Estimation of penetrance in age-dependent genetic disease with sporadic cases from pedigree data

Lucas Ducrot[1], Grégory Nuel[1]

Introduction

In the context of genetic disease with low allele frequency in the general population and high penetrance (i.e. Mendelian disease), family-based approach is convenient as patients are often refered to geneticists due to their strongly affected pedigree. In this context, the estimation of penetrance in age-dependent genetic disease has direct applications in the medical protocol of patient care.

The main issue in these estimations is that genotypes are mostly unknown and must be treated as a latent variable. In the specific case where the disease does not present sporadic cases, the problem is easier as an affected individual is therefore a mutation carrier, the genotype incertainty leans on the unaffected population. In that simple case, methods already exist (Alarcon et al., 2018) based on Expectation-Maximisation (Dempster et al., 1977) and Elston-Stewart algorithms (Elston and Stewart, 1971; Elston et al., 1992).

However, most diseases affect both people with and without known deleterious mutations at different rates. Typical example is breast cancer, as everyone is at risk but especially carriers of mutations (BRCA1/BRCA2 and others) which are affected at a much higher rate (Easton et al., 1993; Stoppa-Lyonnet et al., 1997). The proposed method aims to take into account sporadic cases to generalize previous estimation methods of genetic disease survival.

Objective and Notations

- **Survival mixture of a genetic disease**

  Let consider:
  - a autosomal dominant disorder of one gene and two alleles ("wild-type" 0 and "deleterious" 1), the genotype component $X \in \{00, 01, 10, 11\}$ ($X = 00$ for non-carrier, $X \neq 00$ for carrier);
  - the proportions of carriers in the population π_1 and non-carriers π_0 (with $\pi_0 = 1 - \pi_1$);
  - the specific conditional hazard rates $\lambda_1(t)$ for carriers and $\lambda_0(t)$ for non-carrier;
  - the relative hazard between carriers and non-carriers $\mathrm{RH}(t)$ such as $\lambda_1(t) = \mathrm{RH}(t) \times \lambda_0(t)$;
  - $S(t)$ (resp. $S_0(t)$ and $S_1(t)$) is the survival function (resp. conditional survival functions) associated with hazard $\lambda(t)$ (resp. $\lambda_0(t)$ and $\lambda_1(t)$) such as

$$S(t) = \exp\left(-\int_0^t \lambda(u)\,du\right);\quad S_0(t) = \exp\left(-\int_0^t \lambda_0(u)\,du\right);\quad S_1(t) = \exp\left(-\int_0^t \lambda_1(u)\,du\right).$$

  - a censorship event (which will not be needed) with a distribution function $g(t)$ and a repartition function $G(t)$ such as

$$G(t) = \int_0^t g(u)\,du.$$

- **Objective**

  To estimate this model from pedigree data with a constrained general population incidence $\lambda(t)$.

- **Assumptions**
  - the general population incidence $\lambda(t)$ is known and piecewise constant;
  - the hazard ratio between carriers and non-carriers $\mathrm{RH}(t)$ is unknown but piecewise constant.

Model

The model can be written as followed:

$$\mathbb{P}(T, \delta, X) = \underbrace{\mathbb{P}(X)}_{\text{Genetic Part}} \times \underbrace{\mathbb{P}(T, \delta | X)}_{\text{Survival Part}},$$

where T are ages at diagnostic (or censored ages), δ status (affected or unaffected) and X genotypes (carrier or non-carrier).

- **Genetic Part:** data are pedigrees, so $\mathbb{P}(X)$ can be written as Bayesian network, for each individual i (F set of Founders):

$$\mathbb{P}(X) = \prod_{i \in \mathrm{F}} \mathbb{P}(X_i) \prod_{i \notin \mathrm{F}} \mathbb{P}(X_i | X_{parents_i}).$$

- **Survival Part:** $\delta_i \in \{0, 1\}$ represents the status (affected or not) of individual i
  - if unaffected then

$$\mathbb{P}(T_i = t, \delta_i = 0 | X_i) = \begin{cases} g(t)S_1(t) & \text{if } X_i \neq 00; \\ g(t)S_0(t) & \text{if } X_i = 00; \end{cases} \propto \begin{cases} S_1(t) & \text{if } X_i \neq 00; \\ S_0(t) & \text{if } X_i = 00; \end{cases}$$

  - if affected then

$$\mathbb{P}(T_i = t, \delta_i = 1 | X_i) = \begin{cases} (1 - G(t))S_1(t)\lambda_1(t) & \text{if } X_i \neq 00; \\ (1 - G(t))S_0(t)\lambda_0(t) & \text{if } X_i = 00; \end{cases} \propto \begin{cases} S_1(t)\mathrm{RH}(t) & \text{if } X_i \neq 00; \\ S_0(t) & \text{if } X_i = 00. \end{cases}$$

Typical pedigree data

- Generally 10-40 families in a dataset
- Pedigree data include families' structures, ages or ages at diagnostic and few genotypes

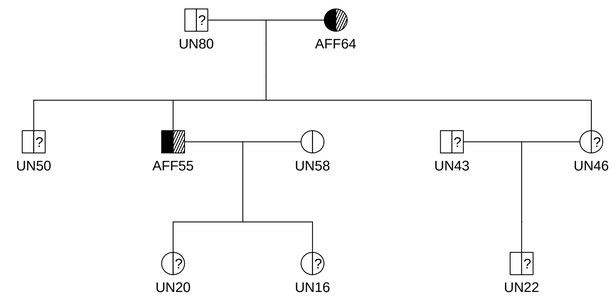

Figure 1: Example of one pedigree.

