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

Lucas Ducrot, G. Nuel

Estimation of penetrance in age-dependent genetic disease with sporadic cases from pedigree data (Temporary version 28/02/2024)

Lucas Ducrot[1], Grégory Nuel[1]

[1] Sorbonne Université, Laboratory of Probability, Statistics and Modeling, Stochastics and Biology Group, Paris, France

Abstract. 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 survival 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 based on Expectation-Maximisation and sum-produt algorithm.

However, most diseases affect both people with and without known deleterious mutations at different rates. The few existing methods in this case generally assume that the incidence of the disease is known in the general population as well as the proportion of mutation carriers. They also assume that the incidence for non-carriers is equal to the incidence for the general population. This is close to reality for mutations with very low allele frequency and very penetrance but falls down in more moderate scenarios.

The proposed method aims to generalize previous estimation methods of genetic disease survival. It relies on two hypothesis: the hazard rate of general population is piecewise constant and known, the hazard ratio between carriers and non-carriers is also piecewise constant.

The model is a survival mixture parameterized by the hazard ratio and the proportion of carriers. At fixed parameters, the hazard rates (incidences) of carriers and non-carriers can be computed under the constrained hazard rate of general population through a fixed point method. With the pedigree data, the likelihood of the model can be computed with a sum-product algorithm and, therefore, the maximum likelihood parameters are estimated using a BFGS optimization algorithm.

The method is tested on 2000 simulated datasets of 744 people (28 families). Standard simulations followed the model with a proportion of carriers at 0.0975, hazard ratio is (RH1=20, RH2=10) with a cut-off at age 50. A robustness analysis is also performed where the dataset are generated with Weibull function as hazard ratio.

Keywords. Survival data, censored data - Mixture models - Biostatistics, genetics, health - Bayesian statistics

1 Introduction

In genetic counselling, risk estimations of genetic disease's onset is generally useful in order to guide medical protocols of patient care. In the context of genetic disease with low allele frequency in the general population and high penetrance (*i.e.* Mendelian disease), family-based approaches are generally used to evaluate that risk as patients are often selected through their strongly affected pedigree.

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 [1] based on Expectation-Maximisation [3] and Sum-product algorithm [10].

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 [4, 9]. The few existing methods [2, 8] in this case generally assume that the incidence of the disease is known in the general population as well as the proportion of mutation carriers. They also assume that the incidence for non-carriers is equal to the incidence for the general population. This is close to reality for mutations with very low allele frequency and very penetrance but falls down in more moderate scenarios.

The proposed method aims to extend the previous estimation methods of genetic disease survival by relaxing some assumptions.

2 Objective and Notations

2.1 Notations

In this article, we consider the following context:

- an autosomal dominant disorder of one gene and two alleles ("normal" 0 and "pathogenic" 1), the genotype component $X \in \{00, 01, 10, 11\}$ ($X = 00$ for non-carrier, $X \neq 00$ for carrier) where the first is the paternal allele and the second the maternal allele;
- the proportions of carriers in the population is denoted π_1 and non-carriers π_0 (with $\pi_0 = 1 - \pi_1$);
- the specific conditional hazard rates are $\lambda_1(t)$ for carriers and $\lambda_0(t)$ for non-carrier;
- We denote 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

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

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

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

2.2 Objective and Assumptions

The objective of this article is to estimate $S_0(t)$, $S_1(t)$ and π_1 from pedigree data with a constrained general population incidence $\lambda(t)$. In order to do so, we make two assumptions :

- the general population incidence $\lambda(t)$ is known and piecewise constant, which is often the case in medical registry (typically for cancer registry with 5-years bins);
- the hazard ratio between carriers and non-carriers $\mathrm{RH}(t)$ is unknown but piecewise constant (the piece-wise constant is not necessary, the main idea is to parameterize the hazard ratio, further extension of the method could study Weibull distribution as parameterization for example).

3 Model

The model describes a group of individuals with potentially family links and the probabilities of their genotypes, ages or ages at diagnosis and status (affected by the disease or unaffected). In mathematical terms, let consider n individuals in set $\mathcal{I} = \{1, ..., n\}$ distributed among N families. The set of founders which are individuals that have no parents in the data is noted $\mathcal{F} \subset \mathcal{I}$. The ages or ages at onset of the disease of all individuals is denoted $T = (T_1, ..., T_n) \in \mathbb{R}^n$ where T_i is the age for individuals i. The genotypes of individuals is denoted $X = (X_1, ..., X_n) \in \{00, 01, 10, 11\}^n$ where 0 represents normal allele and 1 the pathogenic allele and first digit (respectively second) corresponds to the paternal (respectively maternal) allele (i.e. for example $X_i = 01$ means the individual i has a paternal allele 0 and a maternal allele 1). Also $\delta = (\delta_1, ..., \delta_n) \in \{0, 1\}^n$ denotes the status of individuals, δ_i is 1 if the individual i is affected and 0 if unaffected. Therefore this model can be conditioned on the genotype X and decomposed in two subparts, a genetic one, a survival one as follows:

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

- **Genetic Part:** the probability of the genotypes forms a Bayesian network thanks to the family structure as the genotype of one individual only depends on the genotypes of its parents. The set founders of the family $\mathcal{F}$ is the set of individuals that have not parents in the data, the genotypes of these individuals ($\mathbb{P}(X_i)$,$i \in \mathcal{F}$) follow Hardy-Weinberg equilibrium with allelic frequency $f = 1 - \sqrt{1 - \pi_1}$, the non-founders ($\mathbb{P}(X_i|X_{\mathrm{pat}_i}, X_{\mathrm{mat}_i})$,$i \notin \mathcal{F}$) follow Mendelian transmission from parents:

$$\mathbb{P}(X) = \prod_{i \in \mathcal{F}} \mathbb{P}(X_i) \prod_{i \notin \mathcal{F}} \mathbb{P}(X_i|X_{\mathrm{pat}_i}, X_{\mathrm{mat}_i})$$

