Penetrance of interstitial lung disease and lung cancer in *SFTPA1* or *SFTPA2* variants carriers (temporary version 28/02/2024)

Lucas Ducrot[1], Nadia Nathan[2,3], Grégory Nuel[1], Marie Legendre[3,4]

1 Sorbonne Université, Laboratory of Probability, Statistics and Modeling, Stochastics and Biology Group, Paris, France
2 Sorbonne Université, Department of Pediatric Pulmonology and Reference Center for rare lung diseases RespiRare, Armand Trousseau Hospital, APHP, Paris, France
3 Sorbonne Université, Inserm U933 Laboratory of childhood genetic diseases, Armand Trousseau Hospital, Paris, France
4 Sorbonne Université, Department of Molecular Biology, Armand Trousseau Hospital, APHP, Paris, France

Conflict of interest: the authors have no conflict of interest to disclose

Abstract

Interstitial lung disease (ILD) is a chronic condition that affects the lungs, causing progressive damage that can lead to respiratory failure over time. Recent research has identified a link between heterozygous variants in the genes encoding surfactant proteins (SP)-A1 (*SFTPA1*) and SP-A2 (*SFTPA2*) and ILD and lung cancer. The penetrance of those two rare *SFTPA1/2*-associated clinical entities is still unknown while crucial for monitoring and genetic counseling. We have identified pathogenic variants in these two genes in 27 independant families in which at least one individual had ILD and/or lung cancer. The aim of this study was to estimate the penetrance of ILD and lung cancer in heterozygotes for *SFTPA1* or *SFTPA2* variants whose pathogenicity has been confirmed through in vitro functional studies.

Based on extended pedigrees gathering 744 individuals among whom 59 carriers, phenotypic data were retrieved from 328 individuals. Penetrance for ILD and for lung cancer have been assessed by using an existing method based on an EM algorithm of which the E-step is performed through sum-product algorithm (in the Bayesian networks formed by family trees) and the M-step by Kaplan-Meier estimator.

The results show a penetrance to the first event of 50% at the age of 60 years old. The penetrance to the first event is high but not complete reaching 89.3% at the age of 80. The first event is most of the time the ILD and lung cancer typically occurred later. The penetrance to lung cancer is lower than expected (as *SFTPA1* and *SFTPA2* pathogenic variants are linked to increased risks) with a penetrance of 50% at the age of 84 years old and no case before 30.

Key words

Interstitial lung disease, lung fibrosis, lung cancer, surfactant, survival, penetrance, *SFTPA1*, *SFTPA2*

Author summary

Introduction

Interstitial lung diseases (ILD) is a heterogeneous group of rare lung disorders that affect the distal parenchyma. This disease is associated with various degree of lung inflammation and lung remodeling, often leading to lung fibrosis. Monogenic causes represent around 20% of ILD etiologies, mainly including variants in telomerase- and surfactant-related genes. Among the latter, heterozygous variants of *SFTPA1* and *SFTPA2*, encoding the surfactant proteins (SP)-A1 and SP-A2, are associated with various phenotypes ranging from asymptomatic carriers to lung fibrosis and adenocarcinoma of the lung displaying a severe prognosis and leading to lung transplantation or death [1–5]. SP-A1 and SP-A2 are highly autologous proteins that assemble to form oligomers of SP-A. The penetrance of the disease in individuals carrying *SFTPA1* or *SFTPA2* variants, and the reasons for such variability in disease expression remain unknown. Understanding the penetrance of a disease for specific groups of patients has a significant impact on medical protocols especially for risk assessment of individuals who benefit from a pre-symptomatic diagnosis or follow-up, genetic counseling, medical monitoring and prevention.

Thus, this study aims at estimating the penetrance of ILD and lung cancer in *SFTPA1* or *SFTPA2* variant carriers.

Materials and methods

Patients and relatives

In the framework of the french national network for rare lung diseases *RespiFIL*, the families of patients carrying a *SFTPA1* or a *SFTPA2* missense pathogenic variant identified in Trousseau hospital clinical laboratory were included. Pathogenicity of the variants has been confirmed by in vitro functional studies [3]. Pedigrees were analyzed and the following data were collected: age at last follow-up, age at diagnosis of ILD and/or lung cancer, and genotype when available. The study was approved by the relevant ethics committee (*Comité de protection des personnes*) and written informed consent was obtained from all participants or their legal representatives. Clinical information was collected in a legally authorized database (CNIL No. 681248).

Data were retrieved from the standardized form sent by the clinician in charge of the patient. DNA was extracted from whole blood. *SFTPA1* and *SFTPA2* variants were diagnosed by Next Generation Sequencing (NGS) capture targeted panel (SeqCap EZ Choice, Roche diagnostics) or Sanger sequencing (Big Dye V3.1 sequencing kit and 3730XL sequencing machine, Thermo Fisher Scientific). Given the high homology between the *SFTPA1* and *SFTPA2* genes, following a double inhouse-pipeline analysis, NGS data was further analyzed in the IGV viewer by setting the VAF threshold to 5%. For Sanger sequencing, PCR primers were designed to avoid variations with an allelic frequency higher than 1% in the v2.1.1 gnomAD total population. In addition, PCR primers were designed with their most 3' base on a sequence difference between *SFTPA1* and *SFTPA2* to allow specificity.

Survival to an event

Survival to an event refers to the probability that the specified event has not happened up to a certain time (time to event). Especially, it can refer to the probability that a patient has not been diagnosed with a specific disease at a certain age which is called

the survival function S(t). The function of interest that is usually considered instead is the penetrance function F(t) which is linked to the survival as $F(t) = 1 - S(t)$.

