Single Nucleotide Polymorphisms Associated with Fasting Blood Glucose Trajectory and Type 2 Diabetes Incidence: A Joint Modelling Approach *(max 48 characters)*

Mickaël Canouil1,2,3 Philippe Froguel1,2,3,4 and Ghislain Rocheleau1,2,3

1University of Lille, UMR 8199 - EGID, F-59000 Lille, France.

2CNRS, UMR 8199, F-59000 Lille, France.

3Institut Pasteur de Lille, F-59000 Lille, France.

4Department of Genomics of Common Disease, Imperial College London, London, United Kingdom.

###### Corresponding authors

Mickaël Canouil, EGID - UMR 8199, Pôle Recherche - 1er étage Aile Ouest, 1 place de Verdun, 59045 Lille CEDEX, France. E-mail: mickael.canouil@cnrs.fr

Ghislain Rocheleau, Maelstrom Research -The Research Institute of the McGill University Health Centre (RI MUHC), 2155 Guy, 4th Floor, Office 458, Montreal, Quebec, H3H 2R9, Canada. E-mail: grocheleau@maelstrom-research.org

# Abstract *(max 250 words)*

In observational cohorts, longitudinal data are collected with repeated measurements at predetermined time points for many biomarkers, along with other covariates measured at baseline. In these cohorts, time until a certain event of interest occurs is commonly reported and very often, a relationship will be observed between a biomarker repeatedly measured over time and that event. Joint models were designed to efficiently estimate statistical parameters by combining a mixed model for the longitudinal biomarker trajectory and a survival model for the event risk, using a set of random effects to account for the link between the two types of data. First, using genotypes assayed with the Metabochip DNA arrays (Illumina) from close to 4,500 subjects recruited in the French cohort D.E.S.I.R. (Données Épidémiologiques sur le Syndrome d’Insulino-Résistance), we assessed the feasibility of implementing the joint modelling approach in a real high-throughput genomic dataset. Second, we checked model consistency based on different simulation scenarios, varying sample size, minor allele frequency, number of repeated measurements and missing data patterns. In our study, the event of interest was onset of type 2 diabetes (T2D), and the longitudinal biomarker repeatedly measured over time was fasting plasma glucose level. To the best of our knowledge, joint models have never been applied into a genetic epidemiology context and could help identify novel loci sharing effects on both glycaemic traits and T2D.

# Key words

joint modelling; genetic association; longitudinal studies

# Introduction

With the increase availability of longitudinal data and survival data within prospective cohorts, joint models have emerged to account for both type of data, in particular to deal with the informative/non-informative dropouts that occurred in such cohorts. Joint models have been studied and overviewed in the literature (L. M. Chen, Ibrahim, & Chu, 2011; R. Elashoff, li, & Li, 2016; Anastasios A. Tsiatis & Davidian, 2004; M. S. Wulfsohn & Tsiatis, 1997) and software implementation has been proposed within different software and platforms (Diggle & Kenward, 1994; R. M. Elashoff, Li, & Li, 2008; Proust-Lima, Joly, Dartigues, & Jacqmin-Gadda, 2009; Rizopoulos, 2010; Rizopoulos & Ghosh, 2011; J. Sun, Sun, & Liu, 2007).

The main idea behind the joint modelling is 1) to model efficiently the survival process with a time-varying covariate, accounting for missing data and for measurements errors, 2) to account for informative dropouts in the longitudinal data that could arise.

To model the two components of a joint model, a linear mixed effects (LME) model and a Cox proportional hazards model (CoxPH), are classically used to respectively fit the longitudinal component and the survival component. Contrary to a CoxPH, where the time-varying covariate is assumed to be exogenous, meaning without any effects of event at a previous time on the time-varying covariate itself, the joint modelling framework allow to account for en endogenous time-varying covariate , as for example fasting glucose and T2D event. Two approaches can be used for estimation and inference of the parameters: a "naïve" Two-Step (TS) method or a joint likehood method (JM). The first method consists in using the estimation of the random effect or the trajectory, provided by a LME, into a CoxPH, as a time-varying covariate, doing so using partial likelihood from CoxPH for inference of the parameters (Therneau & Grambsch, 2000). The second, is based on a joint likelihood of the two components (longitudinal and survival) at the same time. Comparison of these two approaches showed that the latter offers estimations more consistent and efficient than the former, by means of the maximum likelihood (Albert & Shih, 2010a, 2010b). But JM could be challenging to compute, especially to achieve the convergence of the Expectation-Maximisation (EM) step. Moreover, depending on the number of time points and the sample size, the overall computation time can substantially increase.

In this paper, we examine the sensitivity and robustness of the proposed joint models to assess the two processes (longitudinal and survival) and the link between, in the case where the time-varying covariate of interest is endogenous (Rizopoulos, 2010, 2012), meaning measured with error, incomplete and dropouts are linked to the previous values of the time-varying covariate. We conducted a comprehensive simulation study to compare two joint models methods: JM and TS for joint modelling of the longitudinal and survival components. We included classical approaches to model the two processes separately, as LME and CoxPH models, used in classical genome-wide association studies. The main goal is to improve the statistical power to detect an effect on either or both longitudinal and survival processes, with a reduced bias in the estimations. Using a joint modelling approach compared to separate analyses, thus we also compared JM to a more naïve approach as a TS, combining LME and CoxPH model in a sequential manner. In the context where the highly demanding computation and convergence issues might arise in JM computation, TS might be a good approximation in an application to a genome-scale study in a reasonable time frame, for discovery purposes and refined using JM.

We also investigated and profiled the computation demand of the JM implemented in the R package "JM" and TS implemented using the R packages: "survival" and "nlme".

Lastly, we analysed a real dataset, the prospective cohort D.E.S.I.R. (Données Épidémiologiques sur le Syndrome d’Insulino-Résistance), which includes 5,212 individuals followed-up for 9 year, and extensive phenotypic data has been recorded at 4 different time (visit every 3 years). These individuals were genotyped using the Illumina Metabochip DNA array (includes nearly up to 200,000 SNPs). By exploiting genome-wide association studies (GWAS), the cohort D.E.S.I.R amongst others allowed the identification of novel loci associated to prevalent type 2 diabetes (T2D), on the one hand, and to blood fasting glucose (FG) levels variation, on the other hand (Dupuis et al., 2010; Sladek et al., 2007; Vaxillaire et al., 2014). We focused in our real data application, on prediabetes conditions, such as FG, which is part of the diagnostic definition of T2D (FG>7 mmol/l) and the time-to-onset of T2D, in order to possibly identify loci which might be simultaneously associated to the risk of developing T2D and FG levels variation. Results were then compared to the genetic variants reported in the literature (Vaxillaire et al., 2014; Welter et al., 2014) and to consortium meta-analyses, as DIAGRAM (Morris et al., 2012) and MAGIC (Dupuis et al., 2010).

# Methods

## Models Formulation

### Joint Likelihood Model

Standard formulation of the joint model involves two components: a longitudinal component and a time-to-event component. Let denotes the sample size and , the longitudinal measurements collected for each subjects at times with and , where is the number of measurements on subject . The longitudinal component (measurements) typically consists of a (generalised) linear mixed effect (LME) model, which the correlations between measurements for each subjects are modelled within the random effects parameter .

