frontiers in Genetics

METHODS
published: 14 June 2018
doi: 10.3389/fgene.2018.00210

Jointly Modelling Single Nucleotide Polymorphisms With Longitudinal and Time-to-Event Trait: An Application to Type 2 Diabetes and Fasting Plasma Glucose

Mickaël Canouil[1,2,3]*, *Beverley Balkau*[4,5,6], *Ronan Roussel*[7,8,9], *Philippe Froguel*[1,2,3,10] and *Ghislain Rocheleau*[1,2,3]

[1] Université de Lille, UMR 8199-EGID, Lille, France, [2] Centre National de la Recherche Scientifique, UMR 8199, Lille, France, [3] Institut Pasteur de Lille, Lille, France, [4] Centre for Research in Epidemiology and Population Health, Villejuif, France, [5] Université Paris-Saclay, Université Paris Sud, UVSQ, UMRS 1018, Villejuif, France, [6] Institut National de la Santé et de la Recherche Médicale U1018, Centre de Recherche en Épidémiologie et Santé des Populations, Renal and Cardiovascular Epidemiology, UVSQ-UPS, Villejuif, France, [7] Institut National de la Santé et de la Recherche Médicale U1138 (équipe 2: Pathophysiology and Therapeutics of Vascular and Renal Diseases Related to Diabetes, Centre de Recherches des Cordeliers), Paris, France, [8] Université Paris 7 Denis Diderot, Sorbonne Paris Cité, Paris, France, [9] AP-HP, DHU FIRE, Department of Endocrinology, Diabetology, Nutrition, and Metabolic Diseases, Bichat Claude Bernard Hospital, Paris, France, [10] Department of Genomics of Common Disease, Imperial College London, London, United Kingdom

OPEN ACCESS

Edited by:
Mogens Fenger,
Capital Region of Denmark, Denmark

Reviewed by:
Duncan C. Thomas,
University of Southern California, United States
Paola Sebastiani,
Boston University, United States

***Correspondence:**
Mickaël Canouil
mickael.canouil@cnrs.fr

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *12 February 2018*
Accepted: *25 May 2018*
Published: *14 June 2018*

Citation:
Canouil M, Balkau B, Roussel R, Froguel P and Rocheleau G (2018) Jointly Modelling Single Nucleotide Polymorphisms With Longitudinal and Time-to-Event Trait: An Application to Type 2 Diabetes and Fasting Plasma Glucose. Front. Genet. 9:210. doi: 10.3389/fgene.2018.00210

In observational cohorts, longitudinal data are collected with repeated measurements at predetermined time points for many biomarkers, along with other variables measured at baseline. In these cohorts, time until a certain event of interest occurs is reported and very often, a relationship will be observed between some biomarker repeatedly measured over time and that event. Joint models were designed to efficiently estimate statistical parameters describing this relationship by combining a mixed model for the longitudinal biomarker trajectory and a survival model for the time until occurrence of the event, using a set of random effects to account for the relationship between the two types of data. In this paper, we discuss the implementation of joint models in genetic association studies. First, we check model consistency based on different simulation scenarios, by varying sample sizes, minor allele frequencies and number of repeated measurements. Second, using genotypes assayed with the Metabochip DNA arrays (Illumina) from about 4,500 individuals recruited in the French cohort D.E.S.I.R. (*Data from an Epidemiological Study on the Insulin Resistance syndrome*), we assess the feasibility of implementing the joint modelling approach in a real high-throughput genomic dataset. An alternative model approximating the joint model, called the Two-Step approach (TS), is also presented. Although the joint model shows more precise and less biased estimators than its alternative counterpart, the TS approach results in much reduced computational times, and could thus be used for testing millions of SNPs at the genome-wide scale.

Keywords: joint modelling, survival analysis, longitudinal biomarker, genetic, diabetes, glycaemia

1. INTRODUCTION

With the increased availability of longitudinal and survival data in large cohorts, joint models have emerged as an appropriate approach to account for both types of data, especially when dealing with informative/non-informative dropouts which commonly occur in such cohorts. Joint models have been studied and overviewed in the literature (Wulfsohn and Tsiatis, 1997; Tsiatis and Davidian, 2004; Chen et al., 2011; Elashoff et al., 2016), and their implementation has been proposed in different softwares and platforms (Diggle and Kenward, 1994; Sun et al., 2007; Elashoff et al., 2008; Proust-Lima et al., 2009; Rizopoulos, 2010; Rizopoulos and Ghosh, 2011). Main applications of the joint model approach are: (i) to efficiently model the survival process with a time-varying covariate, accounting for missing data and measurement error; and (ii) to account for informative dropouts in the longitudinal data. To model the two processes 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. Unlike the CoxPH model, in which the time-varying covariate is assumed to be exogenous, i.e., not modified by the occurrence of an event (Kalbfleisch and Prentice, 2002), the joint modelling framework allows to account for an endogenous time-varying covariate. An example of an endogenous covariate is the fasting blood plasma glucose which is irremediably modified due to glucose lowering medication, once T2D is diagnosed.

Two approaches can be used for estimation and inference of the model parameters: a "naive" two-step (TS) method or a joint likelihood method (JM). In the first method, the random effects of the trajectory are estimated by an LME model, then included as a time-varying covariate in a CoxPH model and estimated using the partial likelihood approach (Therneau and Grambsch, 2000). The second method is based on a joint likelihood of the two stochastic components (longitudinal and survival) estimated at the same time. Comparison of these two approaches showed that the latter offers more consistent and efficient estimators than the former (Albert and Shih, 2010a,b). But JM can be challenging to compute, especially when it comes to achieving convergence during the Expectation-Maximisation (EM) step. Moreover, depending on the number of time points and/or the sample size, the overall computational time can substantially increase.

In this paper, we conducted a comprehensive simulation study to compare these two approaches, JM and TS, when jointly modelling the longitudinal and the survival components, under the case of univariate variable for the longitudinal trait. Our main goal is to show that the JM approach, when compared to TS, increases statistical power to detect an effect on either or both the longitudinal and the survival processes, while resulting in bias reduction in parameter estimation. We also showed that, while highly demanding computation and convergence issues might arise during JM computation, TS offers a good alternative to JM in greatly reducing computational time, especially when applied at the genome-scale level.

We also investigated and decomposed the computational time required by R package "JM" (Rizopoulos, 2010, 2017), and by the TS approach which combines R packages "survival" (Therneau, 2017) and "nlme" (Pinheiro et al., 2017).

Finally, we applied both approaches to a real dataset, the D.E.S.I.R. cohort (*Data from the Epidemiological Study on the Insulin Resistance syndrome*), which included 5,212 individuals with extensive phenotypic measurements recorded at four 3-yearly intervals, spanning a 9-year follow up. Individuals were genotyped using the Illumina Metabochip DNA array of nearly 200,000 SNPs (Voight et al., 2012). Relying on the conventional cross-sectional genome-wide association study design, the D.E.S.I.R. cohort was instrumental in identifying novel loci associated with prevalent type 2 diabetes (T2D) and fasting plasma glucose (FPG) level in normoglycaemic individuals (Sladek et al., 2007; Bouatia-Naji et al., 2008; Rung et al., 2009). We specifically focus on time-to-onset of T2D, in order to identify novel loci or to confirm published ones, which could simultaneously be associated with higher risk of developing T2D and/or increased FPG, consequently SNPs are rather analysed one at a time than as clusters. Our results were compared to the genetic variants reported in the literature (Vaxillaire et al., 2014; Welter et al., 2014) and to the meta-analyses published by large consortia, such as DIAGRAM (Morris et al., 2012) and MAGIC (Dupuis et al., 2010) consortia.

