Penetrance estimation of *SORL1* loss-of-function variants using a family-based strategy adjusted on *APOE* genotypes suggest a non-monogenic inheritance

Catherine Schramm,¹ Camille Charbonnier,¹ Aline Zaréa,² Morgane Lacour,² David Wallon,² CNRMAJ collaborators, Anne Boland,³ Jean-François Deleuze,³ Robert Olaso,³ Flora Alarcon,⁴ Dominique Campion,^{1,5} Grégory Nuel,⁶ Gaël Nicolas^{1*}

1 Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, Department of Genetics and CNR-MAJ, 76000 Rouen, France

2 Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, Department of Neurology and CNR-MAJ, 76000 Rouen, France

3 Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine, 91057, Evry, France

4 MAP5, UMR-CNRS 8145, Paris University, 75270 Paris, France

5 Department of Research, Rouvray Psychiatric Hospital, Sotteville-Lès-Rouen, 76681 France.

6 LPSM, CNRS 8001, Sorbonne University, 75005 Paris, France

* corresponding author: gaelnicolas@hotmail.com

Abstract

For complex disorders, estimating the age-related penetrance associated with rare variants of strong effect is essential before a putative use for genetic counseling or disease prevention. However, rarity and co-occurrence with other risk factors make such estimations difficult. In the context of Alzheimer disease, we present a survival model to estimate the penetrance of *SORL1* rare (allele frequency < 1%) Loss-of-Function variants (LoF) while accounting for *APOE*- ϵ 4, the main risk factor (allele frequency $\sim$ 14% in Caucasians). We developed an efficient strategy to compute penetrance estimates accounting for both common and rare genetic variants based on available penetrance curves associated with common risk factors and using incomplete pedigree data to quantify the additional risk conferred by rare variants. Our model combines: (i) a baseline for non-carriers of *SORL1* LoF variants, stratified by *APOE* genotypes derived from the Rotterdam study and (ii) an age-dependent proportional hazard effect for *SORL1* LoF variants estimated from pedigrees with a proband carrying such a variant. We embed this model into an Expectation-Maximisation algorithm to accommodate for missing genotypes. Confidence intervals were computed by bootstraps. To correct for ascertainment bias, proband phenotypes were omitted. We obtained penetrance curves associated with *SORL1* LoF variants at the digenic level. By age 70, we estimate a 100% penetrance of *SORL1* LoF variants only among *APOE*- ϵ 4 ϵ 4 carriers, while penetrance is 56%[40% – 72%] among ϵ 4 heterozygous carriers and 37%[26% – 51%] among ϵ 4 non-carriers. We conclude that rare *SORL1* LoF variants should not be used for genetic counseling regardless of the *APOE* status.

Introduction

In the context of complex disorders, large scale case-control studies based on exome and genome sequencing have enabled the discovery of rare risk variants with strong effects. Such variants could not be caught by genome-wide association studies (GWAS) performed on DNA chips¹. While the variant effects are usually presented as odds ratios or relative risks, absolute penetrance estimates, i.e. the probability of developing the disease according to age for carriers of a given genotype, are much more informative in a clinical setting. Penetrance estimates are indeed essential for genetic counselling, presymptomatic monitoring or preventive strategies. However, both rarity and co-occurrence with other risk factors make it difficult to estimate the penetrance associated with rare variants. Most common complex disorders indeed result from the combined effect of risk factors each associated with diverse degrees of penetrance.

Alzheimer disease (AD) provides a perfect example of this challenge. AD is a common disorder generally affecting elderly patients. In the vast majority of AD cases, the disease is a complex disorder with a high genetic component (for review, see Nicolas et al.²). The $\epsilon 4$ allele of the *APOE* gene (*APOE- $\epsilon 4$*) is the main AD genetic risk factor, both in terms of frequency and of effect size³. However, among early onset AD (EOAD) patients, some families exhibit autosomal dominant penetrance due to a pathogenic variant in the *APP*, *PSEN1* or *PSEN2* genes. In patients negatively screened for these genes, *SORL1* rare protein-truncating variants (PTV) or missense predicted damaging variants were initially identified in 7/29 unrelated EOAD patients with a positive family history of EOAD⁴. In line with *APP*, *PSEN1* and *PSEN2*, *SORL1* is involved in the production of A β peptides, the aggregation of which being a critical triggering event in AD pathophysiology. However, segregation data of *SORL1* variants remains very scarce, not allowing to assess their mode of inheritance. We showed that such variants are enriched in EOAD cases with a positive family history in a case-control study, with exome-wide significance⁵, and such results were subsequently extended to all AD cases with a clear effect on ages at onset^{6–8}. Given the extreme rarity of premature termination codon-introducing variants in controls and very high odds ratios⁶, the question of the penetrance of such variants and a putative use for genetic counseling has been raised. Missense variants with *in silico* predicted deleteriousness have also displayed a strong level of association with the risk of AD, with rarer variants being more strongly associated. However the level of association is of no common measure with what is observed for premature termination introducing variants. Therefore, although some missense variants probably act through a similar loss-of-function effect, evidence suggests that missense variants convey heterogeneous levels of risks. *In vitro* functional analyses are in line with such heterogeneity⁹.

Several approaches have been published to estimate the penetrance of genetic variants. While survival analyses relying on large prospective cohorts with sequencing data available may appear as the gold standard, they find themselves challenged, and thereby impractical, when assessing the penetrance associated with rare variants.

Indeed, they require a large amount of sequencing data and a follow-up period long enough to allow the diagnosis of a sufficient number of cases, which is not yet available. To overcome this limit, methods have been developed to combine data from large prospective cohorts with large case/control studies and evaluate with more accuracy the lifetime risks or penetrance associated with genetic risk variants^{3,10}. Another option is to resort to family-based study designs. Indeed, focusing on families where the proband carries a specific variant of interest enriches the dataset in rare variant carriers. However, the downside is the risk of bias resulting from this ascertainment scheme, that could even be compounded by a selection on age at onset. To overcome this issue, methods based on retrospective¹¹ or prospective^{12,13} likelihood were developed such as to condition the phenotype observation on the ascertainment process. A more “naive” but effective approach consists in computing the likelihood after exclusion of proband phenotypes^{14,15}. All these approaches may be combined with the Elston-Stewart algorithm¹⁶ in order to deal with the occurrence of missing genotypes, a second major limitation in pedigree datasets. However, these methods remain scarcely used, probably because of their lack of flexibility and their complexity. More recently, Alarcon et al.¹⁷ proposed a more flexible approach embedded in an expectation-maximization (EM) framework¹⁸ that alternates between penetrance estimation (M-step) and a belief propagation step to fill in missing genotypes (E-step). It was first used in a context of a monogenic disease, but may be applied to the case of complex diseases and used with parametric or non-parametric survival models combined with proband’s phenotype exclusion. In this paper, we further extend this method to (i) a digenic scenario combining rare and common risk factors and (ii) the integration of previously published data to robustly adjust for the common risk factor. We apply this strategy to assess the penetrance of AD associated with *SORL1* rare loss-of-function (LoF) variants in pedigrees where the AD-affected proband carries such a variant, with a previously published baseline model adjusted for *APOE*- $\epsilon 4$ ¹⁹.