Developed Method

Idea

Considering that the general population incidence $\lambda(t)$ (and by extension $S(t)$) is known, the model is parameterized by π_1 and $RH(t)$. The idea is that with this parametrization, $\lambda_0(t)$ and $\lambda_1(t)$ (as well as $S_0(t)$ and $S_1(t)$) can be computed under the constrained general population incidence $\lambda(t)$ through a fixed point method.
From there, the log-likelihood of the model can be computed with the pedigree data via Elston-Stewart algorithm (Elston and Stewart, 1971; Elston et al., 1992).
Therefore the log-likelihood is a function of π_1 and $\mathrm{RH}(t)$ and computable from pedigree data. The maximum likelihood parameters are estimated using a gradient descent.

- **Fixed point method:**

  Idea:
  $\lambda(t)$ is assumed to be piecewise constant with known cuts (typically for cancer registry with 5-years bins), and $\mathrm{RH}(t)$ also is piecewise constant with known cuts (depend on the model and sometimes on X, e.g. bins $[0, 50]$ and $]50, +\infty[$).
  For a given proportion π_1 and RH, we would like to compute $\lambda_0(t)$ such that:

$$S(t)\lambda(t) = \pi_0 S_0(t)\lambda_0(t) + \pi_1 S_1(t)\lambda_1(t).$$

  To solve this problem, $\lambda_0(t)$ is assumed to be piecewise constant with a thin cutset (e.g. one cut every tenth of a year from 0 to 80) and these following fixed-point iterations are performed:

  - initialize with $\lambda_0(t) = \lambda(t)$;
  - repeat: compute $S_0(t)$ and $S_1(t)$ with current $\lambda_0(t)$ and update

$$\lambda_0(t) = \frac{\lambda(t)S(t)}{\pi_0 S_0(t) + \pi_1 S_1(t)\mathrm{RH}(t)}.$$

  Simple Example:
  Let consider a general population incidence with cuts $20, 40, 60, 80$ and bin-specific yearly incidence $0.000, 0.003, 0.005, 0.010, 0.015$.
  - $\pi_1 = 0.0975$;
  - RH with cuts 50 an bin-specific values $20, 10$.

  Finally, λ_0 cuts are assumed to be every tenth of a year from 0 to 80.

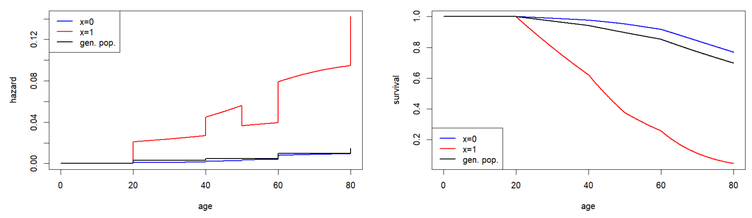

Figure 2: Hazard rates and Survivals after fixed-point convergence in simple example.

- **Log-likelihood computation:**

  For specific parameters $\theta = (\pi_1, \mathrm{RH}(t))$, $\lambda_0(t)$ and $\lambda_1(t)$ (as well as $S_0(t)$ and $S_1(t)$) are computed through the fixed point method. Now the log-likelihood of the model can be written as follows :

$$\mathrm{loglik}(\theta) = \log[\sum_X \prod_i \underbrace{\mathbb{P}(T_i, \delta_i | X_i; \theta)}_{\text{survival component}} \underbrace{\mathbb{P}(X_i | X_{parents_i}; \theta)}_{\text{genetic component}}].$$

  This is computable by method using Elston-Stewart algorithm (Elston and Stewart, 1971; Elston et al., 1992).

- **Maximum Log-likelihood estimation:**

  As previously explained, the log-likelihood of the model can be computed as a function of the parameters π_1 and $\mathrm{RH}(t)$.
  In the simple example,

$$\mathrm{RH}(t) = \begin{cases} \mathrm{RH}_1 & \text{if } t \in [0, 50]; \\ \mathrm{RH}_2 & \text{if } t \in]50, +\infty[. \end{cases}$$

  So here, the model comes down to only 3 parameters $\theta = (\pi_1, \mathrm{RH}_1, \mathrm{RH}_2)$ which are estimated by maximizing the log-likelihood with Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (Nocedal and Wright, 2006).

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta}{arg\,max}\,\mathrm{loglik}(\theta)$$

Results on simulations

- **Simulations**

  2000 datasets are generated:
  - 744 inviduals over 28 families;
  - $\pi_1 = 0.0975$, $\mathrm{RH}_1 = 20$, $\mathrm{RH}_2 = 10$;
  - families' structures based on real families (data APHP);
  - autosomal dominant transmission model with 1 gene and 2 alleles ("wild-type" and "deleterious").

