Improving prevalence estimates from health interview surveys by using a random-forest based multiple imputation to correct for measurement error

I. Pelgrims[1,2,3] • B. Devleesschauwer[3,4] • S. Vandevijvere[3] • E.M. De Clerq[1] • S.Vansteelandt[2] • V.Gorasso[3] • J. Van Der Heyden[3]

1.Risk and Health Impact Assessment, Sciensano, Brussels, Belgium • 2.Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium • 3.Epidemiology and public health, Sciensano, Brussels, Belgium • 4. Department of Veterinary Public Health and Food Safety, Ghent University, Merelbeke, Belgium

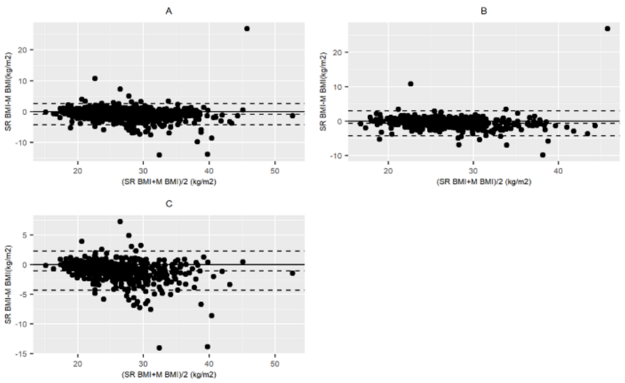
- **Random-forest multiple imputation proves to be a method of choice to correct the bias related to self-reported data in HIS data**
- **Whenever feasible, combined information from HIS and objective measurements should be used in NCD'S risk factors monitoring**

Relying on self-reported (SR) data from Health Interview Surveys often lead to biased prevalence rates of health conditions and NCD's risk factors. Current methods to correct for measurement error related to SR data include regression calibration (RC) or parametric multiple imputations techniques. This study explore a random-forest based multiple imputation by chained equations algorithm (RF MI) for using clinical information from the Belgian health examination survey (BELHES) to improve prevalence estimates of health conditions in the Belgian health interview surveys (BHIS).

Method

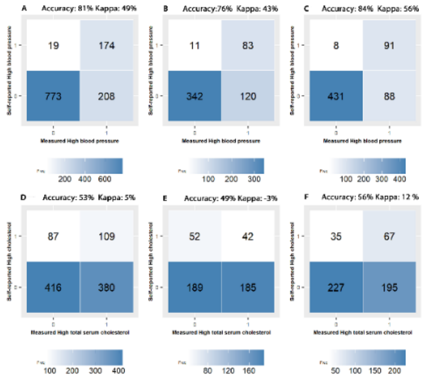
- Health conditions investigated: overweight, obesity, hypertension (HT), hypercholesterolemia (HC).
- 9439 participants of the 2018 BHIS older than 18 years, of which 1184 participated in the 2018 BELHES
- Agreement between SR and measured data: B&A plots, ICC, confusion matrix and Kappa coefficient
- RC: Measured health condition ~SR health condition + age + sex + education
- Classical & RF MI: imputation of missing clinical values for height, weight, HT and HC for every BHIS participant. Imputation model based on SR age, sex, education, height, weight, HT and HC. Prevalence assessed in each completed dataset (10) and results pooled using Rubin's rule.

Figure 1. Bland-Altman plot for analysis of agreement between self-reported and measured Body Mass Index (BMI), for the whole population and by gender



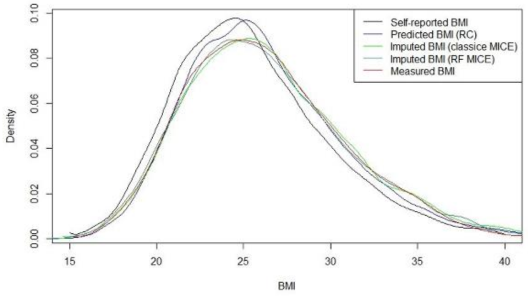
A: Whole study population, B: Men, C: Women. Solid line: mean difference. Dashed lines: upper and lower limits of agreement (mean difference +/- 2 sd)

Figure 2. Confusion matrix comparing self-reported and measured high blood pressure and hypercholesterolemia (for the whole population and by gender)



High blood pressure: A: Whole study population, B: Men, C: Women. Hypercholesterolemia: D : Whole study population, E: Men, F: Women.

Figure 3. BMI distribution using SR, measured and adjusted BHIS data.



Results

- High agreement for height (ICC: 0.96; 95% CI [0.95;0.97]), weight (ICC: 0.95; 95% CI [0.94;0.95]) and BMI (ICC: 0.92; 95% CI [0.86;0.95]). Moderate agreement for HT and poor agreement for HC
- RC: model accuracy was high for height (R^2:93%), and weight (R^2: 95%), moderate for HT (AUC: 86%) and poor for HC (AUC: 65%).
- Adjusted estimation obtained with RC and MI allowed to generate accurate national prevalence estimates, closer to their BELHES clinical counterparts.
- Alll methods provided smaller SE errors than those obtained with clinical data, except for HC for which RF MI was the only approach to provide smaller SE
- Besides its ability to handle data with complex interaction or non-linearity, RF MI does not require to specify an imputation model which is useful to allow secondary analysts to improve their analysis of SR data by using information included in the BELHES

Figure 4. Prevalence estimates of overweight, obesity, hypertension and hypercholesterolemia in Belgium using self-reported, measured and adjusted 2018 BHIS data.

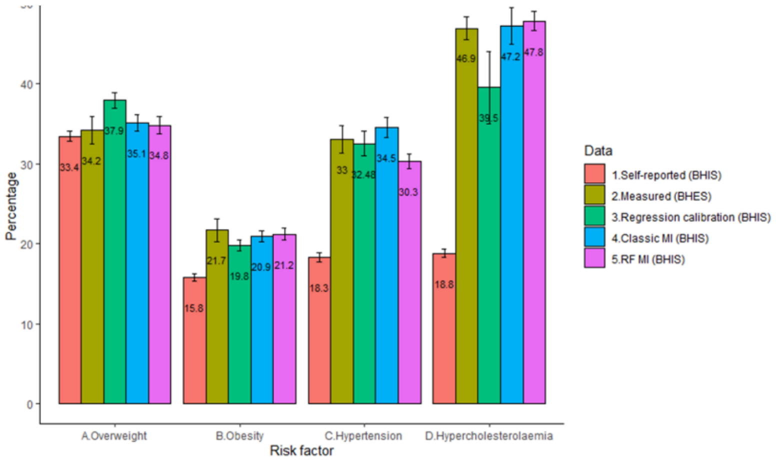


Table 1. Ratio of estimated SE: BELHES 2018 clinical data/adjusted BHIS 2018 data

	Regression calibration	Classic multiple imputation	Random-forest multiple imputation
Overweight	1.77	1.60	1.57
Obesity	2.13	2.07	1.81
Hypertension	1.10	1.34	1.79
Hypercholesterolemia	0.32	0.62	1.21

Sciensano • Pelgrims, Ingrid • Ingrid.pelgrims@sciensano.be • **www.sciensano.be**

This research was funded by the WaIST project