EUROPEAN HEALTH & LIFE EXPECTANCY
EHLEIS
INFORMATION SYSTEM

June 2013

Method to estimate disability prevalence from individual level survey data on disability and diseases: a practical guide

Erasmus MC

Start date of project: 20 April 2012 Duration: 36 months

Table of contents

Abstract ... 3
Acknowledgements ... 4
Introduction ... 5
Part 1: The attribution method ... 7
1. Background of the attribution method ... 7
2. Rationale of the attribution method ... 8
3. General principle of the attribution method ... 8
4. Additive rate regression model as core of attribution method ... 10
4.1 Variations of the disabling impact by age ... 12
5. Calculation of number of disabled by cause across subgroups ... 13
7. Discussion of the decomposition method ... 13
6. Discussion of the attribution method ... 14
Part 2: The attribution software tool ... 16
1. Introduction of the tool ... 16
2. Features of the tool ... 16
2.1 General option 1: Variations across age using Reduced Rank Regression ... 17
2.2 General option 2: Comparison of two populations ... 19
2.3 General option 3: Different models and p-values for model selection ... 20
2.4 Options linked to further application in the decomposition tool ... 21
3. How to use the tool ... 22
3.1 Input preparations and running the model ... 22
3.1.1 Input dataset ... 23
3.1.2 The Excel input file ... 24
3.1.3 Running R ... 25
3.2 Output of the model ... 27
4. Selecting the best model ... 29
4.1 Situation of one population (no population distinguished) ... 29
4.2 Situation of two populations (populations distinguished) ... 29
Part 3: Example ... 33
1. Input ... 33
1.1 Input dataset (ASCII or POR-format) ... 33
1.2 Input specification file (.csv) ... 34
1.3 R syntax file (.R) ... 36
2. Output ... 37
3. Illustration of features of the tool ... 41
3.1 Illustration of model selection ... 41
3.1.1 No distinction between populations ... 41
3.1.2 Distinction between populations ... 45
3.2 Illustrations of additional options of the tool ... 49
3.2.1. Observed vs fitted prevalences ... 49
3.2.2 Machine readilbe output file ... 50
3.2.3 Including data on population in institutions ... 51
3.3 Illustration of model with disease-interactions ... 52
Appendices ... 54
Appendix 1: Analogy of attribution method to crude probabilities of death ... 54
Appendix 2: Numerical examples ... 56
2.2 Numerical example with additive regression ... 56
2.2 Numerical example with additive regression ... 58
2.2.1 Numerical simplified example of attribution of disability to 2 diseases ... 58
Appendix 3: Log-likelihood ratio test ... 61
Appendix 4: Logical expression in selection ... 62
Appendix 5-a: Input specification files and output used in the illustration: no population distinguished ... 63
Appendix 5-b: Input specification files and output used in the illustration: population distinguished ... 64
References ... 66

Abstract

This report focuses on the attribution tool that is developed to estimate disability prevalence by disease from individual-level survey data on disability and diseases [1]. This can be used to obtain disability prevalence by disease data which serves as input for the decomposition tool or for independent analyses on the contribution of specific diseases to disability. The first part of the report is devoted to the attribution method. The second part focuses on the tool, it describes its main features, explains how to use the tool. The third part provides some illustrations. While the first part gives the background of the tool, it is possible to read the more applied second and third part of this document first.

This technical report is an updated version of the technical report of July 2011. The main changes as compared to the prior version are: 1) option to calculate confidence intervals based on bootstrapping and 2) adjusted optimalization so that the tool will still work when negative probabilities of disability are encountered. The extended tool is available on request from the authors (w.nusselder@erasmusmc.nl).

When using the tool, please site:

Nusselder, W. J., C. W. Looman, et al. (2005). "The contribution of specific diseases to educational disparities in disability-free life expectancy." Am J Public Health **95**(11): 2035-2041.

Acknowledgements

This document is delivered as part WP5 of the JA EHLEIS. More information on JA-EHLEIS can be found on the website (http://www.eurohex.eu/index.php?option=welcome). The current work in the JA-EHLEIS builds upon the work in the EHEMU and EHLEIS project. In the EHEMU project (2004-2007) the attribution tool to derive disability data by disease was developed. In the EHEMU project (2007-2010) this tool was developed. In the current JA EHLEIS (2011-2014) the possibility to derive confidence intervals was added and optimalization routine was improved.

This report was prepared by Wilma Nusselder and Caspar Looman. It reflects only its authors' views; the European Commission and those who provided information are therefore not liable for any use that may be made. We thank users of the prior version of the tool for their feedback. We thank Herman van Oyen en Renata Yokota for their comments on an earlier version of this document.

Introduction

Diseases play a major role in the disablement process[2]. The contribution of diseases to disability depends both on the prevalence of the disease and the disabling impact. Both may vary over time and between (sub)populations. To facilitate understanding of disability trends or disparities, it is important to obtain insight in the (varying) contribution of chronic diseases to disability, and in the role of differences in disease prevalence and/or disabling impacts. This report focuses on the attribution tool that is developed to estimate the contribution of diseases to disability from individual-level survey data on disability and diseases [1]. It is an alternative to the well-known population attributable fraction (PAF) approach[3] that does not provide additive contributions. The approach can be used to obtain disability prevalence by disease data which serves as input for the decomposition tool or for independent analyses on the contribution of specific diseases to disability.

This is the second of two technical reports prepared within WP5, focussing on the attribution tool. The first report focuses on the decomposition tool. The decomposition and the attribution tools can be used in concert and independently. The tool can be used to estimate the contribution of specific diseases to disability in a single population. The attribution tool can also provide input for the decomposition of differences in health expectancy between two populations (men vs. women, member states) by cause (diseases), which requires a comparison of disability prevalence by cause between both populations. For both applications, it provides additive cause-specific disability prevalence and provides insight in the role of the prevalence of specific diseases and their disabling impact, and of disability not associated with these diseases. Both disability not associated with the diseases and the disabling impact of specific diseases may vary by age, and if two populations are compared, by age, by population or by age and population.

Application of the attribution tool requires individual level data containing information on presence or absence of disability and presence or absence of selected diseases, as well as age in classes. Apart from this requirement, the tool is flexible. It allows the user to work with different formats of input data and has the option to attribute disability to diseases in a single population, or to compare two populations. The user is free to select diseases and age categories, and sample weighs can be easily applied. The user has the option to obtain additional descriptive output. Several p-values are provided to assist model selection. The model is programmed in R, but the user does not need to have any R knowledge to run the program. All specifications for an analysis have to be entered in a csv or text file, which ensures that later the results can be reproduced easily.

The decomposition and attribution tool are each described in a separate technical report, which can be read independently. The tools are freely available from the author (w.nusselder@erasmusmc.nl).

The current document on the attribution method is organised as follows. In part 1, we explain the attribution method and in part 2 the attribution tool that is based on this method. Within part 1, we give the background, rationale and the general principle of the attribution method. Next we present the additive hazard regression model and the attribution of disability prevalence to specific diseases based on the outcomes of this regression method. This is followed by a short discussion of elements of the approach that may need more explanation. Part 2 focuses on the tool. First we describe the tool and its main features. Part 3 gives examples of its use. Persons mainly interested in using the tool, may first read Part 2 and Part 3, which are more applied, and next read Part 1.

Part 1: The attribution method

1. Background of the attribution method

Data on causes of death are generally available, for instance from the WHO mortality database (http://www.who.int/whosis/mort/download/en/index.html). Causes of death provide additional information on the main disease contributing to death for each individual, which in turn can point at underlying determinants. Hence, looking at causes of death provides useful information to explain differences in mortality between populations (e.g. EU member states), within populations (e.g. between men and women, socio-economic groups) or between different time points.

Similarly, causes of disability may yield important information to understand better differences or changes in observed disability levels, or may point at possible entry points for interventions reducing disability in the population. Additionally, disability prevalence by cause is needed for the decomposition of health expectancy by cause. In contrast with cause-of-death data, however, disability data by cause are not generally available.

In principle, there are several options to obtain data on disability by cause. Ideally one would use prospective data, including information on the time of onset of chronic diseases and disability to reconstruct causes of disability. However, longitudinal data sets of a sufficient size and representative for the population are scarce. Moreover, while longitudinal data provide evidence on the impact of specific diseases on the onset of disability, disability prevalence by cause (in a specific year) also depends on mortality, and hence information on mortality from the different diseases needs to be considered. For instance, if a disease is associated with a high chance of onset of disability, but also with a high chance to die, part of the incident cases will not be represented as prevalent cases of disabled persons later in time. Another option is to use surveys that already include self-reported information on the main condition causing disability, such as the Survey of Disability, Ageing and Carers in Australia or the National Health Interview Survey in the US. However, such survey data on causes of disability are not available in each country and rely on self-report. For the decomposition analyses, an important application of the outcomes of the attribution tool, preferably the same data source and identical disability definition is used as used to calculate health expectancies. Cross-sectional data on diseases and disability are more generally available from national health interview surveys and international surveys, such as Europan Health Interview Survey (EHIS) and SHARE and are commonly used to calculate health expectancies. Therefore, we developed a tool that allows using cross-sectional data on the presence of disability, selected diseases and age to attribute disability to diseases, and hence to estimate disability prevalence by cause.

To estimate cause-specific disability prevalence from cross-sectional data, named here “attribution” the following assumptions are made:

- First, using cross-sectional data we assume that the distribution of disability by cause at the time of the survey is explained entirely by diseases that are (still) present at the time of the survey plus the background risk.
- Second, we assume that the hazards of the diseases have been proporitionally equal during the period of incidence of disabilility. Third, we assume that all persons of a given age are exposed to the same “background” disability risk, i.e. disability risk not associated with any of the diseases.
- Fourth, we assume that causes of disability (diseases and background risk) act as independently competing causes. This assumption is necessary to map disability to single disease groups. However, interactions between causes can be easily added at cost of an extra “disease" (e.g. combination of A and B).
- Fifth, we assume the same start of the time at risk for disability from each cause.

2. Rationale of the attribution method

For each death, one disease is assigned as underlying cause of death (also labelled as “primary”) by the physician who fills in the death certificate. To do so the physician considers all the diseases that are present in the individual at the time of death and follows internationally accepted coding rules to select one underlying cause of death.

For disability data, similar information on underlying causes of disability is not available. When persons with disability report more diseases, it may be unclear which disease caused the disability. And even if a person reports only one disease, this is not necessarily the cause of the disability, since also in persons not reporting any diseases disability occurs. This may reflect that disability can occur without any disease (“old age”), that not all diseases that are present in an individual are reported in the survey, or that diseases/conditions are not present anymore but caused the disability (birth defects, permanent consequences of accidents). We label this disability that is not associated with the included diseases as “background”. An alternative term would be “unexplained”. Since “background” causes disability in persons without any disease, there are no reasons to expect that it does not cause disability in persons who report one or more diseases.

Taking into account that persons can and often do have more diseases and that disability is present in persons without any disease, we aim to attribute disability reported in a survey to either a single disease or to background. Such a “single disease” may also be one specific disease combination (diabetes and heart disease), but for transparency reasons we focus here on single diseases. We use additive regression analysis to attribute disability to diseases, which will be explained in more detail later. First we explain the general principle.

3. General principle of the attribution method

The basic idea is that disability is attributed to a disease by comparing disability in similar persons

who only differ with respect to the presence or absence of the disease. Disability in persons without any diseases is attributed to "background". The background risk, here operationalized as rate or (cumulative) hazard, is the same for all persons. Hence the *difference* in disability between two groups, one group having a certain disease and the other having no diseases, is attributed to the specific disease. We use the term rate and hazard interchangeably for the background and disease risks.

Suppose that we have two groups:
1. No disease: disability prevalence, 30 cases out of population of 100, i.e. 30%, or 0.3
2. With disease: disability prevalence, 60 cases out of population of 100, i.e. 60%, or 0.6

Our approach is similar to the standard approach for multiple causes of death in a multi-decrement life table, where the proportion of the (two) mortality rates in the total mortality rate is used to estimate the rate for each cause [4]. This is explained in table 1. The total disability hazard (rate) in the group with no disease is: -ln(1-0.3) = 0.357. The background hazard (rate) equals the disability hazard (rate) in the group with no disease, and hence is also 0.357. The total disability hazard (rate) for group 2 with the disease is -ln(1-0.6) = 0.916. We assume that this group was exposed to both the background hazard and the disease-specific hazard The disease-specific hazard is the total hazard for group 2 minus the background hazard (0.916-0.357= 0.557). Using the proportions of both hazards in the total hazard, 0.557/0.916 = 61% of the disability in group 2 is caused by the disease and 39% by background, hence disability prevalence due to background in this population is 0.234 (0.6*0.39), and due to the specific disease 0.366 (0.6*0.61). By definition, in the group without any disease 100% of the disability is labelled as background. Note that less than 30% of the group is disabled by background if the disease is present, although the hazard for background is the same in this group. Some of the persons have become disabled by the disease before they would have become disabled by background. So of the total prevalence of disability (0.6) 0.234 is attributed to background and 0.366 to the disease.

Table 1 Illustration of the attribution of disability to specific diseases

		Disability prevalence	Disability hazard (-ln(1-prevalence))	Disability prevalence in diseases population due to
Population without disease		3/100=0.3	0.357	
Population with disease		6/100=0.6	0.916	
	due to background		0.357	0.6*(1-0.61)=0.234
	due to disease		0.916-0.357=0.557	0.6*0.61=0.366
	Proportion of disability hazard in disease due to disease		0.557/0.916=0.610	

Our approach should be contrasted with the population attributable fraction (PAF), which is defined as the proportional reduction in average risk that would be obtained by eliminating a risk factor from the population while the distribution of other risk factors remains unchanged [3]. In this example, this is the fraction of disability among group 2 that would not have occurred if exposure to that disease had not occurred (i.e., (0.6-0.3)/0.6=0.5). That is, after elimination of exposure the fraction disabled would be 0.3 (i.e. 0.5*0.6) in this group. This is slightly higher than the proportion attributed to background in a situation when one other cause (the disease) is present. (0.234). In a situation of competing risks (two risks: background and disease) persons may have become disabled by the disease before they became disabled by background. Eliminating the disease leaves more people at risk for background as there is no competition between the disease and background risks. In a situation of competing risks, the disease risk and background risk compete.

Similar to the situation with two groups, let us now suppose we have four groups (1. no diseases, 2. only A, 3. only B and 4. both A and B). Each of these groups has the same background hazard (rate). If we assume that there is no interaction on an additive scale between diseases A and B, which we have to assume if we want to attribute disability to single diseases, we have to calculate only three values (risk due to A, risk due to B and risk due to background) based on these four groups. This means an over-determined problem that can be optimally solved by regression. Without interaction in the regression model it is assumed that the difference between the risks due to diseases A and B and the sum of risks due to A and due to B are only due to random error.

4. Additive hazard regression model as core of attribution method

The regression method to attribute disability to diseases is based on the multivariate additive regression model [5] and is briefly described before [1]. Clayton describes an additive hazard (rate) model as alternative to the logistic model that cannot be used to obtain additive causes. In the additive regression model of Clayton, the prevalence of a disease is modeled in the presence of two possible risk factors “A” and “B”.

This situation can be compared with a vessel where fluids from two different sources are seeping in with two different flows. The difference in flows is directly related to the fraction at the end that was due to each of the sources.

In the current application, “A” stands for cause A and “B” for cause B, “event 1” for onset of disability because of A and “event 2” because of B and “Disease” stands for disability. Causes can in our case be either diseases or background; they play the same role in the process.

The additive regression model is specified as follows:

$$\hat{y} = 1 \; - \; e^{-\eta}$$

$$\eta = \alpha_a + \sum_d \beta_d X_d$$

where $\hat{y}$ is the estimated probability that the person is disabled, e is the base of the natural logarithm and η the linear predictor. The observed disability (y) follows a binomial distribution. The linear predictor is defined as the sum of the background hazard by age (α_a) and the disease hazards (β_d, labeled as "disabling impact" or "disease effects") for the diseases (*d*) that are present in the respondent (given by the dummy variables X_d).

One of the problems of an additive regression model is that the transformation from the linear predictor (eta) to probability ($\hat{y}$) does not necessarily lead to a positive value: if eta is negative then $\hat{y}$ will also be negative and not valid as a probability. To avoid this, we used penalties in the optimalization routine.

4.1 Variations of the disabling impact by age

The disabling impact of each disease β_d may vary by age. As the full age-interaction term would require n (number of age classes) times m (number of diseases) different parameters, it may be more parsimonious to use the same age pattern for each disease. To do so, Reduced Rank Regression [6-8] can be used. One option is to reduce the rank of the interaction to one. This means that the age-specific disabling impact of each disease, β_{da}, is estimated as the product of an age pattern γ_a which varies by age, but is equal for each disease, and a disease effect δ_d, which varies by disease, but not by age. Next to a one rank solution restricting the age pattern to be the same for all diseases, it is also possible to fit second rank solutions ($\beta_{da} = \gamma_{a1}.\delta_{d1} + \gamma_{a2}.\delta_{d2}$) or even higher (although not with this tool). The full-rank solution gives the same estimates as fitting all β_{da} parameters seperately. The first term of the second rank solution $\ell_{a1}.\delta_{d1}$ equals the first rank solution, so higher ranks can be added one by one to the model. Scaled deviances can be used to test for difference in age pattern for different diseases (likelihood ratio test) and to test whether a second rank is nessecary, i.e. whether there are differences in the age patterns between the diseases.