Under the joint likelihood framework as implemented in "JM" (Rizopoulos, 2010), from the class of "the shared parameter models"(R. Elashoff et al., 2016; Rizopoulos, 2012)

where is the observed value and is the true (unobserved) value of the longitudinal variable. The quantity is a random error term usually assumed to be normally distributed:

The quantity is typically called the trajectory function and is usually specified as a linear (or quadratic) function of time . denotes a covariate of interest and denotes the adjusting covariates according to the model:

For ease of representation, the term will be omitted in the following. Random effects and are assumed to be distributed as a multivariate normal distribution: and assume to be independent from . The coefficient assesses the additive effect on the trajectory function. To account for possibly varying slopes between , an interaction term between and time could be included in the trajectory function. The interaction term was not included in our study. The time-to-event (survival) component usually consists of a parametric (e.g. exponential or Weibull distribution) or semi-parametric (e.g. Cox proportional hazards) model. Let denotes the event time for subject and the right censoring time (e.g. end of the follow-up). With these notations, let be the event indicator, where if and when .

Under the Cox regression model, the hazard of , can be specified with the following equation:

where is the hazard function at time and is the unspecified baseline hazard function assumed to be a piecewise constant with two knots placed at the intermediary times in the following. The coefficient measures the effect of on the time-to-event, whereas measures the association between the time-varying variable and time-to- event. Here we assumed in the model that the subject-specific parameters (intercept and slope) for the trajectory also affect the time-to-event, and so is the parameter that links the two sub-models.

### Two-Step Model

Using the formulation in (A. A. Tsiatis, DeGruttola, & Wulfsohn, 1995) to estimate the joint model defined in Equation (3) and Equation (4), where for any event time , is replaced by its estimation from Equation (3) in Equation (4).

## Simulation Study

Simulation studies were carried out to further examine the sensitivity of the JM estimations within several scenarios. Default settings were based on rs17747324 (TCF7L2) in Table 1. This SNP was selected based on its reported association with T2D (Effect allele: C; ;; (Morris et al., 2012)) and FG (Effect allele: C; ;; (Dupuis et al., 2010)).

Longitudinal data were simulated according to Equation (3), then event times were generated according to the exponential distribution for the CoxPH model (Austin, 2012)

Where was set to achieve the targeted incidence rate in the simulated dataset.  
Datasets were simulated by varying the number of measures (), sample size (, allele frequency ( and the incidence rate (, leading to 240 different scenarios with respect to a full factorial design. Each scenarios were simulated 500 times.

The Root-Mean-Square Error (RMSE) was used to compare the estimations of , and , respectively for the association between and , effect on and effect on from the JM and TS approaches compared to the usual separate approaches, linear mixed effect model and Cox regression with time-varying covariate. Power and type 1 error were explored for all models. The computational burden was also of interest for an application on genome-wide scale and was recorded.

## Computation times

Using our simulations, we provided an approximate computation time for four sample sizes with the default parameters listed in Table 1, used to simulate the datasets, in a UNIX system (Intel® Xeon® CPU E7-8870 v3 @ 2.10GHz) using one CPU (144 available). Time consumptions are given for one model and extrapolate to 100,000, which was the approximate size of our genotyped data, after the quality control of the Illumina Metabochip.

## Real Data

Single Nucleotide Polymorphism (SNP) genotyping was performed with Metabochip DNA arrays (custom iSelect-Illumina genotyping arrays)(B. F. Voight et al., 2012) using the Illumina HiScan technology and GenomeStudio software (Illumina, San Diego, CA, 19 USA) in 5,212 samples from the French cohort D.E.S.I.R. (Données Épidémiologiques sur le Syndrome d’Insulino-Résistance)(B. Balkau, 1996), who have been followed up for 9 years, and extensive phenotypic data has been recorded at 4 different time occasions during follow-up. Quality control was performed using PLINK (1.90 beta)(C. C. Chang et al., 2015; S. Purcell & Chang, 2015).

232 samples were removed due to missing phenotypes, which did not allow to assess the type 2 diabetes (T2D) status. SNPs and samples with a call rate greater or equal to 95 %, with no departures from Hardy-Weinberg equilibrium () and a minor allele frequency (MAF) over 5 % were kept, leading to 101,305 SNPs and 4,426 samples.

A Principal Component Analysis (PCA) was performed in a combined dataset involving the 4,426 samples and the samples from the publicly available 1,000 Genomes project database (The 1,000 Genomes Project Consortium, 2015), and thinned down to only contain SNPs that were present in both dataset and not palindromic. The first two components were sufficient to discriminate ethnic origin. Non-Caucasians samples (62) were pruned out, leading to 4,364 Caucasian samples. In addition, Prevalent T2D (12 samples) were also excluded from the analysis (Figure 1).

The final dataset included 4,352 samples (with 167 T2D incident cases) and 101,305 SNPs.

Using the joint modelling approach implemented in the package JM (Rizopoulos, 2010), within R softwares verions 3.3.1 (R Core Team, 2016), 101,305 SNPs from Metabochip DNA arrays were tested for association to blood fasting glucose and T2D risk simultaneously. Based on the joint modelling formulation in Equation (3) and (4). Let be the trajectory of blood fasting glucose (FG) level, the genotype of a single nucleotide polymorphism (SNP) with covariates as age, gender and BMI. The event of interest here is T2D, so let be the time at which a subject becomes T2D.

As shown in Figure 2, the association between the genotype at a specific SNP and firstly, with FG captured through the parameter , secondly with the time of event (onset of T2D) captured through parameter , by taking into account the link between the dropout process of the measure of FG () and the onset of T2D within the parameter . By definition, a subject is defined as T2D when the measure of FG is over 7 mmol/l (and/or glycaemic treatment) and thus the measure (and the ones after) become missing values, since T2D subjects have to take a treatment to regulate their glycaemia. In this case FG is considered as an endogenous covariate since the dropout process is not independent from the history of .

# Results

## Comparisons of the statistical efficiency

Using a factorial design, we explored the influence of several factors on the performance of estimations using a linear mixed effect model to estimate and using a Cox regression model with time-varying covariate to estimate and compared to the joint modelling approach. For each simulated scenario, we measured the RMSE for all three parameters of interest (, and ).

Due to the complexity of the estimation for JM, convergence could not be obtained ( % of convergence issues in average per scenario) using the algorithm chosen for all of 500 simulations (i.e. "piecewise-PH-aGH" for a time-dependent relative risk model with a piecewise constant baseline risk function using the adaptive Gauss-Hermite quadrature rule to approximate integrals within the Expectation-Maximisation (EM) step).

RSME computed for parameter (Figure 3), the effect of on the trajectory , showed performance quite similar between JM and TS, which could be expected regarding the formulation on the joint model within the "Shared Parameter Models" in which (mean of modelled within LME according to Equation (3)) links the longitudinal data to the time of event.

Regarding the survival sub-model (CoxPH model), RSME for parameter (Figure 4) and for parameter (Figure 5) showed to be better within the joint modelling framework (JM or TS) than within a more classical approach as CoxPH model. When RMSE for was consistent across all scenarios and higher than JM or TS, it decreased with the increasing of either: the sample size, the incidence rate and the allele frequency. However, TS performed as well as JM with regard to .

Differences for parameter were less important, where TS performed as well as CoxPH with time-dependent covariate, due to the partial likelihood inferences used in both approaches. JM estimations were less biased in almost all scenarios or equally biased, where the sample size was higher than , with a type 1 error and power lower for parameter in JM than in TS. Our simulations revealed that JM is more often less (or equally) biased than separate approaches to model the effect of on the longitudinal data , on the one hand, and on the time of event , on the other hand, especially to account for the link between the drop-out process equals to the time-to-event) and the longitudinal trajectory, when separate approaches performed well for or , the bias for was the higher observed in all scenarios.

