addhaz: Contribution of Chronic Diseases to the Disability Burden Using R

by Renata Tiene de Carvalho Yokota, Caspar WN Looman, Wilma Johanna Nusselder, Herman Van Oyen and Geert Molenberghs

Abstract The increase in life expectancy followed by the burden of chronic diseases contributes to disability at older ages. The estimation of how much chronic conditions contribute to disability can be useful to develop public health strategies to reduce the burden. This paper introduces the R package **addhaz**, which is based on the attribution method (Nusselder and Looman, 2004) to partition disability into the additive contributions of diseases using cross-sectional data. The R package includes tools to fit the additive hazard model, the core of the attribution method, to binary and multinomial outcomes. The models are fitted by maximizing the binomial and multinomial log-likelihood functions using constrained optimization. Wald and bootstrap confidence intervals can be obtained for the parameter estimates. Also, the contribution of diseases to the disability prevalence and their bootstrap confidence intervals can be estimated. An additional feature is the possibility to use parallel computing to obtain the bootstrap confidence intervals. In this manuscript, we illustrate the use of **addhaz** with several examples for the binomial and multinomial models, using the data from the Brazilian National Health Survey, 2013.

Introduction

The increase in longevity observed worldwide is usually followed by the burden of chronic diseases, which are among the leading causes of disability late in life (Beard et al., 2016). Disability has become a public health priority due to its adverse effects on health outcomes and quality of life, resulting in increased costs of health care (Yang et al., 2014). Therefore, the identification of which chronic diseases are the main contributors to the disability burden plays an important role in the formulation of public health response to population aging (Klijs et al., 2011).

Although prospective studies are better suited to establish the causal relationship between chronic diseases and disability, they are costly and usually with limited sample size. Alternatively, cross-sectional data has been widely used to investigate the association of chronic diseases and disability. Among the methods based on cross-sectional data, the attribution method proposed by Nusselder and Looman (2004) has the advantage of partitioning the disability prevalence into the additive contributions of chronic diseases, taking into account multimorbidity and that disability can be present even in the absence of chronic diseases. The method was originally proposed for binary outcomes, but it was recently extended to multicategory response variables (Yokota et al., 2017) and it is based on the binomial and multinomial additive hazard models, respectively. The use of non-canonical link functions in the models imposes a constraint on the linear predictor, which limits the use of standard statistical software to fit the models, such as `glm` in R or `proc GLM` in SAS (SAS Institute Inc., 2008). Despite this practical difficulty, the attribution method for binary outcomes has been widely used previously with data from the Netherlands (Nusselder and Looman, 2004; Klijs et al., 2011), Belgium (Nusselder et al., 2005; Yokota et al., 2015b), Germany (Strobl et al., 2013), China (Chen et al., 2013), and Brazil (Yokota et al., 2016), owing to the development of the software in R to fit the binomial model and to estimate the contribution of diseases to the disability prevalence by Nusselder and Looman (2010) for non-R users, which is available upon request to the authors (w.nusselder@erasmusmc.nl).

In this manuscript we present the R package **addhaz**, which is an extension of the R software developed by Nusselder and Looman (2010), offering an open-source implementation of the binomial and multinomial additive hazard models. The R functions can also be used to calculate the contribution of chronic diseases to the disability burden for both models.

This paper is structured as follows. In Section 2, a brief description of the attribution method is presented, followed by the definition of the binomial and multinomial additive hazard models. Section 3 introduces some features and options of **addhaz**. The existing alternative software to fit the binomial and multinomial models is discussed in Section 4. Examples of how to use the R functions to fit the models and to estimate contributions are shown in Section 5, using the data of the 2013 Brazilian National Health Survey (BNHS). The main advantages and disadvantages of the attribution method and **addhaz** are discussed in Section 6. Finally, conclusions and recommendations for future research are outlined in Section 7.

Attribution method

Analogous to the mortality analysis, in which a single disease is assigned as underlying cause of death in the death certificate, the attribution method aims to assess the probability that a single reported disease was the cause of disability in a survey, taking into account that individuals can report more than one disease (multimorbidity) and that disability can be present without any reported diseases (Nusselder and Looman, 2004, 2010).

In the attribution method, the disability that is not associated with any disease included in the analysis is attributed to "background". The background can represent the effect of age-related losses in functioning; underreporting and underdiagnosed diseases; and other causes of disability that were not included in the survey or analysis. More details about the attribution method can be found elsewhere (Nusselder and Looman, 2004, 2010).

The following assumptions are required to fit the binomial and multinomial additive hazard models to cross-sectional data: (i) disability is caused by the diseases that are still present in the time of the survey and the background; (ii) the estimated cross-sectional cumulative rates reflect the transition rates that would have been estimated with longitudinal data (stationarity assumption); (iii) the recovery rate is zero; (iv) the ratio of the cause-specific cumulative rates to the overall cumulative rate is constant over time (proportionality assumption); (v) the start of the time (age) at risk to become disabled is the same for all diseases; (vi) individuals from the same age group are exposed to the same cumulative rate of disability for background; (vii) diseases and background act as independent competing causes of disability (Nusselder and Looman, 2004, 2010).

Binomial additive hazard model

For binary outcomes, the binomial additive hazard model is defined as:

$$\begin{aligned} y_i &\sim Bernoulli(\pi_i) \\ \pi_i &= 1 - \exp(-\eta_i) \\ \eta_i &= \alpha_{a_i} + \sum_{d=1}^{m} \beta_d X_{di} \end{aligned} \tag{1}$$

where y_i is the binary disability outcome; π_i is the disability probability for individual i; η_i is the linear predictor representing the overall cumulative rate (or cumulative hazard) of disability for individual i; a_i denotes the age group of individual i (with f age groups, a_i can only get values between $1, \ldots, f$); α_a is the cumulative rate of disability for background by age group a ($a = 1, \ldots, f$); β_d is the cumulative rate of disability (or disabling impact) for disease $d(1, \ldots, m)$; and X_{di} is the indicator variable for disease d and individual i.

A linear inequality constraint is applied to the linear predictor ($\eta_i \geq 0$) to ensure that π_i lies between $(0, 1)$. The standard errors (SE) for the regression coefficients are estimated based on the inverse of the observed information matrix. The 95% Wald confidence intervals (Wald CI) can be obtained using the standard errors described above (Wald CI) as showed in 2 or via bootstrapping (Efron and Tibshirani, 1994).

$$\begin{aligned} \text{95\% Wald CI} &= \widehat{\alpha}_a \pm 1.96(\widehat{SE}) \\ \text{95\% Wald CI} &= \widehat{\beta}_d \pm 1.96(\widehat{SE}) \end{aligned} \tag{2}$$

Multinomial additive hazard model

The multinomial additive hazard model is an extension of the binomial model:

$$\begin{aligned} y_{ij} &\sim Multinomial(1, \Pi_i) \\ \pi_{ij} &= \left[1 - \exp(-\sum_{q=1}^{c} \eta_{iq})\right] \left(\frac{\eta_{ij}}{\sum_{q=1}^{c} \eta_{iq}}\right) \\ \eta_{ij} &= \alpha_{a_i j} + \sum_{d=1}^{m} \beta_{dj} X_{di} \end{aligned} \tag{3}$$

where y_{ij} is the multinomial response variable (disability) with one independent observation and vector of disability probabilities $\Pi_i = (\pi_{i0}, \ldots, \pi_{ij}, \ldots, \pi_{ic})$ per individual i; π_{ij} is the probability of disability category j for individual i; η_{ij} is the linear predictor (overall cumulative rate) for disability category j and individual i; a_i denotes the age group of individual i (with f age groups, a_i can only

get values between $1, \ldots, f$); α_{aj} is the cumulative rate of disability category j for background by age group a ($a = 1, \ldots, f$); β_{dj} is the cumulative rate of disability category j or disabling impact for disease $d(1, \ldots, m)$; and X_{di} is the indicator variable for disease d and individual i.

Besides the inequality constraint in the linear predictor $\eta_{ij} \geq 0$, an additional constraint is required: $\sum_{j=1}^{c} \pi_{ij} < 1$, to ensure that all $\pi_{ij} > 0$. Similar to the binomial case, the standard errors are estimated by the inverse of the observed information matrix and the 95% Wald CI and bootstrap percentile confidence intervals (bootstrap CI) can be obtained using **addhaz**.

