Published in final edited form as:

Genet Epidemiol. 2013 July ; 37(5): 465–477. doi:10.1002/gepi.21719.

Genetic association analysis and meta-analysis of imputed SNPs in longitudinal studies

Isaac Subirana^{1,2,3} and Juan R González^{*,4,5,1}

¹ CIBER Epidemiology and Public Health (CIBERESP), Spain

² Cardiovascular Epidemiology & Genetics group, Inflammatory and Cardiovascular Disease Programme, IMIM, Parc de Salut Mar, Spain

³ Department of Statistics, University of Barcelona, Spain

⁴ Center for Research in Environmental Epidemiology (CREAL), Spain

⁵ Department of Mathematics, Universitat Autònoma de Barcelona (UAB), Spain

Abstract

In this paper we propose a new method to analyze time-to-event data in longitudinal genetic studies. This method address the fundamental problem of incorporating uncertainty when analyzing survival data and imputed single nucleotide polymorphisms (SNPs) from genomewide association studies (GWAS). Our method incorporates uncertainty in the likelihood function, the opposite of existing methods that incorporate the uncertainty in the design matrix. Through simulation studies and real data analyses, we show that our proposed method is unbiased and provides powerful results. We also show how combining results from different GWAS (meta-analysis) may lead to wrong results when effects are not estimated using our approach. The model is implemented in an R package that is designed to analyze uncertainty not only arising from imputed SNPs, but also from copy number variants (CNVs).

Keywords

imputed SNP; longitudinal studies; GWAS; genetic association

1 Introduction

Genome wide association studies (GWAS) have found a large number of SNPs associated with several complex diseases such as Alzheimer [Rovelet-Lecrux et al., 2006], cancer [Xing et al., 2011], or cardiovascular diseases [Myocardial Infarction Genetics Consortium, 2009], among others. Some of these GWAS are based on cohort studies such as Framingham, [Granada et al., 2011], INMA [Morales et al., 2011], and CHARGE [Fornage et al., 2011]. Therefore, when longitudinal data is available researchers are interested not only in discovering genes related to susceptibility but also in determining genes related to

*Correspondence to: Dr. Juan R. González, Center for research in environmental epidemiology (CREAL) (room 188), Barcelona Biomedical Research Park (PRBB), Plaza Charles Darwin s/n, Barcelona 08003, Spain. jrgonzalez@imim.es.

progression [González et al., 2011, Frey et al., 2006b], survival [Frey et al., 2006a, Heist et al., 2007], time to relapse or recurrence [Wojnar et al., 2009, Xing et al., 2011], or response to treatment [Ueno et al., 2008, Alakus et al., 2009]. Conversely to case-control studies, some of these cohort studies have found genetic variants with very high effects. For example, Frey et al. [2005] discovered that patients diagnosed with colorectal cancer who carried CC genotype of T393C polymorphism were at highest risk for death (hazard ratio, 12.1) compared with TT genotypes. On the other hand, González et al. [2011] estimated that the probability of multiple sclerosis progression for those individuals who carried two copies of the CD24 gene was 3.4 times higher than for those who carried the frequent allele. Probably these high hazard ratios (HR) are overestimated and have to be replicated. Nonetheless, these HR are normally larger than the odds ratio found in case-control studies that lay between 1.1 and 1.5.

Statistical methods to analyze time-to-event data in genetic studies are based on standard survival analysis techniques. The hazard risk of observing the event of interest for a given SNP is normally estimated using either a parametric (Weibull, log-normal, ...) [Park et al., 2010, Del Greco M et al., 2011] or a semi-parametric model (Cox regression) [Takeuchi et al., 2009, Morgan et al., 2011]. To improve power, current GWAS combine information across cohorts. Each study may use different platforms to get genotype information, resulting in different numbers of genotyped SNPs. Therefore, in order to analyze the same number of SNPs for all studies, imputed data are normally obtained from each study. These SNPs are normally estimated with uncertainty; hence, standard statistical techniques cannot be used to assess association between the SNPs and the outcome. Imputing methods provide for each SNP a vector with 3 probabilities, corresponding to each genotype, that indicates the probability of having zero, one, or two copies of the rare allele. The simplest method to estimate the HR for an imputed SNP is to consider the genotype with the highest probability as if it were observed without uncertainty (best guess). Aulchenko et al. [2010] pointed out that this approach, which does not incorporate the uncertainty in the model, is biased and underpowered.

Statistical models to analyze survival data with imputed SNPs should incorporate uncertainty. To our knowledge, the only existing method to do so in cohort studies is the one proposed by Aulchenko et al. [2010], which incorporates the probabilities of each genotype of the imputed SNP in the design matrix. SNPTEST software [Marchini et al., 2007] also implements methods to incorporate uncertainty, but only for case-control studies and does not accommodate time-to-event responses. Parametric or semi-parametric methods can then be used to estimate the HR. The authors argue that this procedure gives unbiased results. However, they did not provide any justification, maybe due to the fact that this method is being widely used in most of the recent papers that use logistic regression models [Kooner et al., 2011, Trompet et al., 2011]. Theoretically, this approach should work in linear models (because it is not biased) but its use cannot be correct when logistic regression is used for case-control studies or with any other non-linear regression, as in survival data models. In order to overcome this difficulty, we propose an alternative approach based on a model that incorporates uncertainty in the likelihood function. The main aim of this paper is to investigate whether a best-guess approach, to incorporate the uncertainty as covariates DOSAGE, and our proposed method can be used to analyze follow-up cohort genetic studies

when analyzing imputed SNPs. We are also interested in assessing the impact of the proposed method (mainly in terms of bias of the pooled HR) when performing meta-analysis by combining HR estimated for imputed SNPs in different studies.

The remainder of the paper is organized as follows: First section, 'Methods', briefly describes the main existing programs to get imputed SNPs. We also provide the likelihood function for the 3 studied methods: best guess, uncertainty as covariates, and our new proposed method. The second section, 'Simulation Study', describes the simulation studies carried out to check whether each method is biased when estimating HRs and to assess power under different scenarios (allele frequency, effect and uncertainty degree). Also we evaluate the impact of combining HR of imputed SNPs in meta-analyses. Simulations are based on imputed SNPs from the Framingham study cohort (<http://www.framingham.com/heart/>) to mimic real situations. The third section, 'Application to real data', presents the results from a real data set using imputed SNPs as well as phenotypes from the Framingham study cohort. Finally, we present our conclusions.