2. METHODS

2.1. Model Formulations

2.1.1. Joint Likelihood Model (JM)

The standard formulation of the joint model involves two components: a longitudinal component and a time-to-event component. Let n denote the sample size, and Y_{ij} the longitudinal measurements collected for each individual i at time points t_{ij}, $i = 1, \cdots, n, j = 1, \cdots, m_i$, where m_i is the number of measurements on individual i. The longitudinal component (i.e., measurements) typically consists of a (generalised) linear mixed effect (LME) model, whose within-subject correlation matrix is modelled using random-effect parameter vector $b_i = \begin{pmatrix} \theta_{0i} \\ \theta_{1i} \end{pmatrix}$.

Under the joint likelihood framework implemented in "JM" (Rizopoulos, 2010, 2017), within the class of "shared parameter models" (Rizopoulos, 2012; Elashoff et al., 2016), we define

$$Y_{ij} = X_{ij} + \epsilon_{ij} \tag{1}$$

where Y_{ij} is the observed value and X_{ij} is the true (unobserved) value of the longitudinal measurement at time t_{ij} for individual i. The quantity ϵ_{ij} is a random error term usually assumed to be normally distributed:

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{2}$$

The quantity X_{ij} is typically called the trajectory function, and is usually specified as a linear or quadratic function of time t_{ij}; for simplicity here, we assume linearity over time. We also define Z_i, a vector denoting the genotype of individual i, and W_i, a set of adjusting covariates:

$$Y_{ij} = X_{ij} + \epsilon_{ij} = \theta_{0i} + \theta_{1i} t_{ij} + \gamma Z_i + \delta W_i + \epsilon_{ij} \tag{3}$$

Again, without any loss of generality, we omit the term δW_i in the following. Random effects θ_{0i} (intercept) and θ_{1i} (slope) are assumed bivariate Normal: $\theta \sim \mathcal{N}_2(\mu, \Sigma)$, and independently distributed from ϵ_{ij}. The coefficient γ assesses the genotypic (additive) effect of variable Z_i on the trajectory function. To account for varying slopes, an interaction term between Z_i and time t_{ij} could be added into the trajectory function; for simplicity, this term was not considered in this 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. T_i denotes the event time for individual i, and C_i the right censoring time (end of the follow-up). Let Δ_i be the event indicator: $\Delta_i = 0$, if $T_i > C_i$, and $\Delta_i = 1$, if $T_i \leq C_i$. Under the Cox proportional hazards model, variable T_i is specified using the following equation:

$$\lambda_i(t) = \lambda_0(t) \exp(\beta X_i(t) + \alpha Z_i + \eta W_i) \tag{4}$$

where $\lambda_i(t)$ is the hazard function at time t for individual i and $\lambda_0(t)$ is the unspecified baseline hazard function, which we assume piecewise constant with two knots placed at intermediate time points in the follow-up. Coefficient α measures the effect of Z_i on the hazard function, while β measures the association between the trajectory function and the hazard function. In this formulation, we suppose that the subject-specific random effect parameters $b_i = \begin{pmatrix} \theta^*_{0i} \\ \theta^*_{1i} \end{pmatrix}$ included in the trajectory $X_i(t)$ could modify the hazard function, which implies that β is the parameter linking the longitudinal and survival components.

2.1.2. Two-Step Model (TS)

As an alternative to JM, and based on the work of Tsiatis et al. (1995), the two-step model estimates parameters of the joint model by first, estimating parameters of the trajectory function $X_i(t)$ in Equation (3), and second, by substituting this estimated trajectory, say $X^*_i(t)$, into Equation (4) before fitting the Cox survival model.

2.2. Simulations

Simulations were carried out to further examine the sensitivity of the JM estimations under several scenarios. Parameters were set based on values estimated (**Table 1**) from the strongest SNPs associated with T2D from the literature, i.e., rs17747324 in gene *TCF7L2* (T2D risk allele: C; $\alpha = 0.358$; $p = 8.5 \times 10^{-55}$ (Morris et al., 2012); FPG increasing allele: C; $\gamma = 0.025$; $p = 6.5 \times 10^{-08}$ Dupuis et al., 2010).

Longitudinal data were simulated according to Equation (3), while event times were generated from an exponential distribution for the CoxPH model (Austin, 2012), with $X_i(t)$ as a linear function.

$$\lambda_0(t) = \lambda \tag{5}$$

$$H_i(T_i) = \int_0^{T_i} \lambda \exp(\beta X_i(t) + \alpha Z_i) dt \tag{6}$$

$$F_i(T_i) = 1 - exp(-H_i(T_i)) = u \tag{7}$$

$$u \sim \mathcal{U}(0, 1) \tag{8}$$

$$T_i = \frac{1}{\beta\theta_{1i}} \log\left(1 - \frac{\beta\theta_{1i} \times \log(1-u)}{\lambda \exp(\beta\theta_{0i} + (\beta\gamma + \alpha)Z_i)}\right) \tag{9}$$

where λ was set to achieve the targeted incidence rate in the simulated dataset.

Datasets were simulated by varying the number of longitudinal measurements $m \in \{2; 3; 4; 5\}$, the number of individuals $n \in \{500; 1,000; 2,500; 5,000; 10,000\}$, the allele frequency $f \in \{0.05; 0.1; 0.25; 0.5\}$ and the incidence rate $d \in \{0.025; 0.05; 0.1\}$, thereby leading to 240 different scenarios. Each scenario was simulated 500 times.

The Root-Mean Square Error (RMSE)

$$\text{RMSE}(\hat{\phi}) = \sqrt{E((\hat{\phi} - \phi)^2)} \tag{10}$$

was used to assess precision of estimators $\phi = (\beta, \gamma, \alpha)$, when testing the association between Y_{ij}, and T_i, the effect of Z_i on Y_{ij} and the effect of Z_i on T_i, respectively. In addition, statistical power and type I error were also estimated. The computational burden of each approach (JM and TS) was also investigated as our goal is to implement these approaches at a genome-wide scale.

2.3. Computational Times

Based on our simulations, we calculated approximate computational times for four sample sizes with parameters as listed in **Table 1**, using a UNIX system with Intel® Xeon® CPU E7- 4870 @ 2.40 GHz (80 such CPUs available computing in parallel). **Table 2** shows computational times for one SNP, and for extrapolating the total computational time for 100,000 SNPs, which is the approximate number of SNPs on the Metabochip, after data cleaning and quality-control for common SNPs (minor allele frequency > 0.05).

To investigate further computational time issues, we profiled the execution of the main function "*jointmodel*" from the R package "JM," which implements the joint likelihood modelling approach as described in this paper. In the "JM" package, the linear mixed effect sub-model is handled by the function "*lme*" from the "nlme" package. One may argue that using a faster approach, e.g., as implemented in the R package "lme4", might decrease the computational time.

2.4. Real Data

SNP genotyping was performed with Metabochip DNA arrays (Voight et al., 2012) using Illumina HiScan technology and GenomeStudio software (Illumina, San Diego, USA) in 5,212 individuals from the French cohort D.E.S.I.R. (Balkau, 1996). These participants have been followed for 9 years, and extensive phenotypic data has been recorded at four different 3-yearly time interval during that follow-up. All participants signed informed consent, and the protocol was approved by the ethics committee of Kremlin Bicetre Hospital, Paris. Quality control was performed using PLINK 1.90 beta version (Chang et al., 2015; Purcell and Chang, 2015). SNPs with call rate of at least 95%, with no significant deviation from Hardy-Weinberg equilibrium at $p > 1 \times 10^{-5}$, and with minor allele frequency (MAF) over

TABLE 1 | Parameters and numerical values used for sensitivity analysis and simulations, based on results from rs17747324 within gene *TCF7L2* in the French cohort D.E.S.I.R.

