Associations between cardiovascular risk factors and audiometric hearing: Findings from the Canadian Longitudinal Study on Aging

Supplementary File: Multiple Imputation (MI) Approach

Table of Contents

Background of multiple imputation

This study focused on the association of cardiovascular risk factors (CV RF) with baseline average pure-tone hearing thresholds and changes in the average pure-tone hearing thresholds (PTA) over three years, and for this purpose, information was collected at two time points (baseline, T_0 and first follow-up, T_1). Repeated measurements of hearing thresholds (at T_0 and T_1) and predictor variables at baseline (T_0) were required to conduct the analyses. During the data collection period, some participants withdrew from the study; in some cases, data were not provided, and a few participants died during the follow-up time. The main analysis considered only individuals with complete data. A secondary analysis was performed in which multivariable regression models were executed using a multiple imputation approach, to determine if the two approaches yielded qualitatively different findings.

Comparisons of baseline characteristics among the group of individuals with complete data versus groups with missing data are described in Table 2 in the main text. We defined the complete group (n=21,492) as individuals who had information for all of our analytic variables. Participants with missing observations were distributed into three groups: (I) participants who were alive but did not have any data at T_1 time point (n=1,770) (II) participants who did not have data at T_1 time point due to death (n=562) and (III) participants who had partially missing value in any of the variables of interest (n=6,273). Table 1 in the main text illustrates that participants in all of the incomplete groups significantly differed from the complete group by being older, less educated, more likely to smoke, more likely to have hypertension, diabetes, obesity, and more likely to have a lower income. They also had worse average pure-tone hearing thresholds and

lower Physical Activity Scale for the Elderly (PASE) scores (indicating less physical activities) at baseline. Health and socioeconomic status generally declined across the groups in the following order: Complete data, data present at both time points but some data missing; no data at T1 but still alive; and no data at T1 and dead. Thus, the complete case analysis results would more likely represent a healthier, better educated and higher income population.

Since we found so many observed characteristics related to the missingness in this study, it is rational to presume that data are missing at random (MAR), which is a basic assumption for multiple imputation (MI) of missing values. To improve the accuracy of our MI procedure, we took the advantages of longitudinal study and used some of the available T_1 information in addition to other variables for imputing missing values of those respective variables.

Missing information on analytic variables

The following table (Table 1) shows that the percentage of missing value for our analytic variables ranged from 0% to 16.0%, and combinedly the total proportion of missing data across all variables was around 30%, thus reducing our sample size from 30,097 to 21,492 as complete cases. As a part of the sensitivity analysis, we decided to impute those missing values using multiple imputation technique.

We performed a Multiple Imputation Chained Equation (MICE) to impute missing values for variables having missing observations, and finally pooled estimates of multiple regression models were obtained through multiple imputation framework.

Table S1: Number of missing values for the variables included in analytic model and auxiliary variables used to improve multiple imputation predictions

Variable	Complete cases	Missing (Imputed)	Missing %	Total
Variables included in the Analytic model				
T_1 mid PTA[1]	25208	4889	16.24	30097
T_0 mid PTA	28715	1382	4.59	30097
Follow-up time	27209	2888	9.6	30097
Education	30047	50	0.17	30097
Income	28156	1941	6.45	30097
PASE	28789	1308	4.35	30097
Obesity	29853	244	0.81	30097
Diabetes	30075	22	0.07	30097
Dyslipidemia	27012	3085	10.25	30097
Hypertension	29963	134	0.45	30097
Smoker	30096	1	0.00	30097
Alcohol	30084	13	0.04	30097
Menopause	15217	103	0.67	15320
Additional Auxiliary Variables included in MI model to improve imputation process				
Auxiliary Variables for hearing				
T_1 low PTA[2]	25248	4849	16.11	30097
T_0 low PTA	28784	1313	4.36	30097
T_1 high PTA[3]	25093	5004	16.63	30097
T_0 high PTA	28569	1528	5.08	30097
T_0 better PTA[4]	29228	869	2.89	30097
T_1 better PTA	25696	4401	14.62	30097
T_0 self-rated hearing[5]	30069	28	0.09	30097
T_1 self-rated hearing	27105	2992	9.94	30097
T_0 hearing device use	30095	2	0.01	30097
T_1 hearing device use	27208	2889	9.6	30097
T_0 problem hearing in noise	30024	73	0.24	30097
T_1 problem hearing in noise	27026	3071	10.2	30097
Auxiliary Variable for Education				
Education change	27750	2347	7.8	30097
Auxiliary Variable for income				
T_1 household income	25994	4103	13.63	30097
Auxiliary Variable for PASE				
T_1 PASE	27763	2334	7.75	30097
Auxiliary Variable for obesity				
T_1 obesity	25976	4121	13.69	30097
Auxiliary Variable for diabetes				