Material and methods

Participants

Recruitment of probands

We first considered all unrelated individuals carrying a *SORL1* rare (allele frequency < 1%) non-synonymous variant identified by whole exome sequencing (WES), with a diagnosis of probable AD²⁰ among patients referred to the National Reference Center for Young Alzheimer Patients (CNRMAJ, Rouen, France) in the context of a nation-wide recruitment. Ages of onset were ≤ 65 years for previously reported patients^{4-6,21} and < 75 for novel patients recruited through the ECASCAD study, whatever their family history and *APOE* genotypes. Diagnoses were based on clinical examination by a neurologist or geriatrician from Memory Clinics and included personal medical and family history assessments, neurological examination, neuropsychological assessment, and neuroimaging. In addition, cerebrospinal fluid biomarkers indicative of AD (as described by Nicolas et al.²²) were available for 80% of the cases who were selected for WES. All variants were detected in the context of WES performed in a research or in a diagnostic setting in the National Center for Research in Human Genomics sequencing center (CNRGH, Evry, France) using Agilent Sureselect all exons human kits and Illumina sequencing technology, as described before²². All patients or legal guardians provided informed written consent for genetic analyses. Patients were recruited following institutional review boards agreements CPP Paris - Ile de France II (RBM-0259) and CPP Ouest 3 (ECASCAD study).

Genetic data processing and *SORL1* variants selection

Sequencing data were processed as previously described in Nicolas et al.²². After quality control, we excluded patients from non Caucasian ancestry as previously performed in our association studies⁵ as well as those carrying a likely pathogenic or a pathogenic variant in a Mendelian dementia gene including *APP*, *PSEN1* and *PSEN2*²². Among the 1295 remaining exomes of unrelated probands, we extracted those including a germline (allelic ratio [0.25-1], all were eventually heterozygous) rare (allele frequency < 1% in gnomAD non-Finish Europeans²³), non synonymous and splice region *SORL1* variants (NM003105.6). We considered, as PTV, all nonsense, frameshift and canonical splice site variants, in addition to splice region variants with a demonstrated effect on splicing (one described in Le Guennec et al.²⁴ and one unpublished patient blood mRNA analyses). We also considered missense variants that were predicted damaging by all three software tools among Polyphen-2²⁵, Mutation Taster²⁶, and SIFT²⁷, thereafter referred to as Mis3 variants, as candidates for analysis. To lessen the heterogeneity of the set of variants under analysis, we restricted our analysis to the selection of *SORL1* loss-of-function variants, namely (i) high confidence PTVs (i.e. not affecting the last coding exon or 50 bp of the penultimate exon) and (ii) candidate *SORL1* Mis3 variants that demonstrated an *in vitro* loss-of-function effect (Mis3-LoF). More precisely, Mis3-LoF were defined as rare Mis3 variants with at least two *in vitro* assays supporting such a loss-of-function effect and including at least one A β peptide

measurement. We retained the following three *Mis3* as non-ambiguously *Mis3*-LoF in this conservative approach: c.994C>T,p.(R332W); c.1960C>T,p.(R654W) (ref nous MedRxiv); and c.1531G>C,p.(G511R)²⁸. Hereafter and for penetrance estimations, we use the term “LoF” when considering PTV plus *Mis3*-LoF as a group.

Family history and genetic investigations in relatives

Overall, we included 27 probands carrying either a *SORL1* PTV or *Mis3*-LoF in this study and we contacted all 27 families, in order to extend pedigree information. Clinical status and age at onset for affected relatives or current age for unaffected relatives were obtained at least for siblings and parents of probands and, when possible for aunts, uncles and grandparents. Both parental sides were systematically investigated even when the presence of an affected parent suggested a unilateral transmission. When clinical examination was not possible, a phone interview was performed. Disease status was set as possible or probable AD according to the NINCDS-ADRDA, missing, or unaffected based on available clinical information from medical charts for affected relatives and from available clinical information or phone interviews for unaffected relatives. In addition, we collected blood samples from affected and unaffected relatives with informed written consent, and performed Sanger sequencing to search for the *SORL1* variant segregating in the family and to determine the *APOE* genotype. At the end, we obtained clinical information for 307 relatives with age ≥ 40 years and genotyping information for 45 relatives including 20 carriers of a *SORL1* LoF variant.

Statistical modeling

To evaluate the age-related penetrance of AD associated with *SORL1* LoF variants, we built a model based on pedigrees. Given the role of common *APOE* alleles in Alzheimer disease, it was essential to account for *APOE* genotypes in this model. However, since the effect of *APOE* alleles has already been deeply studied through large prospective cohort studies, we relied on published results to define the *APOE*-adjusted baseline of our survival model¹⁹. Given the rarity of *SORL1* variants, we make the assumption that this population-based baseline models Alzheimer disease risk for non carriers of *SORL1* LoF variants. Then our model combines this literature-based baseline for non carriers of *SORL1* LoF variants with an age dependent proportional effect for *SORL1* LoF variant carriers computed from our pedigrees (Figure 1). Since the pedigrees include relatives with missing genotypes, we applied an Expectation-Maximisation (EM) algorithm to take full advantage of the available information on the whole pedigree. This algorithm alternates until convergence between replacing unknown genotypes by individual weights (genotype posterior distributions for each individual) based on current age-related penetrances (E-step) and updating the age-related penetrances for each possible genotype based on observations and previously computed individual weights (M-step).

Notations

Let n be the total number of subjects. Let Y_i be the age at disease onset for subject i and C_i the censoring time such that $T_i = \min(Y_i, C_i)$ is the observed time associated with the status δ_i equaling 1 if $T_i = Y_i$ and 0 otherwise. In other words, δ_i indicates if the subject i has developed AD at age T_i .

We want to estimate the penetrance function $F(t | a, s)$, which gives the probability that an individual develops the disease before age t given their *APOE* status a and *SORL1* status s :

$$F(t | a, s) = \mathbb{P}(Y_i \leq t | APOE = a, SORL1 = s)$$

As usually done in survival analysis, the penetrance function is modeled through the survival function $S(t | a, s)$ which models the time to disease onset:

$$S(t | a, s) = 1 - F(t | a, s) = \exp(-\Lambda(t | a, s)) = \exp\left(-\int_0^t \lambda(u | a, s) du\right)$$

where $\lambda(t | a, s)$ denotes the hazard function corresponding to the instantaneous risk of developing AD at time t conditional on not having developed it before for individuals of *APOE* status a and *SORL1* status s ; and similarly, $\Lambda(t | a, s)$ denotes the cumulative risk until time t for these individuals.