Contribution of chronic diseases and background to the disability prevalence

The attribution of disability to chronic diseases depends on the disease prevalence (X_d) and the disabling impacts of the diseases (β_d and β_{dj}) (Nusselder and Looman, 2004, 2010). The contribution of chronic diseases and background to the disability prevalence can be calculated in five steps for both binary and multicategory response variables.

Binary case

For the binary case, the cause-specific disability probabilities for individual i and each chronic condition ($\widehat{D}_{di}$) and the background ($\widehat{B}_i$) are estimated based on the proportionality assumption, analogous to the competing risks setting in mortality analysis (Manton and Stallard, 1984; Chiang, 1991):

$$\begin{aligned} \widehat{D}_{di} &= \widehat{\pi}_i \left(\frac{\widehat{\beta}_d X_{di}}{\widehat{\eta}_i} \right) \\ \widehat{B}_i &= \widehat{\pi}_i \left(\frac{\widehat{\alpha}_{ai}}{\widehat{\eta}_i} \right) \end{aligned} \tag{4}$$

Next, the number of disabled individuals by each disease ($\widehat{N}_d$) and background ($\widehat{N}_b$) are estimated as:

$$\begin{aligned} \widehat{N}_d &= \sum_{i=1}^{n} \widehat{D}_{di} \\ \widehat{N}_b &= \sum_{i=1}^{n} \widehat{B}_i \end{aligned} \tag{5}$$

The absolute contribution of each disease ($\widehat{AC}_d$) and background ($\widehat{AC}_b$) to the total disability prevalence is obtained by:

$$\begin{aligned} \widehat{AC}_d &= \frac{\widehat{N}_d}{n} \\ \widehat{AC}_b &= \frac{\widehat{N}_b}{n} \end{aligned} \tag{6}$$

The absolute contribution of background and diseases defined above sum to the disability prevalence ($\widehat{P}$):

$$\widehat{P} = \widehat{AC}_b + \sum_{d=1}^{m} \widehat{AC}_d \tag{7}$$

Finally, the relative contribution of diseases ($\widehat{RC}_d$) and background ($\widehat{RC}_b$) to the disability prevalence is estimated by:

$$\begin{aligned} \widehat{RC}_d &= \frac{\widehat{AC}_d}{\widehat{P}} \\ \widehat{RC}_b &= \frac{\widehat{AC}_b}{\widehat{P}} \end{aligned} \tag{8}$$

The relative contributions ($\widehat{RC}_d$ and $\widehat{RC}_b$) sum to 1.

Multinomial case

Analogous to the binomial case, the contribution of chronic diseases to the disability prevalence for multinomial outcomes can be obtained in five steps for each category j of the outcome:

1. Cause-specific disability probabilities for each disease ($\widehat{D}_{dij}$) and background ($\widehat{B}_{ij}$) for individual i:

$$\begin{aligned} \widehat{D}_{dij} &= \widehat{\pi}_{ij}\left(\frac{\widehat{\beta}_{dj}X_{di}}{\widehat{\eta}_{ij}}\right) \\ \widehat{B}_{ij} &= \widehat{\pi}_{ij}\left(\frac{\widehat{\alpha}_{a_i j}}{\widehat{\eta}_{ij}}\right) \end{aligned} \tag{9}$$

2. Number of disabled individuals by each disease ($\widehat{N}_{dj}$) and background ($\widehat{N}_{bj}$):

$$\begin{aligned} \widehat{N}_{dj} &= \sum_{i=1}^{n}\widehat{D}_{dij} \\ \widehat{N}_{bj} &= \sum_{i=1}^{n}\widehat{B}_{ij} \end{aligned} \tag{10}$$

3. Absolute contribution of each disease ($\widehat{AC}_{dj}$) and background ($\widehat{AC}_{bj}$) to the total disability prevalence:

$$\begin{aligned} \widehat{AC}_{dj} &= \frac{\widehat{N}_{dj}}{n} \\ \widehat{AC}_{bj} &= \frac{\widehat{N}_{bj}}{n} \end{aligned} \tag{11}$$

4. Total disability prevalence ($\widehat{P}_j$):

$$\widehat{P}_j = \widehat{AC}_{bj} + \sum_{d=1}^{m}\widehat{AC}_{dj} \tag{12}$$

5. Relative contribution of diseases ($\widehat{RC}_{dj}$) and background ($\widehat{RC}_{bj}$) to the disability prevalence:

$$\begin{aligned} \widehat{RC}_{dj} &= \frac{\widehat{AC}_{dj}}{\widehat{P}_j} \\ \widehat{RC}_{bj} &= \frac{\widehat{AC}_{bj}}{\widehat{P}_j} \end{aligned} \tag{13}$$

The absolute contributions defined in equations 11 sum to the prevalence of disability for each category j and the relative contributions defined in equations 13 sum to 1 for each disability category j. The confidence intervals for the absolute and relative contributions for the binary and multinomial cases can be estimated via bootstrapping (Efron and Tibshirani, 1994) in **addhaz**.

Features of addhaz

In this section, a brief explanation about the constrained optimization, the parallel option to obtain the bootstrap CI for the parameter estimates, and the option to perform the likelihood ratio test for model selection is provided.

Constrained optimization

The binomial and multinomial additive hazard models are generalized linear models with non-canonical link functions $\eta_i = \log\left(\frac{1}{1-\pi_i}\right)$ for the binomial model and $\eta_{ij} = [-\log(1-\sum_{q=1}^{c}\pi_{iq})]\left(\frac{\pi_{ij}}{\sum_{q=1}^{c}\pi_{iq}}\right)$ for the multinomial model. For both models, the model parameters are estimated using maximum likelihood. The use of non-canonical link functions requires a constraint in the linear predictors ($\eta_i \geq 0$ and $\eta_{ij} \geq 0$) to ensure that the disability probabilities (π_i and π_{ij}) are valid, i.e., the probabilities lie between 0 and 1. In **addhaz**, this constraint is implemented in the optimization procedure, with an adaptive barrier algorithm (Lange, 2010), by calling `constrOptim` in R.

Parallel computing for the bootstrap CI

Besides the option to estimate the confidence intervals for the parameter estimates based on the standard errors obtained by the inverse of the information matrix for the binomial and multinomial

models (Wald CI), **addhaz** also offers the user the option to obtain the bootstrap CI based on empirical percentiles of the replicates (Efron and Tibshirani, 1994).

To reduce computation time, parallel computing can be used to obtain the bootstrap CI. By default R does not use all the cores available on a computer. The basic idea of parallel computing is to split the work to more than one core of the computer, to execute it in parallel, and then to collect the results. Several R packages can be used to implement parallel computation. In **addhaz** it is implemented by calling the functions `boot` and `boot.ci` in the **boot** package (Canty and Ripley, 2016; Davison and Hinkley, 1997).

Likelihood ratio test

The package also includes a function to perform the likelihood ratio test to compare two binomial or multinomial nested models that can be used for model selection.

The likelihood ratio test is defined as -2*log(likelihood model 1/likelihood model 2).The resulting test statistic is assumed to follow a χ^2 distribution, with degrees of freedom (df) equal to the difference of the df between the models. If the test is statistically significant, the model with more variables fits the data significantly better than the model with less variables.

Alternative software

The original software for the attribution method (Nusselder and Looman, 2004, 2010) was developed in R, but it is not available as an R package, since it focuses on non-R users: an Excel file is used to input the model arguments and this file is called in the R code. This software is restricted to binary outcomes and it is freely available upon request to the authors. Different from **addhaz**, a penalty term is added to the binomial likelihood function when $\pi_i \leq 0$, to ensure that valid probabilities are obtained. The original software also allows the user to obtain the bootstrap CI for the model parameters and contributions. Additionally, it offers the options: (i) reduced rank regression (RRR) (Yee, 2014) to reduce the number of parameters when interactions between diseases and age groups are of interest; and (ii) model selection, using the likelihood ratio test.