Parameters	Values
Number of participants (n)	4,352
Number of measures (m)	4
Diabetes incidence rate (d)	0.0384
Minor allele frequency (f)	0.244
Random effects (θ)	$\sim \mathcal{N}_2\left(\begin{bmatrix} 4.55 \\ 0.0108 \end{bmatrix}, \begin{bmatrix} 0.143 & -0.00109 \\ -0.00109 & 6.8 \times 10^{-04} \end{bmatrix}\right)$
SNP effect on Y_{ij} (γ)	0.0229
SNP effect on T_i (α)	0.265
Association between Y_{ij} and T_i (β)	3.17
Error term (ϵ)	$\sim \mathcal{N}(0, 0.305^2)$

TABLE 2 | Approximate computational times using function "*system.time*" of R software.

	Joint model		Two-step model	
Sample size	**Mean (sd) per SNP in seconds**	**100 K SNPs in days**	**Mean (sd) per SNP in seconds**	**100 K SNPs in days**
500	51 (3.4)	59	0.71 (0.066)	0.82
2,500	100 (11)	120	3.1 (0.092)	3.6
5,000	180 (25)	210	6.3 (0.17)	7.3
10,000	340 (34)	400	9 (0.22)	10

System time was computed ten times per sample size (number of individuals). Extrapolation is displayed for 100,000 SNPs.

5% were kept for analysis, resulting in 101,305 SNPs. Due to missing phenotypes which did not allow to confirm T2D status, 232 individuals were removed. An additional 554 individuals were excluded due to individual call rate lower than 95%, leaving 4,426 individuals for analysis after these quality control steps (**Figure 1**).

Principal component analysis was performed using a combined dataset comprised of the 4,426 D.E.S.I.R. participants, along with participants from the publicly available 1,000 Genomes database (The 1000 Genomes Project Consortium, 2015). SNPs retained for analysis were restricted to those common to both sample sets. The first two components were sufficient to discriminate ethnic origin, which led to exclusion of 62 non Caucasians. A further 12 prevalent cases of T2D at baseline were also removed. As a result, the final dataset included 4,352 individuals, of whom 167 were diagnosed as T2D incident cases. Type 2 diabetes was defined using one of the following criteria: use of glucose lowering medication, and/or fasting plasma glucose [FPG] $\geq$ 7 mmol/L, and/or glycaeted haemoglobin A1c [$HbA1c$] $\geq$ 6.5% (48 mmol/mol).

Using the joint modelling approach implemented in the package "JM" (Rizopoulos, 2010, 2017) within the R software version 3.4.2 (R Core Team, 2017), all 101,305 SNPs were tested for joint association with FPG and T2D. Following the above joint modelling formulation, Y_{ij} denotes the observed values of FPG, Z_i represents the genotype of individual i at each SNP, with W_i being covariates such as age, sex and BMI (**Figure 2**). Finally, T_i gives the time at which an individual is diagnosed with T2D.

In the joint modelling framework, the trajectory of FPG could be viewed as a dropout process, because all FPG values are flagged as missing after T2D diagnosis. In effect, individuals receiving a diabetes diagnostic are immediately placed under treatment to lower and regulate their blood glucose level. Therefore, FPG must be considered as an endogenous covariate, because the dropout process is not independent from the measured glucose values prior to T2D diagnosis.

3. RESULTS

3.1. Comparison of Estimation Accuracy

Due to the complexity of the estimating algorithm within JM, convergence could not be obtained (4.53% of convergence issues on average per scenario, with a standard deviation of 5.81%) for the whole set of 500 simulations (i.e., algorithm "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; Rizopoulos, 2010, 2017).

RMSE for parameter γ (**Figure 3**) showed similar performance for JM and TS. RMSE for parameter β (**Figure 4**) and for parameter α (**Figure 5**) were smaller within the joint modelling framework (either JM or TS) than in the more classical CoxPH model with time-varying fasting plasma glucose. While the RMSE for β remains the same in the CoxPH model across all scenarios, under JM or TS it decreased whenever the sample size, the incidence rate or the allele frequency increases. Differences

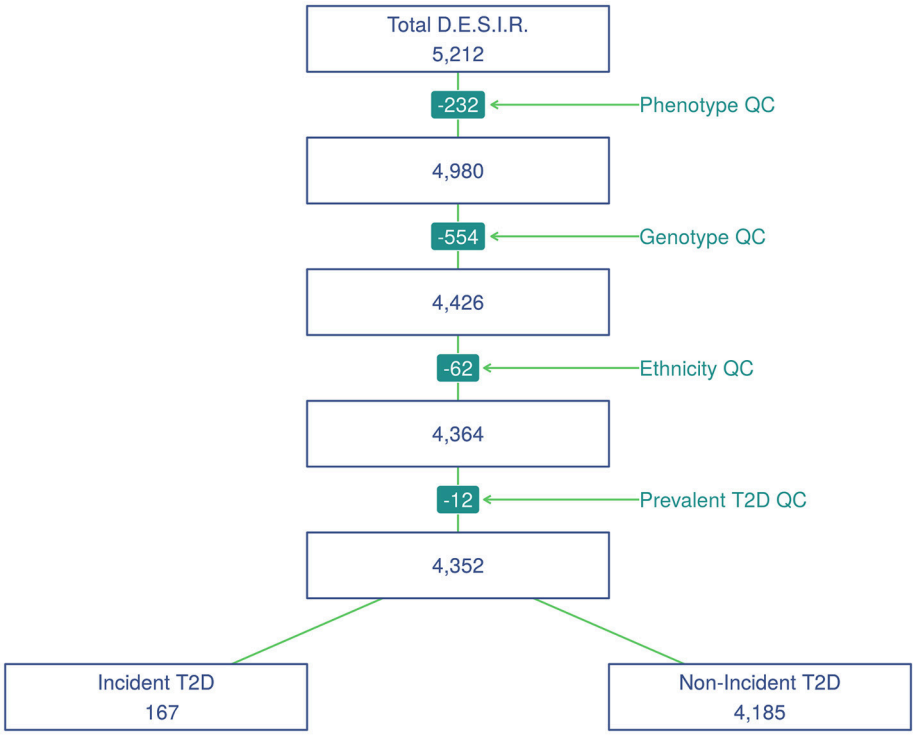

FIGURE 1 | Study flowchart of participants from the French cohort D.E.S.I.R.

in RMSE for parameter α were smaller than for parameter β. Both TS and CoxPH with time-dependent covariate performed similarly, probably because partial likelihood inferences were used in these two approaches. JM estimations, for β and γ, were less biased in almost all scenarios when the sample size was >2,500. The larger bias observed within the extended CoxPH model, especially for β, could be explained by the mischaracterisation of the measurement error in the longitudinal trait.

Overall, our simulations showed that JM is less biased than when separate approaches are used to model the effect of Z_i on the longitudinal process Y_i, and on the time-to-event T_i. While separate approaches performed well for parameters γ and α, the bias for parameter β was the highest across all scenarios.