This hazard function takes the following parametric form:

$$\lambda(t | a, s) = \begin{cases} \lambda_{nc}(t | a) & \text{for non carriers of } SORL1 \text{ LoF variants} \\ \lambda_c(t | a) = \lambda_{nc}(t | a) \exp(\beta(t)) & \text{for carriers of } SORL1 \text{ LoF variants} \end{cases}$$

Estimating $\lambda_{nc}(t, a)$

In this equation, the non carrier hazard rate $\lambda_{nc}(t | a)$ is a piecewise constant function designed to match exactly the *APOE*-specific penetrance estimates derived from age 65 to age 100 in the general population by van der Lee et al.¹⁹ (see supplementary Table 3 of van der Lee et al.¹⁹ and supplemental information for detailed computation). As in the original publication, cut offs are located every 5 years from 65 years-old up to 95 years-old. To account for the accumulation of risk before 65 years-old, the first age category spans from 40 to 65 years-old. To compensate for the rarity of *APOE*- $\epsilon 2$ allele, *APOE* genotypes are modeled through the number of *APOE*- $\epsilon 4$ alleles and therefore averaged into three categories: $\epsilon 4\epsilon 4$ carriers, $\epsilon 4$ heterozygous carriers and $\epsilon 4$ non-carriers.

Estimating $\beta(t)$

Regarding the hazard function for carriers, we also assume $\beta(t)$ is piecewise constant over time and most importantly that this additional age-dependent effect of *SORL1* LoF variants does not depend on *APOE*. Because genotype information is not available for most of the subjects in our cohort, we estimated $\beta(t)$ using an EM algorithm for the estimation procedure¹⁸. This iterative algorithm, alternating E- and M-steps to maximize the likelihood of the model, is efficient when the information is partially missing (here the *APOE* and *SORL1* genotypes). At the E-step, we computed individual weights $w_i(a, s)$ corresponding to the posterior probability that subject i should carry genotype $\{a, s\}$ according to current model estimates. At the M-step, we estimate the $\beta(t)$ coefficients using weights obtained at the E-step. We denote respectively by $w_i^m(a, s)$ and $\hat{\beta}(t)^m$ the values of $w_i(a, s)$ and $\hat{\beta}(t)$ at the m^{th} iteration of the EM algorithm.

E-step Denote by $\mathcal{G}$ the ensemble of the 36 possible *APOE* = $a \times$ *SORL1* = s genotype combinations.

This set accounts for every *APOE* alleles and differentiate paternal from maternal alleles for the purpose of genotype propagation through pedigrees.

The E-step consists in updating for each subject i and each genotype combination $\{a, s\} \in \mathcal{G}$ the posterior probability $w_i(a, s)$ that subject i should carry genotype combination $\{a, s\}$ conditional on **ev** the evidence defined as the probability of observing the data according to the genotype and $\hat{\beta}(t)^m$, the last estimation of $\beta(t)$.

$$w_i^m(a, s) = \mathbb{P}(\text{Genotype combination of } i \text{ is } a, s \mid \text{ev}_i^m(a, s))$$

These probabilities are computed using an *ad hoc* C++ implementation of belief propagation in pedigrees (**bped**) extended from Alarcon et al.¹⁷ to the specific case of *APOE* $\times$ *SORL1* genotypes. Arguments provided to **bped** are the pedigree structure and the $n \times \#\mathcal{G}$ matrix of evidence with:

$$\text{ev}_i^m(a, s) = \mathbb{P}(T_i \mid a, s; \hat{\beta}^m(t)) = \begin{cases} \hat{S}^m(T_i \mid a, s) & \text{if } \delta_i = 0 \\ \hat{S}^m(T_i \mid a, s) \hat{\lambda}^m(T_i \mid a, s) & \text{if } \delta_i = 1 \end{cases}$$

Of note, the evidence and thus posterior probabilities are set to 0 for all genotype combinations $\{a, s\}$ discordant with available information.

M-step We assume that $\beta(t)$ is piecewise constant over J intervals with $\forall j \in \{1, \dots, J\}$, $\beta(t) = \beta_j$ if $t \in [\tau_{j-1}, \tau_j]$, $\tau_0 = 0$ and $\tau_J = +\infty$. The M-step consists in estimating $\beta_{j,j \in \{1, \dots, J\}}$ that maximizes the likelihood of our data. Because of unobservable genotypes, we instead maximized $Q(\beta(t) \mid \hat{\beta}(t)^{m-1})$ the expected value of the log-likelihood function with respect to the current conditional distribution of unobserved genotype combinations $\{a, s\}$ given evidence and the current estimates of the parameters $\hat{\beta}(t)^{m-1}$:

$$\hat{\beta}(t)^m = \arg \max_{\beta(t)} Q(\beta(t) | \hat{\beta}(t)^{m-1}) = \arg \max_{\beta(t)} \sum_{i=1}^n \sum_{\{a,s\} \in \mathcal{G}} w_i^{m-1}(a, s) \left(-\Lambda(T_i | a, s) + \delta_i \log \lambda(T_i | a, s) \right)$$

Besides, note that because the additive *SORL1* effect only appears in the survival and hazard function of *SORL1* carriers, the sum on $\mathcal{G}$ simplifies into a sum on $\mathcal{G}_1$, the subset of genotype combinations that include a *SORL1* LoF variant. In the end the parameters $\hat{\beta}_{j,j \in \{1, \dots, J\}}^m$ that maximize the Q function are straightforwardly obtained by solving for each $j \in \{1, \dots, J\}$, $\frac{\partial Q(\beta(t) | \hat{\beta}(t)^{m-1})}{\partial \beta_j} = 0$ such that:

$$\forall j \in \{1, \dots, J\}, \quad \hat{\beta}_j^m = \log \left(\frac{\sum_{i=1}^n \sum_{T_i \in [\tau_{j-1}; \tau_j[} \sum_{\{a,s\} \in \mathcal{G}_1} w_i^{m-1}(a, s) \delta_i}{\sum_{i=1}^n \sum_{\substack{T_i \geq \tau_{j-1} \\ \{a,s\} \in \mathcal{G}_1}} w_i^{m-1}(a, s) [\Lambda_{nc}^{m-1}(\min(T_i, \tau_j) | a, s) - \Lambda_{nc}^{m-1}(\tau_{j-1} | a, s)]} \right)$$

In other words, the hazard ratio $\exp(\hat{\beta}_j)$ associated with *SORL1* LoF variants on the j^{th} interval may be interpreted as the ratio of the subjects developing the disease within the time-interval j over the time-subjects at risk during this time interval taking into account the *APOE* effect.