In addition, statistical power and type 1 error were also studied (Table ??), for the default simulation settings (Table 1), and showed similar results between JM and TS approaches. Nevertheless, these last simulations highlighted convergence issues that might occur within the joint likelihood approach (19.4 % of the power simulation study).

## Computation times

Computation times are reported in Table 2. We observed that times increased linearly with the sample size in our simulations, but these approximations are very optimistic since the model did not include covariates or complex random parameters.

In order to investigate further the computation time issue, we profiled the execution of the main function "jointmodel" from the R package "JM", which implements the joint likelihood modelling approach described and chosen in this paper. In the "JM" package, the linear mixed effect sub-model is handled by the function "lme" (from "nlme" package), one may argued that using a faster approach (as in the R package "lme4"), the computation time might decrease. As shown in Figure 6, the main issue is within the "jointmodel" function which took over 95 % of the overall computation time, by looking deeper into the call tree diagram, we can see that the more consuming task, within the "jointmodel" function, is the optimisation of the EM algorithm (described in Rizopoulos (2012), Appendix B), despite calculation tricks (i.e. adaptive Gauss-Hermite quadrature for numerical integration).

## Real data application

JM lead to a top 265 SNPs (Figure 7) which were associated with FG and T2D event (pvalue<0.05) through parameters and .

Amongst these 265 SNPs (163 unique genes), 17 genes (Table 3) were already reported to be associated with FG and/or T2D risk.

We focused on and , the effects of a specific SNP in the longitudinal and survival sub-models, the overall results for these parameters are shown in Figure 8. No genetic variants showed highly significant effect (robust to multiple testing correction, i.e. Bonferroni at 5 % is ) for both parameters and at the same time. Only SNPs in the following genes (or within a 100kb window): G6PC2/ABCB11, GCK/YKT6, GCKR and MTNR1B, were significant at a Bonferroni level for testing , with effect per risk allele ranging from an increase of 0.047 mmol/l to 0.10 mmol/l. Focusing on simultaneous associations with the longitudinal and survival processes, revealed well known genes as TCF7L2, which was shown in meta-analysis to be associated with elevated FG and an increase risk of T2D (Table 4). MTNR1B was also found to be nominally associated (34 SNPs within 50kb), with () and () for rs10830963\_G (MTNR1B), the SNP usually reported.

When rs17747324 showed consistent results, for both and with the DIAGRAM meta-analysis (Table 4), rs10830963 showed a reverse effect on T2D compared to the effect reported in MAGIC for FG (, ).

To go further in the comparison of JM and TS, we performed the analysis of the Illumina Metabochip using TS as well. As shown in Figure ??, the approximations of the p-values can be inaccurate, especially for parameter , when for parameter , the approximations are quite similar to the one provided via the joint likelihood framework.

# Discussion / Conclusion

With the increasing number of prospective cohort, the availability of genomics data, through genotyping array and next generation sequencing, the need to develop and use efficient models is important to ensure analysis in a reasonable time frame. In this paper, we proposed a comparison of two approaches, namely Two-Step and joint likelihood methods to infer parameters accounting for an effect on longitudinal and survival processes, without omitting information about the dropouts or the status of the longitudinal variable of interest. As in our real data application, using FG as the longitudinal trait, T2D event as the survival trait which are both linked together by a threshold on FG to define T2D (FG>7 mmol/l). Through simulations over different scenarios’ settings, we showed that joint models are less biased than classical separate approaches, and could provide more insight regarding the event of interest (i.e. T2D), and thus the potential impact of a SNP on the becoming of a patient to T2D status.

By looking at different statistical measures, such as RMSE for bias in the estimations provided by models, type 1 error, power and computation burden, using the available implementation of joints models, revealed that in a genome-wide scale, the use of an approximate method, as TS, might be a good choice for start on a bias/computation time ratio, despite possible lack of accuracy within the estimations provided by such method. TS could be used to overcome the computation burden of the joint likelihood methods, using software already available to perform the two steps: LME and CoxPH, and thus filter out SNPs with low or not detectable associations. Depending on the population settings (sample size, incidence rate, number of measures), the use of a joint likelihood method is highly recommended to get proper estimations of , , the effects of one SNP on the trajectory of FG and time-to T2D, as in our real data. Also, using parallel and grid computing approaches may reduce the computation time in a more suitable time frame for an application to genome-wide scale data (over millions of SNPs).

In our real data application, the results observed for MTNR1B, in the French cohort D.E.S.I.R., even if it seemed to be inconsistent with the literature, it might have unveiled some sub-structure within the T2D population. Moreover, since SNPs in MTNR1B reported for being associated with an increase of FG level in blood and elevated T2D risk, the analyses were performed on different population (via meta-analyses) and never at the same time, through for example a survival model accounting for time-dependent covariate, in this case, FG.

However, MTNR1B results need to be confirmed in a replication cohort, since it might just be cohort specific, in addition the low number of incidence cases in D.E.S.I.R. cohort (167 incident T2D within the follow-up).

# Acknowledgments

This study was supported by grants for funding of scientific research conducted in France and within the European Union: "Centre National de la Recherche Scientifique", "Université de Lille 2", "Institut Pasteur de Lille", "Société Francophone du Diabète", "Lilly", "Contrat de Plan Etat-Région", "Agence Nationale de la Recherche", ANR-10-LABX-46, ANR EQUIPEX Ligan MP : ANR-10-EQPX-07-01, European Research Council GEPIDIAB - 294785.

###### Conflict of interest disclosure

The authors declare that there are no conflicts of interest.

# References

# List of tables

TABLE 1 Default settings for sensitivity analysis and simulation based on rs17747324 (TCF7L2).

|  |  |
| --- | --- |
| Parameters | Values |
| Sample size () | 4,351 |
| Number of measures () | 4 |
| Incidence () | 0.0384 |
| Minor allele frequency () | 0.244 |
| Random effects () |  |
| SNP effect on () | 0.0229 |
| SNP effect on () | 0.265 |
| Association between and () | 3.17 |
| Error term () |  |

TABLE 2 Approximate system time (from the function system.time) using R software. System time is computed ten times per sample sizes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Joint Model |  | Two-Step Model |  |
| Sample Size | mean (sd) / test | 100K test | mean (sd) / test | 100K test |
| 500 | 51 sec(3.4) | 59 days | 0.71 sec(0.066) | 0.82 days |
| 2,500 | 100 sec(11) | 120 days | 3.1 sec(0.092) | 3.6 days |
| 5,000 | 180 sec(25) | 210 days | 6.3 sec(0.17) | 7.3 days |
| 10,000 | 340 sec(34) | 400 days | 9 sec(0.22) | 10 days |