- **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

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

4 Developed method

4.1 Idea

Considering that the general population incidence $\lambda(t)$ (and by extension $S(t)$) is known, the model is parameterized by π_1 and $\mathrm{RH}(t)$. 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 [5, 6] or sum-product algorithm (belief-propagation) [10] in which evidence is based on the calculated $\lambda_0(t)$, $\lambda_1(t)$, $S_0(t)$ and $S_1(t)$.

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 BFGS algorithm [7]. The confidence intervals of the estimated parameters can be computed with the Hessian method.

4.2 Fixed point method

4.2.1 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 $\mathrm{RH}(t)$, 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, $\lambda_0(t)$ which is supposed to be continuous, is discretized with a thin cutset. Therefore, it is assumed to be piecewise constant, with cuts 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

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

4.2.2 Simple Example

Let consider a general population incidence with cuts $20, 40, 60, 80$ and bin-specific yearly incidence $0.000, 0.003, 0.005, 0.010, 0.015$. Fixing the parameters at:

- $\pi_1 = 0.0975$;
- RH with cuts 50 an bin-specific values $20, 10$.

Then, λ_0 cuts are assumed to be every tenth of a year from 0 to 80. From this setup, it is possible to computed λ_0, λ_1, $S_0(t)$ and $S_1(t)$ after convergence to the fixed point as shown in figure 1.

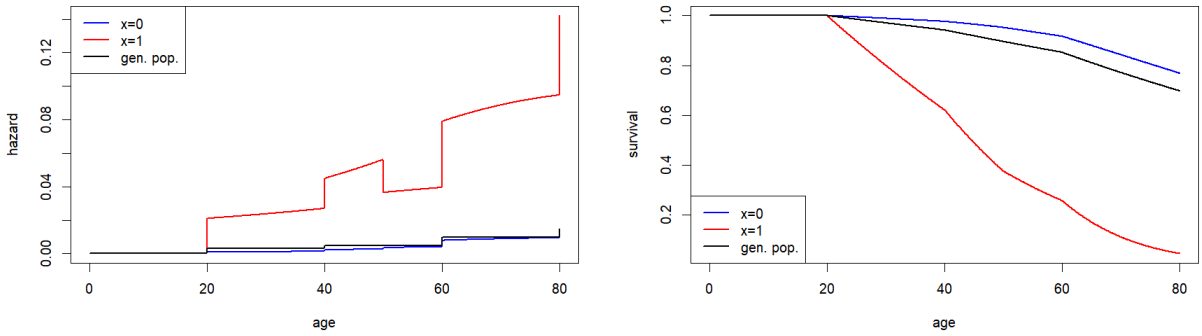

Figure 1: Hazard rates and Survivals after fixed-point convergence in simple example.

4.3 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 :

$$\mathcal{L}(\theta) = \sum_{\text{Families}} \log \left[\sum_X \prod_i \underbrace{\mathbb{P}(T_i, \delta_i | X_i; \theta)}_{\text{survival component}} \underbrace{\mathbb{P}(X_i | X_{\text{pat}_i}; X_{\text{mat}_i}; \theta)}_{\text{genetic component}} \right] .$$

This log-likelihood is compuTable using Elston-Stewart algorithm [5, 6] or sum-product algorithm [10] using $\lambda_0(t)$, $\lambda_1(t)$, $S_0(t)$ and $S_1(t)$ to calculate the evidence. In this article we use *bped*, an C++ implementation of the sum-product algorithm specifically designed for pedigree computation.

4.4 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 $\mathrm{RH}(t)$. $\mathrm{RH}(t)$ being actually a finite number of parameters, for instance in the simple example,

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

The model comes down to a finite number of parameters, here only 3 $\theta = (\pi_1, \mathrm{RH}_1, \mathrm{RH}_2)$ which are estimated by maximizing the log-likelihood with Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [7] implemented in R with the function *optim*.

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} \mathcal{L}(\theta)$$

4.5 Variables substitution

The estimated parameters are $\mathrm{RH}(t)$ and π_1. Actually, it is possible to use a variable substitution in order to constrain the parameters of the model. For instance in the model, the genetic disorder has a low allele frequency in the general population. It is therefore interesting to set $\pi_1 \in [0, 0.2]$ meaning that the proportion of pathogenic variants carriers can not be higher that 0.2. With a similar thinking, the relative risk $\mathrm{RH}(t)$ between carriers and non-carriers is expected to be higher than 1, the pathogenic variants carriers being a priori more at risk than the non-carriers. The rules set for the parameters can be find in the litterature or from knowledge from experts.

To apply these rules, the model includes a variable substitution such that:

- $\pi_1 = 0.2 \times \frac{e^{\theta_1}}{1+e^{\theta_1}}$
- $\mathrm{RH}_1 = 1 + e^{\theta_2}$
- $\mathrm{RH}_2 = 1 + e^{\theta_3}$

Therefore the parameters to estimate are $\{\theta_1, \theta_2, \theta_3\}$ which actually set $\pi_1 \in [0, 0.2]$, $\mathrm{RH}_1 > 1$ and $\mathrm{RH}_2 > 1$. But for the results, the article will present the estimated $\{\pi_1, \mathrm{RH}_1, \mathrm{RH}_2\}$ after substitution of estimated $\{\theta_1, \theta_2, \theta_3\}$ for better practical understanding.