5. Calculation of number of disabled by cause across subgroups

The additive regression model estimates the disabling impact of the disease (β_d or β_{da}). Disability prevalence by cause depends on both the prevalence of the disease (X_d) and the disabling impact of the disease (β_d or β_{da}). Analogous to using the proportional distribution of mortality rates to obtain crude probabilities of death from a single cause in the presence of competing causes [1,4,9], the background and disease-specific hazards (rates) are used to partition the disability by cause. The attribution of disease d is $\frac{\beta_d X_d}{\eta}.\hat{y}$ and of background is $\frac{\alpha_a}{\eta}.\hat{y}$. Stated differently, the part attributed to a specific disease is the fraction of the disease hazard (rate) in the total hazard (rate), being the sum of background and the disease hazards of each disease that is present. The fraction attributed to background is the fraction of the background hazard (rate) in the total hazard (rate).

Applying these formulas gives for every individual the probability of being disabled caused by background *or* disease (if present). Adding the cause-specific probabilities of an individual gives the probability of being disabled for that individual. Adding the cause-specific probabilities of all persons in the dataset, or in a specific age group, gives the total number of disabled by cause in the population, or in that age group. Dividing the number of disabled persons by cause by the total number of persons gives the proportion of disability by cause (prevalence).

The same result can be obtained by first making subgroups of persons having the same age and combination of causes (i.e. specific diseases and background). Applying the same formulas for each subgroup then yields the attributions of specific diseases and background per subgroup. Multiplied with the number of persons in each subgroup, this gives the number of persons with disability caused by each disease and background. Adding all these cause-specific numbers over all subgroups gives the total number of disabled persons by this cause. Dividing the number of disabled persons by cause by the total number of persons in this age group gives the proportion of disability by cause.

6. Calculation of confidence intervals

As with any regression analyses the parameter estimates are provided with estimated standard errors. For all regression parameters it may be assumed that they have a normal distribution, so confidence intervals can be calculated by 1.96 times the standard error. Take care that here the disease parameters (the disabling impacts) are the hazards and do not need to be exponentiated like in proportional hazard regression models. Parameters from the RRR model can be handled in the same way, but when one wants a confidence intervals around the age specific hazard $\beta_{da} = \gamma_{a1}.\delta_{d1} + \gamma_{a2}.\delta_{d2}$ we do not have a standard error for these products. In the same way there is no analytical way to get confidence intervals for the attributions. Therefore we used non-parametric bootstrapping to obtain confidence intervals [10]. In general a bootstrap can be used to find confidence

intervals around parameters that have not been optimized directly. In our case the parameters of the model are the disabling impacts and the optimization routine calculates the (co)- variance of each as a part of the optimalization. For derivatives of them, like the age-specific impacts (gamma*delta) and the attributions the bootstrap is needed. In a non-parametric bootstrap a new dataset ("replica") of the same size is constructed by sampling objects with replacement from the original dataset. This is a way to mimic what would happen if the sampling would have been repeated. Calculating the parameters of interest (age specific impacts and attributions) for this replica gives an idea what the result could have been of repeating the data gathering. This is done several times (number is specified by the user) and the distribution of the results gives a confidence interval for each of the parameters.

7. Discussion of the attribution method

A few issues may merit extra explanation or discussion.

First it is noteworthy that the additive hazard (rate) regression model gives as regression coefficients (β) directly the cumulative hazards (rates). Hence the hazard (rate) of someone having diseases A and B (and zero background risk) is $\beta_A + \beta_B$ and *not* exp($\beta_A + \beta_B$) (the latter equals: exp(β_A)*exp(β_B). The additive regression model for cross-sectional data, is to be distinguished from the multiplicative hazard regression model for cross-sectional data that is called complementary log-log [11,12], from the multiplicative or proportional hazard model for survival data, that is known as the Cox model [13], and from the additive hazard model for survival data, introduced by Aalen [14].

	Cross-sectional data	Survival data
Additive	Additive hazard model (Clayton)	Additive hazard model for survival data (Aalen)
Multiplicative (proportional)	Complementary log-log	Cox proportional hazard model

Second, the rate or cumulative hazard of a specific disease is the disabling impact of that disease. A high rate means a high likelihood that the disease has caused disability. In the situation of only background (that is, no diseases) and thus no competing risks, the background rate can be easily converted into the probability of being disabled from background (probability=1-exp(-background rate)). Note that this formula is the link function from the additive regression model. When diseases are present, there is the situation of competing risks. This requires that first the total rate for this situation (background rate + disease specific rate (s)) is calculated, which can be converted into the total probability (1-exp (-rate)). (Note that this is not the same as 1-exp(-background rate) + 1-exp(-disease rate), which is similar to the situation when transition rates of a multi-state life table are converted into probabilities taking into account the whole set of rates, e.g. by using matrices.)

Third, the approach to obtain additive cause-specific disability prevalence is similar to using the proportionality of rates in the situation of "crude probabilities of death" [4]. This is based on the assumption that if the ratio of the cause-specific force of mortality to the total force is a constant

throughout an interval, this constant should also be equal to the ratio of the corresponding probabilities over the entire interval [4]. For example, this proportionality of rates means that if the rate for A is 0.02, for B is 0.04, and for background is 0.02, then there is 50% chance that it is due to B and 25% is due to A, and 25% is due to background. So out of the 100 disabled cases with similar characteristics, we expect that 50 are due to B and 25 are due to A, and 25 are due to background. For a more detailed description of the similarities with crude probabilities of death, see Appendix 1.

Fourth, our approach based on additive regression should be distinguished from logistic regression. Given that the response variable disability is binomial, logistic regression may seem the first choice, but the logistic model does not allow dividing the total amount of disability over the diseases. While logistic regression is often used to calculate PAFs, by subtracting the proportion with the outcome (disability) after the elimination of the risk factor (disease) from the baseline disability proportion, when more risk factors are present in the same persons, logistic regression does not yield additive contributions of each of these factors. That logistic regression does not yield additive contributions of each factor if more risk factors are present, does not imply that the PAF of combined exposure cannot be obtained. Equations do exist to calculate the PAF of combined exposure [15]. However PAFs of specific causes do not add to this combined exposure. Solutions have been suggested to avoid that the effects of elimination depend on the ordering of the elimination in the presence of more exposures. For instance Eide and Geffelder [16] proposed sequential and average attributable fractions, which requires to do sequential elimination of each risk factor and to repeat this for all possible ordering of elimination. However the authors do not propose what these quantities would mean in practice.

Fifth, it is noteworthy that while we aim to attribute each disabled case to one underlying cause, which can be either background or a specific disease, all the diseases present in the individual are taken into account. If the ratio of the cause-specific hazards to the total hazard is 0.25 for disease B, this means that in a group of similar persons and with the same diseases combinations, in 1 out of 4 cases disease B is labelled as the underlying cause. Hence all prevalent diseases that have a disabling impact larger than zero do contribute to the cause-specific disability prevalence. What is not taken into account in the standard approach is the effect of an interaction between diseases. So if the hazard of the combination of two diseases is more or less than the sum of the hazards of each disease, this is not taken into account given that our purpose was to attribute disability to single diseases. However, it can be easily taken into account by including combinations of diseases as "disease".

Part 2: The attribution software tool

1. Introduction of the tool

The attribution tool allows the user to estimate disability prevalence by disease from individual-level cross-sectional datasets that include information on disability and the presence of specific chronic diseases and age**.** The tool attributes disability to diseases and thus estimates additive cause-specific disability prevalence based on the additive hazard regression model, and provides insight in the role of the prevalence of the diseases and their disabling impact and of disability not associated with these diseases. The attribution tool can be used to attribute disability to diseases in a single population, or to compare two populations. Both disability not associated with the diseases and the disabling impact of specific diseases may vary by age, and if two populations are compared, by age, by population or by age and population. Several p-values are provided to assist model selection.

The attribution tool is flexible as it allows the user to work with different formats for the dataset with individual data. The user is also free to select diseases and age categories, the impact of diseases on disability may vary by age (and population), and sample weighs can be easily applied. The attribution software is programmed in R, but all the user-specified input is communicated to the R program through an Excel-file (saved as "csv-file"), therefore the user does not need to have any R knowledge to use the program. This csv- or txt-file acts as a syntax file: running the job again at any later point in time should produce the same results.

Before we explain in more detail how the tool works, we first explain the features of the tool.

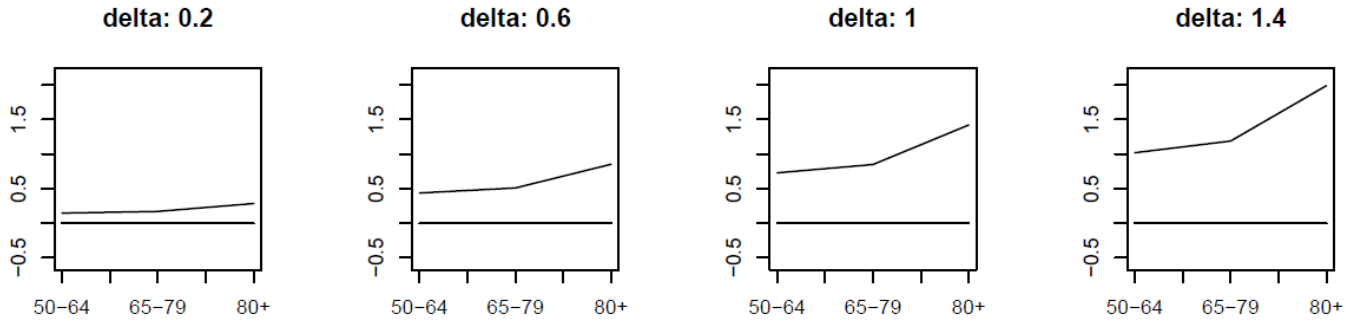
2. Features of the tool

The attribution tool has some features, which are particularly useful for the further application in decomposition tool, such as: 1) the machine readable attribution table, 2) the choice to attribute either the fitted or the observed prevalence of disability by cause, and 3) the possibility to include additional information on the institutionalised population. For more details on the background of these options, see below. More features of the tool were developed with the decomposition tool in mind, but have general applicability. First, the tool offers the possibility to vary both background disability and the disabling impact by age, whereby this disabling impact can be modelled parsimoniously based on Reduced-Rank Regression. Second, the tool offers the possibility to compare two populations in a single regression model. Third, the tool provides different models and p-values based on the likelihood ratio test to assist the user selecting the best model. To provide guidance in model selection, an approach for model selection is developed and illustrated.

2.1 General option 1: Variations across age using Reduced Rank Regression

The tool offers the possibility to vary the disabling impact by age. This option is provided to the user, as it has been shown that the disabling impact of diseases is generally higher at older ages than at younger ages [17]. To avoid the need to estimate parameters for all combinations of diseases by age class, Reduced Rank Regression (RRR) is used. Using a RRR with rank 1 means that all diseases have the same age pattern. The age-specific disability rate for a specific disease is then the product of the age pattern γ_a (gamma) which varies by age, but is equal for each disease, and a disabling impact (disease effect) δ_d (delta). The age specific disability impact β_{ad} then becomes $\gamma_a . \delta_d$. Figure 1 shows the idea of a common age pattern for all diseases (first rank solution). For δ_d we assumed: 0.2, 0.6, 1.0 en 1.4, to illustrate diseases with increasing disabling impacts.

Figure 1: RRR with one axis.



Each of the four figures shows the age specific hazards for a specific disease. The example presented in figure 1 has three broad age groups to model the disabling impact (effect modification). Only the first axis varies, the disabling impacts are proportional between diseases, implying a common age pattern for all four diseases but with different disease impacts (delta (δ)).

To check whether a common age pattern is too much a simplification, also a second rank can be added. This means that deviations from the common age patterns are modelled. While with rank 1 β_{ad}, the age and disease specific effect, is estimated by the product of γ_a and $\delta_{d,}$ with rank 2 this becomes $\gamma_{a1} * \delta_{d1} + \gamma_{a2} * \delta_{d2}$. The estimates for the parameters of rank 1 do not change when adding the second rank. Therefore γ_a and δ_d are the same as γ_{a1} and δ_{d1}.

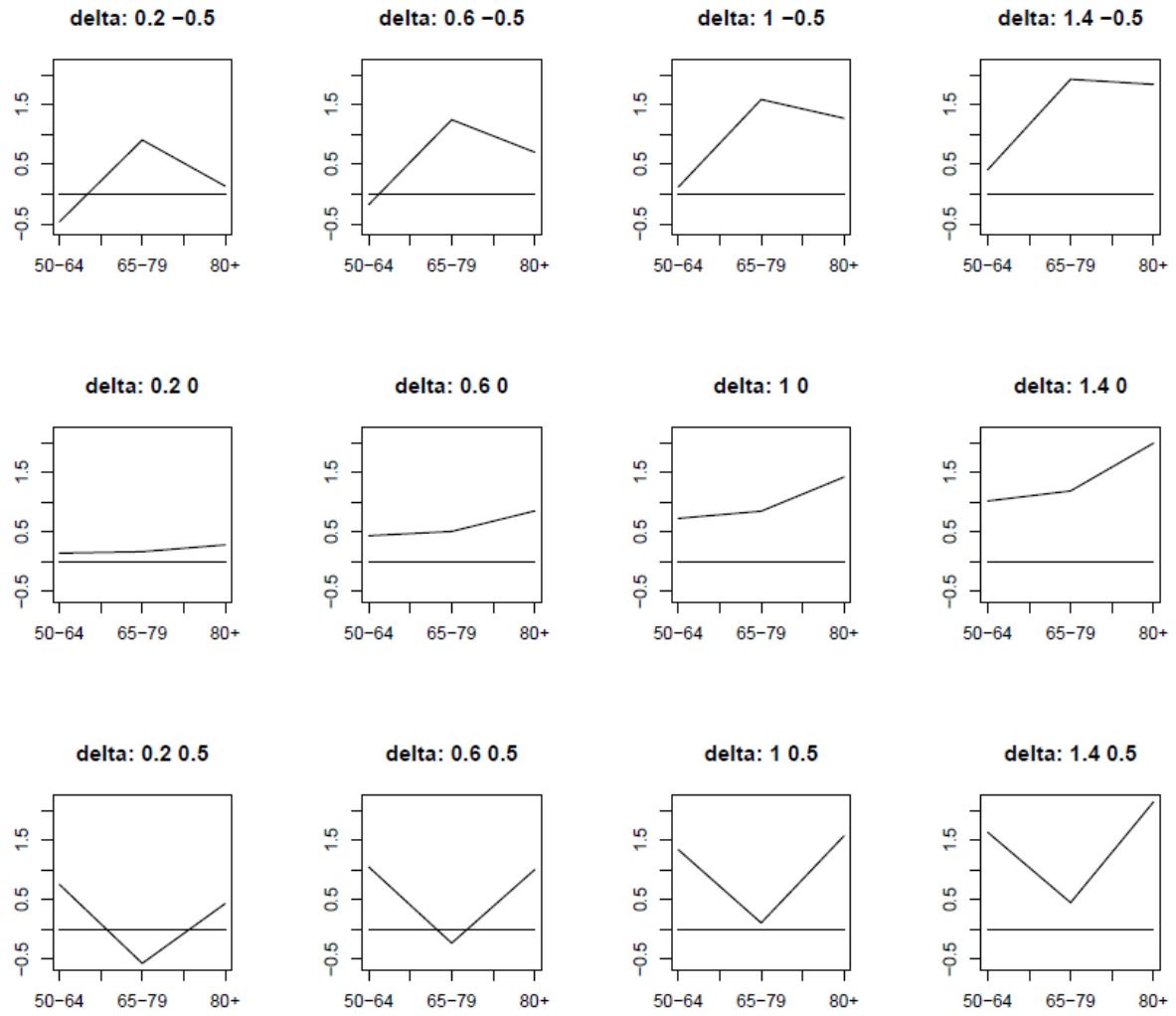
Figure 2 shows the idea of adding a second axis. The disabling impacts by age are shown for 12 hypothetical diseases with different combinations for δ (disease effects) and the age patterns γ. The middle row of pictures shows the situation when $\delta_{d2} = 0$. This is the situation where the second parameter equals 0. For δ_{d1} we assumed the same values as in Figure 1. For each disease there is an increasing disabling impact with "knot" at age 65-69 in this example. In relative sense the increase is the same for all diseases in this row. The upper and lower row of panels show the situation for other

values of δ_{d2}. If δ_{d2} is negative, e.g. -0.5, upper row, the value $-0.5*\delta_d$ is added to the hazards in the middle row, and as a result we see a maximum hazard at age 65-79. If δ_{d2} is positive (e.g. + 0.5, lower) the result is a dip around age 65-79. Note that some of the hazards β_{ad} may become negative; this can be avoided by combining or deleting the specific diseases.