TABLE 3 List of loci found to be associated within the joint modelling framework with both FG and T2D, previously showed as associated with FG and/or T2D in the NHGRI GWAS Catalog (Welter et al., 2014).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SNP | (p-value) | (p-value) | (p-value) | Power() |
| rs6945660\_G (ETV1) | 0.55 () | 0.0352 () | 3.48 () | 69.7% |
| rs1942873\_C (MC4R) | 0.41 () | 0.0234 () | 3.14 () | 69.6% |
| rs55899248\_G (TCF7L2) | 0.292 () | 0.0253 () | 3.49 () | 55.3% |
| rs17301514\_A (ST6GAL1) | -0.657 () | 0.0451 () | 3.65 () | 45.8% |
| rs833425\_C (PTPRD) | 0.321 () | 0.0432 () | 3.51 () | 44.2% |
| rs7072870\_A (C10orf35) | -0.404 () | 0.0248 () | 3.58 () | 39.6% |
| rs61871514\_A (KCNQ1) | 0.425 () | 0.0457 () | 3.18 () | 39.4% |
| rs9883865\_A (ADAMTS9) | -0.598 () | 0.0426 () | 3.2 () | 34.9% |
| rs114508985\_C (HLA) | -0.294 () | 0.0209 () | 3.22 () | 27.1% |
| rs10814856\_T (GLIS3) | -0.265 () | 0.0248 () | 3.2 () | 18.5% |
| rs73025532\_C (SLC22A1) | -0.377 () | 0.0317 () | 3.58 () | 17.3% |
| rs11769484\_C (JAZF1) | -0.254 () | 0.0221 () | 3.21 () | 16.9% |
| rs6450176\_G (ARL15) | -0.291 () | 0.0365 () | 3.54 () | 15.2% |
| rs4712580\_C (CDKAL1) | -0.289 () | 0.0313 () | 3.57 () | 14.0% |
| rs10830963\_G (MTNR1B) | -0.44 () | 0.0991 () | 3.25 () | 10.2% |
| rs853787\_T (ABCB11) | -0.247 () | 0.0831 () | 3.21 () | 3.3% |
| rs560887\_C (G6PC2) | -0.315 () | 0.0992 () | 3.21 () | 2.6% |

TABLE 4 Effects on FG and T2D risk from the joint modelling approach compared to consortium meta-analysis on MTNR1B and MTNR1B loci.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | (p-value) | (p-value) | (p-value) | (p-value) | (p-value) |
| RSID | JM (D.E.S.I.R.) | DIAGRAM | JM (D.E.S.I.R.) | MAGIC | JM (D.E.S.I.R.) |
| rs10830963\_G (MTNR1B) | -0.44 () | 0.104 () | 0.0991 () | 0.079 () | 3.25 () |
| rs17747324\_C (TCF7L2) | 0.265 () | 0.358 () | 0.0229 () | 0.025 () | 3.17 () |

# List of figures

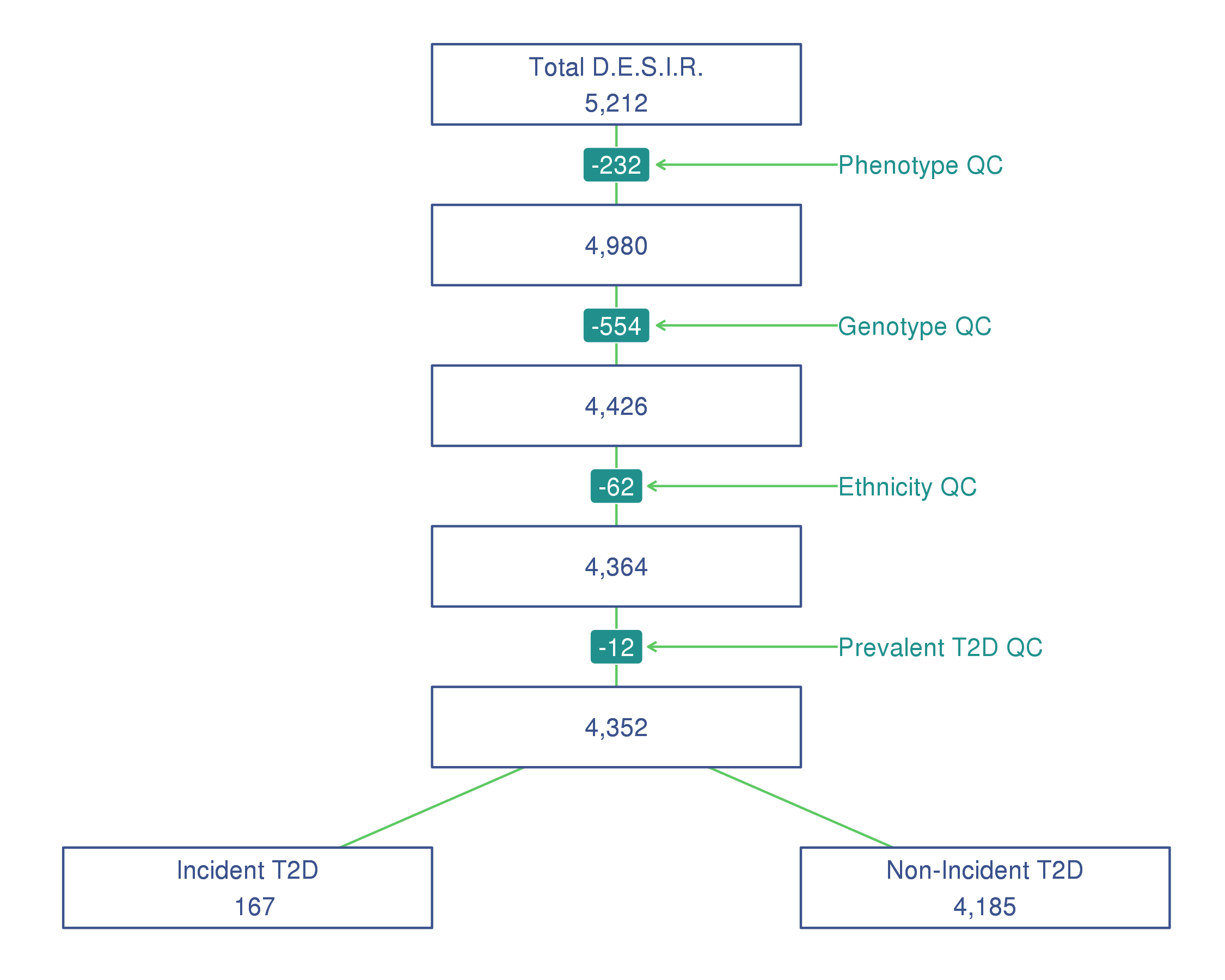


FIGURE 1 Flowchart quality control on samples from the French cohort D.E.S.I.R.