3.2. Computational Time

Computational times are reported in **Table 2**. The time required to complete JM or TS algorithms increased linearly with sample size in our simulations. However, these times are very optimistic since our simulations did not include any covariate or more complex random parameters. The main issue appears within the "*jointmodel*" function which took over 95% of the global computation time. After examination of the call tree diagram, we observe that the more time-consuming task within the "*jointmodel*" function happens during the optimisation of the EM algorithm (described in Rizopoulos, 2012, Appendix B), despite the use of a calculation trick (i.e., adaptive Gauss-Hermite quadrature for numerical integration).

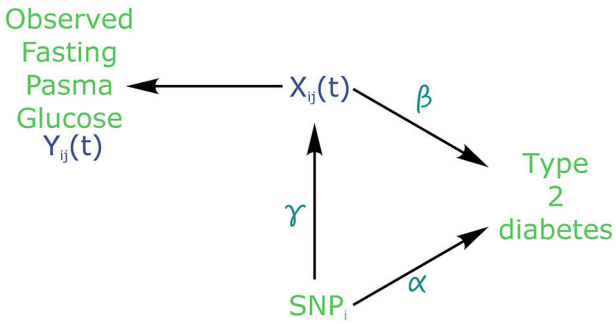

FIGURE 2 | Causal diagram for joint modelling applied to fasting plasma glucose (FPG) and type 2 diabetes (T2D) (adapted from Ibrahim et al., 2010). SNP: Single Nucleotide Polymorphism. X_{ij} is the true (unobserved) FPG trajectory. β measures the association between FPG trajectory and T2D incidence. $\exp(\alpha)$ measures the hazard ratio for T2D. γ measures the association between a SNP_i and FPG trajectory.

3.3. Application in Real Data

Applying the R package "JM" to our D.E.S.I.R. cleaned dataset, 265 SNPs (**Figure 6**) were associated (with $p < 0.05$) with FPG and T2D events through their respective parameters γ and α. Amongst these 265 SNPs (163 unique genes), we identified 17 genes (**Table 3**) which had already been reported to be associated with FPG and/or T2D risk. Parameter β was highly significant (below the genome-wide threshold of 5×10^{-8}) for all these SNPs,

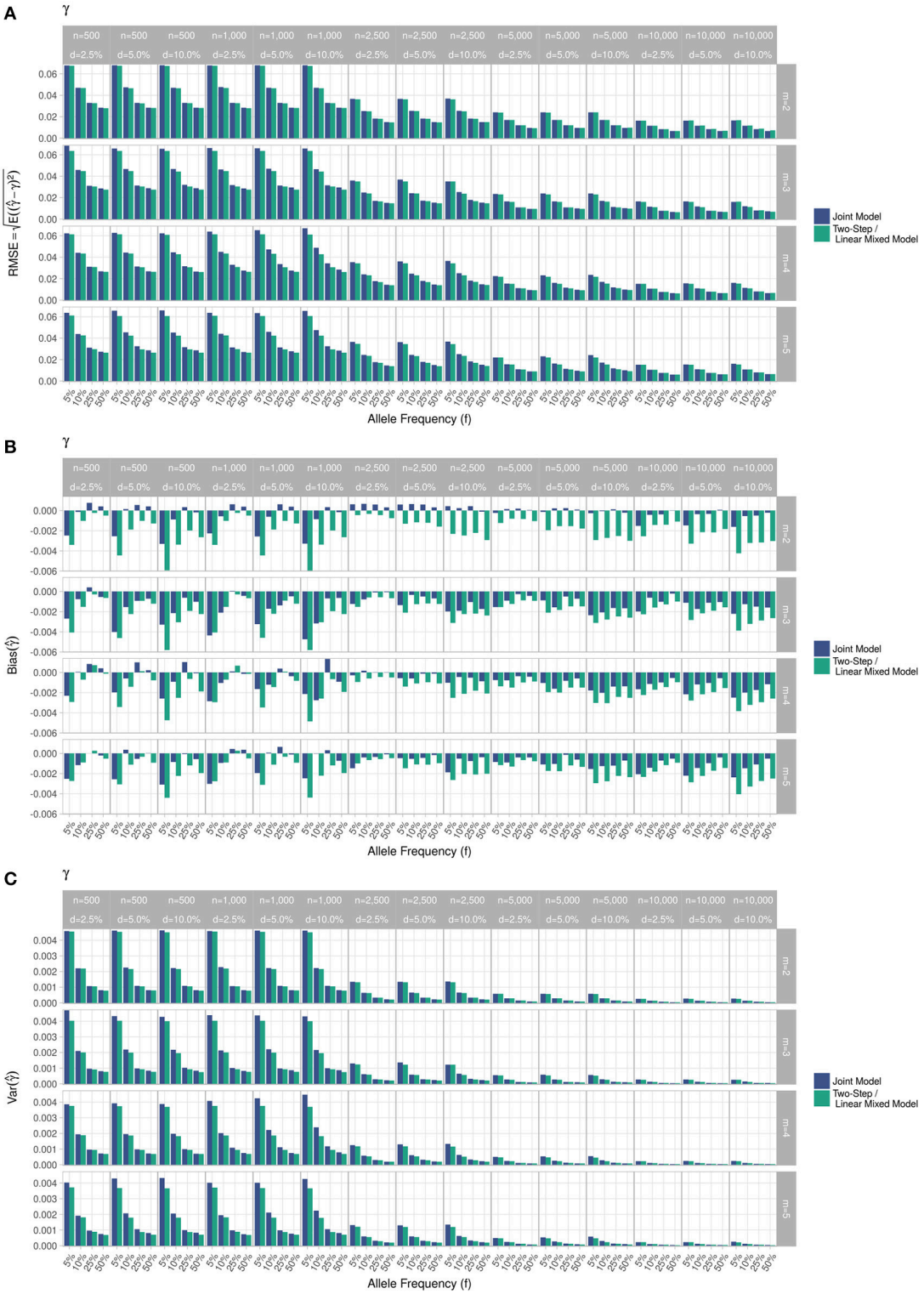

FIGURE 3 | Simulation study for accuracy of estimator $\hat{\gamma}$ provided by the joint model ("JM" package) and by the two-step linear mixed effect model ("nlme" package). **(A)** Displays RMSE, **(B)** Displays bias and **(C)** Displays variance. *m*, number of measures; *n*, number of individuals; *d*, diabetes incidence rate.