The gamma for the first axis is: 0.729, 0.840 en 1.422. This can be seen in the figure with delta is 1 (third column), as the hazards for age 50-64, 65-79 and 80+ have these values. The gammas for the second axis are: 1.224, -1.474 and 0.302. In the figure with delta is 1 (first axis) and -0.5 (second axis) this yields 0.729 - 0.5*1.224 = 0.117 for age group 50-64, 0.840 -0.5*-1.474 = 1.577 for age group 65-79 and 1.422 - 0.5*0.302 = 1.266 for age group 80+. For plus 0.5, 0.5 is added instead of subtracted. For delta is 0.2 (first axis) and 0 (second axis) this yields 0.2 *0.729; 0.2*0.840 and 0.2* 1.422 for the three age groups. A figure showing the age- specific disabling impacts are made automatically by the revised tool.

In most occasions this second axis is not needed, but including a second axis in the regression models allows the user to assess whether one axis sufficiently describes the data. The need for a first rank (common age pattern) as compared to the same disabling impact for all ages can be assessed using the likelihood ratio test. Similarly, the need for the second rank (different age patterns) as compared to first rank can be assessed. For more detail on model selection, see the separate section and the illustration of the model selection in the example.

Figure 2: RRR with second axis. Each of the figures shows the age dependent hazard for a specific hypothetical disease; the heading of the figure for each disease gives values for δ_1 and δ_2. Columns of the figure have the same δ_1, rows of the figure have the same δ_2.



2.2 General option 2: Comparison of two populations

Given that decomposition of differences in health expectancy between two populations (men vs. women, member states) into the additive contribution of different causes (diseases) requires a comparison of disability prevalence by cause, the tool offers the possibility to distinguish two populations in one analysis (run). While of course disability prevalence by disease in each population could be based on separate analyses, running separate analyses (and hence including all potential interactions between the populations in the model) may not always yield the most parsimonious model and does not provide significance for differences. Therefore the regression model, which is the core of the attribution tool model can include two populations. The user can test whether the background hazards differ between both populations. The user can also test whether the disease hazards (disabling impacts) differ between the two populations, both in a situation with and without different background hazards. Distinction for population in RRR in situation of two populations will estimate a model with different background hazards and equal age-specific disabling impacts. A model with both different can be found by running the program for two populations separately. The option to compare

populations has also an added value in other applications, for instance, for the examination of the contribution of diseases to disability it is important to assess to what extent men and women need to be modelled entirely independently.

2.3 General option 3: Different models and p-values for model selection

To select the best model, several decisions have to be made by the user. One decision relates to the modelling of age. P-values based on the likelihood ratio test are provided to assist the user assessing whether the disabling impacts vary by age (reduced rank regression needed or not) and to test whether one age pattern for all diseases decribes the data sufficiently well. The second decision relates to treating the population as one homogeneous population (apart from age), or allowing the background disability, the disabling impacts (disease hazards), or both to differ by population. This issue is not only relevant for the decomposition analyses where two populations are compared, but also, for instance in studies focussing on one country the question arises whether men and women need to modelled seperately. Likelihood ratio tests are available to asses whether there are significant differences between populations in the background hazard, the diseasehazards, and the disease hazards given that the population-specific background hazards differ. Once it is clear that both the background hazards and disease hazards differ by population, each of the populations can be analysed as separate populations (in separate jobs). But if for instance background hazards do differ between populations, but disease hazards not, using separate background hazards, but a common disease hazards yields a more parsimonious model, than fitting two separate models (hence with separate background and separate disease hazards).

Given that the user can specify different models, depending on how the disabling impact is modelled (no age variation, RRR 1-axis, RRR- two axes) and on whether some or all parameters are estimated by population, a whole set of models can be defined and estimated. Model selection is based on the likelihood ratio test, comparing each time two models: one without and one with the variable or interaction in question. If the drop in deviance is large in relation to reduction in the number of degrees of freedom spent (χ^2 distribution) the second model is significantly better. It is up to the user to decide whether the difference is also relevant.

2.4 General option 4: Confidence intervals using bootstrapping

Confidence intervals using bootstrapping are provided for the attributions and for the age-specific disability hazars (rates), i.e. the products of the age pattern γ_a (gamma) disabling impact δ_d (delta). With the distinction by population, bootstrapping is only possible for the variant A (background hazard different) and D (disease hazards different), not for AD because this can be obtained from separate runs of the model. Bootstrapping is not possible for a model with a Reduced Rank Regression with a second axis.

2.5 Options linked to further application in the decomposition tool

For further application in the decomposition tool, the attribution tool offers the users the following options.

First, the tool can attribute the fitted disability prevalence to diseases, based on the regression model, or attribute the **"observed" prevalence** from the survey to diseases. Using observed prevalences might be useful for the decomposition of differences between two health expectancies that are already published. Using observed prevalence then yields the same disability prevalence by age and hence health expectancies, while using fitted values from the regression analysis may yield health expectancies that are slightly different but more stable, though generally differences will be small with a proper fit of the model.

Second, the tool provides a **machine-readable output table**, which can be read by the decomposition tool, so the user can quickly prepare the input for the decomposition tool. Only in case more populations are compared, and hence attribution tables from different models are included in the machine-readable output table, the user has to select the attribution table from the model that fits the data best, and delete the remaining attribution tables.

Third, the tool offers the user the possibility to include in the attribution table information on the **institutionalised population** from a separate data source. This can be relevant, if the dataset with individual level survey data that is used in the regression analyses excludes persons living in institutions. The tool is flexible regarding whether or not, and how to include the institutionalised population, reflecting that in the Sullivan method for calculating health expectancy, the institutionalised population can be included in different ways. There is no need to include separate information on the population in institutions if no (good) data are available on the institutionalised population or if the prevalence of disability in the survey is assumed to represent the entire population, either because 1) the survey covers the institutionalised population, or 2) the institutionalised population is very small and thus would not affect the overall disability prevalence.. In other situations information on the institutionalised population can be included in two ways. First, by including only information on the age-specific proportion of persons living in institutions and assuming that this population can be considered to be all disabled (original Sullivan assumption), and second, by including additional information, based on other sources, about the proportion disabled within the institutionalised population. If the user includes additional information on the institution population, in addition to the dataset with individual level-data used in the additive regression analysis, aggregated data (per age class and if applicable per population characteristic) on the fraction of the total population that lives in institutions, and if applicable, on the fraction of persons living in institutions that is disabled is used in the tool. In the output tables, the number of persons in the population is then inflated by the factor (1/(1-fraction in institutions)). The numbers disabled per disease do not alter, only the total numbers are changed and an additional row is added with numbers disabled in

institutions. These are not attributed to diseases and should be considered as a separate category. In the bootstrap, information on institutionalized population is assumed not to be subject to random error so fractions are taken as deterministic.

3. How to use the tool

The attribution software is programmed in R. All the user-specified input is communicated to the R program through the input specification Excel-file (saved as "csv-file"). Hence, the user does not need to have any R knowledge to use the program.

To use the attribution method in R you need:

I. Input dataset (ASCII or POR-format): example: share1rel2_bothit.txt

II. Input specification file (csv format): example: "input attribit W.csv".

III. R syntax: "attrib.r"

If R is not already installed, do first the following:

- Go to http://lib.stat.cmu.edu/R/CRAN/
- Choose the system you are using, probably 'Windows(95 and later)'
- Follow the instructions

Having R installed on the computer, do the following:

- Put the three files (e.g., script: attrib.r, data file: share1rel2_bothit.txt, and attribit Wit.csv) in one folder.
- Check that the separator is a comma (check the computer), via start\settings\control panel\regional options. If questions pop up your screen (starting with "input file already exists" and "input file may contain features that are not compatible with csv format", the user answers two times "yes".

3.1 Input preparations and how to run the program

First the user has to make a dataset either in txt or in POR format. This can for instance be done in SPSS (saving with EXPORT OUTFILE, and giving extension "por') or in R using sink command and give extension "txt". The user can also use a "sav" file. The program gives a warning, but still runs. The content of this dataset is described in more detail below.

Second, the user has to complete the input specification file. To do so the easiest is to open an existing input specification file. The user should specify the location and name of the data file with individual level data as well as information on the column numbers of column containing information on disability, specific diseases and age. Additionally the user should specify all the user choices when

different options are possible, and the name and location of additional data on the institutionalised population, if applicable.

Third, in the R program the user has to specify the path and name of the Excel input specification file (csv-file). This input specification file may also be in txt format.

Finally, the user has to run the R program. Inspection of the Console Window in R will show whether errors occurred (blue text instead of red). Running a specific Excel input specification file (job) results in an analysis and output-file. The output is written to files (text format), which can be read by e.g., Textpad, Notepad, or MS-Word. Users can opt for additional output. Additionally a machine-readable file with disability by age and cause is provided, which can be read by the decomposition tool.

3.1.1 Input dataset

The cross-sectional dataset should at least include individual-level data on: disability (binary); diseases (binary); age classes (user can choose age classes), and may additionally include: population (binary) and sample weights. The dataset can be in ASCII- format (txt) or SPSS portable (POR) format. The first line of the file in ASCII-format should provide the names of the variables and should be separated by spaces (space-delimited). The file contains one column more than the number of names, as the first column should include row numbers or ID codes. Missing values are not allowed anywhere in the dataset.

The presence of disability is a dichotomous variable (0 or 1). Often disability is measured using multiple questions each having multiple answer categories. The user thus first has to make a dichotomous variable, and this variable is the one to be included in the dataset.

Presence of diseases is a set of dichotomous variables. The user is free to select the number of diseases, though sample size may be a reason to aggregate disease groups. The user may decide to include a category of "other diseases". In principle the tool focuses on attribution of disability to single disease, however if the user wants to add a specific multi-morbidity, a single variable with "1" for the situation that both diseases are present can be used. There is no limit to the length of names of the variables. We suggest naming interactions like "stroke.heart", etc. but other names can also be used.

Age-classes can be defined by the user. The user thus first has to create one or more categorical variables for the age classes (e.g. five year, 10-year or larger) to be included in the dataset. When input based on SPSS is used, take care that one or all the classes have value labels. If only one a f few classes have labels (e.g. 99=missing), only these values will be used and the other records will be deleted. If by accident, the column with weights is indicated to contain the age classes the large number of different values will force the program to exremely long computation time. If this occurs: break of the program (go to the console window and press the red stop icon). It is possible to use different age classes for the age pattern of the disabling impact (if that option is selected), and for the

age pattern of the background effect. We advise not to use broad age groups for background. In case of smaller sample sizes, however, the age pattern of the disabling impact can be summarized to broader age groups. A likelihood ratio test can be used to assess whether smaller age groups improve the fit significantly. This is illustrated in the appendix 2.

Population is a categorical variable with two possible labels. This variable can be used to compare two populations, e.g. men vs. women, or two countries. We recommend stratifying the analyses by sex, either by selecting only one of both sexes, or by using the feature to compare two populations. A likelihood ratio test can then be used to assess whether a distinction by sex in either background hazards or disabling impacts (disease hazards) improves the fit of the model significantly.

The data set can include other variables that can be used to select subpopulations within the dataset. This can be specified in Q4 (see below).

Sample weights can be included to reweigh the sample design. The user can provide normalised weights (with mean weigh of one). If normalised weights are not provided, the program normalises the weight to avoid unwarranted increase of power of the analysis. This normalisation will be applied after selection. Using normalized weights, the sum of the weights remains equal to the number of cases. In case the weights differ extremely (95% range differs more than a factor 10), the analysis will be biased.

The age-specific fractions of persons living in institutions (as part of the total population), and of the fraction disabled among persons living in institutions (as part of the institutionalised population) are not part of the individual-level dataset and are not used in the regression analyses, but can be taken into account in the attribution table. They should be provided in a separate file. If the second set of fractions is not provided, all persons in institutions are considered disabled.

3.1.2 The Excel input specification file

The Excel input file in csv (comma separate variables) provides all information to be used in the R-syntax file. The specification file consists of two columns. The questions in the input specification file are in the right column, the answers should be provided by the user in the left column (range A1..A21). If not applicable, the user can put "NA" in the right column, or leave the column empty, depending on the specification in the right column. The user may only change the content of the first column and should never delete any rows! The file should be saved in excel, choosing "save as", "csv comma delimited". The second column, and more importantly, the rows should never be changed. Except for the obligatory questions, cells may be left empty or NA (not applicable) may be entered. As alternative an input specification file in txt format can be used. It should be organised in two columns, separated by a @.

In the input file the user specifies:

Q1: Name default directory (end with a "\"). All files are expected to be in this directory and all output files will be saved here.

Q2: Name output file

Q3: Name input file, i.e., cross-sectional dataset (extension "por" or "txt")

Q4: Selection of cases or records (e.g. "Crossdat$sex==1&Crossdat$age5>60". Variables to be included in a selection should always start with "Crossdat$" and should include the exact variable name (column name in data file). Be aware that the program is case-sensitive. For more examples of selecting subpopulations, see appendix 4.

Q5: Column number in data set containing population (optional: if no population distinction is used, this can be left empty.

Q6: Column number in data set containing age class (for background hazard)

Q7: Column number in data set containing disability variable

Q8: Column number in data set containing diseases (more diseases can be separated by ",", or can be grouped using ":" (e.g. 5:9,14)

Q9: Column number in data set containing weight (optional: if no weight this can be left blank or "NA" can be entered).

Q10: Disabling impact by age (0 is constant across age (i.e. no RRR), 1 is one axis in RRR model, 2 is two axes in RRR model)

Q11: Column number of age classification of the disabling impact if disabling impact varies by age. This may be the same as Q6.

Q12-1: Name of file with data on institutionalised population for first or only population (optional)

Q12-2: Name of file with data on institutionalised population for second population (the second is optional)

Q13-1: Column numbers with fraction institutionalised and fraction disabled if institutionalised in first or only population (optional); if all persons in institutions disabled, enter NA instead of second column

Q13-2: Column numbers with fraction institutionalised and fraction disabled if institutionalised in second population; if all persons in institutions disabled, enter NA instead of second column

Q14: "F" for fitted numbers of disabled or "O" for observed numbers to construct the table of attributions.

Q15: "F" for no additional output, "T" for additional tables, i.e. tables with relations between diseases, age class and disability. "W" will give this same additional output, but then using weights.

Q16: Name of machine-readable output file (to be used in the decomposition tool).

Q17: Enter number of replicas if you want a bootstrap. If you want a bootstrap for a model with POP, add ,A or ,D to the number of replicas to show whether you want a bootstrap for linmod.a or linmod.d (hence: 1000,A of 1000,D).

3.1.3 Running R

The R-syntax “attrib.r” file is ready to use, users only need to specify the path and name of the input specification file. First open it. This can be done in two ways:

1. In Tinn-R: file / open /, (or double click on name)
2. In R: file/ open script/ open

Then the path and name of the input specification file should be entered or adapted in the line that starts with “filein”.

E.g. filein <- "g:\\doc\\attribution\\input attibWit.csv"

Be careful with the double “\\”

The user can include more input files for different runs. Always the last one will be used by the program. The user can also put “#” before the lines with input files not to be used in the specific run. The program then reads this line as comment.

When running several jobs, we suggest clearing the R console (in Tinn-R: controlling R/clear Console).

The current version of the tool uses the Nelder-Mead optimalization routine. This is slower but more reliable than the BFGS. BFGS is also a quasi-Newton method (also known as a variable metric algorithm), published simultaneously in 1970 by Broyden, Fletcher, Goldfarb and Shanno. BFGS is found to be less reliable in case of an additive hazards regression. Unfortunately it is not possible to use a family of the glm function, which would run much faster: due to the link function negative probabilities can occur, which lead to fatal errors. In the current implementation penalties are used to avoid negative probabilities. Only if problems occur in the optimalization, which become visible by the higher deviance for the model with more parameters, and a p-value of 1.00, the user should change the optimalization to BFGS. This can be done changing the order of “optmethod <- "BFGS" and “optmethod <- "Nelder-Mead" in the first section of the R program.

During the fit of a RRR there is an animation of the fit procedure. For all optimizations an adapted Nelder-Mead scheme is used: After each run of 1000 iterations the result is compared with the result of the last run. If the deviance has dropped more then 1 point a next run is started investigating the same broad range of possibilities as at the start of the procedure,but with the current best fit as starting values. If the deviance dropped less then 1 point a final run is started without a maximum. After reaching the optimum the Hessian is estimated which shows as a repeated spiked pattern sligthly above the optimal value. During a bootstrap the last step is left out and the title of the picture shows the number of the replica.

During the bootstrap a figure will be displayed which shows the progress of the calculations. When a bootstrap of a single linear model (SL) is requested the picture will show on the X-axis the replica numbers. On the Y-axis the attributions (in numbers) of the main attributing diseases are shown for the different age classes. Above the zero are the point estimates and the lines in different colors connect the estimates in the subsequent replicas. It gives an intuitive idea of the certainty of the relative importance of the disease in the different age classes.

3.2 Output of the model

The output file consists of at most 5 parts: 1) data inspection and calculation, 2) descriptive, 3) results from the simple additive regression model, 4) outcomes of reduced-rank regression, and 5) attribution of disability by disease (cause-specific disability prevalence).