Correction for ascertainment bias

To tackle the important issue of ascertainment bias resulting from the recruitment of early onset carriers of *SORL1* LoF variants as probands, we took their genotype information into account at each E-step but discarded proband phenotype information (observed time, disease status) at each M-step as in Alarcon et al.¹⁷.

Moreover, since only few individuals' observation time were greater than 85 years of age, we censored data at 85 years-old.

Model choice

To validate the choice of $\{\tau_0, \dots, \tau_{J-1}\}$ for the piecewise modeling of the effect of *SORL1* LoF variants $\beta(t)$, several cut-offs were envisaged and the different models were compared based on a Bayesian Information Criterion (BIC) measure defined as:

$$\text{BIC} = -2 \times \log(\mathcal{L}) + k \times \log(d)$$

where $\log(\mathcal{L})$ is the log-likelihood of the model returned by **bped** computed using the sum-product algorithm from the evidence probability over all possible genotypes, k is the number of parameters and d the number of subjects who experienced the event of disease onset. The model with the best fit is the one with the lowest BIC.

Computation of confidence intervals

Confidence intervals (CI) were computed using a bootstrap strategy of 500 iterations. At each iteration, families were randomly selected with replacement and model parameters were re-estimated on these observations (same family number) to obtain the associated age-related penetrance curves. Then, for each age, we took the 2.5th and 97.5th quantiles as the boundaries of the 95%CI.

As explained more thoroughly in the supplemental information, uncertainty surrounding *APOE*-adjusted baseline survival curves as reflected by their confidence intervals was taken into account in the bootstrap procedure.

Simulation study

We performed a simulation study in order to assess the robustness of our methodology. We proposed several *scenarii* mimicking all the potential biases of our pedigree dataset: age of probands, missing genotype patterns, unbalanced maternal/paternal ratio information in pedigrees as well as unobserved family history effect. We varied simulation parameters and estimated their impact on bias and variance associated with estimates. For each *scenario*, we compared the methodology with and without discarding proband phenotype information. A detailed description of the simulation study is available in supplemental information.

Results

In order to assess penetrance of rare *SORL1* LoF variants in pedigrees with missing genotypes while adjusting for the *APOE* status, we built a piecewise constant hazard model integrating a literature-based baseline for *APOE*- $\epsilon 4$ effect. The additive effect of *SORL1* variants was estimated from pedigrees through an EM algorithm alternating replacement of unknown genotypes by individual weights (genotype posterior distributions for each individual) based on current age-related penetrance (E-step) and update of the age-related penetrance for each possible genotype based on observations and previously computed individual genotype posterior probabilities (M-step) (Figure 1).

Simulation study

We first aimed at challenging the performances of our algorithm through simulations. We generated artificial datasets under several *SORL1* variant effects as well as several *scenarii* mimicking the potential biases of a pedigree dataset. We varied simulation parameters and investigated their impact on estimation bias and variance (more details provided in Supplemental Information). Our simulation study confirmed that our methodology was robust to various patterns of genotype and phenotype missingness, in particular even in the presence of unbalanced pedigree ascertainment (Figure 2 and Figures S3-S9). Exclusion of the proband phenotype at the M step (hazard ratio computation) reduced efficiently the bias resulting from age-at-onset-based ascertainment. The Bayesian Information Criterion (BIC) was able to discriminate models and identify the best time cut-offs for the piecewise constant function $\beta(t)$ (Figure S10). Based on our simulations study, we concluded that the model might lead to an overestimation of the average variant effect when variant effects are not homogeneous across families.

Description of the pedigrees

27 families carried at least one non-ambiguous LoF variant (21 families with one PTV, one family with 2 PTVs, and 5 families with a Mis3 variant with *in vitro* LoF evidence, Mis3-LoF). Although most of the LoF variants were each private to one family, the p.W804* (c.2412G>A; PTV) and the p.R654W (c.1960C>T; Mis3 with *in vitro* LoF evidence) variants were observed in two and three apparently unrelated families, respectively, even though they were both not observed in the gnomAD database. Among the LoF variants, only 3 were observed at least once in gnomAD (cumulate non neuro AF nfe): p.R332W (c.994C>T, minor allele frequency = 1.1×10^{-5}), p.R866* (c.2596C>T; minor allele frequency = 1.9×10^{-5}) and p.R1207* (c.3619C>T; minor allele frequency = 2.2×10^{-5}). See Table S1 for a detailed list of variants. In LoF families, proband age of onset varies from 48 to 70 years (mean: 58 years). Two probands (7%) were *APOE*- $\epsilon 4\epsilon 4$ carriers (age of onset: 58, 60), 18 (67%) were $\epsilon 4$ heterozygous carriers (age of onset from 50 to 66, mean: 58) and 7 (26%) were $\epsilon 4$ non-carriers (age of onset from 48 to 70, mean: 56). In addition, informative phenotype (disease status and age

≥ 40 years) and genotypes (*APOE* and *SORL1*) were obtained for 307 and 45 relatives, respectively (Table 1). They were collected in both paternal and maternal branches, whatever the apparent disease transmissions (Figure S1). LoF families included a median of 11 informative phenotypes (Q1-Q3: 8.5-17; min-max: 4-24) over 3 generations (Q1-Q3: 3-3; min-max: 2-5) and 2 genotypes (Q1-Q3: 1-4; min-max: 1-6) encompassing data of the proband. Relatives' genotypes were known (genotyped individual or obligate carrier) for 12 affected carriers (age at onset from 55 to 78, median: 68, Q1-Q3: 64.5-74), 12 unaffected carriers (age from 42 to 95, median: 66, Q1-Q3: 52-68.25), 4 affected non carriers (age at onset: 64, 68, 70 and 75) and 21 unaffected non carriers (age from 37 to 86, median: 68, Q1-Q3: 57-70).

Genotype probabilities and penetrance estimates

During the EM procedure, genotyping probabilities were attributed to each participant. Final weights added to already known genotype information led to an expected total number of relatives aged ≥ 40 years carrying a *SORL1* LoF variant of 109.6 (36% of relatives aged ≥ 40 years; Figure 3). Table S3 provides detailed estimated genotypes. Based on BIC (Table S2), we retained a piecewise constant $\beta(t)$ with cut-offs at respectively 60, 65 and 70 years old ($\beta(t) = 3.5$, 95%CI[1.5; 5.0] for $t < 60y$; $\beta(t) = 6.7$, 95%CI[5.1; 8.0] for $60 \leq t < 65$; $\beta(t) = 4.7$, 95%CI[3.9; 5.4] for $65 \leq t < 70$ and $\beta(t) = 3.4$, 95%CI[2.1; 4.7] for $70 \leq t$) leading to an increased age-dependent penetrance for carriers versus non carriers of *SORL1* LoF variants. Following penetrance estimates after adjustment on *APOE*- $\epsilon 4$ status, we observed a full penetrance at the age of 70 years for $\epsilon 4/4$ carriers only, whereas $\epsilon 4$ heterozygous carriers and $\epsilon 4$ non carriers reached respectively 56% (95%CI [40%; 72%]) and 37% (95%CI [26%; 51%]) penetrance (Figure 4). Of note, full penetrance was reached ten years later for $\epsilon 4$ heterozygous carriers (95%CI = [0.91; 1] by age 80) and after 85 years old for $\epsilon 4$ non carriers (95%CI = [0.81; 1] by age 85).