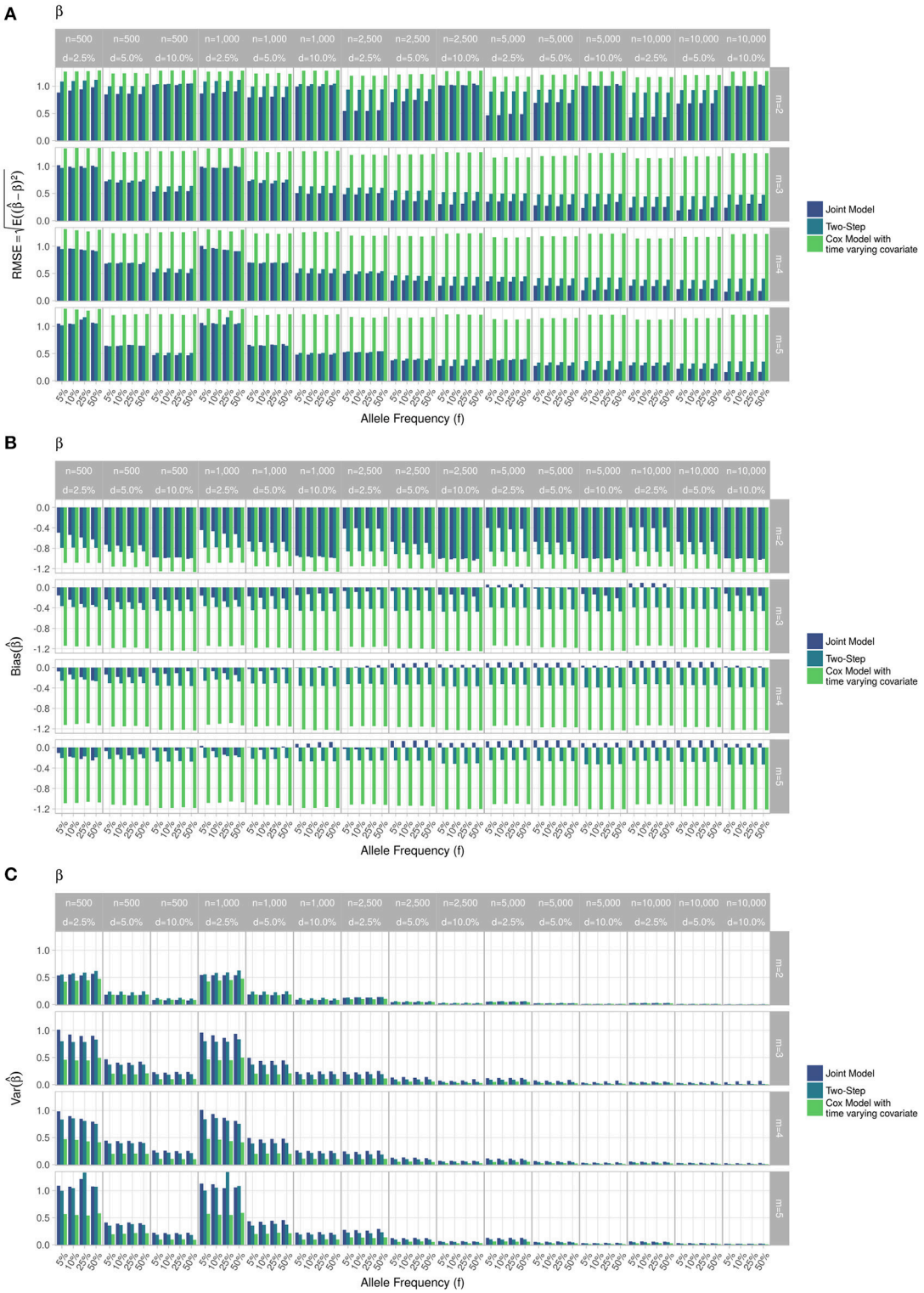

FIGURE 4 | Simulation study for accuracy of estimator $\hat{\beta}$ provided by the joint model ("JM" package), by the two-step linear mixed effect model ("nlme" package) and by the Cox model with time-varying covariate. **(A)** Displays RMSE, **(B)** Displays bias and **(C)** Displays variance. *m*, number of measures; *n*, number of individuals; *d*, diabetes incidence rate.

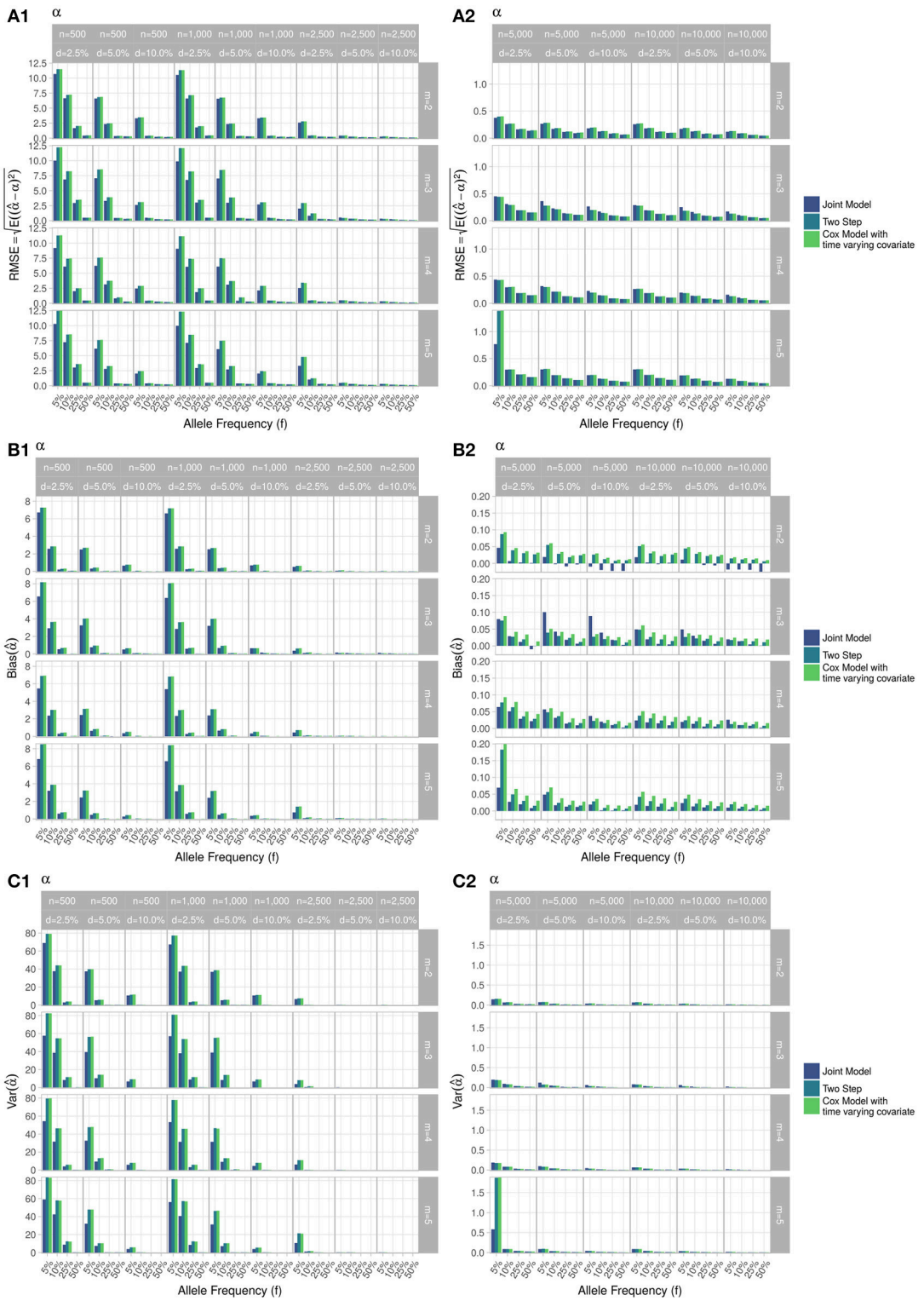

FIGURE 5 | Simulation study for accuracy of estimator $\hat{\alpha}$ provided by the joint model ("JM" package), by the two-step linear mixed effect model ("nlme" package) and by the Cox model with time-varying covariate. *m*, number of measures; *n*, number of individuals; *d*, diabetes incidence rate. **(A)** Displays RMSE, **(B)** Displays bias and **(C)** Displays variance. (1) Displays results for $n = 500$ to $n = 2,500$. (2) Displays results for $n = 5,000$ to $n = 10,000$.