2 Methods

2.1 Imputation programs

Programs to impute SNPs include PLINK, IMPUTE, MACH and fast-PHASE. They have been used in different studies, mainly in case-control settings [Nair et al., 2009, Meyre et al., 2009]. Each program uses its own criteria, data elements, and algorithms to perform the imputation of non-genotyped SNPs. A comprehensive comparison of all these methods can be found in Biernacka et al. [2009]. All of them provide a quality score that gives an idea about how well each genotype is imputed. Most of them return a vector with 3 probabilities for each individual and for each imputed SNP, i.e., the probability that a particular individual has the genotype 'aa', 'aA' or 'AA' (0, 1, or 2 risk alleles, respectively). There also exist other imputation methods that only return the most likely genotype for each individual. This result is less informative than providing the 3 probabilities because uncertainty is not taken into account. It is obvious that, for instance, having a 99% probability of presenting 'AA' genotype is not the same risk as having a 51% probability. In both situations 'AA' will be considered as the most likely genotype but in the second case we are not so sure that it is true, i.e., there is much more uncertainty when assigning the 'AA' genotype. Therefore, we strongly recommend discarding methods that do not return probabilities because uncertainty is important when assessing association [Aulchenko et al., 2010].

2.2 Model to analyze imputed SNPs in longitudinal studies

We consider that each individual is being followed up over a study period $[0, C]$, where C may represent time of study termination or some other right-censoring variable (e.g., another event different from the one under study). The response variable is the time until the event of interest, T , that cannot be observed if the event appears after the censoring time. Therefore, the observable response variable is (Y, δ) where

$$Y = \min(T, C),$$

and

$$\delta = \begin{cases} 1 & \text{if } Y \geq T \\ 0 & \text{if } Y < T \end{cases} \quad (1)$$

The probability of having $k = 0, 1, 2$ copies of the risk allele for a given imputed SNP can be denoted by

$$P(SNP=k) \equiv w_k.$$

Other possible non-genetic covariates (in GWAS we need at least to consider principal components), $\mathbf{Z} = (Z_1, \dots, Z_p)$ can also be observed. Consequently, if the study has n individuals we will observe the following data

$$\mathbf{D} \equiv \{(Y_i, \delta_i), (w_{i0}, w_{i1}, w_{i2}), Z_i, i=1, 2, \dots, n\} \quad (2)$$

where $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{ip})$ is the vector of non-genetic covariates observed for the i -th individual. We propose to model the hazard rate process at time t conditionally on the imputed SNP and the observed covariates via the following Weibull model:

$$\lambda(t|w_{ik}Z) = \phi t^{\phi-1} \exp(\alpha + \beta k + \mathbf{Z}\boldsymbol{\gamma}') \quad (3)$$

where ϕ denotes the Weibull shape parameter. Then, using the total probability theorem, it is straightforward to see that the likelihood function for the observable data is:

$$L(\boldsymbol{\Theta}) = \prod_{i=1}^n \sum_{k=0}^2 P(Y_i, \delta_i | SNP_i = k, \mathbf{Z}_i; \boldsymbol{\Theta}) w_{ik} \quad (4)$$

where $P(Y_i, \delta_i | \cdot)$ is the density function, $f(Y_i | \cdot)$, if $\delta_i = 1$ or the survival function, $S(Y_i | \cdot)$, if $\delta_i = 0$ and $\boldsymbol{\Theta} = (\alpha, \beta, \gamma_1, \dots, \gamma_p, \phi)$ is the vector of parameters, α corresponds to the constant in the linear predictor, β denotes the coefficient (e.g., log-Hazard-Ratio) for the imputed SNP assuming an additive effect (other genetic models can also be considered), $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)$ encode the effects of covariates, and ϕ is the shape parameter.

By maximizing equation (4) we ensure that model parameter estimates are asymptotically non-biased and efficient by maximizing the likelihood function. Notice that our formulation considers that w_{ik} are known. In particular we propose to obtain them from any of the existing imputed methods. This approach was also adopted in the context of case-control studies, with excellent results [Gonzalez et al., 2009]. The Newton-Raphson iterative algorithm can be used to compute the maximum likelihood parameter estimates, $\hat{\boldsymbol{\Theta}}$.

2.2.1 Model parameter estimation and likelihoods—Two existing approaches can be used to maximize (4). Both of them depend on how uncertainty is incorporated into the model. These two methods have some drawbacks mainly due to the fact that they do not maximize the true likelihood function. To overcome this problem, we propose a third

approach to maximize the observed likelihood. The 3 methods can be summarized as follows

- i. **Best guess approach (Naive):** this classical method considers the most probable genotype for each individual and proceeds as if it is the real genotype. This method can be used even if the imputation method only provides the most likely genotype. This approach corresponds to the classical survival models used when analyzing observed SNPs.
- ii. **Probabilities as covariates (DOSAGE):** this method addresses uncertainty by introducing it as a covariate in the design matrix. If an additive inheritance model is assumed, a linear transformation of the 3 probabilities results in a single value per individual, ranging from 0 to 2. This quantity is known as *dosage*.
- iii. **Latent Class (LC):** this method incorporates the uncertainty in the likelihood function (4). Model parameters can be estimated using the Newton-Raphson algorithm using score and Hessian matrices (derived in the present paper - see Appendix). The LC approach has been incorporated to the R software [R Development Core Team, 2008] package CNVassoc [Subirana et al., 2011].

Next, we introduce the likelihood function used to obtain model parameters for the Naive, DOSAGE and LC methods and discuss the adequacy of each one.