Model Description

The function of interest in this article was the penetrance $F(t)$, but for estimation and modeling purposes as well as the usage of specific packages, the estimated function is the survival $S(t)$ which is directly linked to the penetrance as $F(t) = 1 - S(t)$. The model, which is a direct implementation of the article of Alarcon [6] 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). This model can be conditioned on the genotypes and therefore can be decomposed in two subparts, a genetic one and a survival one.

For the genetic part, the joint distribution of the genotypes is given by a Bayesian network thanks to the family structure as the genotype of one individual only depends on the genotypes of its parents. These structures are informative as in pedigree data, most of the genotypes are unknown.

For the survival part, the age and status are independant conditionally to the genotypes, meaning that carriers and non-carriers do not share the same survival. The aim of the article is to estimate the survival of the mutation carriers.

The model is based the following assumptions:

1. One single locus of predisposition following autosomal dominant inheritance with two alleles (one "normal", one "pathogenic") was considered. *SFTPA1* and *SFTPA2* variants were considered in this article as a single locus since: i) they are only 55Mb apart on chromosome 10, ii) they lead to the same range of phenotypes, and iii) they never appear simultaneously in the carriers' families. This is due either to the low frequencies of variants in those genes or to the fact that the occurrence of pathogenic variants in both genes may be non-viable.

2. Genotypes of the families' founders follow Hardy-Weinberg equilibrium with an allelic frequency of the deleterious allele $f = 0.0005$ meaning 1 case over 2000 which is the upper limit for rare genetic disease.

3. Genotypes of descendants follow the Mendelian inheritance principle. The analysis is based on pedigree data, at least one individual of the included families carries a *SFTPA1* or *SFTPA2* variant and these variants run from parents to children with no *de novo* occurrence of the variants.

4. In the included families, only the carriers of the deleterious allele can be affected by the disease (no sporadic cases). ILD and lung cancer are both rare diseases. Therefore, the possibility of sporadic cases in the carrier's family is neglected.

5. The ascertainment bias is treated accordingly to the Proband's phenotype Exclusion Likelihood (PEL). [7–9], meaning that the probands are considered unaffected at age 0 in order to be uninformative toward the disease. However, their genotypes (carrier/non-carrier) were preserved and used in the analysis.

6. There is a possibility of false positives and false negatives throughout genetic testing. A genetic test is assumed to have a probability α (0.0001) of false positives and a probability β (0.02) of false negatives.

Model Fitting

In order to take into account the unknown genotypes in the data, the adopted framework is the same EM framework described by Alarcon [6]]. The objective is to estimate both the a posteriori distribution of being carrier for each individual and the survival function of variant carriers. The EM algorithm is a well-known and used method to compute the maximum likelihood of a model in presence of incomplete data (in this case the genotypes). To do so, the EM algorithm starts with a random initialization of the parameters (i.e. survival) and then alternates two steps:

- Expectation-step: during this step, using the last computed survival function (M-step), the probabilities of being carriers are updated through belief propagation (sum-product algorithm) [10] using *Bped*, a C++ implementation of the algorithm (available on demand to Grégory Nuel). It is similar to Elston-Stewart algorithm [11,12] with an additional backward propagation in order to compute the marginal distribution.
- Maximization-step: during this step, the new survival function is updated using a weighted Kaplan-Meier estimator [13,14]. The weights used are the probabilities of being carrier computed for each individual during the E-step.

These two steps are iterated until convergence or up to 300 iterations.

Statistics

The 95% confidence intervals of the survival functions are computed using the R package Survival [13,14]. The method is used both on an unstratified population and male/female stratified population in order to see if the survival is sex-dependent. The significance of the difference between male and female survival is estimated with a log-rank test. Two methods are used to perform the male/female stratification. The first method is a simple male/female stratification where each group has its own independent survival function. The second method uses Cox proportional hazard model to assess the association between the survival and the sex variable. Both methods are implemented with the R package survival [13,14]. In order to quantify the role of each parameter (f frequency of the allele, α probability of False Positive in genetic testing and β probability of False Negative in genetic testing), a sensitivity analysis is performed. While one parameter is analysed, the others are set at their based values (i.e. $f = 0.0005$, $\alpha = 0.0001$, $\beta = 0.02$). Each parameter is tested over a particular set of values (i.e. $f \in \{0.05, 0.005, 0.0005, 0.00005\}$, $\alpha \in \{0.01, 0.001, 0.0001, 0.00001\}$, $\beta \in \{0.02, 0.002, 0.0002\}$).

Results

Genotypes and phenotypes

A total of 27 families have been included in this study. A *SFTPA1* pathogenic variant was identified in 10 families and a *SFTPA2* pathogenic variant was identified in 17 families. The pedigrees are provided in Supplemental Figure **??**. Among the families, the data of 27 index patients and 717 relatives were analyzed, accounting for a total of 744 included individuals (Table 7). A total of 22 and 37 individuals carried a *SFTPA1* or *SFTPA2* pathogenic variant respectively. An ILD was diagnosed in 64, a lung cancer in 23 and both in 20. Individuals were declared as asymptomatic in 221 cases and the clinical status was unknown in 416 cases. At the study time, 119 individuals were deceased, including 4 from ILD or lung cancer. The median age of the living individuals at the study time was 43 years. The median age at the disease onset was 49 years.