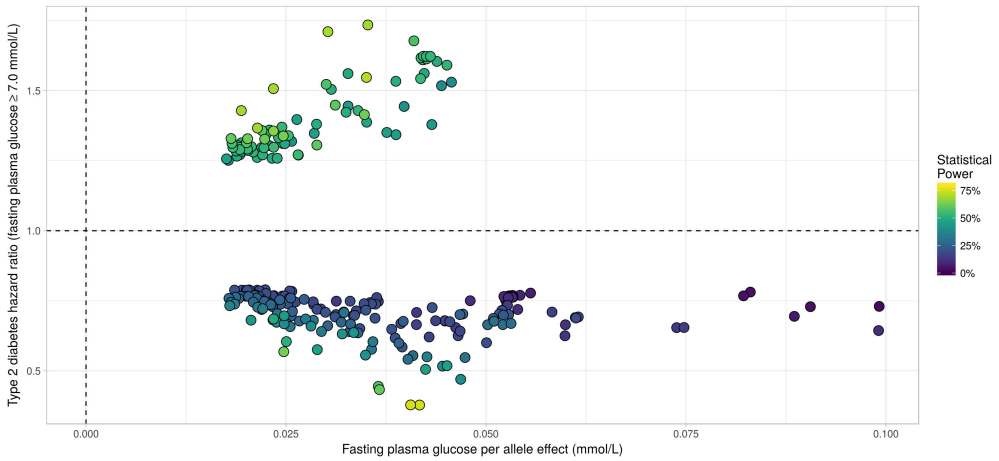

FIGURE 6 | Results from statistical analysis using "JM" (Rizopoulos, 2010, 2017), using fasting plasma glucose increasing allele as effect allele. Estimated effects of γ are displayed on the x-axis, with corresponding estimated hazard ratio $\exp(\alpha)$ on the y-axis. Statistical power reported is the theoretical (retrospective) power to detect a genetic joint effect $\beta\gamma + \alpha$ based on estimated model parameters (Chen et al., 2011).

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

SNP (gene)	α (*p*-value)	γ (*p*-value)	β (*p*-value)	Power($\beta\gamma + \alpha$)%	Risk allele frequency
rs6945660_G (*ETV1*)	0.550 (3.7×10^{-02})	0.035 (2.5×10^{-02})	3.480 (9.6×10^{-45})	69.7	0.91
rs1942873_C (*MC4R*)	0.410 (1.3×10^{-02})	0.023 (3.7×10^{-02})	3.140 (1.9×10^{-41})	69.6	0.81
rs55899248_G (*TCF7L2*)	0.292 (2.7×10^{-02})	0.025 (1.7×10^{-02})	3.490 (1.7×10^{-44})	55.3	0.24
rs17301514_A (*ST6GAL1*)	−0.657 (4.4×10^{-03})	0.045 (3.4×10^{-03})	3.650 (2.9×10^{-45})	45.8	0.09
rs833425_C (*PTPRD*)	0.321 (5.0×10^{-02})	0.043 (4.2×10^{-03})	3.510 (1.3×10^{-43})	44.2	0.10
rs7072870_A (*C10orf35*)	−0.404 (7.5×10^{-03})	0.025 (2.2×10^{-02})	3.580 (1.7×10^{-45})	39.6	0.22
rs61871514_A (*KCNQ1*)	0.425 (4.7×10^{-02})	0.046 (2.0×10^{-02})	3.180 (8.5×10^{-42})	39.4	0.06
rs9883865_A (*ADAMTS9*)	−0.598 (7.5×10^{-04})	0.043 (1.2×10^{-02})	3.200 (5.9×10^{-42})	34.9	0.92
rs114508985_C (*HLA*)	−0.294 (2.1×10^{-02})	0.021 (3.0×10^{-02})	3.220 (8.2×10^{-43})	27.1	0.31
rs10814856_T (*GLIS3*)	−0.265 (4.0×10^{-02})	0.025 (1.5×10^{-02})	3.200 (1.5×10^{-42})	18.5	0.73
rs73025532_C (*SLC22A1*)	−0.377 (4.8×10^{-02})	0.032 (3.6×10^{-02})	3.580 (1.3×10^{-45})	17.3	0.90
rs11769484_C (*JAZF1*)	−0.254 (4.8×10^{-02})	0.022 (3.6×10^{-02})	3.210 (2.1×10^{-42})	16.9	0.77
rs6450176_G (*ARL15*)	−0.291 (1.8×10^{-02})	0.036 (3.0×10^{-04})	3.540 (2.2×10^{-45})	15.2	0.73
rs4712580_C (*CDKAL1*)	−0.289 (4.2×10^{-02})	0.031 (7.4×10^{-03})	3.570 (1.2×10^{-45})	14.0	0.82
rs10830963_G (*MTNR1B*)	−0.440 (9.4×10^{-04})	0.099 (1.3×10^{-23})	3.250 (3.6×10^{-42})	10.2	0.29
rs853787_T (*ABCB11*)	−0.247 (4.3×10^{-02})	0.083 (9.3×10^{-19})	3.210 (1.7×10^{-42})	03.3	0.65
rs560887_C (*G6PC2*)	−0.315 (1.2×10^{-02})	0.099 (9.6×10^{-25})	3.210 (1.3×10^{-42})	02.6	0.70

which was expected considering that β estimates the association between FPG trajectory and T2D risk.

In **Figure 7**, we specifically focused on parameters γ and α. After Bonferroni correction (nominal *p*-value $\simeq 5 \times 10^{-7}$), no genetic variants showed a highly significant association with both parameters γ and α simultaneously; only SNPs in the following genes (or within a 100 kb window) remained significant when testing for γ: *G6PC2*/*ABCB11*, *GCK*/*YKT6*, *GCKR*, and *MTNR1B*, with per-allele increasing effect varying on FPG from 0.100 to 0.047 mmol/L (data not shown). Zooming in on simultaneous associations with the longitudinal and survival processes revealed well known genes, such as *TCF7L2*, which has been shown in many meta-analyses to be associated with elevated FPG and an increased risk of T2D (**Table 4**). *MTNR1B* was also found to be associated (34 SNPs within 30 kb) with estimated $\hat{\alpha} = -0.44$ ($p = 9.37 \times 10^{-04}$) and $\hat{\gamma} = 0.099$ ($p = 1.33 \times 10^{-23}$) for SNP rs10830963, the SNP usually reported in the literature.

To better compare JM and TS, we repeated the analysis on the whole dataset using TS. As shown in **Figure 8**, *p*-values differ, especially when testing parameter α; however for tests on parameter γ, approximations were quite close to the *p*-values provided via the joint likelihood framework.

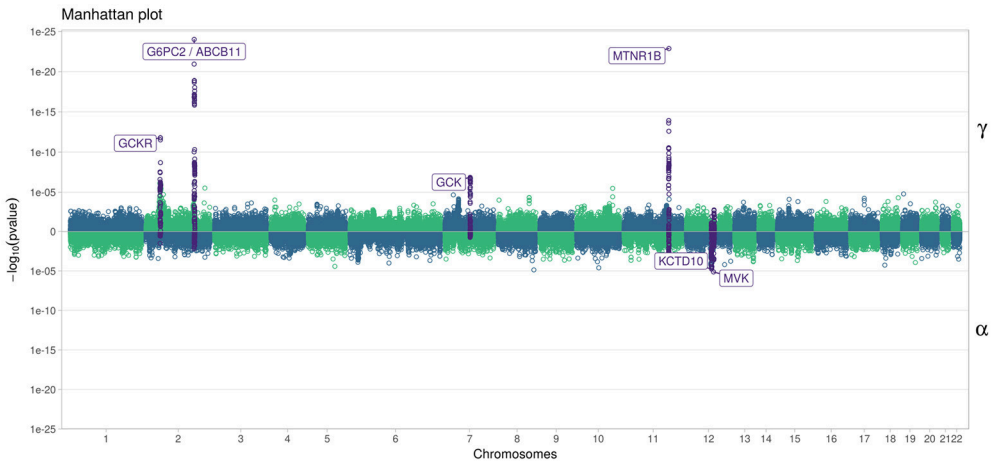

FIGURE 7 | Manhattan plot for estimated effects of γ and α using "JM" (Rizopoulos, 2010, 2017). Results are presented for the cleaned set of 101,305 SNPs.