Discussion

We investigated the age-related penetrance of AD associated with *SORL1* LoF variants, using a family-based approach and adjusting for the common risk factor *APOE*- ϵ 4. We proposed to model the risk conferred by both factors through a piecewise constant hazard model. As expected, our model indicates that *SORL1* LoF variants confer a high risk before 65 years, and thus are strongly associated with EOAD^{6,8}. The risk remained high until 70 years, which is compatible with an association of such a variant with late onset AD (LOAD) among younger patients. However, our results suggested that the penetrance associated with *SORL1* LoF variants should be interpreted in light of the carrier's *APOE* genotype. Indeed, in our model, the penetrance for *SORL1* LoF carriers was complete by age 70 only among ϵ 4 ϵ 4 carriers, whereas the curve for *SORL1* LoF and ϵ 4 heterozygous carriers reached complete penetrance ten years later and even later for ϵ 4 non-carriers.

APOE- ϵ 4 has already been investigated as a modifier of AD age at onset regarding the specific *PSEN1* p.E280A pathogenic variant and in APP V717I families^{29,30}. However, these variants retained full penetrance before the age of 65 regardless to the *APOE* genotype. Because most pathogenic variants in the *APP*, *PSEN1* or *PSEN2* genes confer full penetrance before 65 years by themselves, if not earlier, the *APOE*- ϵ 4 effect is difficult to observe for most of these variants and, thus, *APOE* genotypes are not a clinically meaningful information upon genetic counseling in such families. In contrast, *SORL1* LoF variants are less penetrant and thus *APOE*- ϵ 4 substantially modifies ages at onset for carriers of such a variant, suggesting that presymptomatic testing of *SORL1* LoF variants should not be proposed regardless of *APOE* genotyping even though such digenic presymptomatic testing is not usual in clinical genetics consultations and may be difficult to understand.

Here, we used stringent criteria for restricting our analysis to a set of variants with homogeneous effect (PTV + Mis3-LoF), in accordance with our simulation study. Indeed, since Mis3 variants were defined solely through bioinformatics prediction tools, we can expect that there is some diversity of the variants' effects on SorLA function towards A β secretion^{9,31}. Some of the predicted deleterious variants may remain neutral and some other variants might be associated with an intermediate level of loss of SorLA function regarding A β secretion. In our study, we generated familial data based on our nation-wide recruitment of families that is probably among the largest datasets on *SORL1* LoF families. Despite this huge effort based on the sequencing of a large number of probands with a high level of evidence for an AD diagnosis, first, and individual family extension, second, the number of families remained limited, especially after restriction to the families with a non-ambiguous LoF variant. However, the EM approach allowed us to reach sufficient power to provide meaningful estimates. In this study, we relied on the EM framework for penetrance calculation in incomplete pedigrees, first used by Alarcon et al.¹⁷ and extended to our digenic problematic (E-step extension) and parametric modeling (M-step changes). We finally incorporated it into a 2-step methodology also including a baseline estimation from published data. The major interest for this methodology is the combination of several modules that all may be independently modified and improved for further family-based analyses and should open the door to the analyses of (often incomplete) pedigrees, whatever the disease. Moreover, our model might be further developed to better take

into account the following points and thus help refining estimations and apply to other diseases: controlling for ascertainment biases in other ways than removing probands phenotypes information, applying other strategies for cut-off piecewise constant functions, adjusting for family clusters or other genetic factors and taking death as a competing risk.

The control for ascertainment bias. As probands were selected based on age at onset, we proposed to exclude their phenotypes from model parameters computation (M-step) in order to correct for resulting putative ascertainment bias. Our simulation study showed that this method standardly used in family analyses efficiently helps to correct for this bias, but might lead to an underestimation of parameters at the beginning of the curve. Another approach, based on raking estimators, and more often used in survey analyses³², could be developed to assess if such a strategy could better handle this ascertainment bias by incorporating auxiliary information in the procedure and thus adjust for family recruitment probabilities.

The choice of optimal cut-offs in piecewise constant function. The effect associated with the *SORL1* LoF variants was modeled using a simple but interpretable piecewise constant function. Because we did not know how the instantaneous risk associated with variants changes over time, we compared several models including different cut-offs and selected the best model based on BIC. A possible extension of this method is the repetition of this procedure through bootstrap iterations and a higher quantity of candidate cut-offs resulting in smoother curves³³. However, non parametric and semi-parametric models like for example the Cox's model³⁴ may be also good candidate models conditionally to their ease of interpretation.

Adjustment for family cluster. For the sake of simplicity, we considered that individuals were independent conditionally to their *APOE* $\times$ *SORL1* genotypes. However, as shown by our simulation study, inter-family heterogeneity may lead to misestimation of model parameters. Stratified models and then frailty survival models were developed to handle this kind of data³⁵. The latter proposes to incorporate random term(s) in the model mimicking an intra-familial correlation and thus reduce the part of variability associated with *SORL1* variants effect.

The death as competing risk. Here, we have chosen to consider death as censoring, meaning that death and disease onset are independent. Indeed, there is no demonstrated increased risk of death linked to *SORL1* LoF variants. In contrast, *APOE*- ϵ 4 is related to a higher risk of mortality independently of AD³⁶. Since this was already taken into account in published data we used for baseline risk according to *APOE*, we did not need to adjust a second time for it¹⁹. However, in future applications of this methodology to the case of other diseases, the M-step may easily incorporate models developed in a competing risk framework³⁷.

In our model, we adjusted on the number of *APOE*- ϵ 4 alleles given the frequency of this allele and its moderate to high impact. In addition to *APOE* and *SORL1*, other types of variants can influence AD risk. Common variants identified in GWAS and rare variants identified in exome sequencing studies, mainly in the *ABCA7* and *TREM2* genes, can thus modulate the individual risk as computed using *SORL1* and *APOE* only. Recently, polygenic risk scores (PRS) have been developed, gathering the individual small effect of multiple GWAS SNPs.