	SFTPA1 (n)	SFTPA2 (n)	Total (n)
Families	10	17	27
Individuals :	279	465	744
Males	143	241	384
Females	136	224	360
Dead	46	73	119
Median age at study time	49	41	43
Median age at death	56.5	52.5	54.5
Genotype:			
Heterozygotes for a pathogenic variant	22	37	59
Non-carriers	14	27	41
Unknown	243	401	644
Phenotype:			
Asymptomatic	80	141	221
ILD only	24	40	64
Lung cancer only	5	18	23
Both	6	14	20
Unknown	164	252	416
Median age at ILD and/or lung cancer diagnosis	45	49.5	49

Table 1. Main characteristics, phenotype and genotype of the patients and relatives. Abbreviations: ILD, interstitial lung disease.

Survivals of ILD, lung cancer and to first event for *SFTPA1* or *SFTPA2* variants carriers

The method is applied to compute the survivals for ILD and lung cancer alone and survival to the first event. The survival functions are presented in Figure 5 and in Table 2. The survival to first event at 30 year-old was 0.93. ILD appeared before lung cancer in 15% (3 over 20 diagnosed both with ILD and lung cancer) of cases. The youngest age at lung cancer diagnosis was 30 years. *SFTPA1* and *SFTPA2* pathogenic variant carriers present a high risk of developing either ILD or lung cancer as the penetrance to first event at 80 year-old is 89.4% [74.1-95.7].

Age	ILD	Lung cancer	First event
10	0.983 [0.951 – 1.000]	1.000 [1.000 – 1.000]	0.983 [0.956 – 1.000]
20	0.983 [0.951 – 1.000]	1.000 [1.000 – 1.000]	0.985 [0.956 – 1.000]
30	0.924 [0.856 – 0.998]	0.982 [0.949 – 1.000]	0.933 [0.872 – 0.998]
40	0.795 [0.690 – 0.916]	0.964 [0.917 – 1.000]	0.820 [0.726 – 0.927]
50	0.688 [0.565 – 0.837]	0.939 [0.874 – 1.000]	0.707 [0.593 – 0.843]
60	0.379 [0.254 – 0.566]	0.686 [0.542 – 0.868]	0.374 [0.258 – 0.543]
70	0.126 [0.054 – 0.291]	0.530 [0.355 – 0.791]	0.153 [0.078 – 0.300]
80	0.071 [0.024 – 0.203]	0.530 [0.355 – 0.791]	0.106 [0.043 – 0.259]
90	0.071 [0.024 – 0.203]	0.330 [0.138 – 0.788]	0.056 [0.017 – 0.183]

Table 2. Survival to ILD, lung cancer and to the first event.

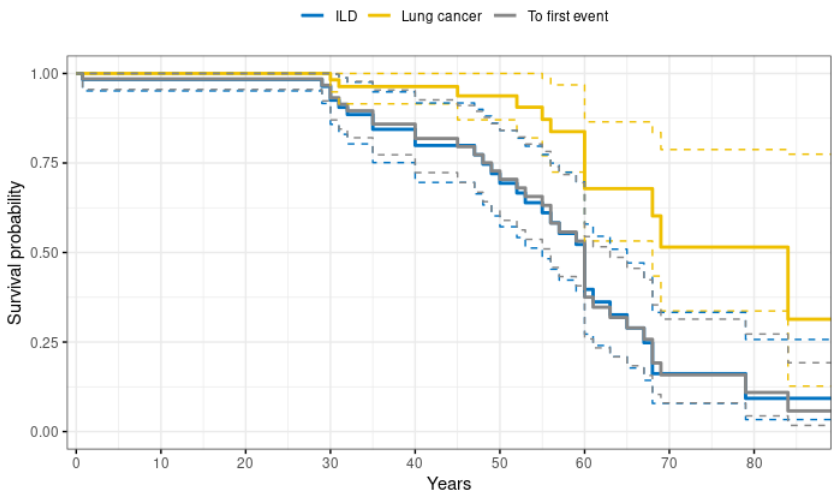


Fig 1. Survival functions (solid lines) and 95% confidence intervals (dotted lines) are provided for interstitial lung disease (ILD) (blue lines), lung cancer (yellow lines) and first event (grey lines).

Sensitivity analysis

The sensitivity analysis showed very low variation dependencies to the different parameters f, α and β for all the computed survivals (Supplementary material and Supplementary Tables 4, 5 and 6 providing survivals to ILD or lung cancer alone and survivals to the first event at 30, 50 and 70 year-old). Considering f, the low dependency probably comes from the fact there is at least one carrier in each family. Therefore, the probability of being carrier relies more on being a relative of a variant carrier than the frequency of the allele in the general population. Considering α and β, the variations are low because there are many variant carriers showcase a disease which, in the model, consolidates the fact they are variant carriers.

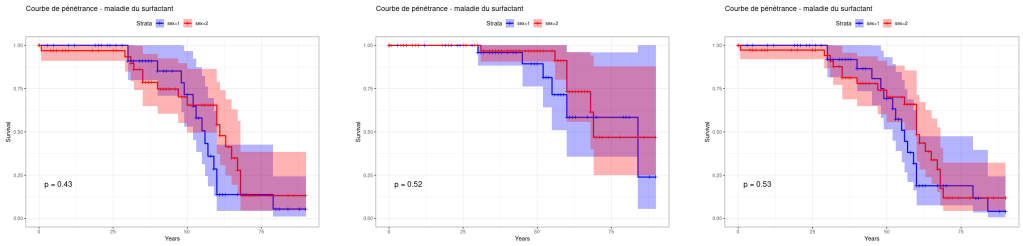
Male/Female stratification

The method was also applied with a stratification male/female, as the previous results, to compute the survivals for both ILD and lung cancer alone and survival to the first event.