Part I serves for data inspection and documentation so the user can easily see which data file, subpopulation, variables and specifications were used. The user should check this in case an error occurs which is recognized by a blue error message in the console window after running R, but also to assess whether the analysis is correctly specified. It provides:

- Path and name of used dataset
- Selection (if any)
- Number of cases and variables after selection
- First five records of dataset
- Information on weights (if used): column, name, mean weight and whether or not weights are normalised to have a mean value of one)
- Columns and original names of selected variables: age, disability, selected diseases
- Specification of Reduced Rank Regression (if used), including rank (1 vs. 2), column and original name of the age variable used for disabling impact by age.
- Whether fitted or observed numbers should be used to construct the attribution table.
- Name of machine-readable output file. This file can be read by the "decomposition program"
- Specifications of the data for the insititutionalized population (if included)
- Number of bootstraps (if included)

Part II provides the first descriptive calculations, preceding the actual additive regression modelling. This part becomes much longer if the user has specified "T" or "W" for additional tables in Q15.

- Prevalence of disability by age class
- Numbers cross-tabulated by disability and age class
- Numbers of persons with diseases
- Overview of relations between diseases, age class and disability (optional)
    - Prevalence of diseases per age class
    - Cross-tabulation of age class by disease (including "no disease" and each selected disease)
    - Prevalence of disability by age class and disease

Part III gives the results from the simple additive regression model (SL), including:

- First five rows from presences of diseases
- Outcomes of simple linear model ("SL")
    - Parameters: estimate, standard error, t-value, Prob (p-value unrounded), CI low, CI high, p-value (rounded)

No t and p-values are given for age groups, as hazards rates for age groups should by definition exceed 0.

- Deviance and degrees of freedom for the simple linear model.

If two populations are distinguished by the user (Q5), the output includes also the result of the population specific models: 1) model "A" background hazards different, model "D", disease effects different, and model "AD" background and disease effects different. In addition, p-values are given for differences in disease effects ("D" vs. "SL"), for differences in background hazards ("A" vs. "SL"), for differences in background hazards given difference in disease effects ("AD" vs. "D"), and for differences in disease effects given difference in background hazards ("AD" vs. "A").

Part IV gives the outcomes of Reduced Rank Regression (only if option 1 or 2 is selected in Q10): including:

- Parameter estimates:
    - Alpha: background hazard (population specific if requested)
    - Gamma (γ_a): disabling impact (for RRR with rank 1 for first axis, for RRR with rank 2 for first and second axis).
    - Delta (δ_d): common age pattern of disabling impact (for RRR with rank 1 for first axis, for RRR with rank 2 for first and second axis).

- Fitted and observed per age class
- Deviance and DF for the RRR model
- P-value for differences in hazards by age group (RR(1) vs. SL), which indicates whether RRR with this rank is necessary.

Summary of the RRR

- Parameters: estimate, standard error, t-value, Prob value, confidence interval low and high, p-value
- Disabling impact (beta=gamma * delta) per disease and age

Part V gives the attribution of disability by disease

First is given:

- nn is number of persons (weighted)
- "disab" is number disabled (weighted). This can be the observed or fitted number.

The attribution to background and specific diseases is presented in three metrics:

1. Attribution as numbers
2. Attribution as fraction of disabled
3. Attribution as fraction of total population

If two populations are distinguished, attribution tables are given for each population. If the user will use the attribution table in the decomposition tool, the user has to split these tables in the machine readable copy in order to make on attribution table for each population. The table with the attribution presented as numbers is used in the decomposition tool. Only this part will be written to the machine readable copy.

4. Selecting the best model

Figure 3 shows possible models with the attributions tool, both for the situation with and without distinction by population.

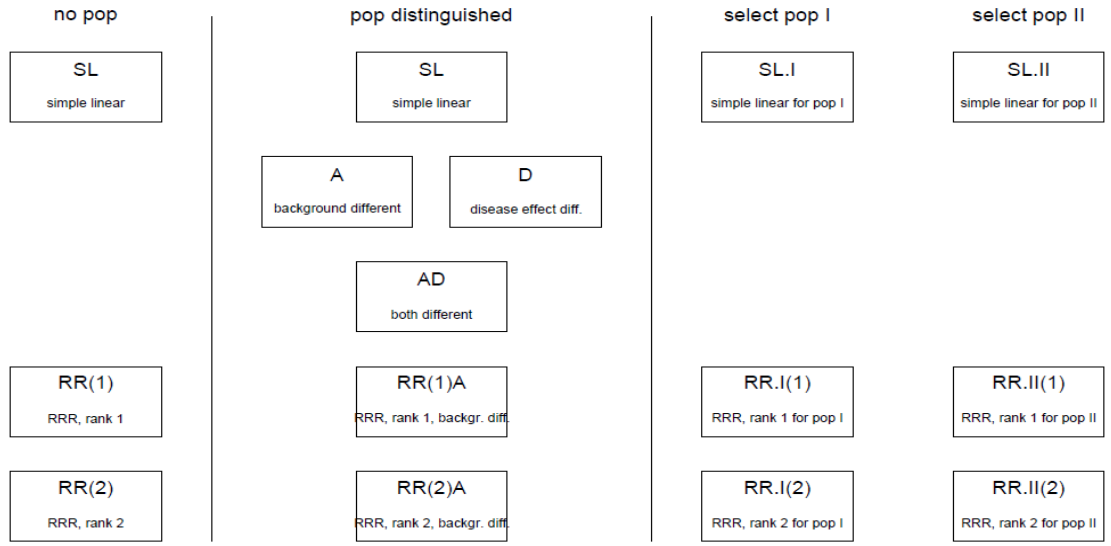


Figure 3: All possible model options with the attribution tool.

4.1 Situation of one population (no population distinguished)

The model selection starts with a simple linear (SL) regression model, where simple means: no differences between populations, and linear means: no interaction between age and diseases, hence one disabling impact for all ages. The next step is to assess whether there is an interaction between age and diseases (disabling impact varies with age). This is based on the comparison between the model with RRR first axis only "RR(1)" vs. simple linear model "SL". If the likelihood ratio test indicates that RRR(1) is needed, the next step is to assess whether one axis in the RRR model is sufficient to summarize the age variation. This is based on comparison of with RRR and second axis "RR(2)" vs. model with RRR with one axis "RR(1)".

4.2 Situation of two populations (populations distinguished)

The model selection again starts with the simple linear model ("SL"). However, in the situation of two populations, the first step is to test whether there are differences between these populations in background and/or disabling impacts. This requires at most four comparisons:

1. "A" vs. "SL": indicating whether background hazards are the same in both populations
2. "D" vs. "SL": indicating whether disabling impacts (disease effects) are the same in both populations

3. “AD” vs. “A”: indicating whether disabling impacts are the same in both populations, given population-specific background hazards (only to be tested if model “A” is significantly better than model “SL”).
4. “AD” vs. “D”: indicating whether background hazards are the same in both populations, given population-specific disabling impacts (only to be tested if model “D” is significantly better than model “SL”).

The second step is to assess possible interaction between age and diseases (disabling impact varies with age), and if so, whether one axis in the RRR model is sufficient to summarize the age variation.

Given all possible models, there are more routes for model selection. We propose the following route (see Table 1).

The user first tests for differences between the two populations in background hazards. To do so the user compares “A” with different background hazards by population with “SL”. Next, the user tests whether the disabling impact of the diseases differ by population. Depending on whether the background hazards differ between the populations, the user has to compare model “D” vs. “SL” (for situation of no differences in background hazards) or model “AD” vs. A” (for situation with differences in background hazards).

If this comparison indicates that the disabling impacts of the diseases are the same in both populations, the user has to test whether the disabling effects differ by age. This can be tested by comparing a model with RRR (first axis) “RR(1)” to the model without RRR, which is “SL” (for situation of no differences in background hazards between population or situation of differences in both background hazards and disabling impacts between populations), or “A” (for situation of differences in background hazards, but not in disabling impacts between populations). For the situation of differences in both background hazards and disabling impacts between populations, the user should use two separate models, one for each population. However, it may be preferable to select the same final model for both populations. This can be achieved by performing the likelihood ratio tests for comparing models of both population I and population II together, hence: “RR.I(1)” + “RR.II(1)” with “SL.I” + “SL.II”. It is noteworthy that model “SL.I”+”SL.II” is identical as “AD”. Both deviances and degrees of freedom may be added.

Table 1 Model selection for situation with two populations										
Step 1	Compare: A vs. SL									
Outcome step 1	Y (p>=0.05), background=same				N (p<0.05), background=different					
Step 2	Compare D vs. SL				Compare AD vs. A					
Outcome step 2	P≥ 0.05: disabling impact does not differ between populations			P< 0.05: disabling impact differs between populations	P≥ 0.05: disabling impact does not differ between populations			P< 0.05: disabling impact differs between populations		
Step 3-a: population	Both populations in one model without distinction for population (column 1 of Figure 3)			Unlikely, see*	Both populations in one model with distinction for population (stay in column 2 of Figure 3)			Each population in separate model (last two columns of Figure 3)		
Step 3-b:	RR(1) vs. SL				RR(1)A vs. A			(RR.I(1)+RR.II(1)) vs. (SL.I+SL.II)		
Outcome step 3	P≥ 0.05: disabling impact does not differ by age	P≥ 0.05: disabling impact differs by age			P≥ 0.05: disabling impact does not differ by age	P≥ 0.05: disabling impact differs by age		P≥ 0.05: disabling impact does not differ by age	P< 0.05: disabling impact differs by age	
	Use: SL	Step 4			Use: A (model with population)	Step 4		Use: SL.I and SL.II (separate model for each population)	Step 4	
Step 4	-	RR(2) vs. RR (1)				RR(2)A vs. RR(1)A		-	RR.I(2)+RR.II(2) vs. RR.I(1)+RR.II(1)	
Outcome step 4	-	P≥ 0.05: disabling impact by age same for all diseases	P < 0.05: disabling impact by age differs between diseases			P≥ 0.05: disabling impact by age same for all diseases	P < 0.05: disabling impact by age differs between diseases	-	P≥ 0.05: disabling impact by age same for all diseases	P < 0.05: disabling impact by age differs between diseases
Final model	SL without population	RR (1) without population	RR(2) [b] without population		A model with population	RR(1).A model with population	RR(2).A [b] model with population	SL.I+SL.II separate model for each population	RR.I(1)+RR.II(1) separate model for each population	RR.I(2)+RR.II(2) [b] separate model for each population

Options: as one population, distinction by populations in one model, two populations in two models.

[b] Tool does not allow testing for higher axes. A model with all age*disease interactions could be fitted.

* First test AD vs. D, if background differs given that the disabling impact differs, the same situation applies with both background and disabling impact differs. If background does not differ, while the disease effect does, no quick testing approach is programmed in the tool.

If this comparison shows that the disabling impacts differ by age, the user still has to assess whether one axis describes sufficiently the variation with age. This can be tested by adding a second axis and comparing the model with RR(2) with the model with RR(1). Depending on the outcomes of the prior steps, this yields a comparison of “RR(2)” vs. “RR(1)” or “RR(2)A” vs. “RR(1)A” or “RR.I(2)”+”RR.II(2)” vs. “RR.I(1)”+”RR.II(1).

In the situation of equal background hazards but different disease hazards, which does not occur frequently, we advice to compare model “AD” with model “D” in order to assess whether, after modelling different disabling impacts by population, the background hazards are still the same. If this is not the case, the situation is similar to one where both background and disease effects differ between the populations. However, if background did not differ after including disabling impacts by population, it is advised to define a new set of diseases, where every original disease is split up into two new diseases, one if the person belongs to population I, the other if the person belongs to population II. This will in fact turn out to be an interaction between disease and population. The same can be done for the age variable to be used in the RR. If then the analysis is run on a dataset with both populations, but without the variable population distinguished, this will yield a model with equal background hazards but disease disabling impacts and age patterns of disabling impacts both different for population.

Part 3: Example

1. Input

1.1 Input dataset (ASCII or POR-format)

In this illustration, we use data from the SHARE survey (wave 1, release 2) for Italy. As disability measure we used the Global Activity Limitation Index (GALI) based on the question "For the past six months at least, to what extent have you been limited because of a health problem in activities people usually do?" We considered "severely limited" and "limited, but not severely" as disabled. We have one dataset including only Italian females to illustrate the situation with one population, and one dataset including Italian males and females to illustrate the situation when two populations are compared. The first is named "share1rel2_WOMENit.txt and the latter dataset is named "share1rel2_bothit.txt". Both are in ASCII format.

"This dataset "share1rel2_bothit.txt" includes the following variables:

1.	SAMPLEID2	Identification number
2.	CVID	Identification number
3.	Population	1 (males) vs. 2 (females)
4.	PH005_5	Disability (yes/no)
5.	Gender	1 (males) 2 (women)
6.	weight	weight to adjust for sampling design and/or non-response
7.	age5	age in five-year age groups, starting at age 50 and ending at age 85+
8.	heart	heart attack
9.	stroke	stroke or cerebrovascular disease
10.	cancer	cancer
11.	copd	asthma, copd
12.	diabetes	diabetes
13.	muskulo	musculoskeletal diseases, including: arthritis + osteoarthritis+hip fracture
14.	other	other diseases including: Stomach or duodenal ulcer, peptic ulcer, cataracts, other
15.	age15	age in 15-year age groups: 55-64; 65-79 and 80+

The dataset “share1rel2_WOMENit.txt is similar but only includes data for women.

1.2 Input specification file (.csv)

The input specification files, fitted in the illustration in part 3 are listed below. First we list for the situation of a single population, next for the situation of two populations. The zip-file including all input specification (csv) and output files is available from the authors.

Single population:

1. AttribW it: example of analyses of one single population (Italian women), hence without distinction in population, and no RRR
    - Main characteristics:
        - single population (Q5 NA or empty)
        - no RRR (Q10 =0, Q11 is NA or empty)

2. AttribRRRit: example of analyses of one single population (Italian women), hence without distinction in population, but with RRR (first axis)
    - Main characteristics:
        - single population (Q5 NA or empty)
        - RRR (Q10 =1, Q11 is 15 (column number of age classes for disabling impact)

3. AttribRRR2it: example of analyses of one single population (Italian women), hence without distinction in population, but with RRR (second axis)
    - Main characteristics:
        - single population (Q5 NA or empty)
        - RRR (Q10 =2, Q11 is 15 (column number of age classes for disabling impact)

In all these three input specifications other options can be changed by the user, including:

1. Q-4: to select cases or records (for more details see appendix 4)
2. Q-9: to include/exclude sample weights. For no sample weights this can be left empty or NA can be filled in
3. Q-12 and Q13: adding information on institutionalised population
4. Q-14 to use observed (O) instead of fitted values (F)
5. Q-15 to obtain additional output (T)

Two populations:

1. Attribit2pop.csv: example of analyses of two populations (Italian men (population 1) and Italian women (population 2).
    - Main characteristics:
        - population included (Q5=5, where 5 is column number for gender)
        - no RRR (Q10 =0, Q11 is NA or empty)

2. Attribit2popRRR.cvs: example of analyses of one single population (Italian women), hence without distinction in population, but with RRR (first axis)
    - Main characteristics:
        - population included (Q5=5, where 5 is column number for gender)
        - RRR (Q10 =1, Q11 is 15 (column number of age classes for disabling impact)

3. Attribit2popRRR2.cvs: example of analyses of one single population (Italian women), hence without distinction in population, but with RRR (second axis)
    - Main characteristics:
        - population included (Q5=5, where 5 is column number for gender)
        - RRR (Q10 =2, Q11 is 15 (column number of age classes for disabling impact)

For first population separately:

4. Attribitpop1RRR.cvs: example of analyses of first population (Italian men), model SL and RRR (first axis)
    - Main characteristics:
        - Single population, but one selected
        - RRR (Q10 =1, Q11 is 15 (column number of age classes for disabling impact)

5. Attribitpop1RRR2.cvs: example of analyses of first population (Italian men), model SL and RRR (second axis)
    - Main characteristics:
        - Single population, but one selected
        - RRR (Q10 =2, Q11 is 15 (column number of age classes for disabling impact)

For second population separately:

6. Attribitpop2RRR.cvs: example of analyses of second population (Italian women), model SL and RRR (first axis)
    - Main characteristics:
        - Single population, but one selected
        - RRR (Q10 =1, Q11 is 15 (column number of age classes for disabling impact)

7. Attribitpop2RRR2.cvs: example of analyses of first population (Italian women), model SL and RRR (second axis)
    - Main characteristics:
        - Single population, but one selected
        - RRR (Q10 =2, Q11 is 15 (column number of age classes for disabling impact)

In all these input specifications also other options can be changed by the user, including:

1. Q-4: to select cases or records (for more details see appendix 4). This is also needed if based on model selection the user has decided that in a situation of two populations, each population is modelled separately. Q-9: to include/exclude sample weights. For no sample weights this can be left empty or NA can be filled in
2. Q-12 and Q13: adding information on institutionalised population
3. Q-14 to use observed (O) instead of fitted values (F)
4. Q-15 to obtain additional output (T)

1.3 R syntax file (.R)

The R program specifies the location (path) and name of the input specification file. Running R creates an output file, specified in Q2 in an input specification file, as well as the machine readable copy (e.g. attribit.mrc)

2. Output

The output starts with data inspection and documentation:

```
[1] "> > > OUTPUT PART 1: DATA INSPECTION AND DOCUMENTATION < < <"
[1] "V:\\PRJCT\\eushare\\caspar\\attribhandleiding\\attribit.out"
[1] "v:\\prjct\\eushare\\caspar\\attribhandleiding\\input attribW it.csv"

[1] "used dataset:"
[1] "V:\\PRJCT\\eushare\\caspar\\attribhandleiding\\share1rel2_WOMENit.txt"
[1] "selection:"
function(){Crossdat[  ,]}


[1] "number of cases and variables after selection:"
[1] 1373   16

[1] "first five records:"
      1:SAMPID2 2:CVID 3:COUNTRY 4:PH005_ 5:gender 6:weight 7:age5 8:heart 9:stroke 10:cancer
13331 1.604e+12      2     italy        1        2   0.5575     80       0        1          0
13332 1.604e+12      1     italy        1        2   0.6756     60       1        0          0
13334 1.604e+12      1     italy        0        2   5.4890     75       0        0          0
13336 1.604e+12      1     italy        1        2   0.7943     50       0        0          1
13337 1.604e+12      2     italy        0        2   1.3472     80       0        0          0
      11:copd 12:diabetes 13:muskulo 14:other 15:age10 16:heart.stroke
13331       0           0          1        0      80+                0
13332       0           0          0        0    50-64                0
13334       0           0          0        0    65-79                0
13336       0           0          0        0    50-64                0
13337       0           0          0        0      80+                0

[1] "weights are in column number:"
[1] 6
[1] "original name:"
[1] "weight"
[1] "weight variable should have 1 as mean value, but is 0.9985202193008"
[1] "weights will be normalised"

[1] "ageclasses are in column number:"
[1] 7
[1] "original name:"
[1] "age5"

[1] "disability is in column number:"
[1] 4
[1] "original name:"
[1] "PH005_"

[1] "column numbers for diseases:"
[1]  8  9 10 11 12 13 14
[1] "list of names:"
[1] "heart"    "stroke"   "cancer"   "copd"     "diabetes" "muskulo"  "other"

[1] "fitted numbers will be used for the attributions table"

[1] "machine readable copy of the result"
[1] "V:\\PRJCT\\eushare\\caspar\\attribhandleiding\\attribit.mrc"
```

Always check the number of cases. Also check the headings of input data in combination with first five records. Finally it is useful to check the data and user-specified choices, including use of weights, selection of age classes, selection of disability measure, selection of diseases and fitted vs. observed numbers.