TABLE 4 | Effect sizes on FPG and T2D risk estimated using JM.

	α (*p*-value)		γ (*p*-value)		β (*p*-value)
SNP (gene)	**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 (9.4×10^{-04})	0.104 (7.3×10^{-07})	0.0991 (1.3×10^{-23})	0.079 (1.3×10^{-68})	3.25 (3.6×10^{-42})
rs17747324_C (*TCF7L2*)	0.265 (4.1×10^{-02})	0.358 (8.5×10^{-55})	0.0229 (3.0×10^{-02})	0.025 (6.5×10^{-08})	3.17 (8.9×10^{-42})

Comparison is shown with effect sizes as reported by consortia meta-analyses in genes MTNR1B *and* TCF7L2.

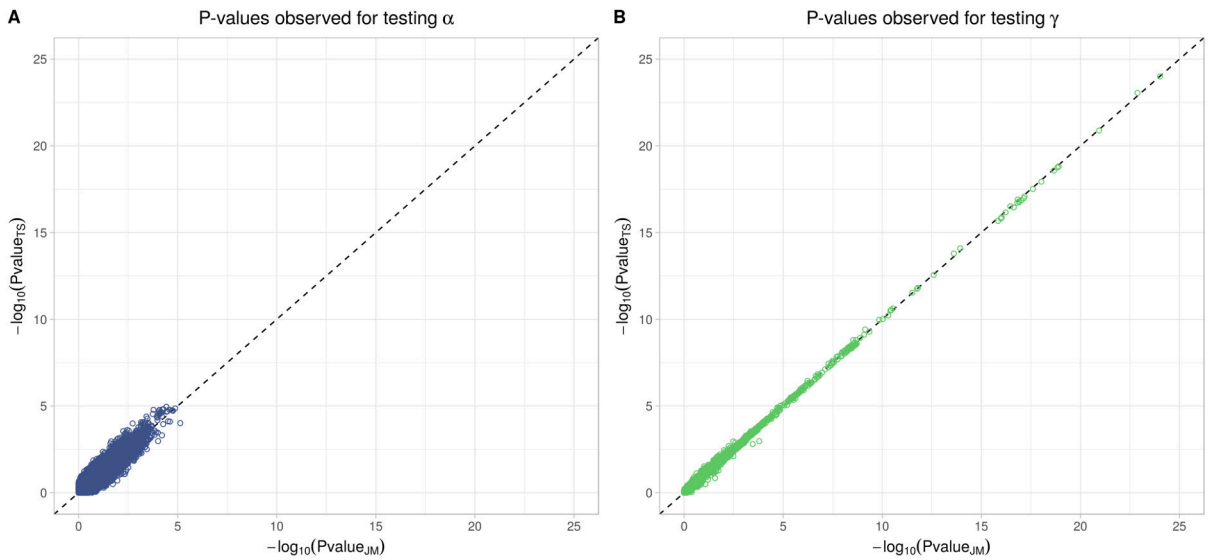

FIGURE 8 | Testing for α (SNP effect on onset of T2D) and γ (SNP effect on the trajectory of FPG) using Two-Step approach compared to Joint Model approach. On the x-axis, $-\log_{10}(p\text{-value})$ from the Joint Model and on the y-axis the corresponding $-\log_{10}(p\text{-value})$ from the approximate Two-Step approach. **(A)** Displays $-\log_{10}(p\text{-value})$ for testing α parameter from Equation (4). **(B)** Displays $-\log_{10}(p\text{-value})$ for testing γ parameter from Equation (3).

4. DISCUSSION

With the ever-increasing availability of genomic data generated by genotyping arrays and next generation sequencing, the need to develop and implement efficient models is important to ensure that statistical analysis will be achieved in a reasonable timeframe. In this paper, we proposed a comparison of two approaches, namely the joint model (JM) and the two-step model

(TS), to estimate parameters accounting simultaneously for a genetic effect on both longitudinal and survival processes without discarding missing values or dropouts commonly generated by the longitudinal measurement process. In our real data application, FPG serves as the longitudinal process, whereas T2D diagnosis generates survival times of interest, both stochastic phenomenon being linked by the fact that a fixed threshold for FPG defines T2D onset (currently, $[FPG] \geq 7$ mmol/L), along with glucose lowering medication use. Through simulations over different scenarios, we showed that joint models are less biased than classical separate approaches. Hopefully, joint models could provide more insight regarding the event of interest, and could assess the potential impact of a genetic marker on incident T2D better than simpler models.

By looking at statistical measures of accuracy such as RMSE for our model estimators, and by estimating the computational time required by the available R implementations of joint models, our study showed that the use of an approximate method at a genome-wide scale, such as TS, might represent a good compromise between accuracy and computational time. TS could be used to overcome the computational burden of current joint likelihood methods by exploiting available R packages performing the two steps, LME and CoxPH, and could help filter out SNPs with low or undetectable associations during a first preliminary scan. However, depending on the parameters of the dataset (sample size, incidence rate, number of measures), a joint likelihood method always has to be preferred over TS when one wants to obtain accurate estimation of parameters γ and α, which describe the SNP effect on the trajectory of FPG and on the time-to-onset of T2D, resp. Although we computed the theoretical statistical power to detect a joint genetic effect $\beta\gamma+\alpha$ (Chen et al., 2011), we did not test this effect at the genome-wide scale due to its computational burden. In this paper, we used the closed-form expression from Chen et al. (2011) to evaluate retrospectively the probability to obtain the same significant results in a similarly designed study to D.E.S.I.R. This closed-form expression could be use to design a new study (e.g., compute the number of samples) which aims at identifying a joint effect. The joint SNP effect can be tested using a likelihood ratio test comparing the full joint model, i.e., with a SNP effect included in both sub-models, to the joint model without a SNP effect in the survival sub-model, as implemented in the package "JM" (Rizopoulos, 2010, 2017).

To fully characterise JM approach, further study needs to be performed, such as missing values distribution according to the usual hypotheses (i.e., Missing At Random, Missing Completely At Random and Missing Not At Random). In this paper, we did not study in our simulations the rates of change effect of the SNPs (i.e., interaction term $SNP \times TIME$) which might also be of interest in the study of a disease such as T2D.

Finally, we would like to reemphasize that using parallel and grid computing approaches will help reduce the global computational time when applied at a genome-wide scale (i.e., with millions of SNPs).