2.2.2 Best guess approach (Naive)—This approach does not maximize (4). Rather, it maximizes the function that results from replacing the probabilities w_{ik} by 0 and 1 depending on the maximum probability observed for each genotype. That is, $w_{ik} = 1$ if w_{ik} is equal to $\max_k(w_{ik})$ and 0 otherwise. Therefore, this method maximizes this *pseudo-likelihood* function:

$$L(\Theta) = \prod_{i=1}^n P(Y_i, \delta_i | SNP_i = k_i^*, Z_i; \Theta), \quad (5)$$

where k_i^* is the most likely genotype for the i -th individual.

It is clear that (5) will be similar to (4) when $\max_k(w_{ik})$ is close to 1 for all individuals (i.e. low uncertainty). However, when uncertainty is moderate or high, (5) will be very different from (4), leading to biased parameter estimates.

2.2.3 Probabilities as covariates (DOSAGE)—DOSAGE approach considers the probabilities w_{ik} as covariates by incorporating them into the design matrix. In this case, it maximizes this *pseudo-likelihood* function:

$$L(\Theta) = \prod_{i=1}^n P(Y_i, \delta_i | d_i, Z_i; \Theta), \quad (6)$$

where d_i is a covariate (or covariates) that depends on the inheritance model as follows:

$$d_i \equiv \begin{cases} 0 * w_{i0} + 1 * w_{i1} + 2 * w_{i2} & \text{for additive models } (dosage) \\ 0 * w_{i0} + 0 * w_{i1} + 1 * w_{i2} & \text{for recessive models} \\ 0 * w_{i0} + 1 * w_{i1} + 1 * w_{i2} & \text{for dominant models} \\ (d_{i1}, d_{i2}) = (w_{i1}, w_{i2}) & \text{for model free} \end{cases}$$

Note that, for the model free, two dummy variables are needed, which in this case are the probability of having 1 copy and the probability of having 2 copies, taking 0 copies as the reference category group. The coefficient of d_i can be interpreted as the effect of the SNP. For the model free, the coefficient of d_{i1} and d_{i2} are the effect of having 1 copy and 2 copies vs. 0 copies, respectively.

The quantity d_i is also known as *dosage* effect when an additive inheritance model is assumed. This quantity is a standard output for most impute software (IMPUTE, PLINK, etc.). It is important to note that while it is easy to compute the *dosage* from the probabilities w_{ik} , it is not possible to obtain the probabilities from the *dosage*.

2.2.4 Latent Class (LC)—The two previous methods have some drawbacks because they do not maximize the proper likelihood (4). Therefore, non-biased and efficient estimates are not guaranteed. Nonetheless, equation (4) can be maximized using an iterative procedure and the analytic score and Hessian function that are derived in the Appendix.

While Naive and DOSAGE approaches maximize their corresponding likelihood functions using standard algorithms, LC requires a more sophisticated one since (4) is a product of a sum that is not simplified when taking the logarithms, as in (5) or (6).

It is important to note that, while Naive and DOSAGE strategies rely on incorporating the imputed SNP probabilities as covariates and fit a standard regression model including a semiparametric Cox regression model, LC cannot only support a fully parametric model such as a Weibull distributed response. However, Weibull distribution is flexible enough to fit on most of real data.

3 Simulation Study

We carried out simulation studies to examine properties of estimators of parameters using Naive, DOSAGE and LC approaches. Our primary aim was to ascertain the robustness of hazard rate estimators of imputed SNPs, $\beta = \log(HR)$. Our second aim was to assess the impact when combining HR estimated in different GWAS using imputed SNPs. We illustrate both simulations in the next sections.

3.1 Accuracy to estimate HR

In order to mimic available imputed SNP data 5,000 datasets were simulated using the following scheme:

- 1. Imputed SNPs:** Imputed SNPs were randomly selected from the Framingham heart cohort study (<http://www.framingham.com/heart/>), a sample with more than 8,477 individuals containing about a million genotyped SNPs and approximately

2.5 million imputed SNPs. Different Minor Allele Frequencies (MAF) and different degrees of imputation uncertainty were considered. The uncertainty measure used was R^2 defined as the correlation between actual genotype and most probable genotype. This definition of R^2 is incorporated by imputation software MACH [Li et al., 2010] which was used to impute SNPs in the Framingham cohort. R^2 ranges from 0 (no information, huge uncertainty) to 1 (complete information, no uncertainty). We selected 100 SNPs from each combination of MAF in [0.01,0.05), [0.05, 0.15), [0.25, 0.35) and [0.45, 0.50) and R^2 in [0.05, 0.15), [0.25, 0.35), [0.45, 0.55), [0.65, 0.75) and [0.85, 0.95). Table 1 shows the total number of imputed SNPs from the Framingham cohort in each of these combination of R^2 by MAF bins.

2. **Non observed genotypes:** Since imputed Framingham genotype data do not include actual genotype, they have been simulated as follows: from each individual and each imputed SNP, one of the 3 possible genotypes were sampled from the imputed probabilities vector. Therefore, we proceeded differently from Zheng et al. [2011]. Of course, our strategy in obtaining the actual genotype assumes that the imputed probabilities are correct. We considered two reasons for adopting this strategy: (i) it saves a lot of time in avoiding the need to perform SNP imputation again (Framingham data contains already imputed SNPs), and (ii) it removes the possible effect of imputation bias in assessing the association test, which is not the aim of our study.
3. **Response (time-to-event) simulation:** Two different strategies have been adopted when simulating the response: (i) assuming a Weibull distribution or (ii) taking the empirical observed time-to-coronary event from Framingham cohort. In the two scenarios, baseline distribution functions has been obtained for the 3 genotypes assuming a proportional hazard:

$$F(t) = 1 - S(t)^{\exp(\beta \text{ SNP})}$$

where F is the distribution function, S is the baseline survival function and SNP is the genotype coded as (0, 1, 2) which represents the number of risk alleles and β is the logarithm of Hazard Ratio (HR).

For (i) scale and shape parameters for S has been obtained from the estimates using the Framingham data, whereas for (ii) S has been obtained as the Kaplan-Meier estimates also from Framingham data. Different scenarios for the SNP effect (i.e. HR) has been performed, ranging from low to very high (1.5, 2, 2.5, 3, 3.5, and 4), also including an scenario with no effect (HR=1).