4.6 Confidence intervals computation

The confidence intervals of the estimated parameters are computed using the Hessian method. The square roots of diagonal elements of the inverted Hessian matrix of the log-likelihood function estimate the standard deviations (SD) of the parameters. If the estimated parameters follow Gaussian distributions (discussed in the appendice), 95% confidence intervals can be calculated adding and substracting $1.96 \times$ SD to the maximum-likelihood parameters.

The variables substitutions used to constrain parameters being strictly monotone, the confidence intervals for substituted variables are calculated by applying the substition to the border of the intervals.

The function *optim* implemented in R proposed an argument (`hessian=TRUE`) which returns an estimation of the Hessian matrix computed during optimization process.

5 Data simulations

5.1 Simulations

The developed method is tested on simulated data. The first set of data is simulated accordingly to the model. A second set of data is generated where the relative hazard is not

piece-wise constant anymore and follows a Weibull function. This set allows to test the robustness of the method when the data do not follow strictly the model but are close enough (the Weibull function being close to a two part piecewise constant model). The family structures are real and taken from a dataset of families with carriers of *SFTPA1* or *SFTPA2* pathogenic variants. The genotypes were determined at theTrousseau hospital molecular genetics laboratory (APHP, Paris), and the loss of function of the variants was demonstrated in vitro (PMID: 32855221).

For both standard simulations and robustness analysis:

- 2000 datasets are generated.
- Each dataset is composed 744 inviduals over 28 families.
- Autosomal dominant transmission model with 1 gene and 2 alleles ("normal" and "pathogenic"). Genotypes of founders follow Hardy-Weinberg equilibrium.
- proportion of carriers is $\pi_1 = 0.0975$

Then the difference is on RH(t) as shown in figure 2 :

- standard analysis: $\text{RH}_1 = 20$, $\text{RH}_2 = 10$ with a cut at 50 years old;
- robustness analysis: RH(t) is a Weibull function (shape = 3 and scale = 1.5).

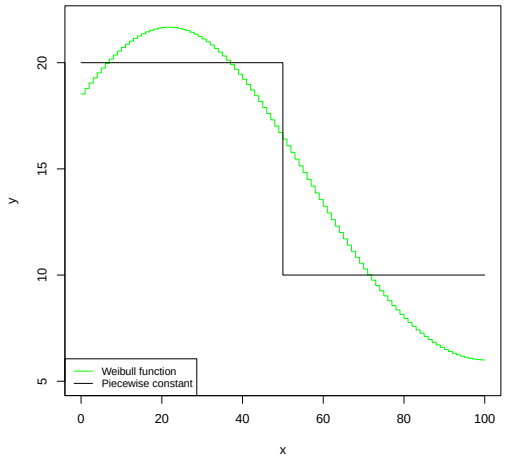

Figure 2: RH(t) for standard (black) and robustness (green) simulations.

5.2 Missing data

Missing data are introduced in the generated dataset in order to mimic real data. Genotypes and phenotypes are randomly considered missing. Four levels of missingness are considered including the oracle:

- Oracle : all the data is known.

- 30% : about 30% of the data is missing. 20% of the phenotypes and 40% of genotypes are missing.
- 50% : about 50% of the data is missing. 35% of the phenotypes and 65% of genotypes are missing.
- 70% : about 70% of the data is missing. 50% of the phenotypes and 90% of genotypes are missing.

5.3 Augmented data

For the standard simulations, augmented datasets are generated to analyse how the developed method scales with dataset's size. To do so, 2000 datasets of 1488 individuals over 56 families (initial size $\times 2$) and 2000 datasets of 2976 individuals over 112 families (initial size $\times 4$) are generated.

6 Results on simulations

6.1 Results on standard and robustness data

The results presented are violin plots of the parameters estimated by the proposed method with Oracle, 30%, 50% and 70% of missing data. The parameters estimated from standard simulations are presented in Figure 3 and those estimated from robustness simulations in Figure 4.

The results show a great fit to the expected values of the parameters both from the standard simulations and the robustness ones. The median values of each parameter for every level of data missingness except the 70% level on robustness analysis which does not match exactly bu remains very close. The variance increases with the level of missing data as expected.

There are few outliers in the estimations that seems to reach the boundaries fixed for our parameters (*i.e.* $\mathrm{RH}_1 > 1$, $\mathrm{RH}_2 > 1$ and $\pi_1 \in [0, 0.2]$).

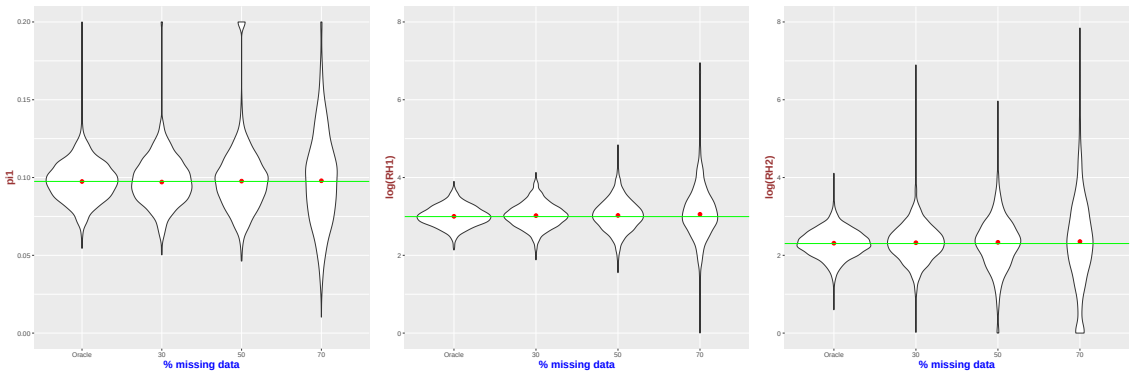

Figure 3: Violin plots of π_1, $\log(\mathrm{RH}_1)$ and $\log(\mathrm{RH}_2)$ estimation on standard simulations. Green line represents the real parameter to estimate.