  In order to mimic real data where missing values are often encountered, each similated dataset is replicated 4 times, each time with less available information. The first replica has 100% of the available data (it is the oracle), the second has 70% of the available data (30% is missing), the third has 50% and the last has 30%.

- **Results**

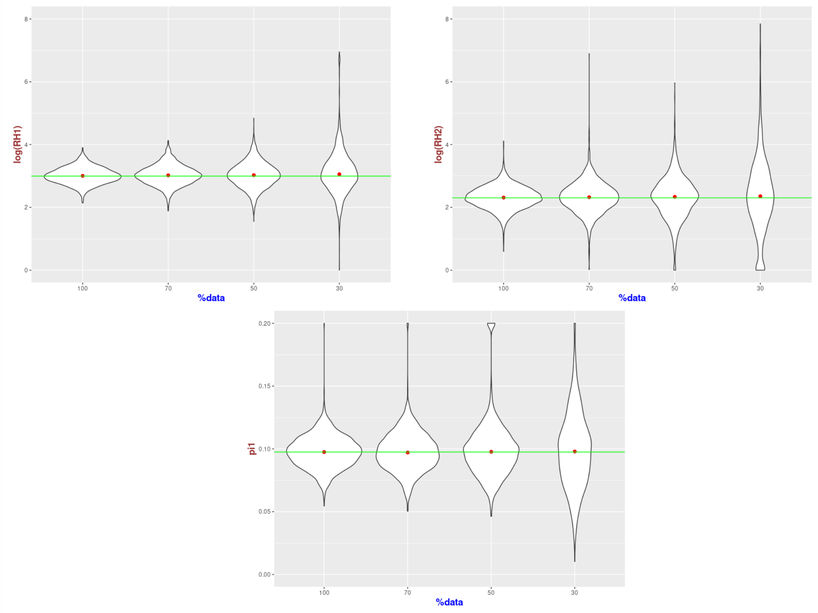

Figure 3: Violin plots of estimated π_1, $log(\mathrm{RH}_1)$ and $log(\mathrm{RH}_2)$ with 100%, 70%, 50% and 30% of available data. The green line on each figure is the real value of each parameter.

Perspectives

- Use bootstrap (by resampling the families) to estimate the parameters from a dataset;
- Take into account the ascertainment bias using statistical adjustment (like raking).

References

Flora Alarcon, Violaine Planté-Bordeneuve, Malin Olsson, and Grégory Nuel. Non-parametric estimation of survival in age-dependent genetic disease and application to the transthyretin-related hereditary amyloidosis. *PLOS ONE*, 13(9):e0203860, September 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0203860.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 0035-9246.

D. F. Easton, D. T. Bishop, D. Ford, and G. P. Crockford. Genetic linkage analysis in familial breast and ovarian cancer: Results from 214 families. The Breast Cancer Linkage Consortium. *American Journal of Human Genetics*, 52(4):678–701, April 1993. ISSN 0002-9297.

R.C. Elston and J. Stewart. A General Model for the Genetic Analysis of Pedigree Data. *Human Heredity*, 21(6):523–542, 1971. ISSN 0001-5652, 1423-0062. doi: 10.1159/000152448.

Robert C. Elston, Varghese T. George, and Forrestt Severtson. The Eiston-Stewart Algorithm for Continuous Genotypes and Environmental Factors. *Human Heredity*, 42(1):16–27, 1992. ISSN 1423-0062, 0001-5652. doi: 10.1159/000154043.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer, New York, 2nd ed edition, 2006. ISBN 978-0-387-30303-1.

D. Stoppa-Lyonnet, P. Laurent-Puig, L. Essioux, S. Pagès, G. Ithier, L. Ligot, A. Fourquet, R. J. Salmon, K. B. Clough, P. Pouillart, C. Bonaïti-Pellié, and G. Thomas. BRCA1 sequence variations in 160 individuals referred to a breast/ovarian family cancer clinic. Institut Curie Breast Cancer Group. *American Journal of Human Genetics*, 60(5):1021–1030, May 1997. ISSN 0002-9297.

Acknowledgment

This work was funded by *ISCD - Sorbonne Universite* (PhD grant). We thank Nadia Nathan and Marie Legendre for the pedigree structures used to test the method. These data were collected by the French national networks for rare lung diseases: Centre de référence des maladies respiratoires rares (RespiRare), Centre de référence des maladies pulmonaires rares (OrphaLung) and Filière de soins pour les maladies respiratoires rares (RespiFIL). The ILD cohort has been developed in collaboration with the Rare Disease Cohort (RaDiCo)-ILD project (ANR-10-COHO-0003), the Clinical research collaboration for chILD-EU and the COST Innovative Grant OpenILD CIG16125.

Correspondance to

lucas.ducrot@sorbonne-universite.fr*
nuel@math.cnrs.fr

*Corresponding author.

[1] Probability and Statistics (LPSM, CNRS 8001), Sorbonne Université, Paris;