For the first strategy (a Weibull distributed response), the following strategy has been followed to simulate censored events: simulated values exceeding the maximum follow-up time observed in Framingham have been censored; in addition, some extra values within the follow-up interval has been censored, too, in order to achieve a similar number of events as observed in the Framingham cohort. In (ii), two Kaplan-Meier curves have been performed,

one taking the observed events and the other taking the censored events. From the first one, time-to-event values, T , have been generated, and censoring times, C , from the second one. The observed times, Y , have been defined as $Y = \min\{C, T\}$. Finally, if $C > T$, the value has been considered censored, and non-censored otherwise (i.e. $C \leq T$). For all strategies, the rate of non-censored events has been similar to the observed coronary events rate in the Framingham cohort (i.e. 8% approximately).

Results from empirical response are shown in the Additional File.

- 4 Estimation:** For each simulated data element (time-to-response, non-observed geno-type, and imputed SNP probabilities), 4 models were estimated: a proportional hazard Cox regression model using the actual (non-observed) genotype (True), the Naive model using the most probable SNP number of alleles, (DOSAGE) with the *dosage* formulation as a predictor, and LC using imputed SNP probabilities. Finally, an additive mode of inheritance was assumed for all the fitted models. The True model was used as the gold standard and can be used as a reference. Note that it is not feasible to use the proportional hazard Cox model for LC approach. Therefore, the first 3 approaches use a semiparametric model (Cox regression) while LC uses a full parametric model (Weibull regression).

In each simulation and model, the following measures were computed: 1) Bias: defined as the difference of the expected and the true HR; 2) MSE: Mean Squared Error, defined as the expected squared difference between estimated log-HR and the true log-HR; 3) Power: Power of detecting an associated SNP; 4) Coverage: Probability that the 95% confidence interval includes the true HR.

3.2 Impact when combining Hazard Ratios for GWAS meta-analysis

Using simulated data from non-censored Weibull distribution, 5 datasets are taken as if they correspond to 5 different GWAS for which the HR was estimated using the true SNP. This will be considered as the gold-standard results. Another 5 datasets were used to compute the HR using the 3 compared approaches (Naive, DOSAGE and LC). Pooled HR was fitted using the *rmeta* package assuming a random-effect model. We repeated this process 5,000 times. For each simulation, pooled HR, heterogeneity p-value and τ^2 estimation were recorded. The heterogeneity p-value tests whether there is an excess of heterogeneity, while τ^2 is a measure of that heterogeneity.

From the same simulated data, we have taken 10 datasets at random, as if they correspond to 10 different GWAS estimated using each of the 3 approaches as well as the true SNP. For each of the 3 approaches and the true SNP, a meta-analysis has been computed to assess the possible bias in estimating the pooled HR.

3.3 Results of simulation study

We computed the bias, power, MSE, and coverage after estimating HR for the imputed SNP. All the results have been plotted in different figures including different panels corresponding to the different scenarios when varying MAF and R-squared index. The figure in each panel

shows the results for different HRs. In the discussion of the simulation results, we first focus on the estimation of HR. Then, we analyze the results to determine the effect of each of the 3 methods when computing pooled HR in meta-analysis settings.

We are illustrating the results for censored Weibull response. For this case, we computed the bias (Figure 1), power (Figure 2), MSE (Supplementary Figure 1), and coverage (Supplementary Figure 2) after estimating HR for the imputed SNP. All these figures include different panels corresponding to the different scenarios when varying MAF and R-squared index. The figure in each panel shows the results for different HRs.

Figure 1 shows that the estimated HR using the LC approach is almost identical to those estimated using the gold-standard method (true SNP) in all scenarios. The DOSAGE method is biased for those HRs larger than 2.5. For this model, we can also observe that the more uncertainty and the higher the MAF, the more bias exists. For instance, when the R-squared index is 0.3 and MAF is equal to 0.5, DOSAGE reports biased results for HR larger than 2. Finally, as one may expect, the Naive approach is completely biased for all scenarios.

Figure 2 shows the power of each approach for a GWAs significance level assuming 2.5 million of imputed SNPs ($\alpha = 2 \cdot 10e - 8$). We can observe that LC and DOSAGE have similar power to detect association between the imputed SNP and the event of interest, while Naive method is less powerful than the other two when uncertainty increases.

Regarding to MSE, Supplementary Figure 1 indicates the LC approach as the best one, although DOSAGE achieve almost the same accuracy. Naive behaves much worse than than LC and DOSAGE methods in almost all scenarios except when uncertainty is low ($R^2 > 85\%$).

Finally, LC also is the best approach with regard to coverage. Supplementary Figure 2 shows that our proposed method has an observed coverage about 95% in all scenarios and for all HRs while the coverage dramatically decreases when HR increases for DOSAGE and Naive approaches.

Very similar results are obtained when simulating empirical distributed response in terms bias (Supplementary Figure 3), power (Supplementary Figure 4), accuracy (Supplementary Figure 5) and coverage (Supplementary Figure 6). It is remarkable to note that LC gives no bias estimates even when using a fully parametric Weibull model (Supplementary Figure 3), and behaves as well as for censored Weibull distributed response. Maybe this is due to empirical data could be fitted satisfactory to a Weibull distribution.

Moving to meta-analysis simulations, Figures 3 and 4 summarize our main findings. Previous simulation studies have mainly shown that Naive and DOSAGE approaches are biased when estimating the risk of observing the event of interest. This has two main limitations when combining HR from different studies that use these approaches: first, the pooled HR is also biased and second, the variance of the pooled HR is inflated because the heterogeneity between studies increases and a random-effect model has to be used to pool the individual HRs. The first problem can be observed in Figure 3, which shows the simulation results when the true HR is equal to 2.5. Each panel corresponds to the meta-

analysis results when combining the individual HRs estimated using each method. We can observe that the pooled HR is biased when using Naive and DOSAGE, while LC is not. The second problem is demonstrated in Figure 4, which shows the false positive rate of declaring that heterogeneity exists when combining 5 HRs estimated from one of the 3 approaches and 5 more HRs estimated from the True model. We simulated individual studies using the same expected HR. Therefore, the null hypothesis of heterogeneity between studies should not be rejected (e.g., false positive rate should be around 5%). However, Figure 4 shows that the type-I error when using Naive and DOSAGE is inflated, and this inflation increases when HR is bigger. In practical terms, when heterogeneity is observed, a random-effects model has to be used to estimate the pooled HR. This may increase the variance of pooled HR, decreasing the power to detect a significant pooled HR.