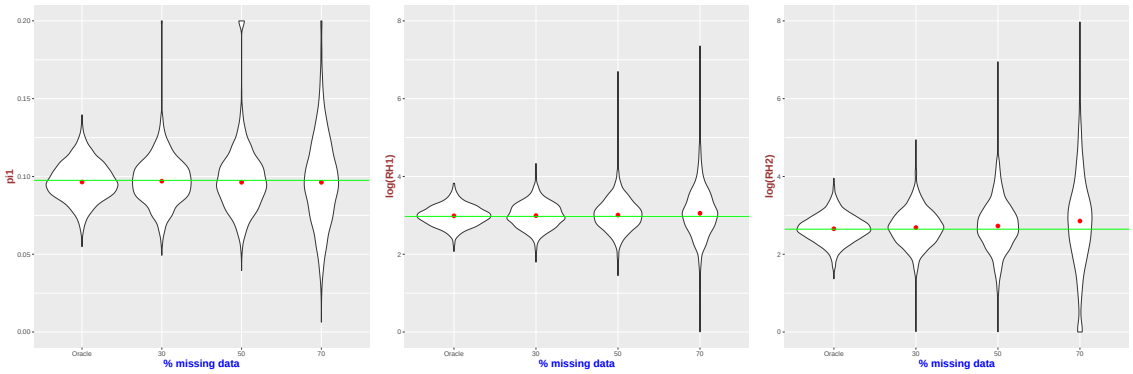

Figure 4: Violin plots of π_1, $\log(\mathrm{RH}_1)$ and $\log(\mathrm{RH}_2)$ estimation on robustness simulations. Green line represents the real parameter to estimate.

6.2 Results on augmented data

The results presented are violin plots of the parameters estimated by the proposed method with Oracle, 30%, 50% and 70% of available data on augmented data. The parameters estimated from datasets of standard simulation size $\times 1$, $\times 2$ and $\times 4$ are presented on the same figures to have a better overview of the results. π_1 estimations are shown in Figure 5, $\log(\mathrm{RH}_1)$ in Figure 6 and $\log(\mathrm{RH}_2)$ in Figure 7.

The results show again a great fit to the expected values of the parameters. The bigger the size, the better the estimations as the variances decrease with datasets size. The variances still increases with the level of missing data as expected.

There are again few outliers in the estimations that seems to reach the boundaries fixed for our parameters on the datasets of size $\times 2$ for π_1.

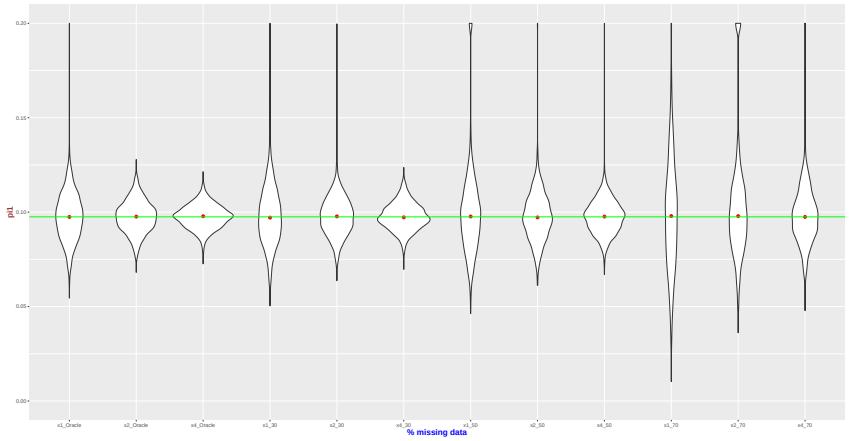

Figure 5: Violin plots of π_1 for various datasets sizes and data missingness. Green line represents the real parameter to estimate.

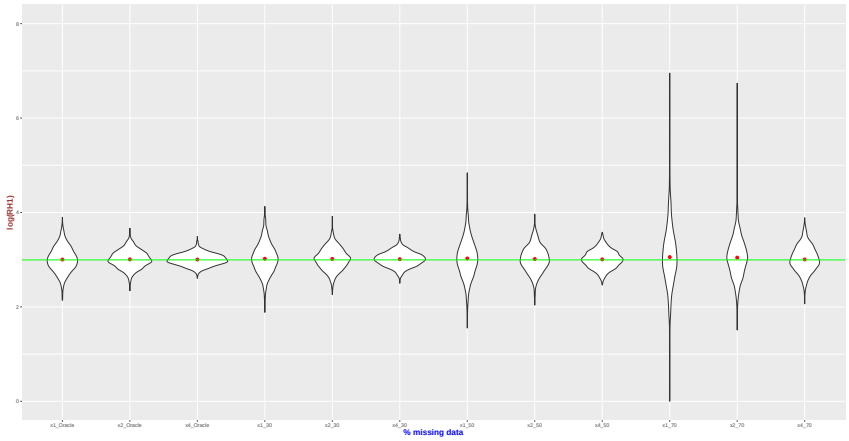

Figure 6: Violin plots of $\log(\mathrm{RH}_1)$ for various datasets sizes and data missingness. Green line represents the real parameter to estimate. Green line represents the real parameter to estimate.