Subsequently, you can inspect the remaining of the output, starting with descriptive tables on prevalence of disability, cross tabulation of population by age and disability and number of persons with specific diseases.

```
[1] "> > > OUTPUT PART 2: DESCRIPTIVE CALCULATIONS < < <"
[1]
"============================================================================================="
[1] "Start of calculations"

[1] "prevalence of disability per age class:"
    50     55     60     65     70     75     80     85
0.2828 0.3504 0.3945 0.4526 0.5714 0.6154 0.7397 0.8286

[1] "numbers by disability and age class:"

     50  55  60  65  70  75  80  85
  0 142 178 175 127  72  40  19   6
  1  56  96 114 105  96  64  54  29

[1] "numbers of diseases:"
   heart   stroke   cancer     copd diabetes  muskulo    other
     115       39       85      125      146      639      309
```

Next you can inspect the regression output:

```
[1] "> > > OUTPUT PART 3: RESULTS FROM THE SIMPLE ADDITIVE REGRESSION MODEL (SL) < < <"

[1] "first five rows of presences of diseases:"
  heart stroke cancer copd diabetes muskulo other
1     0      1      0    0        0       1     0
2     1      0      0    0        0       0     0
3     0      0      0    0        0       0     0
4     0      0      1    0        0       0     0
5     0      0      0    0        0       0     0

[1] "    XXXXX     model SL: simple linear model:     XXXXX     "
glm(formula = disab ~ age5 + Diseases - 1, family = Add.haz(ndis),
    data = Crossdat, weights = Crossdat$wgt)
[1] "     parameters:"
               Estimate Std. Error t value  CIlow    CIhigh p value
age550           0.1202    0.03416      NA  0.06694 0.1996      NA
age555           0.1191    0.03335      NA  0.06694 0.1965      NA
age560           0.1334    0.03793      NA  0.07429 0.2216      NA
age565           0.2329    0.05423      NA  0.14455 0.3558      NA
age570           0.3012    0.07093      NA   0.18593 0.4623     NA
age575           0.3340    0.08540      NA   0.19756 0.5299     NA
age580           0.6659    0.13618      NA  0.43877 0.9701      NA
age585           1.1343    0.25245      NA  0.71959 1.7037      NA
Diseasesheart    0.7641    0.17884   4.273  0.41361 1.1147   0.000
Diseasesstroke   0.7850    0.35844   2.190  0.08243 1.4875   0.029
Diseasescancer   0.4262    0.13336   3.196  0.16484 0.6876   0.001
Diseasescopd     0.1624    0.09712   1.672  0.00000 0.3527   0.095
Diseasesdiabetes 0.6952    0.14880   4.672  0.40355 0.9868   0.000
Diseasesmuskulo  0.4398    0.05388   8.163  0.33424 0.5455   0.000
Diseasesother    0.3800    0.07376   5.152  0.23544 0.5246   0.000
[1] "     deviance and df:"
[1] 1490.97
[1] 1358
```

And finally the attribution tables:

```
[1] "> > > OUTPUT PART 5: ATTRIBUTION OF DISABILITY BY DISEASE < < <"

[1] "attribution of diseases:"
              50      55       60      65      70      75      80      85
nn       183.346 243.013 212.2369 186.645 202.062 140.755 142.944 61.9988
disab     51.968  84.031  78.5461  83.909 115.816  81.486 106.523 50.1649
backgrnd  18.597  23.309  22.3606  32.236  40.080  30.813  50.653 33.5817
heart      1.839   5.660   5.5544   5.306   9.039   6.905  12.065  2.7869
stroke     1.070   1.147   0.6353   1.341   4.242   2.302   2.089  0.0000
cancer     1.708   4.637   3.3836   4.447   3.809   2.169   1.860  1.5533
copd       1.753   2.475   1.4183   1.636   2.153   1.414   1.292  0.5957
diabetes   2.384   5.209   7.9318   7.710  13.142   7.856   8.623  1.0075
muskulo   15.781  28.789  25.6613  25.614  31.451  21.961  18.605  7.4829
other      8.835  12.805  11.6009   5.620  11.900   8.067  11.336  3.1570

[1] "as fraction of disabled:"
              50      55       60      65      70      75      80      85
backgrnd 0.35786 0.27739 0.284681 0.38417 0.34606 0.37814 0.47551 0.66943
heart    0.03539 0.06736 0.070715 0.06323 0.07804 0.08474 0.11326 0.05555
stroke   0.02058 0.01365 0.008088 0.01598 0.03663 0.02825 0.01962 0.00000
cancer   0.03286 0.05518 0.043078 0.05300 0.03289 0.02662 0.01746 0.03096
copd     0.03373 0.02946 0.018056 0.01950 0.01859 0.01736 0.01213 0.01187
diabetes 0.04588 0.06199 0.100983 0.09188 0.11347 0.09641 0.08095 0.02008
muskulo  0.30367 0.34259 0.326704 0.30526 0.27156 0.26950 0.17466 0.14917
other    0.17001 0.15238 0.147696 0.06697 0.10275 0.09899 0.10642 0.06293
[1] "as fraction of total population:"
               50       55       60       65      70      75      80       85
backgrnd 0.101433 0.095918 0.105357 0.172711 0.19835 0.21891 0.35435 0.541651
heart    0.010032 0.023291 0.026171 0.028428 0.04473 0.04906 0.08440 0.044950
stroke   0.005834 0.004719 0.002993 0.007186 0.02099 0.01635 0.01462 0.000000
cancer   0.009314 0.019082 0.015943 0.023827 0.01885 0.01541 0.01301 0.025054
copd     0.009561 0.010186 0.006682 0.008765 0.01066 0.01005 0.00904 0.009608
diabetes 0.013004 0.021435 0.037372 0.041307 0.06504 0.05581 0.06032 0.016251
muskulo  0.086074 0.118466 0.120909 0.137233 0.15565 0.15602 0.13016 0.120694
other    0.048188 0.052691 0.054660 0.030109 0.05889 0.05731 0.07930 0.050920
[1] "  in ( V:\\PRJCT\\eushare\\caspar\\attribhandleiding\\attribit.mrc )"
[1] "V:\\PRJCT\\eushare\\caspar\\attribhandleiding\\attribit.out"
```

The illustration of the model selection describes in more detail the interpretation of the output.

If bootstrap is chosen, the attributions (as fraction of the disabled and as fraction of the total population will be presented as follows.

```
[1] "number of replicas 100"
[1] "   attributions as fraction of the disabled"
$`50`
            backgrnd 50 heart 50 stroke 50 cancer 50   copd 50 diabetes 50 muskulo 50 other 50
original      0.1095 0.008928  0.005693  0.008647  0.008362    0.011689    0.07908  0.04301
bootstrap mean0.1094 0.008739  0.005302  0.009109  0.007755    0.011445    0.07888  0.04495
2.5% lower    0.1063 0.001508  0.000000  0.001795 -0.001757    0.003324    0.05317  0.02872
97.5% upper    .1122 0.021127  0.011145  0.019585  0.019960    0.023533    0.11301  0.06548

$`55`
            backgrnd 55 heart 55 stroke 55 cancer 55   copd 55 diabetes 55 muskulo 55 other 55
original      0.1068  0.02079  0.004651   0.01788  0.008876     0.01919    0.10924  0.04718
bootstrap mean0.1065  0.02204  0.004552   0.01747  0.009595     0.01915    0.10791  0.05006
2.5% lower    0.1029  0.00872  0.001153   0.00704 -0.001386     0.01091    0.07723  0.03015
97.5% upper   0.1099  0.03787  0.009613   0.03420  0.024347     0.03024    0.13576  0.07010

Etc.
[1] "number of replicas 100"
[1] "   attributions as fraction of the population"
$`50`
          backgrnd 50 heart 50 stroke 50 cancer 50   copd 50 diabetes 50 muskulo 50 other 50
original       0.3983  0.03248   0.02071  0.031455  0.030416     0.04252     0.2877   0.1565
bootstrap mean 0.3994  0.03130   0.01942  0.033113  0.027816     0.04130     0.2849   0.1627
2.5% lower     0.3387  0.00543   0.00000  0.006401 -0.006832     0.01224     0.2172   0.1071
97.5% upper    0.4697  0.07231   0.04349  0.071166  0.072864     0.08379     0.3745   0.2292
```

```
$`55`
         backgrnd 55 heart 55 stroke 55 cancer 55   copd 55 diabetes 55 muskulo 55 other 55
original      0.3193  0.06212  0.013899   0.05343  0.026522     0.05734     0.3264   0.1410
bootstrap mean0.3174  0.06500  0.013555   0.05175  0.028093     0.05679     0.3193   0.1481
2.5% lower    0.2711  0.02781  0.003386   0.02197 -0.004082     0.03338     0.2524   0.0996
97.5% upper   0.3716  0.10761  0.027847   0.09520  0.068954     0.09369     0.3901   0.1993


Etc.
```

If bootstrap is chosen, and a RRR model is selected, the bootstrap will also provide confidence intervals for the age-specific disabling impacts (=age specific disease hazards):

```
          CI's of age specific disabling impacts:


              Diseasesheart 50-64 Diseasesstroke 50-64 Diseasescancer 50-64 Diseasescopd 50-64
original                   0.6324               0.7468               0.3896            0.13844
bootstrap mean             0.6978               0.8053               0.4158            0.14717
2.5% lower                 0.4111               0.2417               0.1741           -0.02559
97.5% upper                1.1247               1.4657               0.7208            0.38678
               Diseasesdiabetes 50-64 Diseasesmuskulo 50-64 Diseasesother 50-64
original                   0.5844                0.3952              0.3300
bootstrap mean             0.5946                0.4003              0.3501
2.5% lower                 0.3409                0.2959              0.2114
97.5% upper                0.8490                0.5142              0.4947
              Diseasesheart 65-79 Diseasesstroke 65-79 Diseasescancer 65-79 Diseasescopd 65-79
original                   0.7455               0.8804               0.4593            0.16320
bootstrap mean             0.8064               0.9319               0.4810            0.16580
2.5% lower                 0.4264               0.3335               0.2022           -0.02621
97.5% upper                1.2530               1.7965               0.9200            0.43938
             Diseasesdiabetes 65-79 Diseasesmuskulo 65-79 Diseasesother 65-79 +
original                   0.6889                0.4659              0.3890
bootstrap mean             0.6953                0.4657              0.4110
2.5% lower                 0.3467                0.3007              0.2246
97.5% upper                1.1148                0.6337              0.6676


Etc.
```

3. Illustration of features of the tool

3.1 Illustration of model selection

To illustrate the model selection approach, we first distinguish the situation of no distinction between populations and distinction between two populations. The complete output is given in appendix 5.

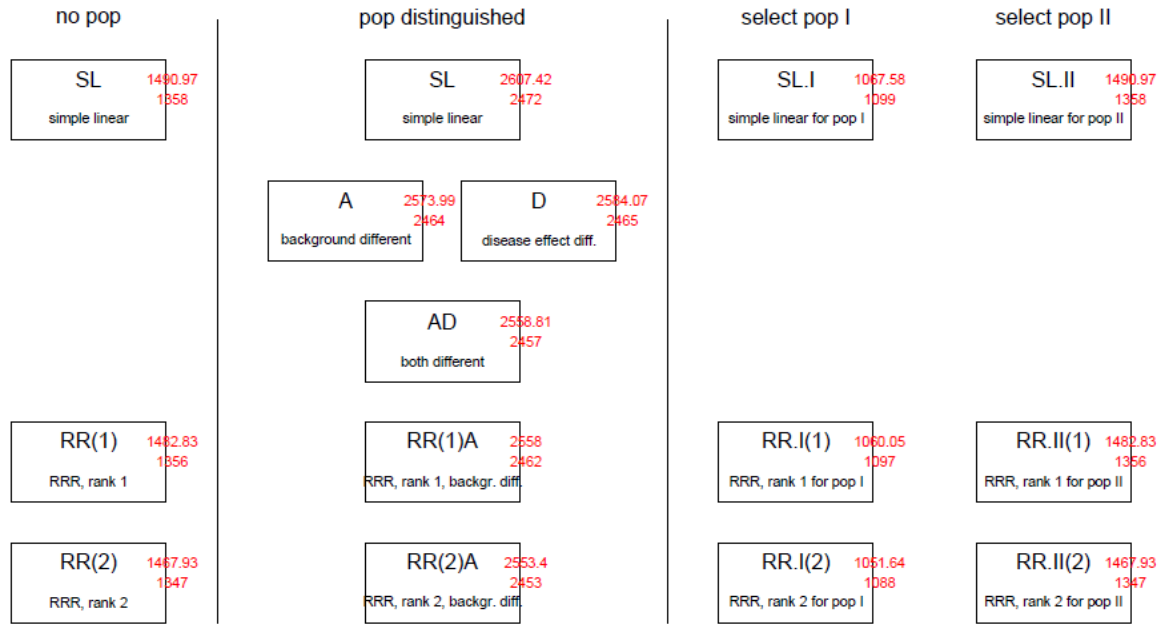


Figure 4: Example of model selection

Figure 4 is similar to Figure 3 with possible model selections, except that now scaled deviances (upper number) and degrees of freedom (lower number) are added that originate from the examples based on Italian women (no population distinguished) and Italian men and women (population distinguished). In the remainder of this section all relevant tests for significance are explained. All p-values can be found in the output and are explained here in more detail only for clarification.

3.1.1 No distinction between populations

This is the situation where the user is interested in analysing disability in a single population. As example we analyse disability in Italian women. The first column presents the scaled deviances and degrees of freedom in the analysis for Italian women. The model selection always starts with a simple linear regression “SL”, i.e., no differences between populations and no interaction between age and diseases (one disabling impact for all ages). For Italian women the deviance of the simple linear model is 1490.97 at 1358 degrees of freedom.