There is a trend for a better survival to the first event in male compared to female before 50 years old, the trend interchanged after 50 (Figures 2 and 3). The differences, however, do not reaching significance (p-values reported in Table 3). More details are presented in Supplementary materials.

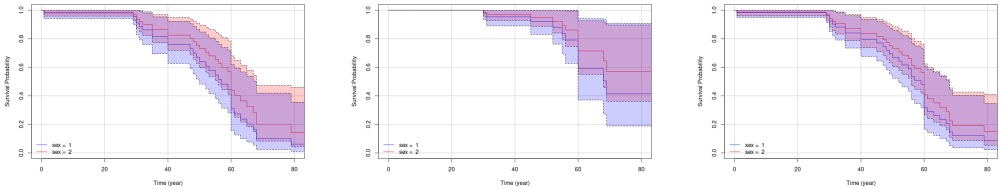
Method	ILD	Lung cancer	To First Event
Standard	0.43	0.52	0.53
Cox	0.34	0.43	0.47

Table 3. P-values for both standard stratification and Cox proportional hazard model and each disease



ILD Lung cancer To first event

Fig 2. Standard Male/Female stratification for ILD, Lung cancer and to the first event survival estimations for *SFTPA1* and *SFTPA2* mutation carriers



ILD Lung cancer To first event

Fig 3. Male/Female Cox proportional hazard method for ILD, Lung cancer and to the first event survival estimations for *SFTPA1* and *SFTPA2* mutation carriers

Discussion

The present study provides the first estimation of penetrance of ILD, lung cancer and of first of these two events for *SFTPA1* and *SFTPA2* variant carriers. These data were obtained using a large cohort of patients carrying *SFTPA1* and *SFTPA2* variants and their relatives. Based on the mathematical model, the computed risk showed that the risk of ILD or lung cancer in *SFTPA1* or *SFTPA2* carriers increases with age (mean age 60) to reach an important – but not complete – penetrance of 89.3 at 80 year-old. Interestingly, despite these variants being previously shown to be associated with a high risk of lung cancer, we prove herein that this event occurs mainly later than ILD, with a penetrance of 50% at 80 year-old, and with no case before 30 years.

Penetrance of dominant diseases is known to be heterogeneous. However, great variations are described depending on the involved gene and the functional consequences of the variants. *SFTPA1* and *SFTPA2* variants are associated with an impaired expression and secretion of the corresponding SP-A1 and SP-A2 proteins. However, the reason(s) why the age at onset of the ILD symptoms vary from a few months to more than 80 years is currently unknown. Viral infections - especially in children - and environmental or occupational exposures in adults may have a role in triggering the disease. Unfortunately, these factors could not be retrieved for a relevant number of patients and relatives for study.

As expected in an autosomal disease, the study did not show any significant difference for penetrance between men and women.

Surfactant disorders are rare causes of ILD in adults. When a *SFTPA1* or *SFTPA2* variant is identified, the result is reported to the patient in the framework of a genetic counseling. In France, the patient has to inform his/her relatives about the genetic results in order to give them the opportunity to ask for a pre-symptomatic diagnosis. This analysis is offered to relatives over the age of majority who are at risk of carrying the variant.

Since 2018, the national network for rare lung disease (*RespiFIL*, www.respifil.fr)

has launched multidisciplinary team meetings for genetic forms of ILD in adults and in children. Despite more and more cases being diagnosed, no guidelines are currently available for the pre-symptomatic management of surfactant diseases. The present study provides crucial information and could help to discuss the following management strategies: (i) the first CT-scan may not be useful before 30 year-old; (ii) if the CT-scan is normal, the timeline between two CT-scans xxxx (iii) *SFTPA1* or *SFTPA2* variant carriers should receive a clear information on environmental factors that could increase risk of lung fibrosis and cancer such as tobacco smoking or occupational exposures.

To validate the model and observe the natural history of the disease, a prospective study including patients with a surfactant-related disease and their adult relatives is currently in progress (*RaDiCo-ILD2*, ClinicalTrials.gov ID NCT06036719). Continuing data collection, including tobacco and occupational exposures is a perspective work to strengthen the results of the study.

The study displays some limits, especially due to the missing data in far relatives, but also because the model is based on the assumption that the disease (lung fibrosis and lung cancer) does not present sporadic cases. This assumption is factually not true but was chosen as the probability of sporadic cases in the observed families is very low and may be neglected. An improvement of the model could be to develop a mathematical model taking sporadic cases into account to compute the penetrance of such disease.

Conclusion

This study estimated the penetrance of interstitial lung disease (ILD) and lung cancer in individuals carrying *SFTPA1* or *SFTPA2* pathogenic variant. The investigation involved 27 independent families with at least one member being *SFTPA1* or *SFTPA2* pathogenic variant carrier.

The penetrance is estimated using an existing method based on an EM algorithm of which the E-step is performed through sum-product algorithm (in the Bayesian networks formed by family trees) and the M-step by Kaplan-Meier estimator.

The results show a penetrance to the first event of 50% at the age of 60 years old. The penetrance to the first event is high but not complete reaching 89.3% [74.0-95.6] at the age of 80. The first event is most of the time the ILD and lung cancer typically occurred later. The penetrance to lung cancer is lower than penetrance to ILD (as *SFTPA1* and *SFTPA2* pathogenic variants are linked to increased risks) with a penetrance of 50% at the age of 84 years old and no case before 30.

Following the guideline, ILD diagnosed after 50 are currently not considered as genetically related. This study shows that for *SFTPA1* and *SFTPA2* pathogenic variant carriers, the median age at ILD diagnosis is 60 [55-65] which means that genetic testing could be appropriate for later forms of ILD.