T_1 diabetes	26715	3382	11.24	30097
Auxiliary Variable for hypertension				
T_1 hypertension	26739	3358	11.16	30097
Auxiliary Variable for smoking				
T_1 smoking status	27209	2888	9.6	30097
Auxiliary Variable for alcohol				
T1 alcohol consumption	27744	2353	7.82	30097

1. Mid PTA: Binaural mid-frequency (1000, 2000, 3000 and 4000 Hz) pure-tone average
2. Low PTA: Binaural low-frequency (500, 1000 and 2000 Hz) pure-tone average
3. High PTA: Binaural high-frequency (3000, 4000, 6000 and 8000 Hz) pure-tone average
4. Better PTA: Mid-frequency PTA in the better ear
5. Self-rated hearing defined as the response to the question, “Is your hearing, using a hearing aid if you use one…” (Excellent, very good, good, fair or poor)

Imputing missing values with Multiple Imputation by Chained Equation (MICE)

Multiple imputation by chained equation (MICE) was employed to deal with missing values for all variables in this analysis. As per Biering and colleagues approach to handling missing data (Biering et al., 2015), we constructed a specific statistical model for each of the variables with missing values and according to the scale of the variable (continuous, nominal and ordinal). For example, we applied linear regression to predict missing continuous variables and ordinal logistic regression to predict missing ordinal variables.

Predictors in each of the chained equations: When a variable with missing data was modelled to impute its missing values, all other analytic variables acted as predictors of the missing values. For baseline variables that were also measured at T_1, we used the T_1 variable (if available) as an auxiliary variable to predict the missing T_0 value.

The following additional auxiliary variables were also used to predict missing mid-frequency PTA values: Low frequency binaural PTA (500 Hz, 1000 Hz, 2000 Hz); high-frequency binaural PTA (4000 Hz, 6000 Hz, 8000 Hz); subjective rating of hearing ability ("Is your hearing, using a hearing aid if you use one excellent, very good, good, fair or poor?") (Likert scale, 1-5); subjective rating of hearing ability in the presence of background noise ("Do you find it difficult to follow a conversation if there is background noise, such as TV, radio or children playing, even if using a hearing aid as usual?" Yes or no); and hearing aid use ("Do you use any aids, specialized equipment, or services for persons hard of hearing, for example, a folume control telephone or TV decoder?" Yes or no). Variables at both time points were used to predict missing values for the binaural mid-frequency PTA (1000 Hz, 2000 Hz, 3000 Hz, 4000 Hz).

The following examples will illustrate the basic statistical approaches to imputing missing values of the binaural mid-frequency PTA at both time points.

$$PTA(mid)_{T1} = PTA(mid)_{T0} + PTA(low)_{T0} + PTA(low)_{T1} + PTA(high)_{T0} + PTA(high)_{T1} + PTA(better)_{T0} + PTA(better)_{T1} + age + sex + race + income + education + obesity + dyslipidemia + smoking + hypertension + alcohol + pase + dcs + self\_rated\ hearing_{T0} + hearingdevice_{to} + problem\ in\ hearing\ in\ noise_{T0} + self\_rated\ hearing_{T1} + hearingdevice_{T1} + problem\ in\ hearing\ in\ noise_{T1} + random\ variation$$

$$PTA(mid)_{T0} = PTA(mid)_{T1} + PTA(low)_{T0} + PTA(low)_{T1} + PTA(high)_{T0} + PTA(high)_{T1} + PTA(better)_{T0} + PTA(better)_{T1} + age + sex + race + income + education + obesity + dyslipidemia + smoking + hypertension + alcohol + pase + dcs + self\_rated\ hearing_{T0} + hearingdevice_{to} + problem\ in\ hearing\ in\ noise_{to} + self\_rated\ hearing_{T1} + hearingdevice_{T1} + problem\ in\ hearing\ in\ noise_{T1} + random\ variation$$

Here, in the first equation, T_0 information of mid PTA and low-high PTA, better PTA (BPTA), hearing device use, hearing problems in background noise, self-rated hearing at both time points were used as an additional set of auxiliary variables in addition to other main-effect model predictors (age, sex, race, income, education, obesity, dyslipidemia, smoking, hypertension, alcohol use, PASE, CLSA data collection site location (DCS)) for imputing the missing value

corresponding to mid pta at T_1 time point. Alternatively, in second equation, the T_1 information for mid pta and low-high pta, bpta, hearing device use, hearing noise, self-rated hearing at both time points along with other covariates were used for imputing T_0 mid PTA variable.

The same approach was applied to the variables where data at both time points are available, even though the number of predictors differed from model to model. In a situation where the variable does not change over time or the variable for which T_1 data were not available, we only used the other analytic variables (i.e., those included in the analytic regression models of our study).