Besides the original software, the parameter estimates of the binomial additive hazard model can be obtained using the R packages **stats** with `glm` and **logbin** with the function `logbin` (Donoghoe, 2016). In both packages, the log-binomial model, i.e., $\pi_i = \exp(\eta_i)$, used to estimate relative risks (Marschner and Gillett, 2012), can be fitted to a transformed version of the response variable $y^* = 1 - y$, with the log link function. The estimated model parameters should be multiplied by -1, since $1 - \pi_i = \exp(-\eta_i)$. However, care should be taken with `glm`: by specifying the option `family = binomial(link = log)` to fit the log-binomial model, convergence failure may occur with the iterative weighted least squares (IWLS) algorithm when the maximum likelihood estimates (MLE) lie on the boundary of the parameter space. In `glm`, the IWLS is modified so that if constraints are violated, step-halving is used iteratively until they are respected. Although this should not result in invalid estimates, it may cause difficulty in convergence. An advantage of **logbin** is that it includes constrained optimization as an option to optimize the binomial log-likelihood function `(method = "ab")`. This is done by calling `constrOptim` in R to constrain the parameter space.

Since **addhaz** was developed with focus on the attribution method, apart from estimating the model parameters for the binomial additive hazard model, it also gives the user the option to obtain the contribution of diseases to the disability prevalence and to obtain bootstrap CI for the parameter estimates and the contributions, using parallel computing to reduce computation time.

To our knowledge, there is no other package available to fit the multinomial additive hazard model. Although it is possible to fit the log-multinomial model (Blizzard and Hosmer, 2007), i.e., $\pi_{ij} = \exp(\eta_{ij})$, with the function `vglm` in **VGAM** (Yee, 2016), different from the binomial case, no simple transformation of the outcome can be applied to obtain the parameter estimates of the multinomial additive hazard model using the log-multinomial model.

Using the package addhaz

In this section, the use of the functions `BinAddHaz`, `MultAddHaz`, and `LRTest` in **addhaz** are illustrated using a subset of the 2013 BNHS available in the package. A selected output of the results is shown.

Data description

The Brazilian National Health Survey (BNHS) ("Pesquisa Nacional de Saúde") was a nationally representative survey of the Brazilian adult population ($\geq$ 18 years) with approximately 60,000 individuals, conducted in 2013. A multistage sampling design with simple random sampling (census tracts) and clustering (households and adults) was used. The response rate was 77%. Survey weights were included to take into account the complex design of the sample. Detailed information about the BNHS can be found in previous publications (Szwarcwald et al., 2014; Yokota et al., 2016).

In **addhaz**, a subset of the BNHS data is available, with women aged $\geq$ 60 years (n = 6,294) and the following variables: disability as binary and multinomial outcomes, survey weight, age, diabetes, arthritis, and stroke (Table 1).

Variable name	Definition	Type	Categories
wgt	survey weight	continuous	-
age	age group	binary	0: 60-79 years 1: $\geq$80 years
diab	diabetes	binary	0: no 1: yes
arth	arthritis	binary	0: no 1: yes
stro	stroke	binary	0: no 1: yes
dis.bin	binary disability outcome	binary	0: no disability 1: disabled
dis.mult	multinomial disability outcome	categorical	0: no disability 1: mild disability 2: severe disability

Table 1: *Description of the variables included in the analysis. Brazilian National Health Survey, 2013.*

The binomial and multinomial disability outcomes were defined based on five instrumental activities of daily living (IADL): shopping, handling finances, taking own medications, going to the doctor, and using transportation. Participants were asked about the degree of difficulty in performing IADL tasks, with possible answers: "1. Unable", "2. A lot of difficulty", "3. Some difficulty", or "4. No difficulty". The definition of the binary and multinomial outcomes is shown in Table 2. The reference category "No disability" represents answer "4" to all IADL questions.

Outcome	Outcome category	Answer to at least one IADL question
Binary	Disabled	1, 2 or 3
Multinomial	Mild disability	3
	Severe disability	1 or 2

Table 2: *Definition of the binary and multinomial disability outcomes. Brazilian National Health Survey, 2013. "IADL" refers to instrumental activities of daily living.*

A summary of the characteristics of the study population is shown in Table 3. A higher proportion of older women ($\geq$80 years), diabetes, arthritis, stroke, and the disease pairs was observed in women with mild and severe disability compared to women without disability.

Examples with binary outcomes

The function `BinAddHaz` fits the binomial additive hazard model and estimates the contribution of diseases to the disability burden for binary outcomes in **addhaz**. To illustrate the use of `BinAddHaz`, five models were fitted with the binary disability outcome: model 1 - with three diseases (no background and diseases by age); model 2 - with only background by age, with bootstrap CI; model 3 - with only diseases by age; model 4 - with background and diseases by age, with bootstrap CI; model 5 - with two-way interaction between diseases. To illustrate how the `LRTest` function can be used for model selection, models 2 and 4 are compared.

Characteristic	Total		No disability		Mild disability		Severe disability	
	N	%	N	%	N	%	N	%
Age (years)								
60-79	5379	85.5	3946	93.5	642	78.3	791	63.0
≥80	915	14.5	273	6.5	178	21.7	464	37.0
Diseases								
Diabetes	1243	19.7	697	16.5	190	23.2	356	28.4
Arthritis	1428	22.7	819	19.4	211	25.7	398	31.7
Stroke	286	4.5	100	2.4	41	5.0	145	11.6
Diabetes and arthritis	298	4.7	135	3.2	53	6.5	110	8.8
Diabetes and stroke	91	1.4	31	0.7	13	1.6	47	3.7
Arthritis and stroke	79	1.3	28	0.7	7	0.9	44	3.5

Table 3: *Characteristics of the study population. Brazilian National Health Survey, 2013. The percentages refer to unweighted proportions, i.e., without taking into account the survey weights.*

Model 1 - Binomial model with three diseases

Before fitting model 1, **addhaz** and the data can be loaded and the names of the variables can be checked using:

```
library("addhaz")
data("disabData")
names(disabData)

[1] "dis.bin" "dis.mult" "wgt" "age" "diab" "arth" "stro"
```

The first binomial model can be fitted with:

```
model1 <- BinAddHaz(dis.bin ~ diab + arth + stro , data = disabData, weights = wgt,
                    attrib = TRUE, type.attrib = "both", collapse.background = FALSE,
                    attrib.disease = FALSE)
```

Since no attribution variable such as age was included in the model, the arguments `collapse.background` and `attrib.disease` were set to `FALSE`. The results of the model were stored as an object called `model1`, which can be checked with the `summary` function:

```
summary(model1)

$`call`
BinAddHaz(formula = dis.bin ~ diab + arth + stro, data = disabData,
          weights = wgt, attrib = TRUE, collapse.background = FALSE,
          attrib.disease = FALSE, type.attrib = "both")

$bootstrap
[1] FALSE

$coefficients
             Estimate      StdErr   t.value       p.value
(Intercept) 0.2970833 0.009426403 31.516082 1.498823e-202
diab        0.1331831 0.023821666  5.590838  2.354697e-08
arth        0.1306445 0.022203662  5.883917  4.212860e-09
stro        0.5927519 0.075697633  7.830521  5.663378e-15

attr(,"class")
[1] "summary.binaddhazmod"
```

The first element of the output `call` is the formula used to fit the model. The `bootstrap`, indicates if the bootstrap CI was requested. Since it was not requested, its value is `FALSE`.

Next, the coefficients are printed, with their estimates, standard errors (calculated based on the inverse of the observed information matrix), the t value (value of the t statistic), and the p value. The intercept represents the background. According to this output, all diseases were significant in the model. To identify the most disabling diseases, i.e., the diseases with highest cumulative rate of disability, the coefficients can be sorted in decreasing order using:

```
sort(model1$coefficients, decreasing = TRUE)

   stro    (Intercept)    diab        arth
0.5927519   0.2970833   0.1331831   0.1306445
```

Stroke was the most disabling disease, while arthritis was the least disabling disease. The 95% Wald CI can be obtained by:

```
model1$ci

                CI2.5    CI97.5
(Intercept) 0.27860754 0.3155590
diab        0.08649261 0.1798735
arth        0.08712532 0.1741637
stro        0.44438455 0.7411193
```

Both the relative and absolute contributions were requested (`attrib = TRUE,type.attrib = "both"`) and can be assessed with:

```
model1$contribution

$`att.rel`
               att.rel
(Intercept) 0.80405374
diab        0.06938567
arth        0.07451155
stro        0.05204903

$att.abs
               att.abs
disab       0.30853338
(Intercept) 0.24807742
diab        0.02140780
arth        0.02298930
stro        0.01605886
```

In the above output, the relative contribution (`att.rel`: the contributions sum to 1) is shown at the top and the absolute contribution (`att.abs`: the contributions sum to the disability prevalence) is presented at the bottom. No confidence intervals are provided for the contributions, as they can only be calculated via bootstrapping and this option was not requested.