While acknowledging certain limitations, such as missing data and assumptions about sporadic cases, the study sets the stage for further research. Ongoing prospective studies, like RaDiCo-ILD2, aim to validate the model and enhance understanding of the natural history of these diseases. Continued data collection, including environmental factors, is crucial for refining and strengthening the outcomes of this study.

Legal and ethical statement

Written informed consents were obtained from the patients.

Acknowledgments

We thank 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.

Authors contribution

LD wrote the manuscript. NN, GN and ML reviewed the manuscript. NN and ML collected and computed the data. LD and GN performed the mathematical analyses.

References

1. Nathan N, Giraud V, Picard C, Nunes H, Dastot-Le Moal F, Copin B, et al. Germline *SFTPA1* mutation in familial idiopathic interstitial pneumonia and lung cancer. Human Molecular Genetics. 2016;25(8):1457–1467. doi:10.1093/hmg/ddw014.

2. Nathan N, Legendre M, Kannengiesser C, Albuisson J, Borie R, Bouvry D, et al. SFTPA mutations in interstitial lung disease (ILD) and lung cancer. In: Diffuse Parenchymal Lung Disease. European Respiratory Society; 2017. p. PA1516. Available from: `http://erj.ersjournals.com/lookup/doi/10.1183/1393003.congress-2017.PA1516`.

3. Legendre M, Butt A, Borie R, Debray MP, Bouvry D, Filhol-Blin E, et al. Functional assessment and phenotypic heterogeneity of *SFTPA1* and *SFTPA2* mutations in interstitial lung diseases and lung cancer. European Respiratory Journal. 2020;56(6):2002806. doi:10.1183/13993003.02806-2020.

4. Wang Y, Kuan PJ, Xing C, Cronkhite JT, Torres F, Rosenblatt RL, et al. Genetic Defects in Surfactant Protein A2 Are Associated with Pulmonary Fibrosis and Lung Cancer. The American Journal of Human Genetics. 2009;84(1):52–59. doi:10.1016/j.ajhg.2008.11.010.

5. Liu L, Liu YJ, Guo T, Luo H. Identification of a Missense Mutation in the *Surfactant Protein A2* Gene in a Chinese Family with Interstitial Lung Disease. DNA and Cell Biology. 2021;40(1):126–131. doi:10.1089/dna.2020.6045.

6. Alarcon F, Planté-Bordeneuve V, Olsson M, Nuel G. Non-parametric estimation of survival in age-dependent genetic disease and application to the transthyretin-related hereditary amyloidosis. PLOS ONE. 2018;13(9):e0203860. doi:10.1371/journal.pone.0203860.

7. Alarcon F, Bourgain C, Gauthier-Villars M, Planté-Bordeneuve V, Stoppa-Lyonnet D, Bonaïti-Pellié C. PEL: an unbiased method for estimating age-dependent genetic disease risk from pedigree data unselected for family history. Genetic Epidemiology. 2009;33(5):379–385. doi:10.1002/gepi.20390.

8. Anheim M, Elbaz A, Lesage S, Durr A, Condroyer C, Viallet F, et al. Penetrance of Parkinson disease in glucocerebrosidase gene mutation carriers. Neurology. 2012;78(6):417–420. doi:10.1212/WNL.0b013e318245f476.

9. Schramm C, Charbonnier C, Zaréa A, Lacour M, Wallon D, CNRMAJ collaborators, et al. Penetrance estimation of *SORL1* loss-of-function variants using a family-based strategy adjusted on *APOE* genotypes suggest a non-monogenic inheritance. Genetics; 2021. Available from: `http://biorxiv.org/lookup/doi/10.1101/2021.06.30.450554`.

10. Totir LR, Fernando RL, Abraham J. An efficient algorithm to compute marginal posterior genotype probabilities for every member of a pedigree with loops. Genetics Selection Evolution. 2009;41(1):52. doi:10.1186/1297-9686-41-52.

11. Elston RC, Stewart J. A General Model for the Genetic Analysis of Pedigree Data. Human Heredity. 1971;21(6):523–542. doi:10.1159/000152448.

12. Elston RC, George VT, Severtson F. The Eiston-Stewart Algorithm for Continuous Genotypes and Environmental Factors. Human Heredity. 1992;42(1):16–27. doi:10.1159/000154043.

13. Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. 2nd ed. Statistics for biology and health. New York Berlin Heidelberg: Springer; 2001.

14. Therneau TM. A Package for Survival Analysis in R; 2023. Available from: `https://CRAN.R-project.org/package=survival`.

Supplementary Material

Sensitivity Analysis

Method

The hyperparameters of the model are α (False Positives probability of genetic testing) set at 0.0001, β (False Negatives probability of genetic testing) set at 0.02 and f (the frequency of the deleterious allele in the general population) set at 0.0005.

This sensitivity analysis investigates the variations of the results for multiple values of α, β and f.

In order to quantify the role of each parameter while one parameter is analysed, the other are set at their based values (i.e. $f = 0.0005$, $\alpha = 0.0001$, $\beta = 0.02$). Each parameter is tested over particular values (i.e. $f \in \{0.05, 0.005, 0.0005, 0.00005\}$, $\alpha \in \{0.01, 0.001, 0.0001, 0.00001\}$, $\beta \in \{0.02, 0.002, 0.0002\}$).

Results

The results are presented in the Tables 1, 2 and 3 which provide survivals to both ILD and lung cancer alone and survivals to the first event at 30, 50 and 70 years old.

The sensitivity analysis shows very low variation dependencies to the different parameters f, α and β for all the computed survivals. Some possible reasons can explain such low dependencies of the results to these parameters.