Table 2 Outcomes SL model Italian women

	Estimate	Std. Error	t value	Pr(>|t|)	Cilow	CIhigh	p value
Age5_50-54	0.1202	0.03416	NA	NA	0.06694	0.1996	NA
Age5_55-59	0.1191	0.03335	NA	NA	0.06694	0.1965	NA
Age5_60-64	0.1334	0.03793	NA	NA	0.07429	0.2216	NA
Age5_65-69	0.2329	0.05423	NA	NA	0.14455	0.3558	NA
Age5_70-74	0.3012	0.07093	NA	NA	0.18593	0.4623	NA
Age5_75-79	0.334	0.0854	NA	NA	0.19756	0.5299	NA
Age5_80-84	0.6659	0.13618	NA	NA	0.43877	0.9701	NA
Age5_85+	1.1343	0.25245	NA	NA	0.71959	1.7037	NA
Diseases heart	0.7641	0.17884	4.273	2.07E-05	0.41361	1.1147	0
Diseases stroke	0.785	0.35844	2.19	2.87E-02	0.08243	1.4875	0.029
Diseases cancer	0.4262	0.13336	3.196	1.43E-03	0.16484	0.6876	0.001
Diseases copd	0.1624	0.09712	1.672	9.47E-02	0	0.3527	0.095
Diseases diabetes	0.6952	0.1488	4.672	3.28E-06	0.40355	0.9868	0
Diseases muskulo	0.4398	0.05388	8.163	7.36E-16	0.33424	0.5455	0
Diseases other	0.38	0.07376	5.152	2.96E-07	0.23544	0.5246	0

The model includes background disability by 5-year age groups. To provide more clarity, we edited the labels of the 5-year age groups. Normally the program gives the values labels. As expected the background disability hazard increases with age. The hazards for the disabling impact of the diseases do not vary by age in this model. Diseases with the highest disabling impacts are: stroke, heart diseases and diabetes in this population. The disabling effect of COPD is not significant in this model.

After running this first model in our model selection approach, the user needs to assess whether the disabling impacts differ by age. In other words, whether there is an interaction between age and diseases. This is operationalized with reduced rank regression (RRR).

The deviance of the RRR model with one axis RR(1) is 1482.83 at 1356 degrees of freedom. The difference between 1358 and 1356 can be explained by the gamma's of the RRR-model (the number of delta's will equal the number of original beta's in the simple linear model). We used the variable AGEC for the age pattern of the disabling which consists of three age classes. As the mean of the gamma's equals 1, adding AGEC costs two degrees of freedom extra.

Table 3 Outcomes model “RR(1)” for Italian women

	Estimate	Std. Error	t value	Cilow	Cihigh	p value
age5_50-54	0.1271	0.03496	NA	0.07217	0.208	NA
age5_55-59	0.1285	0.03425	NA	0.07432	0.2075	NA
age5_60-64	0.1449	0.03907	NA	0.08321	0.2351	NA
age5_65-69	0.229	0.05426	NA	0.14098	0.3524	NA
age5_70-74	0.2886	0.0704	NA	0.17503	0.4492	NA

age5_75-79	0.3245	0.08515	NA	0.18928	0.5205	NA
age5_80-84	0.4141	0.12652	NA	0.22048	0.7112	NA
age5_85+	0.8793	0.23937	NA	0.50234	1.4328	NA
delta.aheart	0.8714	0.20195	4.315	0.47562	1.2672	0
delta.astroke	1.0126	0.43912	2.306	0.15192	1.8733	0.021
delta.acancer	0.5363	0.16422	3.266	0.21441	0.8582	0.001
delta.acopd	0.1885	0.12019	1.568	-0.04708	0.4241	0.117
delta.adiabetes	0.7987	0.17414	4.586	0.45737	1.14	0
delta.amuskulo	0.5481	0.06544	8.375	0.41985	0.6764	0
delta.aother	0.4582	0.08911	5.142	0.28355	0.6329	0
Gamma.ageC50-64	0.7286	0.07811	9.328	0.5755	0.8817	0
Gamma.ageC65-79	0.8493	0.09381	9.054	0.66545	1.0332	0
Gamma.ageC80+	1.4221	0.24249	5.865	0.94682	1.8974	0

5-year age groups are used for background in the model. Now next to background, also the disabling impact varies by age, using RRR with one axis. The gamma (age pattern disabling impact "gamma.ageC50-64", etc.) gives the common age pattern, showing an increase with increasing age. The delta (disabling impact "delta.aheart", etc.) gives the overall disabling impact. The disabling impact by age is the product of gamma and delta (table 4).

Table 4 Parameters of the RR(1) model for Italian women

	Delta	gamma.ageC50-64	gamma.ageC65-79	gamma.ageC80+
Gamma	NA	0.7286	0.8493	1.4221
delta.aheart	0.8714	0.8714*0.7286=0.6349	0.7401	1.2393
delta.astroke	1.0126	0.7378	0.86	1.44
delta.acancer	0.5363	0.3907	0.4555	0.7626
delta.acopd	0.1885	0.1373	0.1601	0.2681
delta.adiabetes	0.7987	0.5819	0.6783	1.1358
delta.amuskulo	0.5481	0.3994	0.4655	0.7795
delta.aother	0.4582	0.3339	0.3892	0.4582*1.4221=0.6516

The likelihood ratio test indicates whether this model with an interaction between AGEC and diseases has a better fit than the model without such an interaction. The difference in deviance is 1490.97-1482.83=8.14, and the Chi^2 distribution for this difference at two degrees of freedom results in a p-value of 0.017. This means that the disabling impacts indeed differ by age. Because by choosing gamma to have a mean of 1 the value of the delta can be compared with the original disabling impacts. Differences are a result of unequal prevalences of diseases in age classes.

The next step is to examine whether the age patterns are equal for all diseases. This can be tested by adding a second axis to the RRR model ("RR(2)"). We compared RR(2) with a scaled deviance of 1467.93 and 1490 degrees of freedom, with model RR(1) with scaled deviance of 1482.83 and 1365 degrees of freedom. As indicated in the output ("p-value for the second axis is (RR(2) vs. RR(1)), this yields a p-value of 0.094, indicating that the age patterns for all diseases do not significantly differ and that the second axis is not necessary.

Table 5 Outcomes model “RR(2)” for Italian women

	Estimate	Std. Error	t value	Cilow	CIhigh	p value
age5_50-54	0.124034	0.0347	NA	0.06975	0.20456	NA
age5_55-59	0.118362	0.0333	NA	0.06634	0.1957	NA
age5_60-64	0.136519	0.03838	NA	0.07654	0.22566	NA
age5_65-69	0.241401	0.05523	NA	0.15114	0.36639	NA
age5_70-74	0.299493	0.07167	NA	0.18341	0.46257	NA
age5_75-79	0.324525	0.08521	NA	0.18923	0.52067	NA
age5_80-84	0.411819	0.12264	NA	0.22291	0.69885	NA
age5_85+	0.872042	0.23728	NA	0.49835	1.42076	NA
delta.aheart	0.871431	0.20195	4.31515	0.47562	1.26725	0
delta.astroke	1.012596	0.43912	2.30596	0.15192	1.87327	0.021
delta.acancer	0.536285	0.16422	3.26565	0.21441	0.85816	0.001
delta.acopd	0.188504	0.12019	1.56832	-0.04708	0.42409	0.117
delta.adiabetes	0.798682	0.17414	4.58647	0.45737	1.13999	0
delta.amuskulo	0.548123	0.06544	8.37536	0.41985	0.6764	0
delta.aother	0.458215	0.08911	5.1419	0.28355	0.63288	0
gamma.ageC50-64	0.72859	0.07811	9.32802	0.5755	0.88168	0
gamma.ageC65-79	0.849312	0.09381	9.054	0.66545	1.03317	0
gamma.ageC80+	1.422097	0.24249	5.86456	0.94682	1.89738	0
delta.a2heart	0.068146	0.13278	0.51324	-0.19209	0.32839	0.608
delta.a2stroke	-0.704	0.0284	-24.7871	-0.75967	-0.64833	0
delta.a2cancer	0.091681	0.10196	0.89915	-0.10817	0.29153	0.369
delta.a2copd	0.13026	0.07764	1.67766	-0.02192	0.28244	0.093
delta.a2diabetes	-0.24179	0.10711	-2.25738	-0.45172	-0.03185	0.024
delta.a2muskulo	0.003168	0.04099	0.07729	-0.07718	0.08351	0.938
delta.a2other	0.061231	0.05843	1.048	-0.05328	0.17575	0.295
gamma2:ageC50-64	1.224149	0.04842	25.28007	1.12924	1.31906	0
gamma2:ageC65-79	-1.47438	0.59622	-2.47289	-2.64297	-0.30579	0.013
gamma2:ageC80+	0.301472	1.35042	0.22324	-2.34536	2.9483	0.823

Table 6 Parameters of the RR(2) model for Italian women

	Delta_AX1	Delta_AX2	Gamma.ageC50-64	Gamma.ageC65-79	Gamma.ageC 80+
Gamma_AX1	NA	NA	0.7286	0.84931	1.4221
Gamma_AX2	NA	NA	1.2241	-1.47438	0.3015
delta.aheart	0.8714	0.068146	0.7286*0.8714+1.2241*0.0681 =0.7183	0.63964	1.2598
delta.astroke	1.0126	-0.704	-0.124	1.89797	1.2278
delta.acancer	0.5363	0.091681	0.503	0.3203	0.7903
delta.acopd	0.1885	0.13026	0.2968	-0.03195	0.3073
delta.adiabetes	0.7987	-0.24179	0.2859	1.03482	1.0629
delta.amuskulo	0.5481	0.003168	0.4032	0.46086	0.7804
delta.aother	0.4582	0.061231	0.4088	0.29889	0.6701

3.1.2 Distinction between populations

This is the situation where the user is interested in making comparisons between populations. We use here Italian men (population I) and Italian women (population II). Again we start with the simplest model without difference between populations: the "SL" model. For this model the deviance is 2707.02 with 2472 degrees of freedom. Note that the number of degrees of freedom is more or less double as compared to the first column, reflecting the fact that the dataset of 1373 women is augmented with almost the same number of men.

First step is to test whether there are differences between populations and second, whether there is an interaction between age and diseases (disabling impact varies with age), and if so, whether one axis is sufficient to summarize the age variation.

To test for differences between populations, the first step is testing for differences in background hazards between the two populations. To do so we compare model "A" with different background hazards for men and women, with model "SL" with the same background disability for men and women.

Table 7 Output model "A" for situation with two populations

	Estimate	Std. Error	t value	Pr(>|t|)	Cilow	Cihigh	p value
age5_I_50-54	0.01317	0.01124	NA	NA	0.002352	0.0427	NA
age5_I_55-59	0.06101	0.02217	NA	NA	0.028751	0.1144	NA
age5_I_60-64	0.08394	0.02906	NA	NA	0.041014	0.1534	NA
age5_I_65-69	0.18126	0.0499	NA	NA	0.10289	0.2968	NA
age5_I_70-74	0.27705	0.06569	NA	NA	0.170453	0.4263	NA
age5_I_75-79	0.15177	0.07447	NA	NA	0.054823	0.3386	NA
age5_I_80-84	0.49367	0.13995	NA	NA	0.275411	0.8191	NA
age5_I_85+	0.13796	0.10591	NA	NA	0.028825	0.4155	NA
age5_II-50-54	0.127	0.03492	NA	NA	0.072147	0.2079	NA
age5_II_55-59	0.12476	0.03365	NA	NA	0.071654	0.2025	NA
age5_II_60-64	0.14048	0.03833	NA	NA	0.080143	0.2292	NA
age5_II_65-69	0.24628	0.05487	NA	NA	0.156152	0.3701	NA
age5_II_70-74	0.34555	0.07244	NA	NA	0.225241	0.5078	NA
age5_II_75-79	0.37706	0.08738	NA	NA	0.234619	0.5751	NA
age5_II_80-84	0.71415	0.13568	NA	NA	0.485071	1.0148	NA
age5_II_85+	1.16508	0.25458	NA	NA	0.745403	1.7381	NA
Diseases_heart	0.82727	0.111	7.453	1.26E-13	0.609708	1.0448	0
Diseases_stroke	0.96924	0.23541	4.117	3.96E-05	0.507828	1.4307	0
Diseases_cancer	0.54506	0.11767	4.632	3.81E-06	0.314434	0.7757	0
Diseases_copd	0.25364	0.07093	3.576	3.56E-04	0.114625	0.3927	0
Diseases_diabetes	0.39258	0.07382	5.318	1.14E-07	0.247898	0.5373	0
Diseases_muskulo	0.39227	0.04224	9.286	3.41E-20	0.309472	0.4751	0
Diseases_other	0.3103	0.04556	6.81	1.22E-11	0.220997	0.3996	0

Background is modelled with variable AGE5, consisting of 8 classes. Using separate background hazards for men (population I) and women (population II) costs 8-1 degrees of freedom (2472-2464).

Together with the drop in deviance of 33.43 this yields a p-value of <0.001 in a chi-square test, indicating that the background hazards for men and women are different. This is reflected in the parameters for background, which show higher background disability risks in population II (women).

The second step is testing whether the disabling impact of the diseases differ by population. Depending on whether there is a difference in background hazards between the populations we have to compare model "D" vs. model "SL" (for situation of no differences in background hazards) or model "AD" vs. model "A" (for situation with differences in background hazards). Given that the likelihood ratio test indicated differences in background hazards, we compare model "AD" vs. "A", showing a drop in deviance of 15.18 at the cost of seven degrees of freedom (p=0.001). This means that we have to differentiate background hazards as well as disease impacts by population.

Table 8 Outcomes Model "AD" for situation with two populations

	Estimate	Std. Error	t value	Pr(>|t|)	Cilow	CIhigh	P value
age5_I_50-54	0.01334	0.01131	NA	NA	0.002405	0.04306	NA
age5_I_55-59	0.06701	0.02322	NA	NA	0.032725	0.12249	NA
age5_I_60-64	0.09093	0.0302	NA	NA	0.045778	0.16268	NA
age5_I_65-69	0.18972	0.05074	NA	NA	0.109491	0.30679	NA
age5_I_70-74	0.30343	0.06795	NA	NA	0.191925	0.4568	NA
age5_I_75-79	0.16557	0.07665	NA	NA	0.063342	0.35638	NA
age5_I_80-84	0.51251	0.1422	NA	NA	0.289599	0.84218	NA
age5_I_85+	0.15885	0.11092	NA	NA	0.037867	0.44791	NA
age5_II-50-54	0.12018	0.03432	NA	NA	0.066762	0.20005	NA
age5_II_55-59	0.11911	0.0335	NA	NA	0.066759	0.19692	NA
age5_II_60-64	0.1334	0.0381	NA	NA	0.074085	0.22207	NA
age5_II_65-69	0.23286	0.05447	NA	NA	0.144232	0.35645	NA
age5_II_70-74	0.30119	0.07125	NA	NA	0.18552	0.46306	NA
age5_II_75-79	0.33397	0.08579	NA	NA	0.197087	0.53089	NA
age5_II_80-84	0.66592	0.13679	NA	NA	0.437926	0.97162	NA
age5_II_85+	1.13433	0.25359	NA	NA	0.718078	1.70659	NA
Diseases_heart.I	0.8488	0.14094	6.022	1.98E-09	0.572552	1.12506	0
Diseases_stroke.I	1.10817	0.31277	3.543	4.03E-04	0.495146	1.7212	0
Diseases_cancer.I	0.71457	0.22346	3.198	1.40E-03	0.276584	1.15256	0.001
Diseases_copd.I	0.36066	0.10406	3.466	5.37E-04	0.156704	0.56461	0.001
Diseases_diabetes.I	0.25696	0.08026	3.202	1.38E-03	0.099657	0.41427	0.001
Diseases_muskulo.I	0.30242	0.06461	4.681	3.01E-06	0.175786	0.42906	0
Diseases_other.I	0.25354	0.05605	4.524	6.37E-06	0.143686	0.3634	0
Diseases_heart.II	0.76415	0.17965	4.254	2.18E-05	0.412043	1.11625	0
Diseases_stroke.II	0.78497	0.36004	2.18	2.93E-02	0.079283	1.49065	0.029
Diseases_cancer.II	0.42623	0.13396	3.182	1.48E-03	0.163669	0.68879	0.001
Diseases_copd.II	0.16239	0.09755	1.665	9.61E-02	0	0.3536	0.096
Diseases_diabetes.II	0.69519	0.14947	4.651	3.48E-06	0.402238	0.98815	0
Diseases_muskulo.II	0.43985	0.05412	8.127	6.90E-16	0.333769	0.54593	0
Diseases_other.II	0.38002	0.07409	5.129	3.14E-07	0.234794	0.52524	0

The outcomes of model AD for instance show that the disabling impact of stroke is higher in men, while that of musculoskeletal is higher in women. These outcomes do not show directly whether

specific disease impacts differ significantly between the two populations. This can be assessed by using a t-statistic, which equals for instance (Estimate Diseases.stroke.I_-Estimate diseases.stroke.IIbeta_V)/sqrt[(Std. Error Diseases.stroke.I)^2+ (Std. Error Diseases.stroke.II) ^2)] based on normal distribution (value > 1.96 means significant difference). Remind that this procedure is only possible when comparing parameters from different populations (e.g. men vs. women). Comparing disabling impacts from the same population requires to take into account the covariance as well.

Given that the likelihood ratio test indicated that both the background and disabling impacts differ between the two populations, we run separate models for men and women.

The third and last step is to assess the significance for possible interaction between age and disease for Italian men and women together, but using separate models for men (population I) and women (population II). Interactions are tested among men and women together using the likelihood ratio test, because otherwise, we may find model specifications to differ for men and women, e.g. disabling impacts do not differ by age for men (RRR not needed), but do differ for women (RRR needed). To avoid such a situation, the models for men and women can be compared together. The deviances and degrees of freedom for men and women together can be calculated by adding the scaled deviances and degrees of freedom from the separate models for men (population I) and women (population II) (Figure 1, column 3). If different interactions between age and disease is not an issue, the user can select the models independently (Figure 3, column 1), for the separate populations.