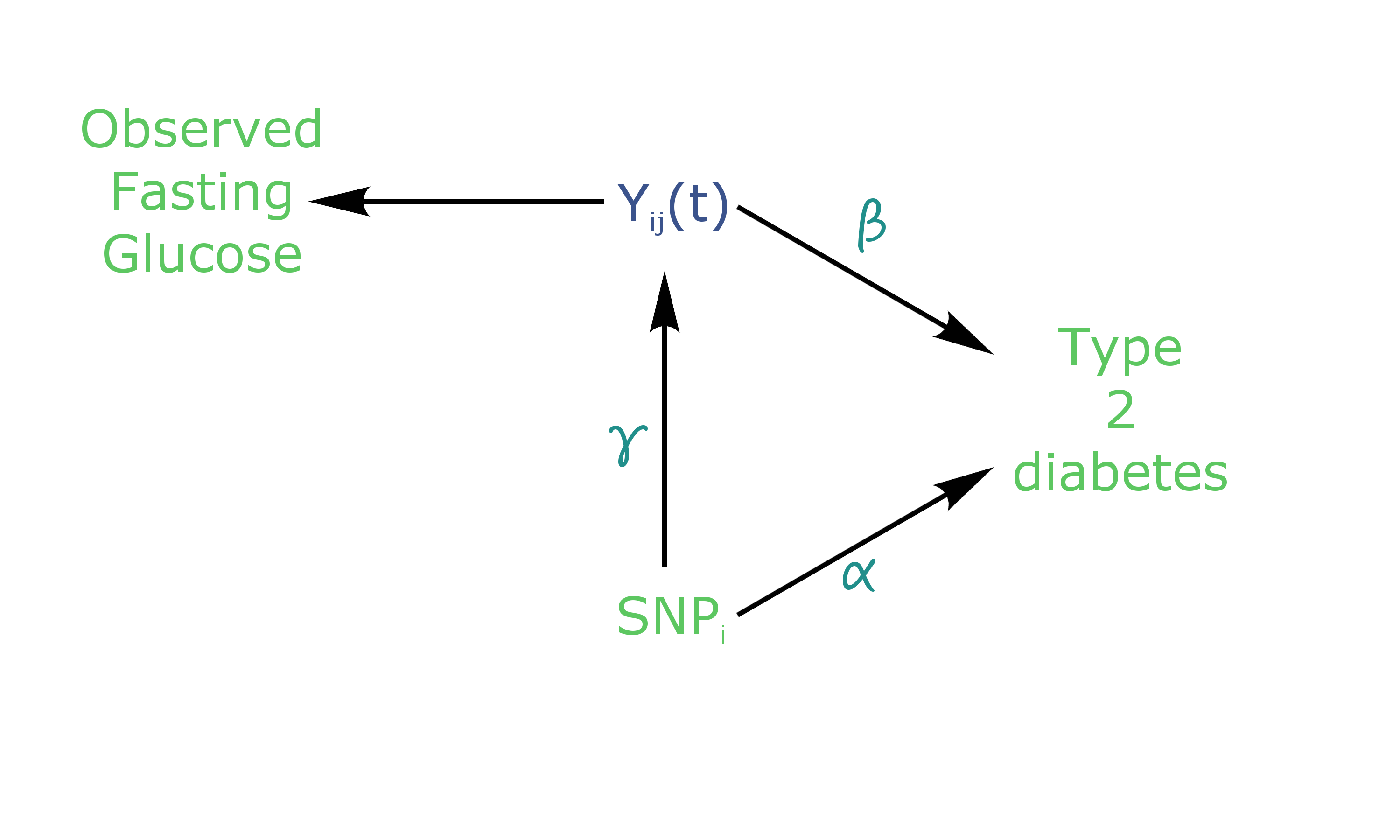


FIGURE 2 Causal diagram for joint modelling applied to T2D (adapted from Ibrahim, Chu, & Chen (2010)).

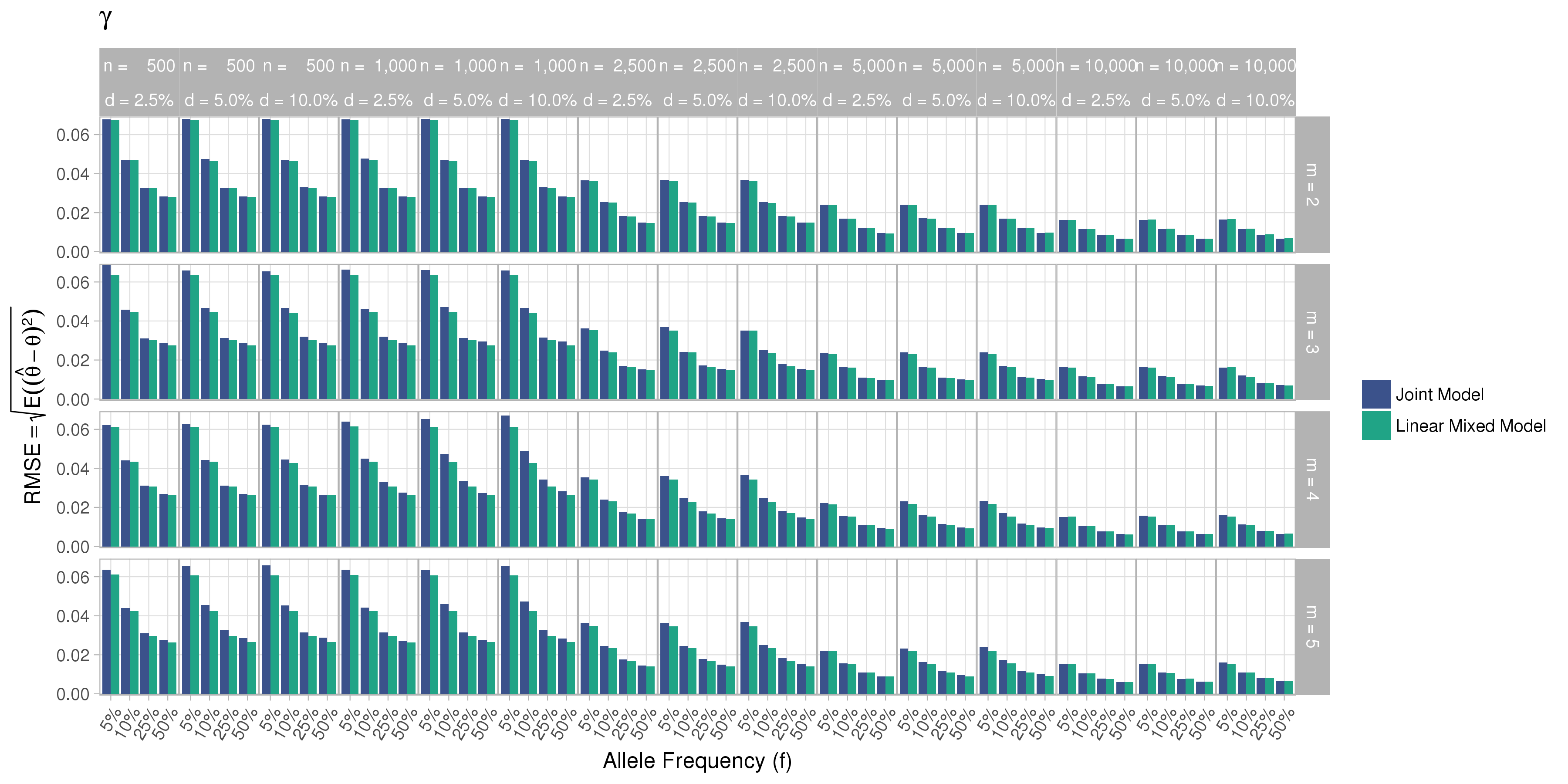


FIGURE 3 Simulation study for sensitivity analysis of the estimation provided by the Joint Model (JM package) and from the Linear Mixed Effect Model (nlme package). , the number of measures; , the sample size; , the incidence rate.