In the output for the absolute contribution, the disability prevalence (`disab`) was 30.9%. The absolute contribution can be sorted in decreasing order using:

```
model1$contribution$att.abs[order(model1$contribution$att.abs[, 1], decreasing = TRUE), ]

    disab   (Intercept)     arth       diab        stro
0.30853338  0.24807742  0.02298930  0.02140780  0.01605886
```

The background (`Intercept`) was the main contributor to the disability burden in this population. In this case, it can represent other causes not included in the model such as dementia or back pain, which are important causes of disability in the older population, but were not included in the analysis. Among the three diseases, arthritis was the main contributor to the disabilitiy burden in older Brazilian women.

It is interesting to note that, although stroke was the most disabling disease, it showed the lowest contribution to the total disability prevalence. This low contribution can be a consequence of the low occurrence of stroke in this population - 4.5% (Table 3) - as the contribution of chronic conditions to the disability prevalence depends on both, the disease occurrence and the disabling impact (Nusselder and Looman, 2010).

Model 2 - Binomial model with only background by age, with bootstrap CI

In model 2, the background is modelled by age group, but the diseases are not. The model can be fitted by:

```
model2 <- BinAddHaz(dis.bin ~ factor(age) -1 + diab + arth + stro , data = disabData,
                    weights = wgt, attrib = TRUE, type.attrib = "both",
                    collapse.background = FALSE, attrib.disease = FALSE, seed = 111,
                    bootstrap = TRUE, conf.level = 0.95, nbootstrap = 1000,
                    parallel = TRUE, type.parallel = "snow", ncpus = 4)
```

Since the background is modelled by age group, `factor(age)` is included in the model. The `-1` is included in the `model.formula` to obtain the coefficients for both age groups, including the reference category. Since the background is modelled by age, it should not be collapsed by age (`collapse.background = FALSE`). As no interaction between diseases and age were included in the model, the argument `attrib.disease` is `FALSE`. The option `seed = 111` allows the user to specify a random number to make the results of the bootstrapping reproducible. In the example above, the random number used was 111. The bootstrap CI for the regression coefficients and the contributions was requested (`bootstrap = TRUE`), with confidence level = 0.95 (`conf.level = 0.95`). The bootstrap CI was based on 1000 replicates (`nbootstrap = 1000`) and it was obatined with parallel computing (`parallel = TRUE`) on Windows (`type.parallel = "snow"`) with 4 CPUS (`ncpus = 4`).

The summary of the model can be assessed with:

```
summary(model2)

$`call`
BinAddHaz(formula = dis.bin ~ factor(age) - 1 + diab + arth + stro, data = disabData,
          weights = wgt, attrib = TRUE, type.attrib = "both",
          collapse.background = FALSE, attrib.disease = FALSE, seed = 111,
          bootstrap = TRUE, conf.level = 0.95, nbootstrap = 1000,
          parallel = TRUE, type.parallel = "snow", ncpus = 4)

$bootstrap
[1] TRUE

$coefficients
               Estimate      CILow    CIHigh
factor(age)0 0.22345184 0.19892365 0.2529054
factor(age)1 1.10472873 0.95279272 1.2705886
diab         0.12935797 0.06191508 0.2044457
arth         0.08531865 0.02375110 0.1513319
stro         0.52453664 0.28688675 0.8509963

$conf.level
[1] 0.95

attr(,"class")
[1] "summary.binaddhazmod"
```

Since the bootstrap CI was requested, `bootstrap` is `TRUE`. The coefficients show that age and all diseases were significant (the null value, i.e. 0, is not included in the bootstrap CI). The `factor(age)0` and `factor(age)1` represents the background cumulative rates for age groups 0 and 1, respectively. The contributions can be checked with:

```
model2$contribution

$att.rel
             Contribution      CILow     CIHigh
factor(age)0   0.53992912 0.51937657 0.56054196
factor(age)1   0.29842186 0.27575265 0.32163433
diab           0.06702577 0.06162503 0.07256112
arth           0.04869951 0.04513817 0.05269298
stro           0.04592374 0.03666781 0.05519789

$att.abs
             Contribution      CILow     CIHigh
disab          0.30902546 0.30227641 0.31623119
factor(age)0   0.16685185 0.16405831 0.16947954
factor(age)1   0.09221995 0.08372162 0.10131428
diab           0.02071267 0.01906935 0.02254047
```

```
arth           0.01504939 0.01396744 0.01628476
stro           0.01419161 0.01127267 0.01727091
```

The contributions and the 95% bootstrap CI are shown. The background is the main contributor to the disability prevalence. Note that by allowing the background to vary by age does not change the disability prevalence (30.9%), as compared to model 1.

Model 3 - Binomial model with only diseases by age

In Model 3, we allow only the diseases to vary by age group by including interactions between age and diseases in the model. Before fitting model 3, a matrix with the diseases to be included in the model can be defined by:

```
disease <- as.matrix(disabData[, c("diab", "arth", "stro")])
```

The first six elements of the matrix created can be checked using:

```
head(disease)
```

```
    diab arth stro
36     1    0    0
98     0    0    0
113    0    1    1
347    1    0    0
352    1    0    0
436    0    0    0
```

The binomial additive hazard model can be fitted with the function:

```
model3 <- BinAddHaz(dis.bin ~ disease:factor(age), data = disabData, weights = wgt,
                    attrib = TRUE, attrib.var = age, type.attrib = "abs",
                    collapse.background = FALSE, attrib.disease = TRUE)
summary(model3)
```

```
$`call`
BinAddHaz(formula = dis.bin ~ disease:factor(age), data = disabData, weights = wgt,
          attrib = TRUE, attrib.var = age, collapse.background = FALSE,
          attrib.disease = TRUE, type.attrib = "abs")

$bootstrap
[1] FALSE

$coefficients
                            Estimate      StdErr    t.value       p.value
(Intercept)              0.29425017 0.009339432 31.5062168 1.991333e-202
diseasediab:factor(age)0 0.07487954 0.022708458  3.2974294  9.811601e-04
diseasearth:factor(age)0 0.01218173 0.020156904  0.6043454  5.456358e-01
diseasestro:factor(age)0 0.44896271 0.072553106  6.1880563  6.474653e-10
diseasediab:factor(age)1 0.83733434 0.128901534  6.4959223  8.884711e-11
diseasearth:factor(age)1 1.32865873 0.143790133  9.2402636  3.299325e-20
diseasestro:factor(age)1 1.60530144 0.373531351  4.2976351  1.752351e-05

attr(,"class")
[1] "summary.binaddhazmod"
```

The `(Intercept)` represents the background, which was not modelled by age. The diseases with `factor(age)0` and `factor(age)1` represent the regression coefficients for age 0 (60-79 years) and age 1 ($\geq$80 years). The output above shows that arthritis was not significant for the reference age category (60-79years) (`diseasearth:factor(age)0`). Only the absolute contribution was requested (`type.attrib = "abs"`) and it can be assessed with:

```
model3$contribution
```

```
                           att.abs
disab.0                0.277835632
backgrnd.0             0.251005540
```

```
diseasediab:factor(age)0.0 0.012575547
diseasearth:factor(age)0.0 0.002220039
diseasestro:factor(age)0.0 0.012034506
disab.1                    0.493678649
backgrnd.1                 0.206353130
diseasediab:factor(age)1.1 0.089832332
diseasearth:factor(age)1.1 0.158378063
diseasestro:factor(age)1.1 0.039115125
```

The attribution is presented by level of the attribution variable (`attrib.var`), which in this example is age. The first five rows show the results for attribution variable level 0, which in this case represents age group 60-79 years and the last five rows represent the results for attribution variable level 1 ($\geq$80 years). The results indicate that the disability prevalence in the older women (49.4%) was much larger than in the younger age group (27.8%). While the three diseases included in the model showed a low contribution to the disability prevalence in women aged 60-79 years (<1.5%), arthritis was by far the most important disease contributing to the disability prevalence in the oldest women (15.9%).