For men, population I, we run the SL.I model and the model with RRR, 1 axis (RR.I(1). We do the same for women, population II, and run SL.II, (which is the same as SL model in the situation where we assessed only women separately)) and RR.II(1). We then compare the sum of the scaled deviance and degrees of freedom based on the two SL models vs. the sum based on the two RR (1) models. This last calculation of a p value cannot be found in the output and is something to be executed by the user, for instance in Excel. Based on the output, it can be seen that the scaled deviance of RR.I(1) is 1060.05 with 1097 degrees of freedom. For RR.II(1) this is 1482.83 and 1356, yielding a total of 2542.88 and 2453, as compared to the total scaled deviance of 2558.55 with 2457 degree of freedom for SL.I and SL.II added. This yields a p-value of 0.0035. Hence the disabling impacts are not the same across age.

Next, we ran and compared (RR.I(2)+ (RR.II(2)). The sum of the scaled deviances was: 2519.57 with 2435 degrees of freedom. As compared to (RR.I(1)+ (RR.II(1)), this yielded a p-value of 0.18. The conclusion is that a second axis is not needed. So the final model is: separate model for men and women, and in each of these models disabling impact based on RRR with one axis.

If the dataset is large, it could be that a more complex model has a significant better fit, but that the improvement is judged not to be relevant, because differences between the parameters of the two populations are small. As for such situations no generally applicable rules can be provided, decisions hereabouts are left to the investigator.

3.1.2 Confidence intervals using bootstrapping

For the illustration we used 100 replicas but higher numbers should be used. We suggest 1000.

If bootstrap is chosen and a RRR model is selected, the bootstrap will also provide confidence intervals for the age-specific disabling impacts (=age specific disease hazards):

```
          CI's of age specific disabling impacts:


              Diseasesheart 50-64 Diseasesstroke 50-64 Diseasescancer 50-64 Diseasescopd 50-64
original                   0.6324               0.7468               0.3896            0.13844
bootstrap mean             0.6978               0.8053               0.4158            0.14717
2.5% lower                 0.4111               0.2417               0.1741           -0.02559
97.5% upper                1.1247               1.4657               0.7208            0.38678
               Diseasesdiabetes 50-64 Diseasesmuskulo 50-64 Diseasesother 50-64
original                   0.5844                0.3952              0.3300
bootstrap mean             0.5946                0.4003              0.3501
2.5% lower                 0.3409                0.2959              0.2114
97.5% upper                0.8490                0.5142              0.4947
              Diseasesheart 65-79 Diseasesstroke 65-79 Diseasescancer 65-79 Diseasescopd 65-79
original                   0.7455               0.8804               0.4593            0.16320
bootstrap mean             0.8064               0.9319               0.4810            0.16580
2.5% lower                 0.4264               0.3335               0.2022           -0.02621
97.5% upper                1.2530               1.7965               0.9200            0.43938
             Diseasesdiabetes 65-79 Diseasesmuskulo 65-79 Diseasesother 65-79
original                   0.6889                0.4659              0.3890
bootstrap mean             0.6953                0.4657              0.4110
2.5% lower                 0.3467                0.3007              0.2246
97.5% upper                1.1148                0.6337              0.6676


Etc.
```

The last parameter Diseaseother 65-79 shows the danger of choosing the number of bootstraps too low. The bootstrap mean should be almost equal to the original estimate, but here is 0.4110 instead of 0.3890. For any model with bootstrapping the attributions (as fraction of the disabled and as fraction of the total population) are presented as follows:

```
[1] "number of replicas 100"
[1] "   attributions as fraction of the disabled"
$`50`
            backgrnd 50 heart 50 stroke 50 cancer 50   copd 50 diabetes 50 muskulo 50 other 50
original      0.1095 0.008928  0.005693  0.008647  0.008362    0.011689    0.07908  0.04301
bootstrap mean0.1094 0.008739  0.005302  0.009109  0.007755    0.011445    0.07888  0.04495
2.5% lower    0.1063 0.001508  0.000000  0.001795 -0.001757    0.003324    0.05317  0.02872
97.5% upper    .1122 0.021127  0.011145  0.019585  0.019960    0.023533    0.11301  0.06548

$`55`
            backgrnd 55 heart 55 stroke 55 cancer 55   copd 55 diabetes 55 muskulo 55 other 55
original      0.1068  0.02079  0.004651   0.01788  0.008876     0.01919    0.10924  0.04718
bootstrap mean0.1065  0.02204  0.004552   0.01747  0.009595     0.01915    0.10791  0.05006
2.5% lower    0.1029  0.00872  0.001153   0.00704 -0.001386     0.01091    0.07723  0.03015
```

```
97.5% upper   0.1099  0.03787  0.009613   0.03420  0.024347     0.03024    0.13576  0.07010

Etc.


[1] "number of replicas 100"
[1] "   attributions as fraction of the population"
$`50`
          backgrnd 50 heart 50 stroke 50 cancer 50   copd 50 diabetes 50 muskulo 50 other 50
original       0.3983  0.03248   0.02071  0.031455  0.030416     0.04252     0.2877   0.1565
bootstrap mean 0.3994  0.03130   0.01942  0.033113  0.027816     0.04130     0.2849   0.1627
2.5% lower     0.3387  0.00543   0.00000  0.006401 -0.006832     0.01224     0.2172   0.1071
97.5% upper    0.4697  0.07231   0.04349  0.071166  0.072864     0.08379     0.3745   0.2292

$`55`
          backgrnd 55 heart 55 stroke 55 cancer 55   copd 55 diabetes 55 muskulo 55 other 55
original       0.3193  0.06212  0.013899   0.05343  0.026522     0.05734     0.3264   0.1410
bootstrap mean0.3174  0.06500  0.013555   0.05175  0.028093     0.05679     0.3193   0.1481
2.5% lower     0.2711  0.02781  0.003386   0.02197 -0.004082     0.03338     0.2524   0.0996
97.5% upper    0.3716  0.10761  0.027847   0.09520  0.068954     0.09369     0.3901   0.1993


Etc.
```

3.2 Illustrations of additional options of the tool

3.2.1. Observed vs. fitted prevalences

First we show the outcomes when fitted values are used in the attribution table. It is noteworthy that the fitted values are not the same as those in the attribution table because we used sample weights.

```
[1] "fitted numbers will be used for the attributions table"


[1] "> > > OUTPUT PART 5: ATTRIBUTION OF DISABILITY BY DISEASE < < <"

[1] "attribution of diseases:"
              50      55       60      65      70      75      80      85
nn       183.346 243.013 212.2369 186.645 202.062 140.755 142.944 61.9988
disab     51.968  84.031  78.5461  83.909 115.816  81.486 106.523 50.1649
backgrnd  18.597  23.309  22.3606  32.236  40.080  30.813  50.653 33.5817
heart      1.839   5.660   5.5544   5.306   9.039   6.905  12.065  2.7869
stroke     1.070   1.147   0.6353   1.341   4.242   2.302   2.089  0.0000
cancer     1.708   4.637   3.3836   4.447   3.809   2.169   1.860  1.5533
copd       1.753   2.475   1.4183   1.636   2.153   1.414   1.292  0.5957
diabetes   2.384   5.209   7.9318   7.710  13.142   7.856   8.623  1.0075
muskulo   15.781  28.789  25.6613  25.614  31.451  21.961  18.605  7.4829
other      8.835  12.805  11.6009   5.620  11.900   8.067  11.336  3.1570
```

Next we show the outcomes when observed values are used in the attribution table. The observed values are not the same as those in the attribution table because we used sample weights.

```
[1] "observed numbers will be used for the attributions table"
```

```
              50       55       60       65       70       75       80       85
nn       183.346 243.013 212.2369 186.645 202.062 140.755 142.944 61.9988
disab     52.018  79.647  75.9581  80.210 118.652  83.538 108.005 51.0436
backgrnd  18.615  22.093  21.6238  30.814  41.061  31.589  51.358 34.1699
heart      1.841   5.365   5.3713   5.072   9.260   7.079  12.233  2.8357
stroke     1.071   1.087   0.6143   1.282   4.346   2.360   2.119  0.0000
cancer     1.709   4.395   3.2721   4.251   3.902   2.224   1.886  1.5805
copd       1.755   2.346   1.3715   1.564   2.206   1.450   1.310  0.6061
diabetes   2.387   4.937   7.6704   7.370  13.464   8.054   8.743  1.0252
muskulo   15.797  27.287  24.8158  24.485  32.221  22.514  18.864  7.6139
other      8.844  12.137  11.2187   5.372  12.191   8.270  11.494  3.2123
```

3.2.2 Machine readable output file

Automatically a machine readable output file, with extension “mrc” is produced that can be read by the decomposition program.

```
              50       55       60       65       70       75       80       85
nn       183.346 243.013 212.2369 186.645 202.062 140.755 142.944 61.9988
disab     51.968  84.031  78.5461  83.909 115.816  81.486 106.523 50.1649
backgrnd  18.597  23.309  22.3606  32.236  40.080  30.813  50.653 33.5817
heart      1.839   5.660   5.5544   5.306   9.039   6.905  12.065  2.7869
stroke     1.070   1.147   0.6353   1.341   4.242   2.302   2.089  0.0000
cancer     1.708   4.637   3.3836   4.447   3.809   2.169   1.860  1.5533
copd       1.753   2.475   1.4183   1.636   2.153   1.414   1.292  0.5957
diabetes   2.384   5.209   7.9318   7.710  13.142   7.856   8.623  1.0075
muskulo   15.781  28.789  25.6613  25.614  31.451  21.961  18.605  7.4829
other      8.835  12.805  11.6009   5.620  11.900   8.067  11.336  3.1570
```

nn is the number of persons, disab is the number of disabled persons and background up to other is the number of disabled cases attributed to background up to other.

For example, table 9 gives a few persons classified by disease presence, age 50-54, as well as linear predictor (linp), fitted value (fv) and wgt (sample weight). This table can be made by entering maketable9() in the console window after running the program.

Table 9 Persons classified by disease presence, age 50-54, as well as linear predictor (linp) and fitted value (fv) and wgt (sample weight)

heart	stroke	cancer	copd	diabetes	muskulo	Other	age5	Linp	fv	wgt
1	0	0	0	1	0	0	50	1.5795	0.7939	1.7717
1	0	0	0	0	0	0	50	0.8843	0.587	0.5118
1	0	0	0	0	0	0	50	0.8843	0.587	0.7352
1	0	0	0	0	0	0	50	0.8843	0.587	0.5366

1	0	0	1	1	0	0	50	1.7419	0.8248	0.7023

Table 10 Parameters of additive regression model (SL)

age550	age555	age560	age565	age570	age575	age580	age585
0.1202	0.1191	0.1334	0.2329	0.3012	0.334	0.6659	1.1343
Diseasesheart	Diseasesstroke	Diseasescancer	Diseasescopd	Diseasesdiabetes	Diseasesmuskulo	Diseasesother	
0.7641	0.785	0.4262	0.1624	0.6952	0.4398	0.38	

Here we will explain the calculation of the estimated prevalence of disability by heart of 1.839 in detail, based on table 9 and 10. Table 9 shows that all five persons presented had heart disease in the age class 50-54. Three of them had no other diseases, so their hazard was 0.1202+0.7641=0.8843 (Table 10). The first person in Table 9 had next to heart disease also diabetes, which results in a hazard of 1.5795 (=0.1202+0.7641+0.6952) and the last person had also COPD: resulting in a hazard of 1.7419. The estimated probabilities of being disabled were 0.5870 (e.g. 1-(exp(-0.8843))=0.5870), 0.7939 and 0.8248 respectively. For the second, third and fourth women we now know that the probability of the disability being caused by heart is 0.7641/0.8843, thus 86.4%. For the other two this becomes 48.4% and 43.9%. The weighted sum of these probabilities based on all persons age 50-54 is 1.839, indicating that 1.839 persons aged 50-54 are disabled from heart disease.

It is important to realize that the attribution table always gives the weighted numbers. When sample weights are used, the number of disabled persons in this table will differ from the numbers given in the table with observed vs. expected numbers, which is also given in the output.

3.2.3 Including data on population in institutions

One option of the tool is to include information on the **institutionalised population** from a separate data source in the attribution table.

This example uses data for Dutch women, as we did not have information on institutionalization for Italian women. The "file instellingen NL.csv" includes (for each age and sex) the proportion institutionalized (column 3) and proportion disabled, given institutionalized (column 5). The input specification file and output file are given in appendix 6a. In this example we have one population only (Dutch women), hence we enter the name and the column numbers (as c(3,5)) for the FIRST population only.

```
[1] "file with data about institutions per age class:"
[1] "V:\\PRJCT\\eushare\\caspar\\attribhandleiding\\instellingen NL.csv"
[1] "name of column with fractions institutionalised per age class:"
```

```
[1] "inst.women"
[1] 0.006584 0.006581 0.008968 0.014050 0.038441 0.081068 0.232380 0.389685
[1] "name of column with fractions disabled for people in institution per age class:"
[1] "ong.women"
[1] 1.0000 1.0000 0.9230 0.8045 0.7236 0.6648 0.7279 0.7170
```

This gives the following result:

```
[1] "> > > OUTPUT PART 5: ATTRIBUTION OF DISABILITY BY DISEASE < < <"


[1] "attribution of diseases:"
              50      55      60      65      70      75      80      85
nn       307.021 352.319 253.271 191.693 179.916 125.145 89.8883 96.6713
disab    134.725 163.040 133.973  92.663  90.245  81.057 60.4272 64.0339
insti      2.021   2.319   2.096   2.167   5.004   6.744 15.2039 27.0089
backgrnd  74.425  69.476  58.865  28.644  29.681  28.016 16.5044 11.3208
HEART      2.121   5.763   3.382   4.034   2.919   5.192  2.0211  2.0732
STROKE     1.240   2.834   1.955   2.295   3.059   2.641  1.2043  2.5925
CANCER     2.363   3.106   3.501   2.264   1.816   1.549  0.9695  0.8185
COPD      10.072  15.454   7.852   7.432   3.768   5.098  2.6250  1.8959
DIABETES   3.523   3.834   5.156   4.491   4.584   4.251  1.7456  1.1511
MUSKULO   17.454  31.035  27.057  21.406  22.254  15.893 10.9917  9.9782
OTHER     21.507  29.219  24.108  19.931  17.160  11.672  9.1618  7.1947
[1] "as fraction of disabled:"
```

The numbers of disabled by diseases have not changed, but an extra line is added for the institutionalized. The result without institutionalized starts as follows:

>

[1] "attribution of diseases:"

```
              50      55      60      65      70      75      80      85
nn       305.000 350.000 251.000 189.000 173.000 115.000 69.0000 59.0000
disab    132.704 160.721 131.877  90.496  85.240  74.312 45.2233 37.0250
```

For the last age class (0.3897 institutionalized, from which 0.717 are disabled) the calculations are as follows:

new nn = old nn/(1- fraction institutionalized) : 59/(1-0.3897) = 96.67

insti = new nn * fraction institutionalized * fraction disabled : 96.67*0.3897*0.7170 = 27.01

new disab = old disab + insti : 37.03 + 27.00 = 37.03

3.3 Illustration of model with disease-interactions

The user can test whether hazards for a specific pair of diseases are additive, as assumed in the additive hazard model without interactions, and for instance not multiplicative (proportional). This can be tested by adding an interaction term to the model. The user should then first create a new variable, with 1 indicating a situation when both diseases are present. Here, an illustrative example is given for a situation with interaction of heart disease and stroke:

1: SAMPID2 8:heart 9:stroke 16:heart.stroke

13331	1604200070200	0	1	0
13332	1604200190600	1	0	0
13367	1604201507700	1	1	1
13368	1604201516900	0	0	0

Note that not all columns are shown.

The new variable is named heart.stroke (in R it is conventional to put dots into long names). In the input csv file we add the column number 16 to Q8: 8:14,16 so an extra "disease" will be included to the analysis.

The relevant part of the output is:

	Estimate	Std. Error	t value	Pr(>|t|)	CIlow	CIhigh	p value
age550	0.1202	0.03418	NA	NA	0.06694	0.1997	NA
age555	0.1191	0.03337	NA	NA	0.06694	0.1966	NA
age560	0.1334	0.03795	NA	NA	0.07429	0.2217	NA
age565	0.2330	0.05426	NA	NA	0.14459	0.3560	NA
age570	0.3015	0.07098	NA	NA	0.18612	0.4626	NA
age575	0.3343	0.08547	NA	NA	0.19779	0.5304	NA
age580	0.6657	0.13621	NA	NA	0.43850	0.9699	NA
age585	1.1347	0.25258	NA	NA	0.71974	1.7044	NA
Diseasesheart	0.7578	0.18045	4.1997	2.847e-05	0.40415	1.1115	0.000
Diseasesstroke	0.7557	0.37435	2.0186	4.373e-02	0.02193	1.4894	0.044
Diseasescancer	0.4265	0.13344	3.1961	1.425e-03	0.16495	0.6880	0.001
Diseasescopd	0.1625	0.09717	1.6721	9.474e-02	0.00000	0.3529	0.095
Diseasesdiabetes	0.6951	0.14885	4.6698	3.313e-06	0.40337	0.9869	0.000
Diseasesmuskulo	0.4400	0.05392	8.1605	7.536e-16	0.33431	0.5457	0.000
Diseasesother	0.3797	0.07379	5.1456	3.058e-07	0.23506	0.5243	0.000
Diseasesheart.stroke	0.2778	1.23906	0.2242	8.226e-01	0.00000	2.7064	0.823

The hazard for someone with both heart and stroke is estimated as 0.7578+0.7557+0.2778, of which the last figure represents the interaction effect. In this example, the interaction is not significant, hence, additivity of these two hazards can be assumed without interaction effects.