Although there is not a unique method to build PRS, including the selection of variants that can range from a restricted number of SNPs associated with AD after Bonferroni correction to a large number SNPs with an association p-value below 0.5, the effect of PRS as modulators of the *APOE* genotype has been assessed¹⁹. Adjusting the effect of rare *SORL1* variants on that PRS-adjusted *APOE*- ϵ 4 is thus an interesting perspective, conditional to the feasible extension of the E-step to such a multigenic model, i.e. allowing for the imputation of multiples SNPs for PRS estimation in relatives without drastically increasing computational cost. However, given the large risk load already conferred by *APOE* and *SORL1*, we do not expect that PRS may play a clinically-relevant role in *SORL1*+*APOE*- ϵ 4 carriers. This might however be different in case of co-occurrence of an *ABCA7* or a *TREM2* risk variant. Indeed, odds ratios of rare deleterious variants of either gene are in the order of magnitude of 2.5-4, and even higher for the extremely rare *TREM2* PTVs²¹, and lower for the R62H *TREM2* variant³⁸. Thus, an oligogenic model taking into account these genes would be welcome. However, the extreme rarity of carriers of unambiguously deleterious/associated *TREM2*, *ABCA7* and *SORL1* variants make such an oligogenic model challenging to assess with sufficient power.

In conclusion, we propose here digenic penetrance estimations for *SORL1* LoF variants together with *APOE*- ϵ 4 alleles. We used a family-based approach that demonstrated to be powerful to provide meaningful estimations. Our recruitment was likely closer to that of genetic counseling requests in families than to the general population, and our estimates may thus be taken into account when an asymptomatic relative may request information following the identification of a *SORL1* LoF variant found in a proband. We consider that *SORL1* presymptomatic testing should not be performed regardless of *APOE* genotyping, given the large modifying effect in these families. Our estimations may be also useful for elaborating preventive clinical trials based on similar recruitment criteria. Further work remains necessary to estimate the penetrance of other *SORL1* missense variants as well as rare variants in other risk genes, such as *TREM2* and *ABCA7*. Finally, an integrative model is warranted in exceptional cases of co-occurrence of rare deleterious variants in multiple genes, together with the *APOE* genotype, that may additionally be PRS-adjusted.

Supplemental Data

Supplemental Data include supplemental methods, detailed results from the simulation study, eleven figures and three tables.

Declaration of Interests

The authors declare no competing interests.

Acknowledgments

This work was supported by grants from Fondation pour la Recherche Médicale (Equipe FRM DEQ20170336711), Fondation Alzheimer (ECASCAD study), France Alzheimer association (AAPSM2019 – grant #1957) and CNRMAJ. CS is supported by the Fondation pour la Recherche Médicale (ARF201909009263). This work is a collaboration between CEA-DRF-Jacob-CNRGH-CHU de Rouen.

CNRMAJ collaborators, by city:

ALES: Julia NIVELLE; AMIENS: Daniela ANDRIUTA, Mélanie BARBAY, Olivier GODEFROY, Alexandre PERRON; ANGERS: Valérie CHAUVIRE, Frédérique ETCHARRY-BOUYX, Virginie PICHON; ARLES: Marie DE VERDAL; BAYONNE: Guillaume BALLAN; BEAUNE: Fabienne CONTEGAL-CALLIER; BESANÇON: Sophie HAFFEN, Eloi MAGNIN, Alice VOILLY; BORDEAUX: Sophie AURIACOMBE, Chloé GREGOIRE, Brice LAURENS, Vincent PLANCHE; BREST: Amélie LEBLANC; CAEN: Pierre BRANGER, Julien COGEZ, Sophie DAUTRICOURT, Olivier MARTINAUD, Vincent DE LA SAYETTE; CLERMONT-FERRAND: Didier DEFFOND, Elsa DIONET; COLMAR: Pierre ANTHONY, Geoffroy HAUTECLOQUE, François SELLAL; DIJON: Yannick BEJOT, Benoit DELPONT, Mathilde GRABER, Julien GUENIAT, Catia KHOUMRI, Sophie MOHR, Christel THAUVIN; GRENOBLE: Olivier MOREAUD, Mathilde SAUVÉE; LILLE: Pascaline CASSAGNAUD, Yaohua CHEN, Charlotte CRINQUETTE, Vincent DERAMECOURT, Céline DEROLLEZ, Thibaud LEBOUVIER, Marie-Anne MACKOWIAK, Aurélien MAUREILLE, Florence PASQUIER, Adeline ROLLIN-SILLAIRES; LIMOGES: Leslie CARTZ-PIVER, Philippe COURATIER, Benjamin DAURIAT; LYON: Bernard CROISILLE, Maïté FORMAGLIO, Hélène MOLLION, Elisabeth OLLAGNON ROMAN; MARSEILLE: Mathieu CECCALDI, Lea CORNEILLE, Mira DIDIC, Boris DUFOURNET, Stephan GRIMALDI, Claude GUEROT, Lejla KORIC; MONTPELLIER: Karim BENNYS, Giulia DIEMERT, Audrey GABELLE, Pierre LABAUGE, Anaïs LIPPI; Cécilia MARELLI, Cédric TURPINAT; MULHOUSE: Emmanuelle GINGLINGER; NANCY: Thérèse JONVEAUX; NANTES: Claire BOUTOLEAU BRETONNIÈRE, Hélène COURTEMANCHE, Nathalie WAGEMANN; NICE: Annabelle CHAUSSENOT; NIMES: Giovanni CASTELNOVO, Camille HEITZ, Lila SIRVEN VILLAROS; ORLÉANS: Hélène-Marie LANOISELÉE, Christophe TOMASINO; PARIS: Stéphanie BOMBOIS, Emmanuel COGNAT, Astrid DE LIÈGE, Julien DUMURGIER, Lorraine HAMELIN, Claire HOURREGUE, Julien LAGARDE, Isabelle LE BER, Camille

NOIRAY, Claire PAQUET, Pauline ROD-OLIVIERI, Carole ROUÉ-JAGOT, Dario SARACINO, Marie SARAZIN; PER-
 PIGNAN: Anaïs DUTRAY, Snejana JURICI, Laurène VAN DAMME; POITIERS: Gwenaël LE GUYADER; QUIMPER:
 Philippe DIRAISON; REIMS: Martine DOCO FENZY, Céline POIRSIER; RENNES: Serge BELLIARD, Florence
 DEMURGER, Anne GAINCHE-SALMON, Cezara HANTA; ROUEN: Aude DOAN, Didier HANNEQUIN, Clémence
 HARDY, Morgane LACOUR, Alexandre MORIN, Gaël NICOLAS, David WALLON, Aline ZAREA; SAINT BRIEUC:
 Olivier VERCRUYSSSE; SAINT ETIENNE: Jean-Claude GETENET; STRASBOURG: Frédéric BLANC, Benjamin
 CRETIN, Hélène DURAND, Nathalie PHILIPPI; TOULOUSE: Marie BENAITEAU, Jasmine CARLIER, Emilie MI-
 LONGO RIGAL, Jérémie PARIENTE, Marie RAFIQ, Camille TISSERAND; TOURS: Anna-Chloé BALAGEAS, Emilie
 BEAUFILS; VILLEURBANNE: Christine CHAMPION, Pierre KROLAK-SALMON.