In our real data application, rs17747324 showed consistent results with the DIAGRAM and MAGIC (for FPG) consortia for both α and γ (**Table 4**), but rs10830963 showed an opposite effect on T2D risk compared to the effect reported in the DIAGRAM consortium ($\hat{\alpha} = 0.104$, $p = 7.3 \times 10^{-07}$). Results observed for *MTNR1B* (rs10830963) in the French cohort D.E.S.I.R., albeit inconsistent with previous studies, may uncover some interesting peculiarities pertaining to T2D incident cases in this population. In the literature, SNPs in *MTNR1B* were reported as being associated with higher FPG and T2D risk, but meta-analyses were performed on populations with different genetic backgrounds, and the two traits have never been jointly co-analysed. However, we realize that *MTNR1B* associations identified in our study need to be confirmed and replicated in other cohorts, as they might be cohort-specific. Finally, a major limitation in our study comes from the low number of incident T2D cases in the D.E.S.I.R. cohort (167 incident T2D cases in 4,352 individuals followed over 9 years), resulting in (retrospective) power, no higher than 70%, as shown in **Figure 6** and **Table 3**.

DATA AVAILABILITY

The datasets for this manuscript are not publicly available due to consideration of intellectual property, ongoing active collaborations and to continuing analyses by the study investigators.

Requests to access the datasets should be directed to PF (p.froguel@imperial.ac.uk).

AUTHOR CONTRIBUTIONS

GR and MC: contributed to the study conception and design; BB, PF, and RR: made the genomic and phenotypic data available; MC: conducted the data and statistical analyses; GR and MC: interpreted the data and results; MC: drafted the first version of the manuscript; BB, GR, PF, and RR: edited and provided critical revisions to the manuscript. All authors contributed to manuscript revision, read and approved the submitted version.

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.

The D.E.S.I.R. study has been funded by INSERM contracts with Caisse nationale de l'assurance Maladie des Travailleurs Salariés (CNAMTS), Lilly, Novartis Pharma, and Sanofi-Aventis; INSERM (Réseaux en Santé Publique, Interactions entre les déterminants de la santé, Cohortes Santé TGIR 2008); the Association Diabète Risque Vasculaire; the Fédération Française de Cardiologie; La Fondation de France; Association de Langue Française pour l'Etude du Diabète et des Maladies Métaboliques (ALFEDIAM)/Société Francophone de Diabétologie (SFD); l'Office National Interprofessionnel des Vins (ONIVINS); Ardix Medical; Bayer Diagnostics; Becton Dickinson; Cardionics; Merck Santé; Novo Nordisk; Pierre Fabre; Roche; Topcon.

The D.E.S.I.R. Study Group. INSERM U1018: B. Balkau, P. Ducimetière, E. Eschwège; INSERM U367: F. Alhenc-Gelas; CHU D'Angers: Y Gallois, A. Girault; Centre de Recherche des Cordeliers, INSERM U1138, Bichat Hospital: F. Fumeron, M. Marre, R. Roussel; CHU de Rennes: F. Bonnet; CNRS UMR8090, Lille: A. Bonnefond, S. Cauchi, P. Froguel; Centres d'Examens de Santé: Alençon, Angers, Blois, Caen, Chateauroux, Chartres, Cholet, Le Mans, Orléans, Tours; Institut de Recherche Médecine Générale: J. Cogneau; General practitioners of the region; Institut inter-régional pour la Santé: C. Born, E. Caces, M. Cailleau, O Lantieri, J.G. Morea.

REFERENCES

Albert, P. S., and Shih, J. H. (2010a). An approach for jointly modeling multivariate longitudinal measurements and discrete time-to-event data. *Ann. Appl. Stat.* 4, 1517–1532. doi: 10.1214/10-AOAS339

Albert, P. S., and Shih, J. H. (2010b). On estimating the relationship between longitudinal measurements and time-to-event data using a simple two-stage procedure. *Biometrics* 66, 983–987. doi: 10.1111/j.1541-0420.2009.01324_1.x

Austin, P. C. (2012). Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Stat. Med.* 31, 3946–3958. doi: 10.1002/sim.5452

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. *Rev. D'épidémiologie Et De Santé Publique*, 44, 373–375.

Bouatia-Naji, N., Rocheleau, G., Van Lommel, L., Lemaire, K., Schuit, F., Cavalcanti-Proença, C., et al. (2008). A polymorphism within the G6PC2 gene is associated with fasting plasma glucose levels. *Science*, 320, 1085–1088. doi: 10.1126/science.1156849

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4:7. doi: 10.1186/s13742-015-0047-8

Chen, L. M., Ibrahim, J. G., and Chu, H. (2011). Sample size and power determination in joint modeling of longitudinal and survival data. *Stat. Med.* 30, 2295–2309. doi: 10.1002/sim.4263

Diggle, P., and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *J. R. Stat. Soc. Ser. C*, 43, 49–93.

Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A. U., et al. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.* 42, 105–116. doi: 10.1038/ng.520

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

Elashoff, R. M., Li, G., and Li, N. (2008). A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics* 64, 762–771. doi: 10.1111/j.1541-0420.2007.00952.x

Ibrahim, J. G., Chu, H., and Chen, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *J. Clin. Oncol.* 28, 2796. doi: 10.1200/JCO.2009.25.0654

Kalbfleisch, J. D., and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons, Inc.

Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segrè, A. V., Steinthorsdottir, V., et al. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* 44, 981–990. doi: 10.1038/ng.2383

Pinheiro, J., Bates, D., and R-core (2017). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-131.

Proust-Lima, C., Joly, P., Dartigues, J.-F., and Jacqmin-Gadda, H. (2009). Joint modelling of multivariate longitudinal outcomes and a time-to-event: a nonlinear latent class approach. *Comput. Stat. Data Anal.* 53, 1142–1154. doi: 10.1016/j.csda.2008.10.017

Purcell, S., and Chang, C. (2015). PLINK v1.90b3.36. Available online at: https://www.cog-genomics.org/plink/1.9/general_usage#cite

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rizopoulos, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. *J. Stat. Softw.* 35, 1–33. doi: 10.18637/jss.v035.i09

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. London, UK: CRC Press.

Rizopoulos, D. (2017). *JM: Joint Modeling of Longitudinal and Survival Data*. R package version 1.4-7.

Rizopoulos, D., and Ghosh, P. (2011). A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Stat. Med.* 30, 1366–1380. doi: 10.1002/sim.4205

Rung, J., Cauchi, S., Albrechtsen, A., Shen, L., Rocheleau, G., Cavalcanti-Proença, C., et al. (2009). Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. *Nat. Genet.* 41, 1110–1115. doi: 10.1038/ng.443

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., et al. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 445, 881–885. doi: 10.1038/nature05616

Sun, J., Sun, L., and Liu, D. (2007). Regression analysis of longitudinal data in the presence of informative observation and censoring times. *J. Am. Stat. Assoc.* 102, 1397–1406. doi: 10.1198/016214507000000851

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Therneau, T. M. (2017). *survival: Survival Analysis*. R package version 2.41-3.

Therneau, T. M., and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model. Statistics for Biology and Health*. New York, NY: Springer New York.

Tsiatis, A. A., and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Stat. Sin.* 14, 809–834. Available online at: http://www.jstor.org/stable/10.2307/24307417

Tsiatis, A. A., DeGruttola, V., and 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. *J. Am. Stat. Assoc.* 90, 27–37. doi: 10.1080/01621459.1995.10476485

Vaxillaire, M., Yengo, L., Lobbens, S., Rocheleau, G., Eury, E., Lantieri, O., et al. (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, 1601–1610. doi: 10.1007/s00125-014-3277-x

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., et al. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* 8:e1002793. doi: 10.1371/journal.pgen.1002793

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, 42, D1001–D1006. doi: 10.1093/nar/gkt1229

Wulfsohn, M. S., and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* 53, 330. doi: 10.2307/2533118

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.