Appendices

Appendix 1: Analogy of attribution method to crude probabilities of death

The basic principles of the attribution method that was explained in this manual are analogous to the principles of crude probabilities of death. Among others, we assumed an additive nature of cause-specific disability hazards. As an example to further explain this idea of additive cause-specific disability hazards, we first distinguish only two disease groups (A and B), and one age group. An example for the procedure for more diseases and age groups, using additive regression analysis is given later on.

We divide subjects into subgroups, defined by the combination of causes, that is (1) no disease (i.e. only background risk), (2) only disease A, (3) only disease B and (4) both disease A and B. All persons are exposed to the same background risk of disability (within one age-sex group). In group 1, we assume exposure to the background risk only. In addition to this background, in groups 2 to 4, persons are exposed to A (group 2), B (group 3) or A and B (group 4). In groups 2-4, competing risks of disability exist and, consequently, the number of persons disabled from one specific cause depends on the risk from multiple cause(s) as well.

Analogous to the multi-decrement life table, which is used to analyze observed distributions of deaths by cause in similar situations of competing causes of death [9] we used (hazard) rates to obtain 'crude probabilities'. These are probabilities in the situation in which several causes are acting simultaneously. Under the assumption of additivity of rates, the total rate can be decomposed into cause-specific rates. Thus, adding all cause-specific rates again gives the total rate [9]

Several steps/characteristics/properties can be distinguished. First, to obtain disability rates, the probability of disability in the sample(q), which is an estimate of the probability of being disabled (π), is converted into a (hazard) rate, using [9]:

$$\text{rate (tot)} = -\ln(1-q) \qquad \text{(A1)}$$

The total rate equals the sum of cause-specific rates (additivity of rates), where both diseases and background (risk from other factors than diseases considered) are considered as causes, or:

$$\text{rate (tot)} = \text{rate (bg)} + \text{rate (A)}.X_A + \text{rate (B)}.X_B \qquad \text{(A2)}$$

where, X_A is a dummy for the presence of cause A and *bg* is background.

Second, the cause-specific rates are obtained by comparing the groups with and without the cause. The rates for the different groups can be decomposed into:

$rate(tot_{bg})$ = rate(bg) (A3)

$rate(tot_{A})$ = rate(bg) + rate (A) (A4)

$rate(tot_{B})$ = rate(bg) + rate (B) (A5)

$rate(tot_{AB})$ = rate(bg) + rate (A) + rate (B) (A6)

These relationships are used to calculate the contributions of background and the different diseases. A separate rate could be obtained for the combination of disease A and B in the same way, but in a situation with multiple causes, calculating rates for all combinations would be impractical and lead to unstable results (because of small numbers). Moreover, in that situation, it would not be possible to attribute disability to single diseases only, but disability would also be attributed to combinations of diseases. Regression analysis (see next section) gives the best estimates of the contributions of each cause under the assumption of independent risks.

Third, the proportional distribution of each cause-specific rate is used to obtain crude cause-specific probabilities for each group. For instance, for group 3 (exposed to background and B) this gives:

q_{bg} = q x (rate (bg) / rate (tot_B)) (A6)

q_B = q x (rate (B / rate (tot_B)) (A7)

By definition, for group 3 the sum of the three crude probabilities gives the total probability.

$Q(tot) = q_{bg} + q_B$ (A8)

This equals:

$Q(tot) = 1 - exp(-(rate(bg) + rate(B)) = exp(-rate(tot))$ (A8C)

Appendix 2: Numerical examples

2.1 Numerical example

This example illustrates approach with one cause of disability, here one disease, no background. In total of the 508 persons, 101 persons are disabled. Disability prevalence is thus: 0.1988. The total population (one age group) consists of a group that is exposed to the disease (n=100) and a group who is not exposed (n=408). In the group with the disease 50 persons are disabled (out of 100 persons). In the group without the disease 51 persons are disabled (out of 408 persons). Based on this information the number of disabled and prevalence of disability can be decomposed by cause as follows:

	# disabled	N (sample)	P=prob (prevalence)	Total rate disab	cause-spec rate (disabling impact)	Part due to background	Part due to Disease	Crude prob. due to background	Crude prob. due to disease	# disabled due to backgr.	# disabled due to exposure
No disease (unexposed)	51	408	=51/408 0.125	=-ln(1-0.125) = 0.133531	0.133531	=rate background/rate total =0.13353/0.133531 =1	0	=Prob*part due to background =0.125*1	0	=0.125*408 =51	0
Disease (exposed)	50	100	=50/100 0.5	=-ln(1-0.5) =0.693147	=0.693147-0.133531 0.559616	=rate background/rate total =0.13353/0.693147 =0.193	=rate disease/rate total =0.559616/0.693147 =0.807	=0.193*0.5 =0.096 =(0.1335/0.693)*1-exp(-0.693)	=Prob*part due to disesase =0.8073*0.5 =0.403677 =(0.5596/0.693)*1-exp(-0.693)	=0.096*100 9.63	=0.4037*100 40.37
Total	101	508	0.198819							60.63	40.37

small differences due to rounding

or: (shortcut)

	Prob disab	hazard disab	cause-spec hazard	% haz due to background $=D_1/D_{tot}$	% haz due to exposure $=D_2/D_{tot}$	# total disabled D_{tot}	# disabled due to backgr. D_1	# disabled due to exposure D_2
unexposed	0.125	0.133531	0.133531	100%	0%	51	51 =100%	0
exposed	0.5	0.693147	0.559616	19.3% =01335/0.693	80.7% =0.5335/0.639	50	9.63 =0.193* 50	40.37 =0.807* 50
total	0.198819	0.220765				101	60.63	40.37

2.2 Numerical example with additive regression

Here we give a simplified numerical example based on the additive regression model. We assume a situation with two diseases, as this allows showing how the attribution is calculated.

2.2.1 Numerical example of attribution of disability to 2 diseases

All ages (50+)	Disease A	
Disease B	Absent	Present
Absent	963	375
Present	125	83

Example of regression output, based on all ages:

Background age 50-54	0.3228
Background age 55-59 – 80-84	Included in model but not in this table
Background age 90+	0.9673
Disabling impact (rate) Disease A	0.7348
Disabling impact (rate) Disease B	0.7395

Here we show how we attribute disability to diseases in the age group 90+. We start from the disease combinations:

90+	Disease A	
Disease B	Absent	Present
Absent	9	6
Present	2	4

Based on this table, and the background and two disease hazards, we can calculate cause-specific disability proportions:

	N	Background hazard	Hazard A	Hazard B	Total hazard (calculation)	Total hazard (result)	Part due to background (calculation)	Part due to A (calculation)	Part due to B (calculation)
Neither A nor B	9	0.9673			=0.9673	=0.9673	=0.9673/0.9673		
Dis A, not B	6	0.9673	0.7348		=0.9673+ 0.7348	=1.721	= 0.9673/1.721	= 0.7348/1.721	
Dis B, not A	2	0.9673		0.7395	=0.9673 +0.7395	=1.768	= 0.9673/1.768		= 0.7395/1.768
Both A and B	4	0.9673	0.7348	0.7395	=0.9673+ 0.7348+0.7395	=2.4416	= 0.9673/2.4416	= 0.7348/2.4416	= 0.7395/2.4416

	N	Part due to background (Result)	Part due to A (result)	Part due to B (result)	Total proportion disabled	Total proportion disabled (result)	Proportion disabled due to background	Proportion disabled due to A	Proportion disabled due to B
Neither A nor B	9	1.000	0	0	=1-exp(-0.9673)	0.61989	=0.61989*1 =0.61989	0	0
Dis A, not B	6	0.562	0.427		=1-exp(-1.721)	0.82111	=0.82111*0.562 =0.46146	=0.82111*0.427 =0.3506	0
Dis B, not A	2	0.547		0.418	=1-exp(-1.768)	0.82933	=0.82933*0.547 =0.4536	0	=0.82933*0.418 =0.34666
Both A and B	4	0.396	0.301	0.303	=1-exp(-2.4416)	0.91298	=0.91298*0.396 =0.3615	=0.91298*0.301 =0.27481	=0.91298*0.303 =0.27663

	N	Number disabled due to background	Number disabled due to A	Number disabled due to B	Number Disabled due to background (Result)	Number Disabled due to A (Result)	Number Disabled to B (result)	Total disabled (estimated)
Neither A nor B	9	=0.61989*9	0	0	5.58	0.00	0.00	5.58
Dis A, not B	6	=0.46146*6	=0.35061*6	0	2.77	2.10	0.00	4.87
Dis B, not A	2	=0.45364*2	0	=0.34666*2	0.91	0.00	0.69	1.60
Both A and B	4	=0.36154*4	=0.27481*4	=0.27663*4	1.45	1.10	1.11	3.65
Total					10.71	3.20	1.80	15.7

Appendix 3: Likelihood ratio test

The likelihood ratio test to compare two models can be performed as follows:

- subtract the log-likelihood of model without the additional parameter from the log-likelihood of the model with the additional parameters
- subtract the degrees of freedom of model without the additional parameter from the deviance of the model with the additional parameter
- You can use for instance Excel to do a chi-square test to test for significance =CHIDIST(dif log-likelihood, dif degrees of freedom). This yields the p-value.

Appendix 4: Logical expressions in selections

In the input file, specific selections of the data can be specified in Q4. This table shows which logical expressions can be used for selecting subgroups in the data. Depending on whether a variable in the selection is a numeric variable “” should be used. For a numeric variable no “” is used.

	Gender is numeric variable, or no selection based on gender	Gender is not numeric variable
Older or equal 30	Crossdat$age5 >= 30	
Women	Crossdat$gender == 2	Crossdat$gender == "F"
Women younger than 75	Crossdat$gender == 1 & Crossdat$age5< 75	Crossdat$gender == "M” & Crossdat$age5< 75
Older than 30 or women	Crossdat$age > 30 \| Crossdat$gender == 1	Crossdat$age > 30 \| Crossdat$gender == "M”

Appendix 5-a: Illustration of various models: no population distinguished

The models fitted in the illustration in part 3 for the situation of a single population are listed below. A zip-file including all cvs and output files is available from the authors. The file with the SHARE data is not included. For those who are registered users of SHARE, this file is available on request. For more details on using SHARE, see http://www.share-project.be/access.htm.

1. AttribW it: example of analyses of one single population (Italian women), hence without distinction in population, and no RRR

   Main characteristics:

   - single population (Q5 NA or empty)
   - no RRR (Q10 =0, Q11 is NA or empty)

2. AttribRRRit: example of analyses of one single population (Italian women), hence without distinction in population, but with RRR (first axis)

   Main characteristics:

   - single population (Q5 NA or empty)
   - RRR (Q10 =1, Q11 is 15 (column number of age classes for disabling impact)

3. AttribRRR2it: example of analyses of one single population (Italian women), hence without distinction in population, but with RRR (second axis)

   Main characteristics:

   - single population (Q5 NA or empty)
   - RRR (Q10 =2, Q11 is 15 (column number of age classes for disabling impact)

In all these three input specifications other options can be changed by the user, including:

- Q-4: to select cases or records (for more details see appendix 4)
- Q-9: to include/exclude sample weights. For no sample weights this can be left empty or NA can be filled in
- Q-12 and Q13: adding information on institutionalised population (See: input attribW nlinst.csv)
- Q-14 to use observed (O) instead of fitted values (F) (See: input attribW it O.csv)
- Q-15 to obtain additional output (T) (See input: input attribitvar.csv)

An example with interactions between diseases is given in: input attribitint.csv

Appendix 5-b: Input specification files and output used in the illustration: population distinguished

The models fitted in the illustration in part 3, wherein two populations are included, are listed below. A zip-file including all cvs and output files is available from the authors. The file with the SHARE data is not included. For those who are registered users of SHARE, this file is available on request. For more details on using SHARE, see http://www.share-project.be/access.htm.

1. Attribit2pop.csv: example of analyses of two populations (Italian men (population 1) and Italian women (population 2).

   Main characteristics:

   - population included (Q5=5, where 5 is column number for gender)
   - no RRR (Q10 =0, Q11 is NA or empty)

2. Attribit2popRRR.cvs: example of analyses of one single population (Italian women), hence without distinction in population, but with RRR (first axis)

   Main characteristics:

   - population included (Q5=5, where 5 is column number for gender)
   - RRR (Q10 =1, Q11 is 15 (column number of age classes for disabling impact)

3. Attribit2popRRR2.cvs: example of analyses of one single population (Italian women), hence without distinction in population, but with RRR (second axis)

   Main characteristics:

   - population included (Q5=5, where 5 is column number for gender)
   - RRR (Q10 =2, Q11 is 15 (column number of age classes for disabling impact)

For first population separately:

4. Attribitpop1RRR.cvs: example of analyses of first population (Italian men), model SL and RRR (first axis)

   Main characteristics:

   - Single population, but one selected
   - RRR (Q10 =1, Q11 is 15 (column number of age classes for disabling impact)

5. Attribitpop1RRR2.cvs: example of analyses of first population (Italian men), model SL and RRR second axis)

Main characteristics:

- Single population, but one selected
- RRR (Q10 =2, Q11 is 15 (column number of age classes for disabling impact)

For second population separately:

6. Attribitpop2RRR.cvs: example of analyses of second population (Italian women), model SL and RRR (first axis)

   Main characteristics:

   - Single population, but one selected
   - RRR (Q10 =1, Q11 is 15 (column number of age classes for disabling impact)

7. Attribitpop2RRR2.cvs: example of analyses of first population (Italian women), model SL and RRR (second axis)

   Main characteristics:

   - Single population, but one selected
   - RRR (Q10 =2, Q11 is 15 (column number of age classes for disabling impact)

In all these input specifications other options can be changed by the user, including:

- Q-4: to select cases or records (for more details see appendix 4). This is also needed if based on model selection the user has decided that in a situation of two populations, each population is modelled separately. Q-9: to include/exclude sample weights. For no sample weights this can be left empty or NA can be filled in
- Q-12 and Q13: adding information on institutionalised population
- Q-14 to use observed (O) instead of fitted values (F)
- Q-15 to obtain additional output (T)

References

1. Nusselder WJ, Looman CW. Decomposition of differences in health expectancy by cause. *Demography* 2004; **41**(2): 315-34.
2. Verbrugge LM. The twain meet: empirical explanations of sex differences in health and mortality. *Journal of health and social behavior* 1989; **30**(3): 282-304.
3. Rockhill B, Newman B, Weinberg C. Use and misuse of population attributable fractions. *Am J Public Health* 1998; **88**(1): 15-9.
4. Chiang CL. On the probability of death from specific causes in hte presence of competing risks. *Proc Fourth Berkeley Symp on Math Statist and Prob* 1961; **4**: 169-80. .
5. Clayton D, Hills M. Statistical models in epidemiology. Oxford: Oxford University Press; 1993.
6. Davis P, Tso M-S. Procedures for reduced rank regression. *Appl Stat* 1982; **31**: 244-55.
7. Van den Brink P, ter Braak C. Principal response curves: analysis of time-dependent multivariate responses of biological community to stress. . *Environmental Toxicology and Chemistry* 1999; **18**: 138-48.
8. Yee T, Hastie T. Reduced-rank vector generalized linear models. *Statistical modelling* 2003; **3**(1): 15-41.
9. Manton KG, Stallard E. Recent Trends in Mortality Analysis. . Orlando, Fl:: Academic Press; 1984.
10. Efron B, Tibshirani RJ. An introduction to the bootstrap Palto Alto, California; 1994.
11. Aitkin M, Anderson D, Francis B, Hinde J. Statistical Modelling in GLIM Oxford; 1989.
12. Selvin S. Modern Applied Biosta. Oxford: Oxford University Press; 1998.
13. Cox DR. Regression models and life table (with Discussion) *J Roy StatistSoc* 1972; **34**: 187-220.
14. Aalen OO. Nonparametric inference in connection with multiple decrement models. . *Scandinavian Journal of Statistics* 1976; (3): 15-27.
15. Ezzati M, Hoorn SV, Rodgers A, Lopez AD, Mathers CD, Murray CJ. Estimates of global and regional potential health gains from reducing multiple major risk factors. *Lancet* 2003; **362**(9380): 271-80.
16. Eide GE, Gefeller O. Sequential and average attributable fractions as aids in the selection of preventive strategies. *J Clin Epidemiol* 1995; **48**(5): 645-55.
17. Nusselder WJ, Looman CW, Mackenbach JP, et al. The contribution of specific diseases to educational disparities in disability-free life expectancy. *Am J Public Health* 2005; **95**(11): 2035-41.