Considering f, the low dependency probably may come from the fact there is at least one carrier in each family. Therefore the probability of being carrier relies more on being a relative of a mutation carrier than the frequency of the allele in the general population.

Considering α and β, these parameters are here to take into account potential genetic testing errors, such as a possible non-carrier affected by the disease which is not possible in the model (consdering disease with no sporadic cases). The variations may be low because there is no such cases in the data.

ILD			
α	**30**	**50**	**70**
1e-05	0.926 [0.859,0.998]	0.696 [0.576,0.842]	0.165 [0.081,0.336]
1e-04	0.925 [0.858,0.998]	0.693 [0.571,0.840]	0.161 [0.078,0.332]
1e-03	0.923 [0.853,0.998]	0.684 [0.560,0.836]	0.152 [0.071,0.322]
1e-02	0.920 [0.849,0.998]	0.677 [0.552,0.832]	0.146 [0.067,0.315]
Lung cancer			
α	**30**	**50**	**70**
1e-05	0.982 [0.948,1.000]	0.938 [0.871,1.000]	0.518 [0.340,0.790]
1e-04	0.982 [0.948,1.000]	0.937 [0.870,1.000]	0.515 [0.336,0.787]
1e-03	0.981 [0.946,1.000]	0.934 [0.864,1.000]	0.498 [0.320,0.775]
1e-02	0.979 [0.941,1.000]	0.927 [0.850,1.000]	0.466 [0.291,0.746]
To First Event			
α	**30**	**50**	**70**
1e-05	0.932 [0.871,0.998]	0.707 [0.593,0.842]	0.160 [0.081,0.316]
1e-04	0.932 [0.870,0.998]	0.704 [0.589,0.841]	0.158 [0.080,0.313]
1e-03	0.930 [0.867,0.998]	0.697 [0.580,0.837]	0.152 [0.075,0.306]
1e-02	0.928 [0.863,0.998]	0.690 [0.572,0.833]	0.144 [0.070,0.294]

Table 4. Survivals at 30, 50 and 70 years old for different values of α with $\beta = 0.02$ and $f = 0.0005$ fixed.

ILD			
β	**30**	**50**	**70**
2e-04	0.924 [0.855,0.998]	0.688 [0.566,0.838]	0.154 [0.0739,0.322]
2e-03	0.924 [0.856,0.998]	0.689 [0.566,0.838]	0.155 [0.0746,0.323]
2e-02	0.925 [0.858,0.998]	0.693 [0.571,0.840]	0.161 [0.0786,0.332]
Lung cancer			
β	**30**	**50**	**70**
2e-04	0.981 [0.947,1.000]	0.936 [0.868,1.000]	0.507 [0.330,0.780]
2e-03	0.981 [0.947,1.000]	0.936 [0.868,1.000]	0.509 [0.331,0.781]
2e-02	0.982 [0.948,1.000]	0.937 [0.870,1.000]	0.515 [0.336,0.787]
To First Event			
β	**30**	**50**	**70**
2e-04	0.931 [0.868,0.998]	0.700 [0.585,0.839]	0.153 [0.076,0.306]
2e-03	0.931 [0.868,0.998]	0.701 [0.585,0.839]	0.154 [0.077,0.307]
2e-02	0.932 [0.870,0.998]	0.704 [0.589,0.841]	0.158 [0.080,0.313]

Table 5. Survivals at 30, 50 and 70 years old for different values of β with $\alpha = 0.0001$ and $f = 0.0005$ fixed.

Male/Female Stratification

Method

The method is used both on an unstratified population and male/female stratified population in order to see if survivals to both ILD and lung cancer alone and survival to the first event are sex-dependant as it is standard to test in clinical statistics. For the male/female stratification, the significance of the difference between male and female survivals is estimated with a log-rank test.

To stratified the population into male and female, two methods are used :

- A standard stratification where male and female survivals are estimated separatively

ILD			
f	**30**	**50**	**70**
5e-05	0.924 [0.856,0.998]	0.690 [0.567,0.839]	0.163 [0.078,0.341]
5e-04	0.925 [0.858,0.998]	0.693 [0.571,0.840]	0.161 [0.078,0.332]
5e-03	0.924 [0.855,0.998]	0.686 [0.563,0.836]	0.136 [0.063,0.296]
5e-02	0.924 [0.856,0.998]	0.688 [0.565,0.837]	0.126 [0.054,0.291]
Lung cancer			
f	**30**	**50**	**70**
5e-05	0.981 [0.946,1.000]	0.934 [0.865,1.000]	0.502 [0.324,0.780]
5e-04	0.982 [0.948,1.000]	0.937 [0.870,1.000]	0.515 [0.336,0.787]
5e-03	0.982 [0.948,1.000]	0.937 [0.870,1.000]	0.515 [0.338,0.785]
5e-02	0.982 [0.949,1.000]	0.939 [0.874,1.000]	0.530 [0.355,0.791]
To First Event			
f	**30**	**50**	**70**
5e-05	0.931 [0.868,0.998]	0.701 [0.585,0.839]	0.158 [0.079,0.316]
5e-04	0.932 [0.870,0.998]	0.704 [0.589,0.841]	0.158 [0.080,0.313]
5e-03	0.932 [0.870,0.998]	0.704 [0.589,0.841]	0.155 [0.078,0.308]
5e-02	0.933 [0.872,0.998]	0.707 [0.593,0.843]	0.153 [0.078,0.300]

Table 6. Survivals at 30, 50 and 70 years old for different values of f with $\alpha = 0.0001$ and $\beta = 0.02$ fixed.

during the M-step of the EM aglorithm.