Finally, it must be emphasized that the failure rate (i.e. the number of iterations in which models have not converged) has been very low or null (data not shown). Only in very extreme scenarios (MAF below 5% and R-squared below 15%) has the failure rate been high (up to about 10% and 20% for DOSAGE and LC), respectively, and up to about 45% for Naive method. Therefore, in this scenario, the obtained results from the simulation study and meta-analysis study may not be reliable.

4 Application to real data

We illustrate our proposed method by analysing SNPs and non-genetic variables from the Framingham study data. We obtained access to phenotype and genotype (imputed SNPs) data under the Framingham Share initiative via the Database of Genotypes and Phenotypes (dbGaP, ncbi.nlm.nih.gov/dbgap; Project number 1534). From all imputed SNPs existing in the database we discarded all the rare ones (with a minor allele frequency $< 1\%$) and the ones imputed with almost no uncertainty ($R^2 < 99\%$). Also, those SNPs with a very poor imputation quality ($R^2 < 10\%$) were removed. Finally, a set of 1,464,340 imputed SNPs were analyzed.

For each imputed SNP, the 3 approaches (Naive, DOSAGE and LC) were fitted taking the time to coronary heart disease as the response. For the Naive and DOSAGE approaches a proportional Cox regression was adjusted. As covariates, we considered age, sex and the 5 first principal components to take into account family structure data. The first 5 principal components capture more than 80% of the variability. Finally, we analyzed 3,008 individuals from the Framingham offspring cohort with response information and all covariate data available.

To address an excess of familial or ethnic data structure of the data (not captured by the first 5 principal components), p-values were corrected by the genomic inflation factor λ (1.062, 1.069 and 1.058 for Naive, DOSAGE and LC approaches, respectively). Note that genomic inflation factor is very close to one. One explanation could be that the incorporation of the 5 principal components captures well the familiar structure in all 3 approaches.

LC or DOSAGE approaches gave very similar results, with a pairwise Spearman correlation $> 99\%$. The correlation between the Naive approach and any other two is smaller but also high (approximately 93%).

None of the 3 approaches produced any significant results, except that one SNP achieved a significance level slightly beyond the Bonferroni-corrected threshold using the DOSAGE approach (see Manhattan plots represented in Figure 5). This is probably an artefact since the rest of SNPs are far from the significance level. On another hand, there seems to be no inflation of type I error rate according to QQ-plots represented in Figure 6 for none of the 3 approaches.

The failure rate (i.e. the number of imputed SNPs for which models have not converged have achieved a non reliable estimation) has been very low for all 3 methods: less than 0.4%.

4.1 Computational time

Although Aulchenko et al. [2010] cites that the LC approach is computationally demanding, observed times required to analyze Framingham imputed SNPs on real data phenotypes did not confirm this. The number of iterations required to achieve convergence is low for the vast majority of SNPs (5 or 6 iterations). This makes the computation time consumed in fitting the LC approach model similar to that of a standard Cox regression model. To analyze 1,464,340 imputed SNPs on 3,008 individuals and adjusting by age, sex and five genetic principal components, it took 16.7, 17.4 and 29.8 hours for Naive, DOSAGE and LC methods, respectively, using a Linux platform x86 64-redhat-linux-gnu (64-bit) CPU with 2GHz and 16G of RAM Memory, and using version 2.15.1 of R software. Therefore, we have shown that, once first and second derivatives expressions from the likelihood functions are specified, LC is not more computationally demanding than Naive or DOSAGE approaches, and it is feasible to perform GWAS analysis using the LC approach using a multicore processor; for example, with just 10 cores, analysis of a GWAS may take less than 3 hours.

5 Discussion

We studied how to incorporate uncertainty when analyzing imputed SNP in cohort design studies and conducted a comprehensive analysis comparing two existing methods with our proposed one. The simulations demonstrated that DOSAGE is biased. This bias is negligible for small HRs, but for moderate-large effects ($HR > 2 - 3$) the results are not correct. This may explain why DOSAGE (and other similar methods like SNPTTEST) can be a good option for analyzing case-control studies, where ORs are between 1.1 and 1.5. However, longitudinal studies have shown that the effect of a given SNP may be very strong. For instance, González et al. [2011] and Frey et al. [2005] reported a Hazard Ratio of 3.1 and 12.1, respectively. Our proposed method, which truly incorporates uncertainty in the likelihood function properly, showed excellent results in all scenarios, in particular for moderate-high risk estimates. What is clear from our simulation studies is that the Naive method (e.g., best guess) should not be used any more, at least not when uncertainty exists.

Using real data, we also showed that none of the 3 methods inflates the type I Error, i.e., they produce no excess of false positives. Indeed, we have shown that our proposed method takes comparable amount of time to fit the data than standard regression models such as Cox, making feasible to analyze a GWA study with 1,5 million of imputed SNPs.

While LC and DOSAGE have very similar power for all scenarios, the latter is less accurate (it is more biased and the coverage rate is lower when the effect increases). LC and DOSAGE achieve similar power in all scenarios, while Naive is much less powerful when uncertainty increases. Good power is important to discover associated SNP. However, accurate estimations can be crucial in practical terms, such as in the case of developing a genetic risk model function. A risk model requires knowledge of the effect of a given SNP. It can be estimated using data from a single study or a better risk estimation can be obtained from a meta-analysis using published papers. In both cases a non-biased estimation is important to achieving an accurate genetic risk model.

Our approach would also be useful in comparing the effect observed in our cohort with a published study. Let us assume that the interested SNP is not genotyped in our cohort because we use a different or less dense platform. After imputing, the LC method guarantees that both effects will be comparable; using the DOSAGE or Naive approaches may obtain a different risk estimate and we will not be able to say whether it is a real difference or is biased.