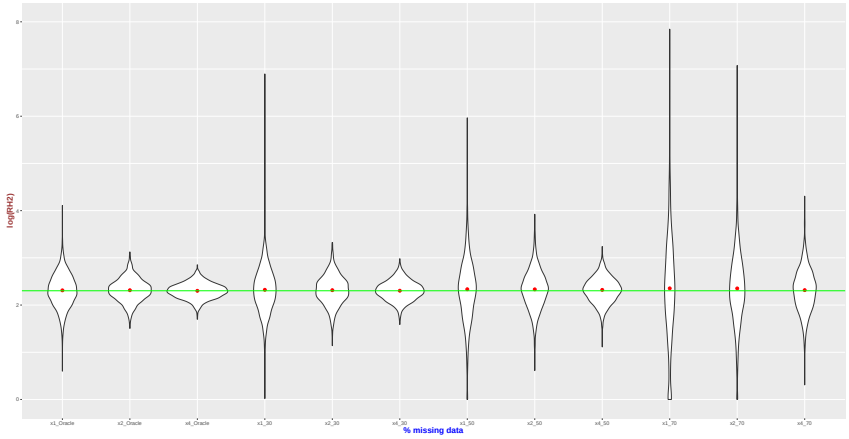

Figure 7: Violin plots of $\log(\mathrm{RH}_2)$ for various datasets sizes and data missingness. Green line represents the real parameter to estimate. Green line represents the real parameter to estimate.

6.3 Confidence intervals dependance on dataset's size

The coverage probability of confidence intervals computed for 100 datasets with the Hessian methods are presented in Table 1. The distributions of confidence intervals' sizes are presented for each parameter, level of missingness and datasets' size in Figure 8. Green represents the standard simulations, purple the datasets of size $\times 2$ and in brown the datasets of size $\times 4$.

The coverage probabilities decrease with higher level of missingness in the data. It seems that the coverage probability increases with the size of the datasets but it is not the case for the dataset size $\times 2$ at Oracle and 50% missing data level. The 30% level of missingness showcases the worst coverage probability overall for every parameter and datasets' size.

The size of the confidence intervals decreases the higher the dataset size is. Similarly the size of the confidence intervals decreases with low level of missingness. It is expected to perform better, the more information are known.

	×1, Oracle	×2, Oracle	×4, Oracle	×1, 30%	×2, 30%	×4, 30%
π_1	0.95 [0.90,0.99]	0.96 [0.92,0.99]	0.96 [0.92,0.99]	0.94 [0.89,0.98]	0.95 [0.90,0.99]	0.96 [0.92,0.99]
RH_1	0.94 [0.89,0.98]	0.91 [0.85,0.96]	0.97 [0.93,1.0]	0.93 [0.88,0.98]	0.94 [0.89,0.98]	0.99 [0.97,1.0]
RH_2	0.95 [0.90,0.99]	0.90 [0.84,0.95]	0.91 [0.85,0.96]	0.96 [0.92,0.99]	0.89 [0.83,0.95]	0.94 [0.89,0.98]
	×1, 50%	×2, 50%	×4, 50%	×1, 70%	×2, 70%	×4, 70%
π_1	0.95 [0.90,0.99]	0.92 [0.86,0.97]	0.97 [0.93,1.0]	0.89 [0.83,0.95]	0.91 [0.85,0.96]	0.97 [0.93,1.0]
RH_1	0.94 [0.89,0.98]	0.91 [0.85,0.96]	0.96 [0.92,0.99]	0.90 [0.84,0.95]	0.91 [0.85,0.96]	0.95 [0.90,0.99]
RH_2	0.95 [0.90,0.99]	0.93 [0.88,0.98]	0.95 [0.90,0.99]	0.82 [0.74,0.89]	0.89 [0.83,0.95]	0.93 [0.88,0.98]

Table 1: Coverage probability for each parameter and for each dataset size and missing data type.

7 Discussion

According to the results, the proposed method seems to estimate correctly the model parameters. The more data are available, the better are the estimations.

When applied to simulations generated from a sligthly different model, the method still estimates correctly the parameters (as observed according to the model on the simulated datasets).

The confidence intervals are very large for the 70% level of missingness but narrow with larger datasets and less missing data. However, the coverage probabilities do not fit to the expected 95% confidence intervals. The main reason probably being that the estimated parameters $\{\theta_1, \theta_2, \theta_3\}$ are mainly not Gaussian according the Shapiro-Wilk test as shown in the appendice. The confidence intervals remain useful with bigger datasets and less missing data. It is still possible to use a bootstrap method to estimate the confidence intervals also shown in the Appendice, the downside being increased computing cost.

8 Conclusion and perspectives

In conclusion, this article proposes a new method to estimate the penetrance/survival of a genetic disease with sporadic cases from pedigree data. Previous published methods generally make three assumptions. The first one is that the proportion of pathogenic variant carriers in the general population is known. The second one is that the incidence of the disease for the general population is also known. Finally these methods approximate the incidence of the non-carriers by the incidence of the general population. This last assumption is not far from reality in the case of very rare pathogenic allele and high penetrance but falls down as the allele is more and more common and the disease moderatly penetrant.