Model 4 - Binomial model with background and diseases by age, with bootstrap CI

The same matrix of diseases defined for model 3 will be used in model 4. This model can be fitted with the function:

```
model4 <- BinAddHaz(dis.bin ~ factor(age) - 1 + disease:factor(age), data = disabData,
                    weights = wgt, attrib = TRUE, attrib.var = age, type.attrib = "abs",
                    collapse.background = FALSE, attrib.disease = TRUE, seed = 111,
                    bootstrap = TRUE, conf.level = 0.95, nbootstrap = 1000,
                    parallel = TRUE, type.parallel = "snow", ncpus = 4)
```

The `-1` in the `model.formula` is used to obtain a different parameterization than the default: here we obtain the parameter estimates for all the age groups, including the reference category. Since we want to estimate the contributions for background by age, the argument `collapse.background` is set to `FALSE`. If this argument would be set to `TRUE`, only one background, common to all age groups, would be estimated. Since the contributions of diseases should be estimated by age group, the argument `attrib.disease` was set to `TRUE`. The parallel option for the bootstrap CI was used (`parallel = TRUE`) on a Windows computer (`type.parallel = "snow"`) with 4 cores (`ncpus = 4`). The option `seed = 111` allows the user to specify a random number (in this case 111) to make the results of the bootstrapping reproducible. The summary of the model can be checked with:

```
summary(model4)

$call
BinAddHaz(formula = dis.bin ~ factor(age) - 1 + disease:factor(age), data = disabData,
          weights = wgt, attrib = TRUE, attrib.var = age, collapse.background = FALSE,
          attrib.disease = TRUE, type.attrib = "abs", seed = 111, bootstrap = TRUE,
          conf.level = 0.95, nbootstrap = 1000, parallel = TRUE,
          type.parallel = "snow", ncpus = 4)

$bootstrap
[1] TRUE

$coefficients
                           Estimate        CILow    CIHigh
factor(age)0             0.22661055  0.201218693 0.2568179
factor(age)1             0.94910725  0.784733478 1.1292937
factor(age)0:diseasediab 0.12749849  0.056516120 0.2036685
factor(age)1:diseasediab 0.24124648 -0.181208625 0.7471999
factor(age)0:diseasearth 0.06884402  0.009035558 0.1345238
factor(age)1:diseasearth 0.75879140  0.349882903 1.2234047
factor(age)0:diseasestro 0.48788899  0.255772550 0.8331633
factor(age)1:diseasestro 1.13333515  0.426477637 2.2240599

$conf.level
[1] 0.95
```

```
attr(,"class")
[1] "summary.binaddhazmod"
```

The output with the results of the model is shown for `factor(age)0`, which represents the age group of 60-79 years, and `factor(age)1`, representing the age group of ≥80 years. Diabetes was not significant for age group 1, as the bootstrap CI includes the null value, i.e. 0. To identify the most disabling diseases, two objects (`coef.age0` and `coef.age1`) with the coefficients for each age group can be created and sorted in decreasing order using:

```
coef.age0 <- model4$coefficients[seq(1, length(model4$coefficients), 2)]
coef.age1 <- model4$coefficients[seq(0, length(model4$coefficients), 2)]

sort(coef.age0, decreasing = TRUE)
factor(age)0:diseasestro             factor(age)0 factor(age)0:diseasediab
             0.48788899               0.22661055               0.12749849
factor(age)0:diseasearth
             0.06884402

sort(coef.age1, decreasing = TRUE)
factor(age)1:diseasestro             factor(age)1 factor(age)1:diseasearth
              1.1333352                0.9491072                0.7587914
factor(age)1:diseasediab
              0.2412465
```

Stroke was the most disabling disease in both age groups. Arthritis and diabetes showed the lowest disabling impact in women aged 60-79 years and ≥80 years, respectively.

Since only the absolute contribution was requested (`type.attrib = "abs"`), the results for the absolute contribution can be assessed with:

```
model4$contribution

                              att.abs      CILow     CIHigh
disab.0                    0.24442949 0.24102940 0.24813627
backgrnd.0                 0.19743946 0.19694283 0.19790210
factor(age)0:diseasediab.0 0.02141352 0.01956557 0.02342391
factor(age)0:diseasearth.0 0.01253887 0.01153068 0.01363192
factor(age)0:diseasestro.0 0.01303764 0.01003208 0.01636211
disab.1                    0.69484947 0.68537515 0.70574268
backgrnd.1                 0.54868976 0.53993448 0.55643174
factor(age)1:diseasediab.1 0.02637877 0.02080951 0.03249654
factor(age)1:diseasearth.1 0.09137348 0.07746954 0.10759760
factor(age)1:diseasestro.1 0.02840746 0.01932950 0.03894183
```

The CILow and CIHigh refers to the 2.5th and 97.5th percentiles of the 1,000 bootstrap replicates, since the bootstrap CI was requested (`bootstrap = TRUE`) with `conf.level = 0.95`. To identify the main contributors to the disability burden, two objects (one for each age group) can be defined with the absolute contribution and bootstrap CI using:

```
cont.age0 <- model4$contribution[c(1:5), ]
cont.age1 <- model4$contribution[c(6:10), ]
cont.age0[order(cont.age0[, 1], decreasing = TRUE), ]

                              att.abs      CILow     CIHigh
disab.0                    0.24442949 0.24102940 0.24813627
backgrnd.0                 0.19743946 0.19694283 0.19790210
factor(age)0:diseasediab.0 0.02141352 0.01956557 0.02342391
factor(age)0:diseasestro.0 0.01303764 0.01003208 0.01636211
factor(age)0:diseasearth.0 0.01253887 0.01153068 0.01363192

cont.age1[order(cont.age1[, 1], decreasing = TRUE), ]

                              att.abs      CILow     CIHigh
disab.1                    0.69484947 0.68537515 0.70574268
backgrnd.1                 0.54868976 0.53993448 0.55643174
factor(age)1:diseasearth.1 0.09137348 0.07746954 0.10759760
```

```
factor(age)1:diseasestro.1 0.02840746 0.01932950 0.03894183
factor(age)1:diseasediab.1 0.02637877 0.02080951 0.03249654
```

According to the results, the disability prevalence in the oldest women (69.5%) was 2.8 times larger than in women aged 60-79 years (24.4%). The background was the main contributor to the disability prevalence in both age groups. Among the chronic conditions, diabetes was the main contributor to the disability prevalence in women aged 60-79 years (2.1%) while arthritis contributed most to the disability burden in older women (9.1%).

Model 5 - Binomial model with two-way interaction between diseases

In model 5, the independence assumption (assumption vii) is violated and two-way interactions between diseases are included in the model. In total, 6 parameters and the intercept will be estimated in model 5. The model can be fitted by:

```
model5 <- BinAddHaz(dis.bin ~ (diab + arth + stro)^2, data = disabData, weights = wgt,
                    attrib = TRUE, collapse.background = FALSE, attrib.disease = FALSE,
                    type.attrib = "both")

summary(model5)

$`call`
BinAddHaz(formula = dis.bin ~ (diab + arth + stro)^2, data = disabData, weights = wgt,
          attrib = TRUE, collapse.background = FALSE, attrib.disease = FALSE,
          type.attrib = "both")

$bootstrap
[1] FALSE

$coefficients
              Estimate      StdErr    t.value       p.value
(Intercept)  0.2988163 0.009586541 31.1703950 1.803828e-198
diab         0.1178815 0.026104065  4.5158287  6.422107e-06
arth         0.1101293 0.023712731  4.6443118  3.481543e-06
stro         0.8145107 0.116867992  6.9694938  3.505379e-12
diab:arth    0.1563155 0.067097034  2.3296932  1.985387e-02
diab:stro   -0.5072111 0.154344770 -3.2862213  1.020980e-03
arth:stro   -0.1494752 0.176853232 -0.8451937  3.980349e-01

attr(,"class")
[1] "summary.binaddhazmod"
```

The main effects of all the diseases were significant, but only the interaction between diabetes and arthritis and between diabetes and stroke were significant. The negative disabling impact of the interaction term between diabetes and stroke should be carefully interpreted, as it is based on a small sample size ($n = 91$) (Table 3).