In conclusion, when analyzing longitudinal data from cohort studies, the LC approach is the only method (of the 3 considered) that maximizes the proper likelihood function. Therefore, it is the only one that guarantees the good properties of the maximum likelihood estimates and yields unbiased and powerful results. In practical terms, we have observed that when the effect of the SNP is high ($HR > 2$) only LC has fully satisfactory estimates in terms of bias, coverage and power regardless of the degree of uncertainty. In other situations (low HRs) DOSAGE and LC behave similarly.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278.

Authors want to thank Elaine Lilly for her contribution in the English revision of the manuscript, and Mikel Esnaola for his great contribution in translating the algorithms code syntax to C language.

APPENDIX

A Likelihood, score and Hessian functions

The logarithm of the likelihood function (log-likelihood) for the Latent Class model is

$$\log L(\Theta) = \sum_{i=1}^n \log \left\{ \sum_{k=1}^2 P(Y_i, \delta_i | SNP_i = k, Z_i; \Theta) w_{ik} \right\}$$

where $\Theta = (\alpha, \beta, \gamma_1, \dots, \gamma_p, \varphi)$ is the vector of parameters. For convenience, we will denote θ_s as the s -th component of Θ vector, which can be either the constant α , the SNP coefficient β , the covariates coefficients γ_j ($j = 1, \dots, p$) or the scalar parameter φ .

The first derivatives of the Log-likelihood function with respect to the s -th component of Θ :

$$\frac{\partial \log L(\Theta)}{\partial \theta_s} = \sum_{i=1}^n \frac{\sum_{k=0}^2 \frac{\partial h_{ik}}{\partial \theta_s}}{g_i}$$

and the second derivatives of the log-likelihood function:

$$\frac{\partial^2 \log L(\Theta)}{\partial \theta_s \partial \theta_{s'}} = \sum_{i=1}^n \frac{\sum_{k=0}^2 \frac{\partial h_{ik}}{\partial \theta_s} \frac{\partial h_{ik}}{\partial \theta_{s'}} g_i - \sum_{k=0}^2 \frac{\partial h_{ik}}{\partial \theta_s} \sum_{k=0}^2 \frac{\partial h_{ik}}{\partial \theta_{s'}}}{g_i^2}$$

$$h_{ik} \equiv P(Y_i, \delta_i | SNP_i = k, \mathbf{Z}_i; \Theta) w_{ik},$$

$$g_i \equiv \sum_{k=0}^2 h_{ik}$$

where

$$h_{ik} \equiv P(Y_i, \delta_i | SNP_i = k, \mathbf{Z}_i; \Theta) w_{ik},$$

$$g_i \equiv \sum_{k=0}^2 h_{ik}$$

Finally, we define the linear predictor as

$$\eta_{ik} \equiv \alpha + \beta k + \mathbf{Z}_j \boldsymbol{\gamma}'$$

where k is the number of risk alleles (i.e. 0, 1 or 2 copies), and $\mathbf{Z}_j$ is the covariate vector for the i -th individual.

Two situations must be considered: whether the time to response for the i -th individual ($\delta_i = 1$) has been observed or not ($\delta_i = 0$), where δ is called the censor indicator or variable.

A.1 If $\delta_i = 0$

h_{ik} function takes the form:

$$h_{ik} = w_{ik} e^{-\lambda_{ik} Y_i^\theta}$$

where $\lambda_{ik} = \exp(\eta_{ik})$.

- First derivatives

– with respect to α, β or γ_j ($\theta_s \in \{\alpha, \beta, \gamma_j\}$):

$$\frac{\partial h_{ik}}{\partial \theta_s} = h_{ik} x_{isk} \left(-Y_i^\phi \lambda_{ik} \right)$$

where x_{isk} takes the value of 1, k or Z_{ij} depending on whether the derivative is with respect to α, β or γ_j , respectively, and Z_{ij} is the value of j -th covariate for the i -th individual.

– with respect to the shape parameter ϕ :

$$\frac{\partial h_{ik}}{\partial \phi} = -h_{ik} \lambda_{ik} Y_i^\phi \log(Y_i)$$

• Second derivatives:

– with respect to α, β or γ_j ($\theta_s, \theta_{s'} \in \{\alpha, \beta, \gamma_j\}$):

$$\frac{\partial^2 h_{ik}}{\partial \theta_s \partial \theta_{s'}} = x_{isk} \left[-Y_i^\phi \lambda_{ik} \left(x_{is'k} h_{ik} - \frac{\partial h_{ik}}{\partial \theta_{s'}} \right) \right]$$

where x_{isk} takes the value of 1, k or Z_{ij} depending on whether θ_s is equal to α, β or γ_j , respectively. Similarly, $x_{is'k}$ takes the value of 1, k or Z_{ij} depending on whether $\theta_{s'}$ is equal to α, β or γ_j respectively.

– with respect to shape parameter ϕ :

$$\frac{\partial^2 h_{ik}}{\partial \phi^2} = -\lambda_{ik} \log(Y_i) Y_i^\phi \left(\frac{\partial h_{ik}}{\partial \phi} + h_{ik} \log(Y_i) \right)$$

• Cross-derivatives between the constant/coefficients ($\theta_s \in \{\alpha, \beta, \gamma_j\}$) and the shape parameter ϕ

$$\frac{\partial^2 h_{ik}}{\partial \theta_s \partial \phi} = x_{isk} \left[-\lambda_{ik} Y_i^\phi \left(\frac{\partial h_{ik}}{\partial \phi} + h_{ik} \log(Y_i) \right) \right]$$

where x_{isk} takes the value of 1, k or Z_{ij} depending on whether θ_s is equal to α, β or γ_j respectively.

A.2 If $\delta_i = 1$

h_{ik} function takes the form:

$$h_{ik} = w_{ik} \lambda_{ik} \phi Y_i^{\phi-1} e^{-\lambda_{ik} Y_i^\phi}$$

• First derivatives

– with respect to α, β or γ_j ($\theta_s \in \{\alpha, \beta, \gamma_j\}$):

$$\frac{\partial h_{ik}}{\partial \theta_s} = h_{ik} x_{isk} \left(1 - Y_i^\phi \lambda_{ik} \right)$$

where x_{ij} takes the value of 1, k or Z_{ij} depending on whether the derivative is with respect to α, β or γ_j respectively and Z_{ij} is the value of j -th covariate for the i -th individual.