The proposed method generalises previous methods, relying only on the known incidence in general population assumption. To do so, the method incorporates the proportion of carriers in the population as a parameter of the model and use a fixed-point method to compute the incidences of carriers and non-carriers constrained by the incidence in the general population. The cost of this generalization is a parametrization of model.

The method performed well at estimating the parameters of the model on a simple example and on a robustness analysis. The method was also tested on a different set of simulations for which the results are presented in the appendice.

The main perspective of this work is to test the method on biased data which are the norms for collected pedigree in genetics. Indeed, patients are generally selected to be addressed to genetic counselling through specific sets of rules depending on countries/hospitals, this

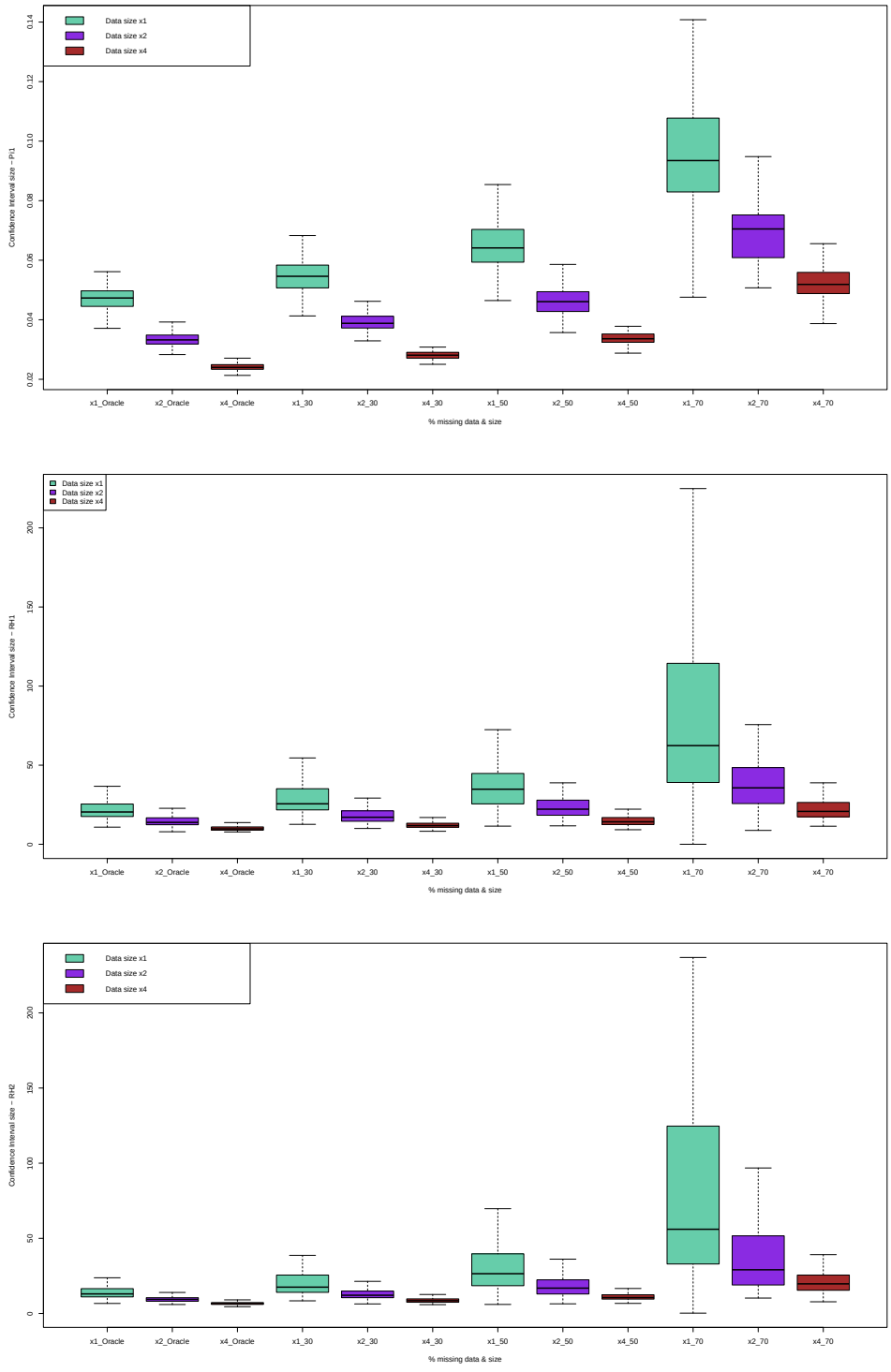

Figure 8: Boxplots of the confidence intervals sizes for parameters π_1, RH_1 and RH_2.

selection induced a first bias. Then, amongst these selected patients, only the carriers are generally followed and their family tested, which includes a second layer of selection.

Moreover, it would be interesting is to relax the assumption on the relative hazard $\mathrm{RH}(t)$

which is currently piecewise constant. For instance, the robustness analysis is performed using a Weibull function as relative hazard, which is a function parameterized by only two parameters (scale and shape). It would interesting to implement more diverse relative hazard options. This perspective also leads to another question which would be to study model selection with this method. From an unspecify model, it would be interesting to determine the optimal numbers and positions of cuts in $\mathrm{RH}(t)$.

Finally this model makes the assumption that the phenotypes are independant conditionally to the genotypes. This is a strandard assumption which has its limits espacially when the genetic disease presents major environmental (smoking for lung cancer for instance) and/or polygenic risk factors. It would be interesting to add exposition variable or frailty to the model which it is not straight forward because of the general population incidence constraint.