References

1. Ezewudo, M. and Zwick, M. E. (2013). Evaluating rare variants in complex disorders using next-generation sequencing. *Current psychiatry reports* *15*, 349.
2. Nicolas, G., Charbonnier, C., and Campion, D. (2016). From common to rare variants: the genetic component of Alzheimer disease. *Human heredity* *81*, 129–141.
3. Genin, E., Hannequin, D., Wallon, D., Sleegers, K., Hiltunen, M., Combarros, O., Bullido, M. J., Engelborghs, S., De Deyn, P., Berr, C., et al. (2011). APOE and Alzheimer disease: a major gene with semi-dominant inheritance. *Molecular psychiatry* *16*, 903–907.
4. Pottier, C., Hannequin, D., Coutant, S., Rovelet-Lecrux, A., Wallon, D., Rousseau, S., Legallic, S., Paquet, C., Bombois, S., Pariente, J., et al. (2012). High frequency of potentially pathogenic SORL1 mutations in autosomal dominant early-onset Alzheimer disease. *Molecular psychiatry* *17*, 875–879.
5. Nicolas, G., Charbonnier, C., Wallon, D., Quenez, O., Bellenguez, C., Grenier-Boley, B., Rousseau, S., Richard, A.-C., Rovelet-Lecrux, A., Le Guennec, K., et al. (2016). SORL1 rare variants: a major risk factor for familial early-onset Alzheimer’s disease. *Molecular psychiatry* *21*, 831–836.
6. Campion, D., Charbonnier, C., and Nicolas, G. (2019). SORL1 genetic variants and Alzheimer disease risk: a literature review and meta-analysis of sequencing data. *Acta neuropathologica* *138*, 173–186.
7. Holstege, H., Hulsman, M., Charbonnier, C., Grenier-Boley, B., Quenez, O., Grozeva, D., van Rooij, J. G., Sims, R., Ahmad, S., Amin, N., et al. (2020). Exome sequencing identifies novel AD-associated genes. medRxiv.
8. Raghavan, N. S., Brickman, A. M., Andrews, H., Manly, J. J., Schupf, N., Lantigua, R., Wolock, C. J., Kamalakaran, S., Petrovski, S., Tosto, G., et al. (2018). Whole-exome sequencing in 20,197 persons for rare variants in Alzheimer’s disease. *Annals of clinical and translational neurology* *5*, 832–842.
9. Rovelet-Lecrux, A., Feuillette, S., Miguel, L., Schramm, C., Pernet, S., Segalas-Milazzo, I., Guilhaudis, L., Rousseau, S., Riou, G., Frebourg, T., et al. (2021). Impaired SORLA maturation and trafficking as a new mechanism for SORL1 missense variants in Alzheimer disease. bioRxiv.
10. Satagopan, J. M., Offit, K., Foulkes, W., Robson, M. E., Wacholder, S., Eng, C. M., Karp, S. E., and Begg, C. B. (2001). The lifetime risks of breast cancer in Ashkenazi Jewish carriers of BRCA1 and BRCA2 mutations. *Cancer Epidemiology and Prevention Biomarkers* *10*, 467–473.
11. Carayol, J. and Bonaïti-Pellié, C. (2004). Estimating penetrance from family data using a retrospective likelihood when ascertainment depends on genotype and age of onset. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* *27*, 109–117.

12. Le Bihan, C., Moutou, C., Brugières, L., Feunteun, J., and Bonaïti-Pellié, C. (1995). ARCAD: A method for estimating age-dependent disease risk associated with mutation carrier status from family data. *Genetic epidemiology* *12*, 13–25.
13. Plante-Bordeneuve, V., Carayol, J., Ferreira, A., Adams, D., Clerget-Darpoux, F., Misrahi, M., Said, G., and Bonaiti-Pellie, C. (2003). Genetic study of transthyretin amyloid neuropathies: carrier risks among french and portuguese families. *Journal of medical genetics* *40*, e120–e120.
14. Weinberg, W. (1912). Methode und fehlerquellen der untersuchung auf mendelsche zahlen beim menschen. *Arch Rass Ges Biol* *9*, 165–174.
15. Alarcon, F., Bourgain, C., Gauthier-Villars, M., Planté-Bordeneuve, V., Stoppa-Lyonnet, D., and Bonaïti-Pellié, C. (2009). PEL: an unbiased method for estimating age-dependent genetic disease risk from pedigree data unselected for family history. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* *33*, 379–385.
16. Elston, R. C. and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human heredity* *21*, 523–542.
17. Alarcon, F., Planté-Bordeneuve, V., Olsson, M., and Nuel, G. (2018). Non-parametric estimation of survival in age-dependent genetic disease and application to the transthyretin-related hereditary amyloidosis. *PloS one* *13*, e0203860.
18. Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* *39*, 1–22.
19. van der Lee, S. J., Wolters, F. J., Ikram, M. K., Hofman, A., Ikram, M. A., Amin, N., and van Duijn, C. M. (2018). The effect of APOE and other common genetic variants on the onset of Alzheimer’s disease and dementia: a community-based cohort study. *The Lancet Neurology* *17*, 434–444.
20. Jack Jr, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., Holtzman, D. M., Jagust, W., Jessen, F., Karlawish, J., et al. (2018). NIA-AA research framework: toward a biological definition of Alzheimer’s disease. *Alzheimer’s & Dementia* *14*, 535–562.
21. Bellenguez, C., Charbonnier, C., Grenier-Boley, B., Quenez, O., Le Guennec, K., Nicolas, G., Chauhan, G., Wallon, D., Rousseau, S., Richard, A. C., et al. (2017). Contribution to Alzheimer’s disease risk of rare variants in TREM2, SORL1, and ABCA7 in 1779 cases and 1273 controls. *Neurobiology of aging* *59*, 220–e1.
22. Nicolas, G., Wallon, D., Charbonnier, C., Quenez, O., Rousseau, S., Richard, A.-C., Rovelet-Lecrux, A., Coutant, S., Le Guennec, K., Bacq, D., et al. (2016). Screening of dementia genes by whole-exome