Likelihood ratio test for model selection

To illustrate the use of the function `LRTest` to perform the likelihood ratio test (LRT) for model selection, models 2 (`model2`) and 4 (`model4`) are compared. The LRT can be performed with:

```
LRTest(model4, model2)

Likelihood ratio test
Model 1:
dis.bin ~ factor(age) - 1 + disease:factor(age)
Model 2:
dis.bin ~ factor(age) - 1 + diab + arth + stro
  Res.df  Res.Dev df Deviance   Pr(>Chi)
1   6286 6916.818
2   6289 6946.224  3   29.405 1.8407e-06
```

The output shows the models that are being compared: `Model 1` is the model with the interactions with diseases and age (previous model 4) and `Model 2` is the model without the interactions between

diseases and age (previous model 2). The degrees of freedom for each model (`Res.df`), the residual deviance, i.e. the 2*log-likelihood of each model (`Res.Dev`), the difference in the degrees of freedom between the models (`df`), the difference between the 2*log-likelihood of the models, i.e. the value of the likelihood ratio test statistic (`Deviance`), and the p-value of the test statistic, based on the χ^2 distribution (`Pr(>Chi)`) are presented. Since the test was statistically significant at 0.05 significance level, model 4, which includes interaction between diseases and age, fits the data better than model 2.

Examples with multinomial outcomes

To fit the multinomial additive hazard model and to estimate the contribution of chronic conditions to the disability burden for multinomial outcomes, the function `MultAddHaz` can be used. As an illustration, two models were fitted: model 6 - with only background by age; and model 7 - with background and diseases by age, with bootstrap CI.

Model 6 - Multinomial model with only background by age

Model 6 can be fitted with the function:

```
model6 <- MultAddHaz(dis.mult ~ factor(age) - 1 + diab + arth + stro, data = disabData,
                     weights = wgt, attrib = TRUE, seed = 111, collapse.background = FALSE,
                     attrib.disease = FALSE,  type.attrib = "both")
```

The results of the model can be visualized using:

```
summary(model6)

$`call`
MultAddHaz(formula = dis.mult ~ factor(age) - 1 + diab + arth +
           stro, data = disabData, weights = wgt, attrib = TRUE, seed = 111,
           collapse.background = FALSE, attrib.disease = FALSE, type.attrib = "both")

$bootstrap
[1] FALSE

$coefficients
                   Estimate      StdErr    t.value      p.value
factor(age)0.y1 0.117959944 0.006083174 19.3911842 2.109290e-81
factor(age)1.y1 0.285541582 0.022330639 12.7869869 5.585601e-37
diab.y1         0.002717701 0.011329960  0.2398685 8.104400e-01
arth.y1         0.015602747 0.011534798  1.3526675 1.762105e-01
stro.y1         0.024923984 0.025498946  0.9774515 3.283833e-01
factor(age)0.y2 0.107643802 0.005969496 18.0323096 6.503330e-71
factor(age)1.y2 0.817043178 0.043161129 18.9300697 9.103853e-78
diab.y2         0.124326766 0.017245323  7.2093036 6.283854e-13
arth.y2         0.063767963 0.014095689  4.5239336 6.181719e-06
stro.y2         0.499262854 0.064799754  7.7047030 1.514581e-14

attr(,"class")
[1] "summary.multaddhazmod"
```

In the above output, the results identified with $y1$ refer to the outcome (dis.mult) = 1 (mild disability) and the results with $y2$ refer to the outcome (dis.mult) = 2 (severe disability). For mild disability only the background (`factor(age)0.y1`) and (`factor(age)1.y1`) was significant while all the diseases and the background were signifcant for women with severe disability. Similar to the binomial model, the most disabling diseases can be identified by:

```
coef.mild <- model6$coefficients[1:5, ]
coef.sev <- model6$coefficients[6:10, ]

sort(coef.mild, decreasing = TRUE)
factor(age)1.y1 factor(age)0.y1         stro.y1         arth.y1         diab.y1
    0.285541582     0.117959944     0.024923984     0.015602747     0.002717701

sort(coef.sev, decreasing = TRUE)
```

```
factor(age)1.y2         stro.y2         diab.y2 factor(age)0.y2         arth.y2
     0.81704318      0.49926285      0.12432677      0.10764380      0.06376796
```

Background and stroke showed the highest disabling impact for mild and severe disability. The relative and absolute contributions can be checked with:

```
model6$contribution

$`att.rel`
                    att.rel
factor(age)0.y1 0.804099194
factor(age)1.y1 0.164283693
diab.y1         0.003665546
arth.y1         0.023317949
stro.y1         0.004633619
factor(age)0.y2 0.426874738
factor(age)1.y2 0.336655536
diab.y2         0.105566151
arth.y2         0.058984332
stro.y2         0.071919244

$att.abs
                     att.abs
disab.y1        0.1265087428
factor(age)0.y1 0.1017255781
factor(age)1.y1 0.0207833234
diab.y1         0.0004637236
arth.y1         0.0029499244
stro.y1         0.0005861933
disab.y2        0.1901766667
factor(age)0.y2 0.0811816147
factor(age)1.y2 0.0640240276
diab.y2         0.0200762187
arth.y2         0.0112174436
stro.y2         0.0136773621
```

It is interesting to note that the severe disability prevalence (19.0%) was 1.5 times higher than the mild disability prevalence (12.7%). The results for the relative contribution can be sorted in decreasing order by:

```
rel.cont.mild <- model6$contribution$att.rel[1:5, ]
rel.cont.sev <- model6$contribution$att.rel[6:10, ]

sort(rel.cont.mild, decreasing = TRUE)
factor(age)0.y1 factor(age)1.y1         arth.y1         stro.y1         diab.y1
    0.804099194     0.164283693     0.023317949     0.004633619     0.003665546

sort(rel.cont.sev, decreasing = TRUE)
factor(age)0.y2 factor(age)1.y2         diab.y2         stro.y2         arth.y2
     0.42687474      0.33665554      0.10556615      0.07191924      0.05898433
```

The background was the main contributor to the disability burden, representing 96.8% (0.80 + 0.16) and 76.4% (0.43 + 0.34) of the mild and severe disability prevalence, respectively. Arthritis (2.3%) was the main contributor to the mild disability prevalence while diabetes (10.6%) contributed most to the severe disability prevalence.

Model 7 - Multinomial model with background and diseases by age, with bootstrap CI

The matrix with the diseases (`disease`) defined for model 3 is used to fit model 7:

```
model7 <- MultAddHaz(dis.mult ~ factor(age) -1 + disease:factor(age),
                     data = disabData, weights = wgt, attrib = TRUE, attrib.var = age,
                     seed = 111, collapse.background = FALSE, attrib.disease = TRUE,
                     type.attrib = "both", bootstrap = TRUE, conf.level = 0.95,
                     nbootstrap = 1000, parallel = TRUE, type.parallel = "snow",
                     ncpus = 4)
```

The -1 was added to the model.formula to obtain the parameter estimates for the background for all age groups, including the reference category. Since the background should be estimated by age, collapse.background is set to FALSE. Additionally, attrib.disease is set to TRUE, as interactions between age and diseases were included in the model and the contribution of diseases should be estimated by age. The seed argument in MultAddHaz is used to obtain reproducible results for the starting values used in the constrained optimization, which are randomly generated, and for the bootstrap CI. Besides the summary function, the disabling impacts and the bootstrap CI can also be assessed with:

```
cbind(model7$coefficients, model7$ci)
```

```
                            Coefficients       CILow     CIHigh
factor(age)0.y1              0.117255379  0.10064630 0.13689281
factor(age)1.y1              0.277818866  0.20558349 0.37304023
factor(age)0:diseasediab.y1  0.005926107 -0.03440926 0.04770883
factor(age)1:diseasediab.y1 -0.027900849 -0.16964726 0.15898841
factor(age)0:diseasearth.y1  0.012814735 -0.02467113 0.05156243
factor(age)1:diseasearth.y1  0.110733103 -0.08447433 0.30975351
factor(age)0:diseasestro.y1  0.032163873 -0.03657685 0.14572706
factor(age)1:diseasestro.y1 -0.028504716 -0.22807053 0.22952796
factor(age)0.y2              0.109165655  0.09146724 0.13167070
factor(age)1.y2              0.672941316  0.53792263 0.83185332
factor(age)0:diseasediab.y2  0.121508443  0.06618176 0.17864913
factor(age)1:diseasediab.y2  0.282986455 -0.09742139 0.80712949
factor(age)0:diseasearth.y2  0.054335292  0.01095538 0.09957643
factor(age)1:diseasearth.y2  0.635463641  0.31671319 1.05424237
factor(age)0:diseasestro.y2  0.456594023  0.24959719 0.72863913
factor(age)1:diseasestro.y2  1.233578243  0.49988818 2.27216717
```

In the output, factor(age)0 and factor(age)1 refers to the age groups 60-79 years and $\geq$ 80 years, respectively. y1 refers to disability category 1, which here represents mild disability and y2 refers to disability category 2, representing severe disability.