Acknowledgments

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.

References

[1] F. Alarcon, V. Planté-Bordeneuve, M. Olsson, and G. Nuel. Non-parametric estimation of survival in age-dependent genetic disease and application to the transthyretin-related hereditary amyloidosis. *PLOS ONE*, 13(9):e0203860, Sept. 2018.

[2] B. Bonaïti, V. Bonadona, H. Perdry, N. Andrieu, and C. Bonaïti-Pellié. Estimating penetrance from multiple case families with predisposing mutations: extension of the 'genotype-restricted likelihood' (GRL) method. *European Journal of Human Genetics*, 19(2):173–179, Feb. 2011.

[3] 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.

[4] 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, Apr. 1993.

[5] R. Elston and J. Stewart. A General Model for the Genetic Analysis of Pedigree Data. *Human Heredity*, 21(6):523–542, 1971.

[6] R. C. Elston, V. T. George, and F. Severtson. The Eiston-Stewart Algorithm for Continuous Genotypes and Environmental Factors. *Human Heredity*, 42(1):16–27, 1992.

[7] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer series in operations research. Springer, New York, 2nd ed edition, 2006.

[8] C. Schramm, C. Charbonnier, A. Zaréa, M. Lacour, D. Wallon, CNRMAJ collaborators, A. Boland, J.-F. Deleuze, R. Olaso, F. Alarcon, D. Campion, G. Nuel, and G. Nicolas. Penetrance estimation of *SORL1* loss-of-function variants using a family-based strategy adjusted on *APOE* genotypes suggest a non-monogenic inheritance. preprint, Genetics, July 2021.

[9] 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.

[10] L. R. Totir, R. L. Fernando, and J. Abraham. An efficient algorithm to compute marginal posterior genotype probabilities for every member of a pedigree with loops. *Genetics Selection Evolution*, 41(1):52, Dec. 2009.

9 Appendix

9.1 Normality of estimated parameters

9.1.1 Method

The normality of the estimated parameters $\{\theta_1, \theta_2, \theta_3\}$ is tested with a Shapiro-Wilk test which tests the null hypothesis "the tested distribution is Gaussian". Therefore, if the p-value is less than 0.05 (data is likely to occur less than 5% of the time under the null hypothesis), the null hypothesis is rejected and the tested distribution does not follow a Gaussian distribution.

9.1.2 Results

The Shapiro-Wilk test results for each parameter, each level of missingness are presented in Table 2 for standard dataset size, Table 3 for dataset size $\times 2$ and Table 4 for dataset size $\times 4$.

$\times 1$	Oracle	30%	50%	70%
θ_1	1.36e-63	1.73e-67	5.23e-66	1.30e-49
θ_2	0.269	0.000223	0.00161	2.19e-53
θ_3	2.99e-12	8.97e-31	1.31e-45	8.52e-48

Table 2: Shapiro-Wilk test p-values for each parameter and each level of missingness on dataset of standard size.

$\times 2$	Oracle	30%	50%	70%
θ_1	0.842	9.12e-50	1.07e-68	2.75e-63
θ_2	0.360	0.761	0.0708	1.78e-26
θ_3	0.00105	8.27e-07	1.15e-08	4.45e-49

Table 3: Shapiro-Wilk test p-values for each parameter and each level of missingness on dataset of size $\times 2$

×4	Oracle	30%	50%	70%
θ_1	0.417	0.000697	4.17e-64	5.89e-69
θ_2	0.0747	0.903	0.136	0.249
θ_3	0.131	0.509	0.000413	7.37e-14

Table 4: Shapiro-Wilk test p-values for each parameter and each level of missingness on dataset of size ×4

According to the results, the estimated parameters $\{\theta_1, \theta_2, \theta_3\}$ are mostly not Gaussian. However, it seems that the bigger the data (increased dataset size), the closer to Gaussian the distributions are. Similarly, the p-values increase as the missingness decrease (with highest p-values for the Oracle).

9.2 Results on modified simulations

This section contains the results obtained with the proposed method on a different set of simulations.

9.2.1 Simulations

The set of data is simulated accordingly to the model. The familial structures are real and taken from an AH-HP dataset of families with *SFTPA1* and *SFTPA2* pathogenic variants carriers. The families with more than 30 individuals are removed. The remaining data represent 206 individuals over 17 families, each family is then copied to obtain 412 individuals over 32 families. From this set of families' structures, the genotypes and phenotypes are generated the same way as the previous simulations.

- 100 datasets are generated.
- Each dataset is composed 412 individuals over 32 families.
- Autosomal dominant transmission model with 1 gene and 2 alleles ("normal" and "pathogenic"). Genotypes of founders follow Hardy-Weinberg equilibrium.
- proportion of carriers is $\pi_1 = 0.0975$
- the relative hazard $\text{RH}_1 = 20$, $\text{RH}_2 = 10$ with a cut at 50 years old;

9.2.2 Missing data

In this scenario, only the genotypes can be missing, the phenotypes are all known. These data represent more condensed families, with less individuals but with more information on the phenotypes. Therefore, the data showcase only 412 people (compared to the 744 previously), over 32 families (28 on the other dataset).

Different levels of missingness are generated:

- Oracle: all the genotypes are known.
- 50% : 50% of the genotypes are missing completely at random.
- 66% : 66% of the genotypes are missing completely at random.
- 75% : 75% of the genotypes are missing completely at random.
- 90% : 90% of the genotypes are missing completely at random.