- sequencing in early-onset Alzheimer disease: input and lessons. *European Journal of Human Genetics* *24*, 710–716.
23. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* *581*, 434–443.
24. Le Guennec, K., Tubeuf, H., Hannequin, D., Wallon, D., Quenez, O., Rousseau, S., Richard, A.-C., Deleuze, J.-F., Boland, A., Frebourg, T., et al. (2018). Biallelic loss of function of SORL1 in an early onset Alzheimer’s disease patient. *Journal of Alzheimer’s Disease* *62*, 821–831.
25. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature methods* *7*, 248–249.
26. Schwarz, J. M., Cooper, D. N., Schuelke, M., and Seelow, D. (2014). Mutationtaster2: mutation prediction for the deep-sequencing age. *Nature methods* *11*, 361–362.
27. Ng, P. C. and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome research* *11*, 863–874.
28. Caglayan, S., Takagi-Niidome, S., Liao, F., Carlo, A.-S., Schmidt, V., Burgert, T., Kitago, Y., Füchtbauer, E.-M., Füchtbauer, A., Holtzman, D. M., et al. (2014). Lysosomal sorting of amyloid- β by the SORLA receptor is impaired by a familial Alzheimer’s disease mutation. *Science translational medicine* *6*, 223ra20–223ra20.
29. Pastor, P., Roe, C. M., Villegas, A., Bedoya, G., Chakraverty, S., García, G., Tirado, V., Norton, J., Ríos, S., Martínez, M., et al. (2003). Apolipoprotein ϵ 4 modifies Alzheimer’s disease onset in an E280A PS1 kindred. *Annals of neurology* *54*, 163–169.
30. Zhang, G., Xie, Y., Wang, W., Feng, X., and Jia, J. (2017). Clinical characterization of an APP mutation (V717I) in five Han Chinese families with early-onset Alzheimer’s disease. *Journal of the neurological sciences* *372*, 379–386.
31. Rovelet-Lecrux, A., Charbonnier, C., Wallon, D., Nicolas, G., Seaman, M., Pottier, C., Breusegem, S., Mathur, P., Jenardhanan, P., Le Guennec, K., et al. (2015). De novo deleterious genetic variations target a biological network centered on A β peptide in early-onset Alzheimer disease. *Molecular psychiatry* *20*, 1046–1056.
32. Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American statistical Association* *88*, 1013–1020.

33. Bouaziz, O. and Nuel, G. (2016). L0 regularisation for the estimation of piecewise constant hazard rates in survival analysis. arXiv preprint arXiv:1609.04595.
34. Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* *34*, 187–202.
35. Govindarajulu, U. S. and D’Agostino Sr, R. B. (2020). Review of current advances in survival analysis and frailty models. *Wiley Interdisciplinary Reviews: Computational Statistics* *12*, e1504.
36. Ewbank, D. C. (2002). Mortality differences by APOE genotype estimated from demographic synthesis. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* *22*, 146–155.
37. Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association* *94*, 496–509.
38. Sims, R., Van Der Lee, S. J., Naj, A. C., Bellenguez, C., Badarinarayan, N., Jakobsdottir, J., Kunkle, B. W., Boland, A., Raybould, R., Bis, J. C., et al. (2017). Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer’s disease. *Nature genetics* *49*, 1373–1384.

Figures

Figure 1: Summary of the method.

$\lambda(t | a, s)$ refers to the instantaneous risk to develop the disease depending on age t and genotype $(a, s) \in APOE \times SORL1$. $\lambda_{nc}(t | a)$ refers to the specific instantaneous risk associated with non carriers of *SORL1* variant of interest stratified on *APOE* genotype and derived from the Rotterdam study¹⁹. $\beta(t)$ refers to the additional effect of the *SORL1* variant. E/M-steps refer to Expectation/Maximization steps. $w_i(a, s)$: individual weight updating at each E-step iteration referring to posterior probability distribution of individual i for combined genotype (a, s) ; y: years; *SORL1* +: carrier of the variant of interest in *SORL1* gene; *SORL1* WT: wild type for *SORL1* (non carriers of the variant of interest); ?: unknown genotype. The red arrow indicates the proband.

Figure 2: Estimation of bias in our simulation study

Bias was estimated for each of the four constant parameters (in columns) of the piecewise constant $\beta(t)$ referring to the additional effect of the *SORL1* variant of interest through 4 *scenarii* of simulation (in rows). Results are provided for probands included and excluded from the analysis during the Maximization-step of the EM algorithm. For the baseline *scenario*, we generated 27 families mimicking what we observed in our dataset in terms of *SORL1* LoF variant effect, ascertainment and available genotypes. Then the model was challenged through three additional *scenarii*: (i) Unbalanced phenotype information: if one of the parents was affected, we removed all the phenotypic information of the parental branch with the unaffected parent. (ii) Relatives' genotypes 100% missing: we removed all information about relatives' genotypes. (iii) Heterogeneous variant effect: instead of generating age at onset based on a constant variant effect, we generated age at onset based on a normal distribution of the variant effect with a variance equaling to 1.

Figure 3 : Number of affected and unaffected individuals according to their age and their *SORL1* status after genotyping imputation.

This graph was obtained for all individuals with age of onset or censoring above 40 years. It represents the number of carriers (upper part) and non carriers (lower part) of *SORL1* LoF variant according to their disease status and age intervals (age of onset for probands and affected relatives and censoring for unaffected relatives). Transparency differentiates available genotypes (already known, including those of probands) from those estimated at the end of the algorithm.

Figure 4: Age dependent penetrance for carriers and non carriers of a *SORL1* LoF variant

The penetrance is displayed with its 95% confidence interval according to the number of *APOE*-ε4 allele from 65 to 85 years of age. Curves for non carriers and their confidence intervals were obtained from our estimation of $\lambda_{nc}(t | a)$ based on the Rotterdam study¹⁹. Data from pedigrees were censored at 85 years. Confidence intervals for *SORL1* LoF variant carriers were obtained from 2.5th and 97.5th quantiles of 500 bootstrap iterations. Penetrance values at 65, 70, 75, 80 and 85 years of age, for carriers of *SORL1* LoF variant are displayed below the figure.

Tables

	Probands	Affected relatives	Unaffected relatives
Informative phenotypes, N	27	61	246
Males/Females, N (%) ^a	11 (41%)/16 (59%)	16 (26%)/45 (74%)	121 (50%)/120 (50%)
Age, y, median (Q1-Q3)	58 (52.5-61)	67 (63-74)	68 (55.25-78)
Available genotypes, N	27	13	32
<i>SORL1</i> carriers, N (%)	27 (100%)	9 (69%)	11 (34%)
<i>APOE</i> × <i>SORL1</i> (+ vs WT)	+	+ WT	+ WT
ε4 non-carriers, N (%)	7 (26%)	6 (46%) 3 (23%)	7 (22%) 14 (44%)
ε4 heterozygous carriers, N (%)	18 (67%)	3 (23%) 1 (8%)	2 (6%) 6 (19%)
ε4ε4 carriers, N (%)	2 (7%)	0 (0%) 0 (0%)	2 (6%) 1 (3%)

Table 1: Description of informative phenotypes and genotypes available in the 27 families with a LoF variant.

N: number; y: years; Q1: first quartile; Q3: third quartile; WT: wild type. We considered as informative individuals for the phenotype, all those having well-established disease status as well as age at onset or censoring above 40 years. Genotypes were available for 71 (21%) of individuals with informative phenotype as well as for one unaffected relative (censoring age: 37 years, genotype: *APOE*-ε4 heterozygous, *SORL1* WT).

a) percentages are given over known gender.