- A Cox proportional hazard model where the male and female survivals are estimated jointly during the M-step and share an exponential coefficient.

Results

With the two stratification methods, the results show differences between male and female survivals but non-significant (P-value reported in Table 1).

With the two stratification methods, the results show differences between male and female survival functions but the log-rank test is not significant (P-value reported on the graphs). It is still possible that the sex plays a role in the survival to the ILD or lung cancer for *SFTPA1* or *SFTPA2* mutation carriers but the available data are not currently sufficient to assess that properly.

Method	**ILD**	**Lung cancer**	**To First Event**
Standard	0.43	0.52	0.53
Cox	0.34	0.43	0.47

Table 7. P-values for both standard stratification and Cox proportional hazard model and each disease

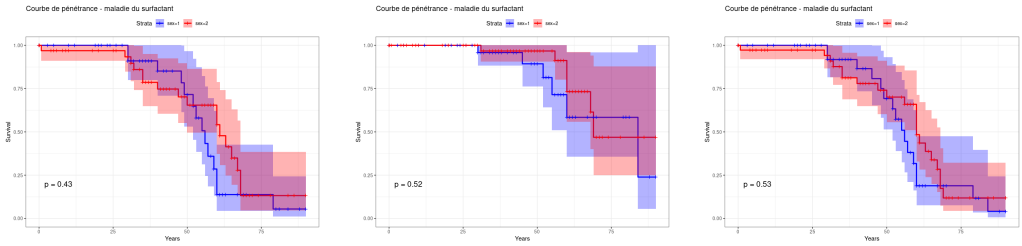
Model

Model description

In the context of genetic diseases with age dependencies, the function of interest is generally the penetrance:

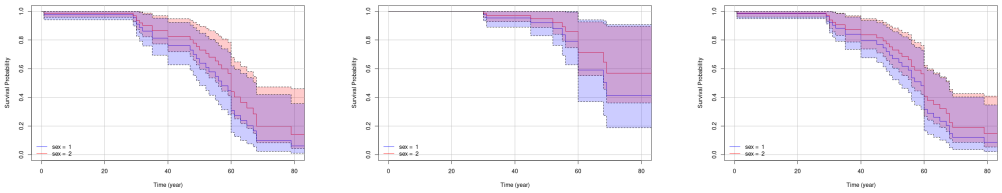
$$F(t) = \mathbb{P}(\text{disease diagnosed before age t})$$

,



ILD Lung cancer To first event

Fig 4. Standard Male/Female stratification for ILD, Lung cancer and to the first event survival estimations for *SFTPA1* and *SFTPA2* mutation carriers



ILD Lung cancer To first event

Fig 5. Male/Female Cox proportional hazard method for ILD, Lung cancer and to the first event survival estimations for *SFTPA1* and *SFTPA2* mutation carriers

but in this article, the estimated function is the survival function:

$$S(t) = \mathbb{P}(\text{disease not diagnosed before age t})$$

.

Though it is easy to retrieve the penetrance as it is directly linked to the survival:

$$S(t) = 1 - F(t)$$

.

The model, which is a direct implementation of the method of Alarcon [6], 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 $\mathcal{I} = \{1, ..., n\}$, 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 wild-type allele and 1 the deleterious 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. Finally $G = (G_1, ..., G_n) \in \{0, 1\}^n$ represent the genetic test where G_i is the result of the genetic testing for individual i (0 for non-carriers and 1 for carriers). (T, δ) and G are considered independant conditionnally to X meaning that genetic testing process is independant from the age and status of individual. Therefore this model can be conditioned on the genotype X and decomposed in two subparts, a genetic one, a survival one and a genetic testing one as follows:

$$\mathbb{P}(T, \delta, X, G) = \mathbb{P}(X) \times \mathbb{P}(T, \delta, G|X) = \underbrace{\mathbb{P}(X)}_{\text{Genetic Part}} \times \underbrace{\mathbb{P}(T, \delta|X)}_{\text{Survival Part}} \times \underbrace{\mathbb{P}(G|X)}_{\text{Testing Part}} \quad (1)$$

- **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, they follow Hardy-Weinberg equilibrium with allelic frequency f, the non-founders follow Mendelian transmission from parents:

$$\mathbb{P}(X) = \prod_{i \in \mathcal{F}} \mathbb{P}(X_i) \prod_{i \notin \mathcal{F}} \mathbb{P}(X_i | X_{\text{pat}_i}, X_{\text{mat}_i}) \tag{2}$$

- **Survival Part:** $S(t)$ and $\lambda(t)$ represent survival and hazard rate for mutation carriers

$$\mathbb{P}(T_i = t, \delta_i = 0 | X_i) = \begin{cases} S(t) & \text{if } X_i \neq 00 \\ 1 & \text{if } X_i = 00 \end{cases} \tag{3}$$

$$\mathbb{P}(T_i = t, \delta_i = 1 | X_i) = \begin{cases} S(t)\lambda(t) & \text{if } X_i \neq 00 \\ 0 & \text{if } X_i = 00 \end{cases} \tag{4}$$

- **Testing Part:** α represents the probability of False Positive (being genotyped as carrier while being non-carrier) and β represents the probability of False Positive (being genotyped as non-carrier while being carrier).
  - True negative: $\mathbb{P}(G_i = 0 | X_i = 00) = 1 - \alpha$
  - False negative: $\mathbb{P}(G_i = 0 | X_i \neq 00) = \beta$
  - False positive: $\mathbb{P}(G_i = 1 | X_i = 00) = \alpha$
  - True positive: $\mathbb{P}(G_i = 1 | X_i \neq 00) = 1 - \beta$

From this model description, the proposed model is based on some assumptions:

1. One single locus of predisposition following autosomal dominant inheritance with two alleles (one "wild-type", one "deleterious").
2. Genotypes of the families' founders follow Hardy-Weinberg equilibrium with an allelic frequency of the deleterious allele f.
3. Genotypes of descendants follow the Mendelian inheritance principle.
4. Only the carriers of the deleterious allele can be affected by the disease (no sporadic cases).
5. The ascertainment bias is treated accordingly to PEL standard (Proband's phenotype Exclusion Likelihood) [7].
6. Possibility of False Positives and False Negative throughout genetic testing. The hypothesis allows to better take into account genetic testing. A genetic test is assumed to have a probability α of False Positives and a probability β of False Negatives.