– with respect to the shape parameter ϕ :

$$\frac{\partial h_{ik}}{\partial \phi} = -h_{ik} \left(\lambda_{ik} Y_i^\phi \log(Y_i) - \log(Y_i) - \frac{1}{\phi} \right)$$

• Second derivatives:

– with respect to α, β or γ_j ($\theta_s, s' \in \{\alpha, \beta, \gamma_j\}$):

$$\frac{\partial^2 h_{ik}}{\partial \theta_s \partial \theta_{s'}} = x_{isk} \left[\frac{\partial h_{ik}}{\partial \theta_{s'}} \left(1 - \lambda_{ik} Y_i^\phi \right) - h_{ik} \lambda_{ik} Y_i^\phi x_{is'k} \right]$$

where x_{isk} takes the value of 1, k or Z_{ij} depending on whether θ_s is equal to α, β or γ_j respectively. Similarly, $x_{is'k}$ takes the value of 1, k or Z_{ij} depending on whether $\theta_{s'}$ is equal to α, β or γ_j respectively.

– with respect to shape parameter ϕ :

$$\frac{\partial^2 h_{ik}}{\partial \phi^2} = -\frac{\partial h_{ik}}{\partial \phi} \left(Y_i^\phi \log(Y_i) \lambda_{ik} - \log(Y_i) - \frac{1}{\phi} \right) - h_{ik} \left(Y_i^\phi \log^2(Y_i) \lambda_{ik} + \frac{1}{\phi^2} \right)$$

• Cross-derivatives between the constant/coefficients ($\theta_s \in \{\alpha, \beta, \gamma_j\}$) and the shape parameter ϕ

$$\frac{\partial^2 h_{ik}}{\partial \theta_s \partial \phi} = x_{isk} \left[\frac{\partial h_{ik}}{\partial \phi} \left(1 - \lambda_{ik} Y_i^\phi \right) - h_{ik} \lambda_{ik} Y_i^\phi \log(Y_i) \right]$$

where x_{isk} takes the value of 1, k or Z_{ij} depending on whether θ_s is equal to α, β or γ_j respectively.

References

- Alakus H, Warnecke-Eberz U, Bollschweiler E, Mönig SP, Vallböhmer D, Brabender J, Drebbler U, Baldus SE, Riemann K, Siffert W, et al. GNAS1 T393C polymorphism is associated with histopathological response to neoadjuvant radiochemotherapy in esophageal cancer. The Pharmacogenomics Journal. 2009; 9(3):202–207. [PubMed: 19274060]
- Aulchenko Y, Struchalin M, van Duijn C. ProbABEL package for genome-wide association analysis of imputed data. BMC Bioinformatics. 2010; 11:134. [PubMed: 20233392]

- Biernacka J, Tang R, Li J, McDonnell S, Rabe K, Sinnwell J, Rider D, Andrade M, Goode E, Fridley B. Assessment of genotype imputation methods. *BMC Proceedings*. Dec.2009 3(Suppl 7):S5. [PubMed: 20018042]
- Del Greco M F, Pattaro C, Luchner A, Pichler I, Winkler T, Hicks AA, Fuchs-berger C, Franke A, Melville SA, Peters A, et al. Genome-wide association analysis and fine mapping of nt-probnp level provide novel insight into the role of the mthfrcln6-nppa-nppb gene cluster. *Human molecular genetics*. 2011; 20(8):1660. [PubMed: 21273288]
- Fornage M, Debette S, Bis JC, Schmidt H, Ikram MA, Dufouil C, Sigurdsson T, Lumley S, DeStefano AL, Fazekas F, Vrooman HA, Shibata DK, Maillard P, Zijdenbos A, Smith AV, Gudnason H, de Boer R, Cushman M, Mazoyer B, Heiss G, Vernooij C, Enzinger MW, Glazer NL, Beiser A, Knopman DS, Cavalieri M, Niessen WJ, Harris TB, Petrovic K, Lopez OL, Au R, Lambert JC, Hofman A, Gottesman RF, Garcia SR, andHeckbert M, Atwood LD, Catellier DJ, Uitterlinden AG, Yang Q, Smith NL, AspelundT Romero JR. Rice K, Taylor KD, Nalls MA, Rotter JJ, Sharrett R, van P, Amouyel Duijn CM. Wolf PA, Gudnason V, van der Lugt A. Boerwinkle E, Psaty BM, Seshadri S, Tzourio C, Breteler MM, Mosley TH, Schmidt R, Longstreth WT, DeCarli C, Launer LJ. Genome-wide association studies of cerebral white matter lesion burden. *Annals of neurology*. Jun; 2011 69(6):928–39. [PubMed: 21681796]
- Frey UH, Alakus H, Wohlschlaeger J, Schmitz KJ, Winde G, van Calker HG, Jöckel KH, Siffert W, Schmid KW. Gnas1 t393c polymorphism and survival in patients with sporadic colorectal cancer. *Clinical cancer research*. 2005; 11(14):5071–5077. [PubMed: 16033819]
- Frey UH, Lümmer G, Jäger T, Jöckel KH, Schmid KW, Rübber H, Müller N, Siffert W, Eisenhardt A. The gnas1 t393c polymorphism predicts survival in patients with clear cell renal cell carcinoma. *Clinical cancer research*. 2006a; 12(3):759–763. [PubMed: 16467086]
- Frey UH, Nüchel H, Sellmann L, Siemer D, Küppers R, Dürig J, Dührsen U, Siffert W. The gnas1 t393c polymorphism is associated with disease progression and survival in chronic lymphocytic leukemia. *Clinical cancer research*. 2006b; 12(19):5686–5692. [PubMed: 17020971]
- Gonzalez JR, Subirana I, Escaramis G, Peraza S, Caceres A, Estivill X, Armengol L. Accounting for uncertainty when assessing association between copy number and disease: a latent class model. *BMC Bioinformatics*. 2009; 10:172. [PubMed: 19500389]
- González SJ, Rojas JJ, Redal MA, Patrucco L, Correale J, Argibay PF, Cristiano E. Cd24 as a genetic modifier of disease progression in multiple sclerosis in argentinean patients. *J Neurol Sci*. Aug; 2011 307(1-2):18–21. [PubMed: 21641619]
- Granada M, Wilk JB, Tuzova M, Strachan DP, Weidinger S, Albrecht E, Gieger C, Heinrich J, Himes BE, Hunninghake GM, Celedón JC, Weiss ST, Cruikshank WW, Farrer LA Center DM. O'Connor GT. A genome-wide association study of plasma total ige concentrations in the framingham heart study. *J Allergy Clin Immunol*. Nov.2011 41(3):762.
- Heist RS, Zhou W, Chirieac LR, Cogan-Drew T, Liu G, Su L, Neuberg D, Lynch TJ, Wain JC, Christiani DC. Mdm2 polymorphism, survival, and histology in early-stage non-small-cell lung cancer. *Journal of clinical oncology*. 2007; 25(16):2243–2247. [PubMed: 17538168]
- Kooner JS, Saleheen D, Sim X, Sehmi J, Zhang W, Frossard P, Been LF, Chia KS, Dimas AS, Hassanali N, Jafar T, Jowett JB, Li X, Radha V, Rees SD, Takeuchi R, Young F, Aung T, Basit A, Chidambaram M, Das D, Grundberg E, Hedman AK, HydrieZI Islam M. Khor CC, Kowlessur S, Kristensen MM, Liju S, Lim WY, Matthews J, Liu DR, Morris AP, Nica AC, Pinidiyapathirage JM, Prokopenko I, Rasheed A, Samuel M, Shah N, Shera AS, Small KS, Suo C, Wickremasinghe AR, Wong TY, Yang M, DIAGRAM; MuTHER. ZhangF. Abecasis GR, Barnett AH, Caulfield M, Deloukas P, Frayling TM Froguel P. Kato N, Katulanda P, Kelly MA, Liang J, Mohan V, Sanghera J, Scott DK, Seielstad M, Zimmet PZ, Elliott P, Teo YY, McCarthy MI, Danesh J, Tai ES, Chambers JC. Genome-wide association study in individuals of south asianancestry identifies six new type 2 diabetes susceptibility loci. *Nat Genet*. Aug; 2011 43(10):1984–9.
- Li, Willer CJ.; Ding, Y.J.; Scheet, P.; Abecasis, GR. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*. 2010; 34(8):816–834. [PubMed: 21058334]
- Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies via imputation of genotypes. *Nature Genetics*. 2007; 39:906–913. [PubMed: 17572673]