Two coefficients (for diabetes and stroke in women aged $\geq$ 80 years with mild disability) were negative. This suggests a “protective" effect of these conditions. However, these results should be carefully interpreted as they were not statistically significant.

To identify the most disabling diseases for mild and severe disability by age group, the following code can be used:

```
mild.age0 <- model7$coefficients[seq(1, length(model7$coefficients), 2), ][1:4]
sev.age0 <- model7$coefficients[seq(1, length(model7$coefficients), 2), ][5:8]
mild.age1 <- model7$coefficients[seq(0, length(model7$coefficients), 2), ][1:4]
sev.age1 <- model7$coefficients[seq(0, length(model7$coefficients), 2), ][5:8]

mild.age0[order(mild.age0, decreasing = TRUE)]
            factor(age)0.y1 factor(age)0:diseasestro.y1
                0.117255379                 0.032163873
factor(age)0:diseasearth.y1 factor(age)0:diseasediab.y1
                0.012814735                 0.005926107

sev.age0[order(sev.age0, decreasing = TRUE)]
factor(age)0:diseasestro.y2 factor(age)0:diseasediab.y2
                 0.45659402                  0.12150844
            factor(age)0.y2 factor(age)0:diseasearth.y2
                 0.10916565                  0.05433529

mild.age1[order(mild.age1, decreasing = TRUE)]
            factor(age)1.y1 factor(age)1:diseasearth.y1
                 0.27781887                  0.11073310
factor(age)1:diseasediab.y1 factor(age)1:diseasestro.y1
                -0.02790085                 -0.02850472

sev.age1[order(sev.age1, decreasing = TRUE)]
factor(age)1:diseasestro.y2             factor(age)1.y2
                  1.2335782                   0.6729413
```

```
factor(age)1:diseasearth.y2 factor(age)1:diseasediab.y2
                  0.6354636                   0.2829865
```

Stroke was the most disabling disease in women with severe disability in both age groups and in women aged 60-79 years with mild disability while arthritis was the most disabling disease in women aged $\geq$ 80 years with mild disability.

The main contributors to the disability burden, based on the absolute contribution can be assessed with:

```
cont.mild.age0 <- model7$contribution$att.abs[1:5, ]
cont.sev.age0 <- model7$contribution$att.abs[6:10, ]
cont.mild.age1 <- model7$contribution$att.abs[11:15, ]
cont.sev.age1 <- model7$contribution$att.abs[16:20, ]

cont.mild.age0[order(cont.mild.age0[, 1], decreasing = TRUE), ]
                                  att.abs        CILow      CIHigh
disab.0.y1                   0.1146399721 0.1142911352 0.115027470
backgrnd.0.y1                0.1103973843 0.1103736506 0.110418787
factor(age)0:diseasearth.0.y1 0.0024598331 0.0022352202 0.002706063
factor(age)0:diseasediab.0.y1 0.0010238914 0.0009230407 0.001137036
factor(age)0:diseasestro.0.y1 0.0007588633 0.0005256459 0.001035586

cont.sev.age0[order(cont.sev.age0[, 1], decreasing = TRUE), ]
                                   att.abs       CILow     CIHigh
disab.0.y2                   0.137612800 0.133986991 0.14160126
backgrnd.0.y2                0.095139459 0.094884399 0.09537052
factor(age)0:diseasediab.0.y2 0.020450103 0.018538859 0.02259582
factor(age)0:diseasestro.0.y2 0.012179369 0.009234356 0.01571648
factor(age)0:diseasearth.0.y2 0.009843869 0.008984070 0.01077563

cont.mild.age1[order(cont.mild.age1[, 1], decreasing = TRUE), ]
                                    att.abs        CILow       CIHigh
disab.1.y1                    0.2532173709  0.248985442  0.257917633
backgrnd.1.y1                 0.2408954310  0.240163632  0.241550122
factor(age)1:diseasearth.1.y1  0.0170621273  0.012596001  0.022104725
factor(age)1:diseasestro.1.y1 -0.0006391061 -0.001067613 -0.000299713
factor(age)1:diseasediab.1.y1 -0.0041010813 -0.005510841 -0.002757878

cont.sev.age1[order(cont.sev.age1[, 1], decreasing = TRUE), ]
                                  att.abs      CILow     CIHigh
disab.1.y2                   0.52797382 0.51474229 0.54101616
backgrnd.1.y2                0.38747404 0.38073529 0.39414511
factor(age)1:diseasearth.1.y2 0.07581147 0.06237721 0.09017923
factor(age)1:diseasestro.1.y2 0.03293044 0.02088364 0.04734430
factor(age)1:diseasediab.1.y2 0.03175788 0.02394332 0.04002170
```

The severe disability prevalence (60-79 years = 13.8%; $\geq$80 years = 52.8%) was larger than the mild disability prevalence (60-79 years = 11.5%; $\geq$80 years = 25.3%) in both age groups. Arthritis was the main contributor to the mild disability prevalence in both age groups and to the severe disability prevalence in women aged $\geq$80 years, while diabetes was the main contributor to the severe disability prevalence in women aged 60-79 years.

Discussion

In this paper we introduced the R package **addhaz** developed to fit the binomial and multinomial additive hazard models to estimate the contribution of diseases to the disability prevalence using cross-sectional data.

The R package **addhaz** was developed based on the R functions developed by Nusselder and Looman (Nusselder and Looman, 2010) for binomial disability outcomes and for non-R users. The main advantages of **addhaz** compared to the original R functions are: (i) option to use the attribution method for multinomial responses using the function `MultAddHaz`; and (ii) option to do parallel computing for the calculation of the bootstrap percentile confidence intervals. However, the possibility to use reduced rank regression (Yee, 2014) to estimate the cause-specific disability rates by age group,

for example, which is available in the original R functions (Nusselder and Looman, 2010), is not available in **addhaz**. Nonetheless, in **addhaz** these interactions can be estimated by including full interaction terms between chronic conditions and age groups.

Although the parameter estimates of the binomial additive hazard model can also be obtained with the R package **logbin**, the contribution of diseases to the disability prevalence is not provided, since **logbin** was not developed with focus on the attribution method. Therefore, for analysis aimed at the attribution of disability to diseases, we recommend the use of **addhaz**. For multinomial outcomes, no other software is available to fit the multinomial additive hazard model and to calculate the contributions, to our knowledge.

One could argue that instead of using the multinomial model, two binomial models could be fitted: (i) no x mild disability; and (ii) no x severe disability. Although this is indeed possible, with a minor loss of precision (larger standard errors) for the parameter estimates when the reference category ("no disability", in our example) is the most frequent category in the population (which is the case for the subset of the BNHS used here, as 67% were not disabled, 13% reported mild disability, and 20% were severely disabled) (Agresti, 2002), the prevalence of the various disability categories do not sum to 100%, as can be observed in a previous study that assessed the difference in the mild and severe disability burden using two binomial models (Yokota et al., 2015a).

Different models with different options were presented using **addhaz**, showing a wide possibility of application to the users. One example is the investigation of the role of multimorbidity on the disability prevalence, which was assessed by the inclusion of two-way interactions between diseases in the model, as presented in model 2. Even though the examples only included the combination of two diseases, higher order interactions can also be included in the models, with the sample size being the limiting factor. In addition, since the prevalence of chronic conditions tends to increase over age, the model parameterization to include interactions between diseases and age groups was also shown for the binary (models 3 and 4) and multinomial disability outcomes (model 7). Although age group was used as the stratification variable to estimate the disabling impacts of the diseases and background, other variables can be used, such as education attainment and sex.

Furthermore, we illustrated how the likelihood ratio test (LRT) can be performed for model selection using the function `LRTest`. The LRT can be also performed for model selection with the multinomial additive hazard model.