Model Fitting with EM algorithm

The model fitting is performed using the EM framework on the pedigree data as described in the article of Alarcon [6]. The objective is to estimate both the a posteriori distribution of being carrier for each individual and the survival function of mutation carriers. The EM algorithm is a well-known and used method to compute the maximum likelihood of a model in presence of incomplete data (in this case the genotypes). To do so, an auxiliary Q function need to be introduced. Here is a brief description of EM algorithm:

Idea: EM Algorithm is an iterative algorithm used to find parameters of the maximum log-likelihood of probabilistic models with latent variables

Model: T, X random variables following a distribution of parameter θ, X is unobserved

Maximum Likelihood Estimator: $\hat{\theta} = \arg\max_{\theta} \sum_{X} \mathbb{P}(T, X|\theta)$

Auxiliary function:

$$Q(\theta|\theta_{\text{old}}) = \int \mathbb{P}(X|T;\theta_{\text{old}}) \log \mathbb{P}(T, X|\theta) dX$$

Algorithm:

- **Expectation-step:** compute Expectation of $Q(\theta|\theta_{\text{old}})$
- **Maximization-step:** maximization of Q to find $M(\theta) = \arg\max_{\theta'} Q(\theta'|\theta)$

Application to the model

Applied to the model described in this article, the auxiliary function Q can be written as follows:

$$Q(\theta|\theta_{\text{old}}) = \text{cst.} + \sum_{i} \mathbb{P}(X_i \neq 00|\text{ev};\theta_{\text{old}}) \log \mathbb{P}(T_i, \delta_i|X_i \neq 00;\theta) \tag{5}$$

Starting from arbitrary θ, then the two steps of the EM algorithm are done as follows:

- Expectation-step: during this step, using the last computed survival function $\theta_{\text{old}} = \theta$ (M-step), the probabilities of being carriers $w_i = \mathbb{P}(X_i \neq 00|\text{ev}, \theta_{old})$ are updated through belief propagation.
- M-step: during this step, the new survival function θ is computed using a weighted Kaplan-Meier estimator [13, 14] which maximizes the Q function. The weights used are the probabilities of being carrier w_i computed for each individual during the E-step.

These two steps are iterated until convergence or up to 300 irations.

E-step

The E-step is performed using Bped, an implementation of belief propagation (sum-product algorithm) [10] in C++ (available on demand to Grégory Nuel), as used in Alarcon's article [6]. It is similar to Elston-Stewart algorithm [11, 12] with an additional backward propagation in order to compute the marginal distribution. Bped requires an evidence file to compute the a posteriori law of genotypes. The initial evidence can be written as follows:

- For individuals that are unaffected ($\delta = 0$):

$$\mathbb{P}(T_i = t, \delta_i = 0|X_i) \propto \begin{cases} S(t) & \text{if } X_i \neq 00 \\ 1 & \text{if } X_i = 00 \end{cases} \tag{6}$$

- For individuals that are affected ($\delta = 1$):

$$\mathbb{P}(T_i = t, \delta_i = 1|X_i) \propto \begin{cases} 1 & \text{if } X_i \neq 00 \\ 0 & \text{if } X_i = 00 \end{cases} \tag{7}$$

While taking into account the possibility of genotyping errors, the evidence can be modified as such (only for genotyped individuals which represent a fraction of the total population in the data):

- For individuals that are genotyped as non-carriers ($G = 0$):

$$\mathbb{P}(T_i = t, \delta_i, G_i = 0|X_i) = \begin{cases} \mathbb{P}(T_i = t, \delta_i|X_i \neq 00) \times \beta & \text{if } X_i \neq 00 \\ \mathbb{P}(T_i = t, \delta_i|X_i = 00) \times (1 - \alpha) & \text{if } X_i = 00 \end{cases} \tag{8}$$

- For individuals that are genotyped as carriers ($G = 1$):

$$\mathbb{P}(T_i = t, \delta_i, G_i = 1|X_i) = \begin{cases} \mathbb{P}(T_i = t, \delta_i|X_i \neq 00) \times (1 - \beta) & \text{if } X_i \neq 00 \\ \mathbb{P}(T_i = t, \delta_i|X_i = 00) \times \alpha & \text{if } X_i = 00 \end{cases} \tag{9}$$

M-step

The M-step is performed with weighted Kaplan-Meier survival estimator [13, 14] which maximizes exactly the defined Q auxiliary function using as weights the $w_i = \mathbb{P}(X_i \neq 00|ev, \theta_{old})$.

$$Q(\theta|\theta_{\text{old}}) = \text{cst.} + \sum_i \underbrace{\mathbb{P}(X_i \neq 00|\text{ev}; \theta_{\text{old}})}_{\text{weights } w_i} \log \underbrace{\mathbb{P}(T_i, \delta_i|X_i \neq 00; \theta)}_{\text{survival}} \tag{10}$$