Disabil.* 2019;8:681.
- [7] Osam JA, et al. The use of assistive technologies among children with disabilities: the perception of parents of children with disabilities in Ghana. *Disabil Rehabil Assist Technol.* 2019.
- [8] D’Orazio M, Zio MD, Scanu M. *Statistical matching: theory and practice.* Hoboken (NJ): John Wiley & Sons; 2006.
- [9] D’Orazio M. *Statistical matching and imputation of survey data with StatMatch.* 2016.
- [10] Rubin DB. Statistical matching using file concatenation with adjusted weights and multiple imputations. *J Bus Econ Stat.* 1986;4:87–94.
- [11] Simonson J, Gordo LR, Kelle N. Statistical matching of the German aging survey and the sample of active pension accounts as a source for analyzing life courses and old age incomes. *Hist Soc Res.* 2012;37:185–210.
- [12] Marcello D, Zio MD, Scanu M. Statistical matching for categorical data: displaying uncertainty and using logical constraints. *J Off Stat.* 2006;22:137.
- [13] Zhang L-C. On proxy variables and categorical data fusion. *J Off Stat.* 2015;31:783–807.
- [14] Leulescu A, Agafitei M. Statistical matching: a model based approach for data integration. *Eurostat-Methodologies and Working papers.* 2013.
- [15] D’Orazio M. Statistical matching and imputation of survey data with the Package StatMatch for the R Environment. R package vignette. 2011. [http://www.cros-portal.eu/sites/default/files/Statistical\\_Matching\\_with\\_StatMatch.pdf](http://www.cros-portal.eu/sites/default/files/Statistical_Matching_with_StatMatch.pdf)
- [16] Ebuenyi ID, Smith EM, Kafumba J, et al. Implementation of the Assistive Product List (APL) in Malawi through development of appropriate policy and systems: an action research protocol. *BMJ Open.* 2020;10:e040281
- [17] Eide AH, Munthali A. Living conditions among persons with disabilities in Malawi. A national, representative survey. 2017.
- [18] WHO. *Towards a common language for functioning, disability, and health: ICF. The international classification of functioning, disability and health,* 2002.
- [19] Nsonm I. *Malawi demographic and health survey 2015–16.* Zomba (Malawi): National Statistical Office; Rockville (MD): ICF; 2017.
- [20] World Health Organization. *World report on disability 2011.* Geneva (Switzerland): World Health Organization; 2011.
- [21] World Health Organization. *Joint position paper on the provision of mobility devices in less-resourced settings: a step towards implementation of the Convention on the Rights of Persons with Disabilities (CRPD) related to personal mobility.* Geneva (Switzerland): World Health Organization; 2011.
- [22] Allen S, Resnik L, Roy J. Promoting independence for wheelchair users: the role of home accommodations. *Gerontologist.* 2006;46:115–123.
- [23] Weber R, Weber D. Statistical matching of EU-SILC and the Household Budget Survey to compare poverty estimates using income, expenditures and material deprivation. *Eurostat Working Papers and Methodologies.* 2013.
- [24] Agresti A. *An introduction to categorical data analysis.* 3rd ed. New York (NY): Wiley; 2007.
- [25] Field A. *Discovering statistics using IBM SPSS statistics.* London (UK): Sage; 2013.
- [26] Reynolds HT. *Analysis of nominal data.* Vol. 7. 1984. Newburypark (CA): Sage.
- [27] D’Orazio M, Zio MD, Scanu M. Old and new approaches in statistical matching when samples are drawn with complex survey designs. *Proceedings of the 45th “Riunione Scientifica della Societa’Italiana di Statistica”;* Padova, Italy. 2010. p. 16–18.
- [28] D’Orazio M, Di Z, Scanu M. Statistical matching of data from complex sample surveys. *Proceedings of the European Conference on Quality in Official Statistics-Q2012.* 2012.
- [29] D’Orazio M. *Statistical matching: metodological issues and practice with R-StatMtach.* *Proceedings of the EUSTAT 55th International Statistical Seminar;* 2013 Nov 21–22; Vitoria-Gasteiz, Spain. 2013.
- [30] Office NS. *2018. Malawi population and housing census. Preliminary report.* 2018. Zomba (Malawi): National Statistical Office.
- [31] Jylhä M. What is self-rated health and why does it predict mortality? Towards a unified conceptual model. *Soc Sci Med.* 2009;69:307–316.
- [32] Jylhä M, et al. Walking difficulty, walking speed, and age as predictors of self-rated health: the women’s health and aging study. *J Gerontol A Biol Sci Med Sci.* 2001;56:M609–M617.
- [33] Mwaijande VT. *Access to education and assistive devices for children with physical disabilities in Tanzania.* Oslo and Akershus University College; 2014.
- [34] MacLachlan M, Banes D, Bell D, et al. Assistive technology policy: a position paper from the first global research, innovation, and education on assistive technology (GREAT) summit. *Disabil Rehabil Assist Technol.* 2018;13:454–466.
- [35] Tebbutt E, Brodmann R, Borg J, et al. Assistive products and the sustainable development goals (SDGs). *Global Health.* 2016;12:79.
- [36] Serrat O. Political economy analysis for development effectiveness. In: Serrat O, editor. *Knowledge solutions.* Singapore: Springer; 2017. p. 207–222.