The attribution method has some limitations that should be considered. The main limitation of the method is the causality assumption. Although a causal relationship between diseases and disability is plausible and has been proposed in several disability models (Verbrugge and Jette, 1994), it cannot be assessed with cross-sectional data. As a consequence, disability is incorrectly attributed to diseases when disability occurred before the diseases. Although the parallel option reduces significantly computation time for calculating bootstrap confidence intervals, fitting the multinomial model to high dimensional data can still be time-consuming. For example, the computational time to fit model 7, in a Windows computer Intel(R) Core (TM) i7-4600 CPU with 2.1GHz and 2.7GHz, 8GB (RAM), using the parallel option with 4 cores, was 23.04 hours.

Summary

In conclusion, **addhaz** is a publicly available tool to assess the contribution of chronic conditions to the disability prevalence, using cross-sectional data. The results produced by the tool can be used by policymakers to reduce the disability burden. Future areas of interest to improve the package include the extension of the multinomial model to ordinal responses and alternatives to reduce computation time for high dimensional data.

Bibliography

A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, 2002. ISBN 0-471-36093-7. [p92]

J. R. Beard, A. Officer, I. A. de Carvalho, R. Sadana, A. M. Pot, J.-P. Michel, P. Lloyd-Sherlock, J. E. Epping-Jordan, G. G. Peeters, W. R. Mahanani, et al. The world report on ageing and health: A policy framework for healthy ageing. *The Lancet*, 387(10033):2145–2154, 2016. URL http://dx.doi.org/10.1016/S0140-6736(15)00516-4. [p75]

L. Blizzard and D. Hosmer. The log multinomial regression model for nominal outcomes with more than two attributes. *Biometrical Journal*, pages 889–902, 2007. URL https://doi.org/10.1002/bimj.200610377. [p79]

A. Canty and B. Ripley. **boot**: *Bootstrap R (S-PLUS) Functions*, 2016. URL https://CRAN.R-project.org/package=boot. R package version 1.3-18. [p79]

H. Chen, H. Wang, E. M. Crimmins, G. Chen, C. Huang, and X. Zheng. The contributions of diseases to disability burden among the elderly population in china. *Journal of Aging and Health*, pages 261–82, 2013. URL https://doi.org/10.1177/0898264313514442. [p75]

C. Chiang. Competing risks in mortality analysis. *Annual Review of Public Health*, pages 281–307, 1991. URL https://doi.org/10.1146/annurev.pu.12.050191.001433. [p77]

A. Davison and D. Hinkley. *Bootstrap Methods and Their Applications*. Cambridge University Press, 1997. ISBN 0-521-57391-2. [p79]

M. W. Donoghoe. **logbin**: *Relative Risk Regression Using the Log-Binomial Model*, 2016. URL https://CRAN.R-project.org/package=logbin. R package version 2.0.1. [p79]

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC Press, 1st edition, 1994. ISBN 0412042312. [p76, 78, 79]

B. Klijs, W. J. Nusselder, C. W. Looman, and J. P. Mackenbach. Contribution of chronic disease to the burden of disability. *PLoS One*, 6(9):e25325, 2011. URL https://doi.org/10.1371/journal.pone.0025325. [p75]

K. Lange. *Numerical Analysis for Statisticians*. Springer-Verlag, 2nd edition, 2010. ISBN 978-1-4419-5944-7. [p78]

K. Manton and E. Stallard. *Recent Trends in Mortality Analysis*. Academic Press, 1984. ISBN 0124700209. [p77]

I. Marschner and A. Gillett. Relative risk regression: Reliable and flexible methods for log-binomial models. *Biostatistics*, pages 179–192, 2012. URL https://doi.org/10.1093/biostatistics/kxr030. [p79]

W. Nusselder and C. Looman. WP7: Decomposition tools - Technical report on attribution tool. Technical report, 2010. URL http://www.eurohex.eu/pdf/Reports{_}2010/2010TR7.2{_}TRonattributiontool.pdf. [p75, 76, 77, 79, 82, 91, 92]

W. J. Nusselder and C. W. Looman. Decomposition of differences in health expectancy by cause. *Demography*, 41(2):315–334, 2004. URL https://doi.org/10.1353/dem.2004.00. [p75, 76, 77, 79]

W. J. Nusselder, C. W. Looman, J. P. Mackenbach, M. Huisman, H. Van Oyen, P. Deboosere, S. Gadeyne, and A. E. Kunst. The contribution of specific diseases to educational disparities in disability-free life expectancy. *American Journal of Public Health*, 95(11):2035–2041, 2005. URL http://doi.org/10.2105/AJPH.2004.054700. [p75]

SAS Institute Inc. The sas system, version 9.2, 2008. URL http://www.sas.com/. [p75]

R. Strobl, M. Müller, R. Emeny, A. Peters, and E. Grill. Distribution and determinants of functioning and disability in aged adults-results from the german kora-age study. *BMC Public Health*, 13(1):1, 2013. URL http://doi.org/10.1186/1471-2458-13-137. [p75]

C. Szwarcwald, D. Malta, C. Pereira, M. Vieira, W. Conde, P. Souza Júnior, G. Damacena, L. Azevedo, G. Silva, M. Theme Filha, C. Lopes, D. Romero, W. Almeida, and C. Monteiro. Pesquisa nacional de saúde no brasil: Concepção e metodologia de aplicação. *Ciência & Saúde Coletiva*, 19(2):333–342, 2014. URL http://dx.doi.org/10.1590/1413-81232014192.14072012. [p80]

L. Verbrugge and A. Jette. The disablement process. *Social Science & Medicine*, pages 1–14, 1994. URL http://dx.doi.org/10.1016/0277-9536(94)90294-1. [p92]

M. Yang, X. Ding, and B. Dong. The measurement of disability in the elderly: A systematic review of self-reported questionnaires. *Journal of the American Medical Directors Association*, 15(2):150–e1, 2014. URL https://doi.org/10.1016/j.jamda.2013.10.004. [p75]

T. Yee. Reduced-rank vector generalized linear models with two linear predictors. *Computational Statistics & Data Analysis*, pages 889–902, 2014. URL https://doi.org/10.1016/j.csda.2013.01.012. [p79, 91]

T. W. Yee. **VGAM**: *Vector Generalized Linear and Additive Models*, 2016. URL https://CRAN.R-project.org/package=VGAM. R package version 1.0-2. [p79]

R. Yokota, J. Van der Heyden, S. Demarest, J. Tafforeau, W. Nusselder, P. Deboosere, and H. Van Oyen. Contribution of chronic diseases to mild and severe disability burden in Belgium. *Archives of Public Health*, page 37, 2015a. URL http://doi.org/10.1186/s13690-015-0083-y. [p92]

R. T. C. Yokota, N. Berger, W. J. Nusselder, J.-M. Robine, J. Tafforeau, P. Deboosere, and H. Van Oyen. Contribution of chronic diseases to the disability burden in a population 15 years and older, Belgium, 1997-2008. *BMC Public Health*, 15(1):1, 2015b. URL https://doi.org/10.1186/s12889-015-1574-z. [p75]

R. T. C. Yokota, L. de Moura, S. S. C. de Araújo Andrade, N. N. B. de Sá, W. J. Nusselder, and H. Van Oyen. Contribution of chronic conditions to gender disparities in disability in the older population in Brazil, 2013. *International Journal of Public Health*, 61(9):1003–1012, 2016. URL https://doi.org/10.1007/s00038-016-0843-7. [p75, 80]

R. T. C. Yokota, H. Van Oyen, C. W. N. Looman, W. J. Nusselder, M. Otava, Y. W. Kifle, and G. Molenberghs. Multinomial additive hazard model to assess the disability burden using cross-sectional data. *Biometrical J.*, 2017. URL https://doi.org/10.1002/bimj.201600157. [p75]

Renata Tiene de Carvalho Yokota
Epidemiology and Public Health, Sciensano
Interface Demography, Vrije Universiteit Brussel
Belgium
Renata.Yokota@sciensano.be

Caspar W. N. Looman
Department of Public Health, Erasmus Medical Center
Netherlands
c.looman@erasmusmc.nl

Wilma Johanna Nusselder
Department of Public Health, Erasmus Medical Center
Netherlands
w.nusselder@erasmusmc.nl

Herman Van Oyen
Epidemiology and Public Health, Sciensano
Department of Public Health, Universiteit Gent
Belgium
Herman.VanOyen@sciensano.be

Geert Molenberghs
Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Universiteit Hasselt and Katholieke Universiteit Leuven
Belgium
geert.molenberghs@uhasselt.be