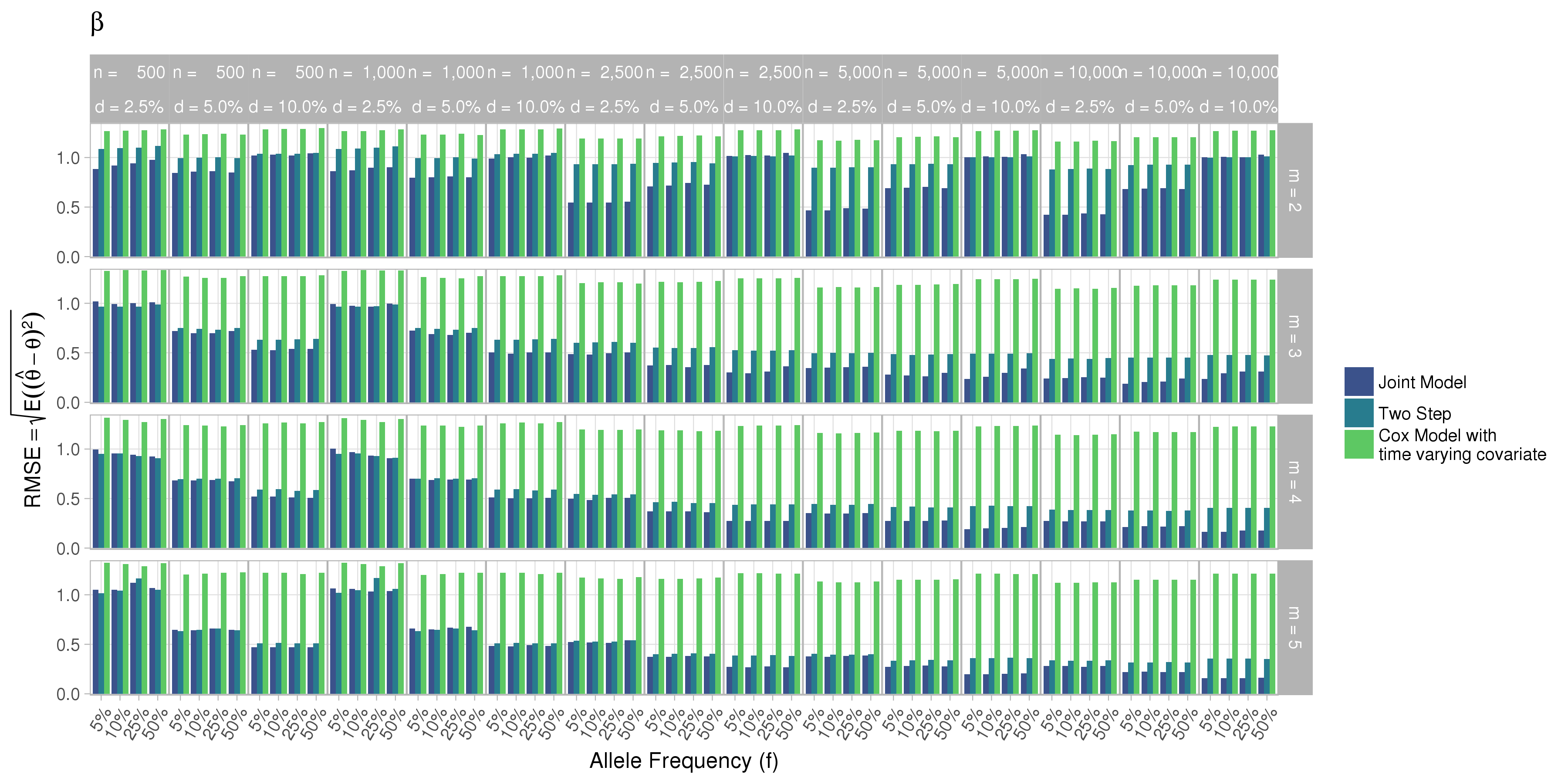


FIGURE 4 Simulation study for sensitivity analysis of the estimation provided by the Joint Model (JM package) and from the Linear Mixed Effect Model (nlme package). , the number of measures; , the sample size; , the incidence rate.

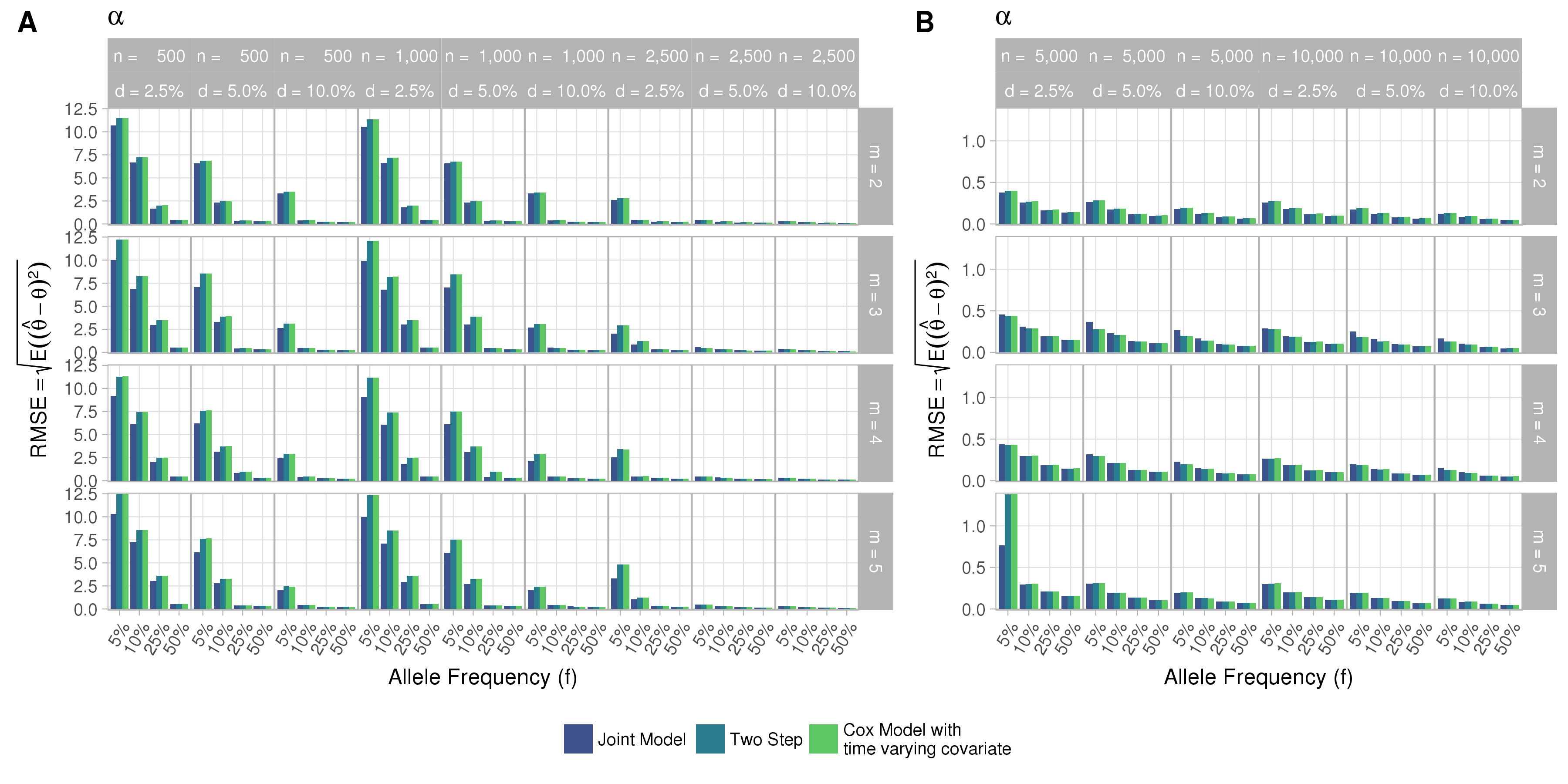


FIGURE 5 Simulation study for sensitivity analysis of the estimation provided by the Joint Model (JM package) and from the Linear Mixed Effect Model (nlme package). , the number of measures; , the sample size; , the incidence rate.