9.2.3 Results

The results on modified simulations showcase the same trends as the results on standard simulations. The Violin plots of the estimated π_1, RH_1 and RH_2 are presented on Figure 9.

The coverage probabilities for each parameter and each level of missingness are presented on Table 5 and the distribution of the size of confidence intervals are shown on boxplot in Figure 10.

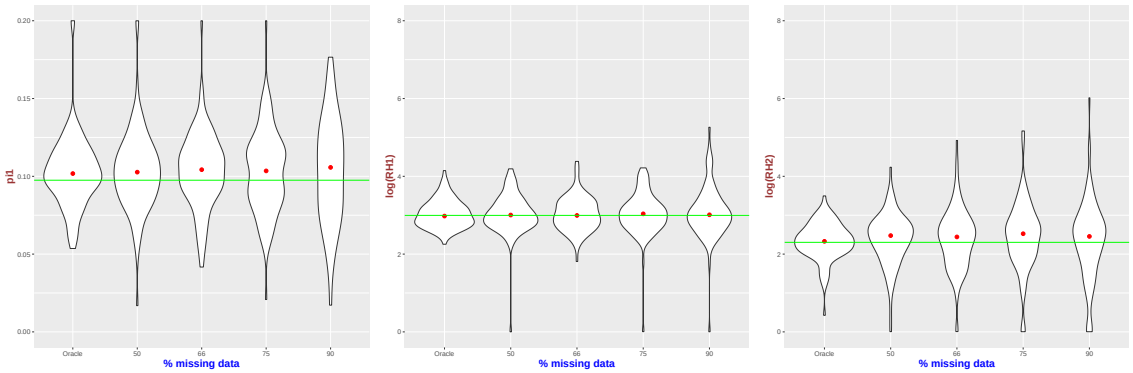

Figure 9: Violin plots of π_1, $\log(\mathrm{RH}_1)$ and $\log(\mathrm{RH}_2)$ estimation on modified standard simulations. Green line represents the real parameter to estimate.

	Oracle	50%	66%	75%	90%
π_1	0.94 [0.89,0.98]	0.89 [0.83,0.95]	0.93 [0.88,0.98]	0.93 [0.88,0.98]	0.94 [0.89,0.98]
RH_1	0.96 [0.92,0.99]	0.89 [0.83,0.95]	0.95 [0.90,0.99]	0.91 [0.85,0.96]	0.93 [0.88,0.98]
RH_2	0.97 [0.93,1.0]	0.90 [0.84,0.95]	0.91 [0.85,0.96]	0.89 [0.83,0.95]	0.87 [0.80,0.93]

Table 5: Coverage probability for each parameter and for each dataset size and missing data type.

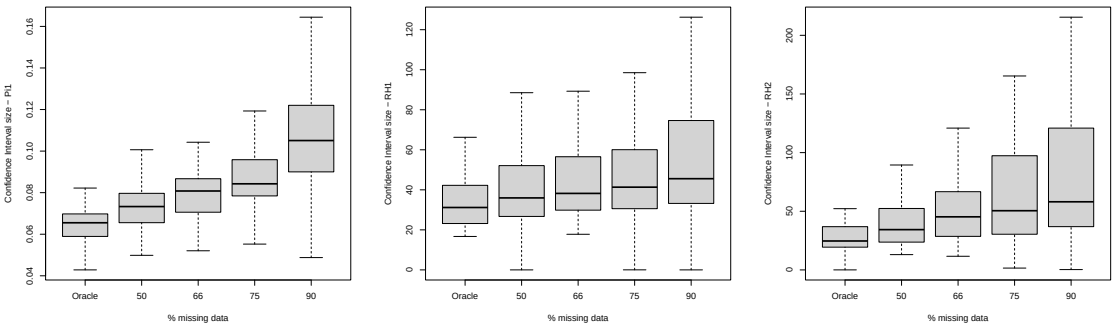

Figure 10: Boxplots of the confidence intervals sizes for parameters π_1, RH_1 and RH_2 on modified simulations.

9.3 Bootstrap for confidence interval estimation

9.3.1 Method

The bootstrap method is a resampling technique used in statistics to estimate the distribution of parameters by repeatedly resampling with replacement from the observed data.

Here is how it is performed in the context of this article:

- If the original dataset is composed of N families, randomly draw N families with replacement from the original dataset. This means that some observations may be repeated in the resampled dataset, while others may be omitted.
- Estimate the parameters with the method from the dataset generated with resampling.
- Repeat this procedure in order to generate 100 (for instance) values of estimated parameters.
- Take the 2.5 and 97.5 percentiles for the parameters which are the lower and upper bounds of the confidence interval for the estimated parameter.

9.3.2 Results

The bootstrap method is applied to 50 datasets from the standard simulations at all the levels of missingness. Each dataset is resampled 100 times. The coverage probability results are shown in Table 6.

	Oracle	30%	50%	70%
π_1	0.94 [0.86, 1.0]	0.92 [0.84, 0.98]	0.96 [0.90, 1.0]	0.98 [0.94, 1.0]
RH_1	0.94 [0.86, 1.0]	0.96 [0.90, 1.0]	0.90 [0.82, 0.98]	0.88 [0.78, 0.96]
RH_2	0.94 [0.86, 1.0]	0.90 [0.82, 0.98]	0.78 [0.66, 0.88]	0.92 [0.84, 0.98]

Table 6: Coverage probability for each parameter and for each dataset size and missing data type with the bootstrap strategy.