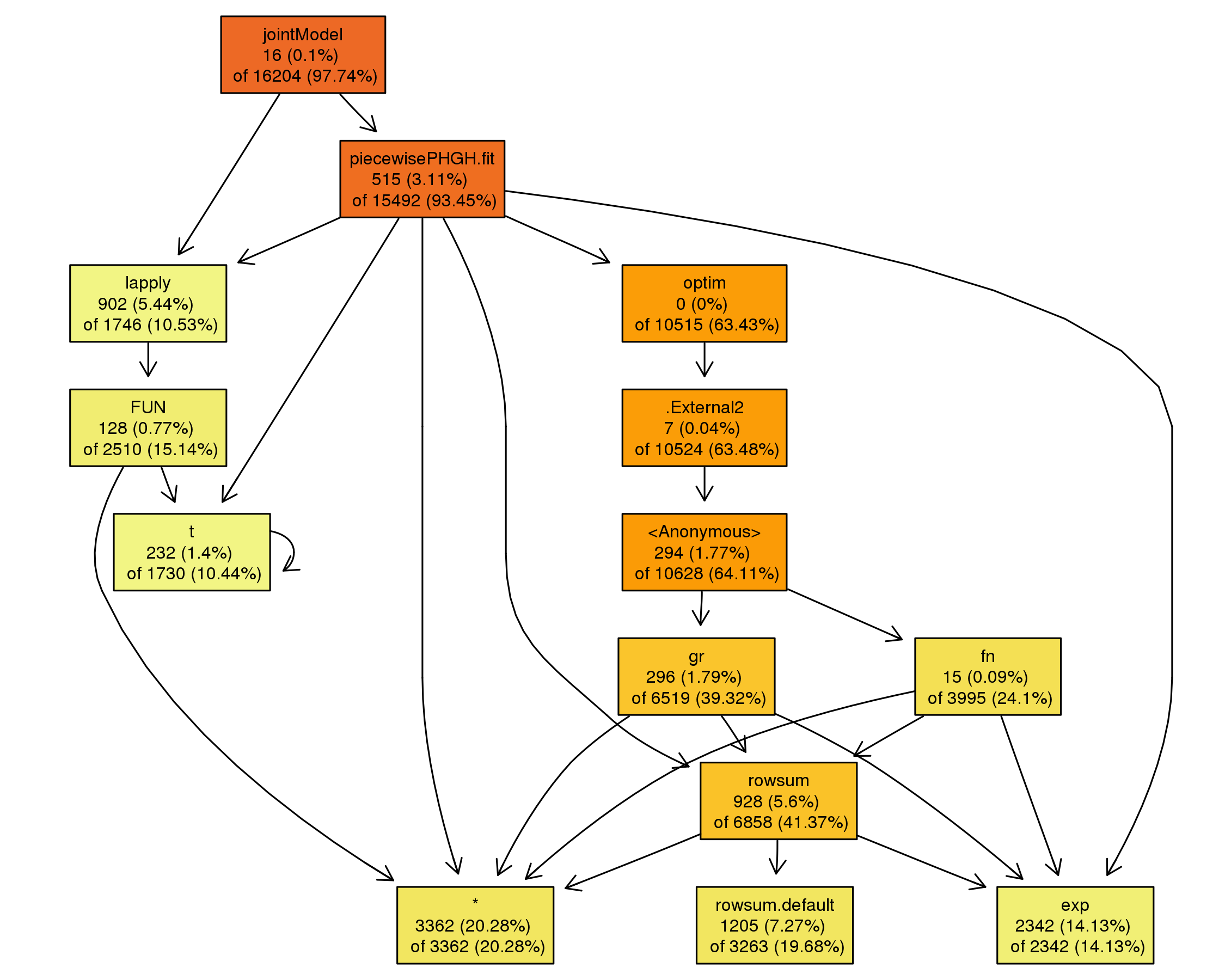


FIGURE 6 Call tree diagram of the main function jointmodel in the R package JM. Call based on a simulated dataset with three measures and five thousand subjects (default settings from Table 1).

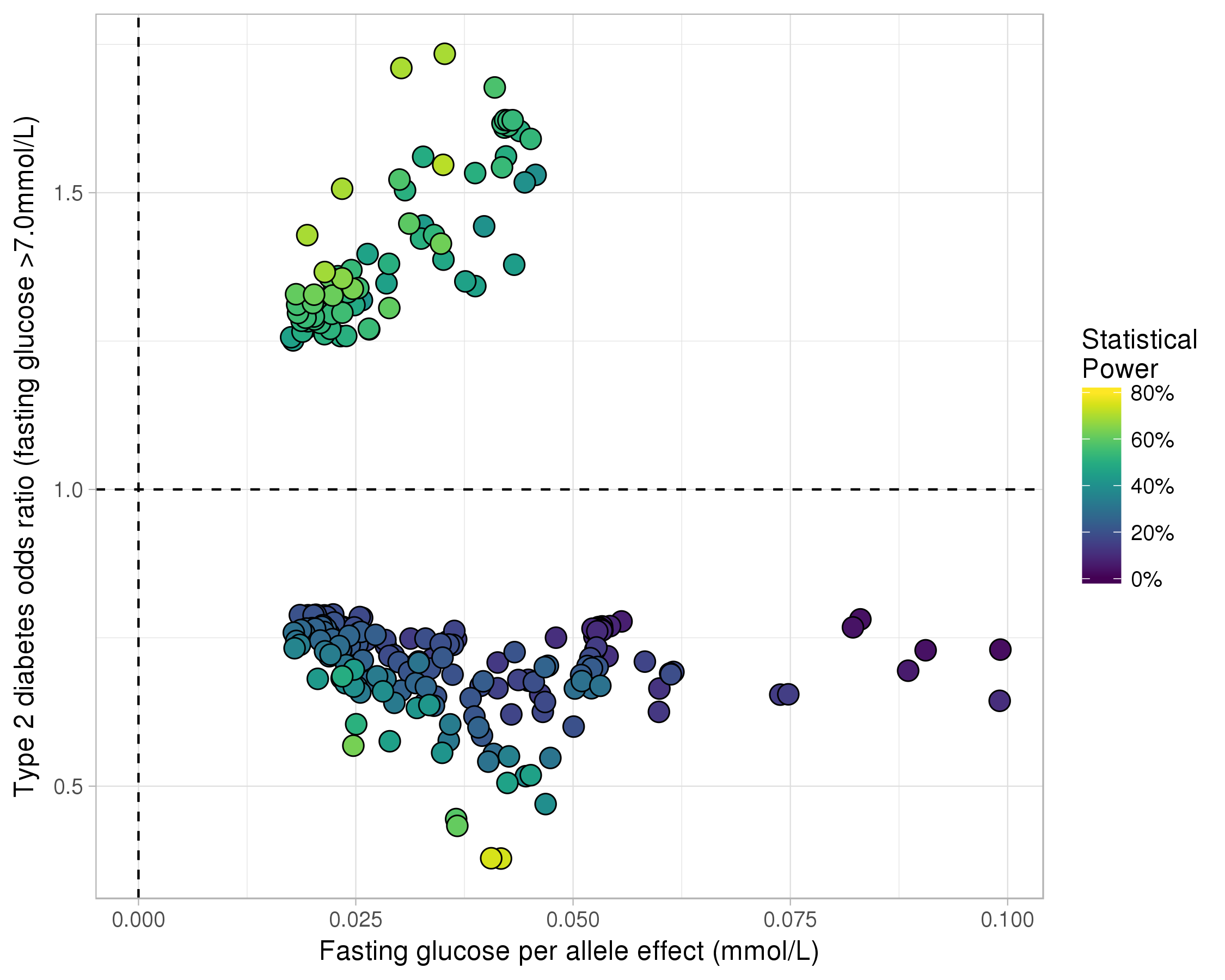


FIGURE 7 Results for effects in x-axis and in y-axis from Joint Model (Rizopoulos, 2010). Power reported is the theoretical power to detect a joint effect (L. M. Chen et al., 2011).

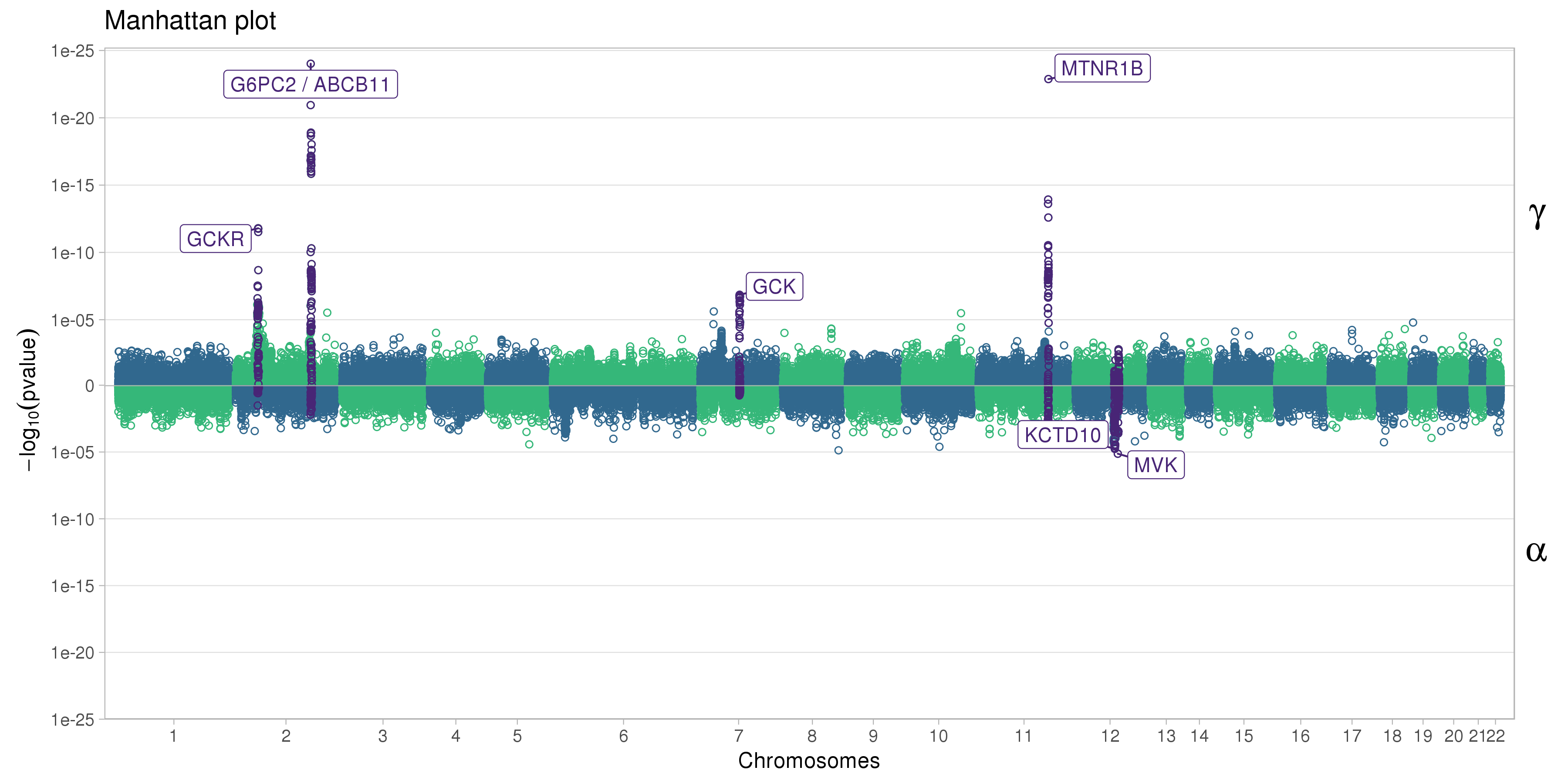


FIGURE 8 Manhattan plot for effects and from the joint model analysis for 101,305 SNPs of the Illumina Metabochip.

Albert, P. S., & Shih, J. H. (2010a). An approach for jointly modeling multivariate longitudinal measurements and discrete time-to-event data. *The Annals of Applied Statistics*, *4*(3), 1517–1532.

Albert, P. S., & Shih, J. H. (2010b). On Estimating the Relationship between Longitudinal Measurements and Time-to-Event Data Using a Simple Two-Stage Procedure. *Biometrics*, *66*(3), 983–987.

Austin, P. C. (2012). Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Statistics in Medicine*, *31*(29), 3946–3958.

Balkau, B. (1996). An epidemiologic survey from a network of French Health Examination Centres, (D.E.S.I.R.): Epidemiologic data on the insulin resistance syndrome. *Revue D’épidémiologie Et De Santé Publique*, *44*(4), 373–375.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*, 7.

Chen, L. M., Ibrahim, J. G., & Chu, H. (2011). Sample size and power determination in joint modeling of longitudinal and survival data. *Statistics in Medicine*, *30*(18), 2295–2309.

Diggle, P., & Kenward, M. G. (1994). Informative Drop-Out in Longitudinal Data Analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *43*(1), 49–93.

Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A. U., … Barroso, I. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genetics*, *42*(2), 105–116.

Elashoff, R. M., Li, G., & Li, N. (2008). A Joint Model for Longitudinal Measurements and Survival Data in the Presence of Multiple Failure Types. *Biometrics*, *64*(3), 762–771.

Elashoff, R., li, G., & Li, N. (2016). *Joint Modeling of Longitudinal and Time-to-Event Data* (1st ed.). Chapman; Hall/CRC.

Ibrahim, J. G., Chu, H., & Chen, L. M. (2010). Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data. *Journal of Clinical Oncology*, *28*(16), 2796.

Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segrè, A. V., Steinthorsdottir, V., … McCarthy, M. I. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, *44*(9), 981–990.

Proust-Lima, C., Joly, P., Dartigues, J.-F., & Jacqmin-Gadda, H. (2009). Joint modelling of multivariate longitudinal outcomes and a time-to-event: A nonlinear latent class approach. *Computational Statistics & Data Analysis*, *53*(4), 1142–1154.

Purcell, S., & Chang, C. (2015). PLINK v1.90b3.36.

R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rizopoulos, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, *35*(9), 1–33.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press.

Rizopoulos, D., & Ghosh, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine*, *30*(12), 1366–1380.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., … Froguel, P. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, *445*(7130), 881–885.

Sun, J., Sun, L., & Liu, D. (2007). Regression Analysis of Longitudinal Data in the Presence of Informative Observation and Censoring Times. *Journal of the American Statistical Association*, *102*(480), 1397–1406.

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. (K. Dietz, M. Gail, K. Krickeberg, J. Samet, & A. Tsiatis, Eds.). New York, NY: Springer New York.

Tsiatis, A. A., & Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, *14*, 809–834.

Tsiatis, A. A., DeGruttola, V., & Wulfsohn, M. S. (1995). Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS. *Journal of the American Statistical Association*, *90*(429), 27–37.

Vaxillaire, M., Yengo, L., Lobbens, S., Rocheleau, G., Eury, E., Lantieri, O., … Froguel, P. (2014). Type 2 diabetes-related genetic risk scores associated with variations in fasting plasma glucose and development of impaired glucose homeostasis in the prospective DESIR study. *Diabetologia*, *57*(8), 1601–1610.

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., … Boehnke, M. (2012). The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. *PLoS Genetics*, *8*(8), e1002793.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., … Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, *42*(D1), D1001–D1006.

Wulfsohn, M. S., & Tsiatis, A. A. (1997). A Joint Model for Survival and Longitudinal Data Measured with Error. *Biometrics*, *53*(1), 330.