- Meyre D, Delplanque J, Chèvre JC, Lecoœur C, Lobbens S, Gallina S, Durand E, Vatin V, Degraeve F, Proença C, et al. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in european populations. *Nature genetics*. 2009; 41(2):157–159. [PubMed: 19151714]
- Morales E, Bustamante M, Gonzalez JR, Guxens M, Torrent M, Mendez M, Garcia-Esteban R, Julvez J, Forns J, Vrijheid M, et al. Genetic variants of the FADS gene cluster and ELOVL gene family, colostrums LC-PUFA levels, breastfeeding, and child cognition. *PloS one*. 2011; 6(2):e17181. [PubMed: 21383846]
- Morgan T, House J, Cresci S, Jones P, Allayee H, Hazen S, Patel Y, Patel R, Eapen D, Waddy S, et al. Investigation of 95 variants identified in a genome-wide study for association with mortality after acute coronary syndrome. *BMC Medical Genetics*. 2011; 12(1):127. [PubMed: 21957892]
- Myocardial Infarction Genetics Consortium. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet*. Mar.2009 41(3):762.
- Nair RP, Duffin KC, Helms C, Ding J, Stuart PE, Goldgar D, Gudjonsson JE, Li Y, Tejasvi T, Feng BJ, et al. Genome-wide scan reveals association of psoriasis with il-23 and nf-kb pathways. *Nature genetics*. 2009; 41(2):199–204. [PubMed: 19169254]
- Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature genetics*. 2010; 42(7):570–575. [PubMed: 20562874]
- R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2008. URL <http://www.R-project.org>. ISBN 3-900051-07-0
- Rovelet-Lecrux A, Hannequin D, Raux G, Le Meur N, Laquerriere A, Vital A, Dumanchin C, Feuillet S, Brice A, Vercelletto M, Dubas F, Frebourg T, Campion D. App locus duplication causes autosomal dominant early-onset alzheimer disease with cerebral amyloid angiopathy. *Nat Genet*. 2006; 38(1):24–6. [PubMed: 16369530]
- Subirana I, Diaz-Uriarte R, Lucas JR, González G. Cnvassoc: Association analysis of cnv data using r. *BMC Medical Genomics*. May.2011 24(4):47. [PubMed: 21609482]
- Takeuchi F, McGinnis R, Bourgeois S, Barnes C, Eriksson N, Soranzo N, Whittaker P, Ranganath V, Kumanduri V, McLaren W, et al. A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS genetics*. 2009; 5(3):e1000433. [PubMed: 19300499]
- Trompet S, de Craen A, Postmus I, Ford I, Sattar N, Caslake M, Stott D, Buckley B, Sacks F, Devlin J, et al. Replication of LDL GWAs hits in PROSPER/PHASE as validation for future (pharmaco) genetic analyses. *BMC medical genetics*. 2011; 12(1):131. [PubMed: 21977987]
- Ueno H, Sato Y, Okusaka T, Furuse J, Ishii H, Sekine A, Nakamura Y, Kaniwa N, Sawada J, Saijo N. Association of snps in abcc1 gene with overall survival in stage iv pancreatic adenocarcinoma patients treated with gemcitabine monotherapy. *J Clin Oncol*. May.2008 26(suppl) abstr 14504.
- Wojnar M, Brower KJ, Strobbe S, Ilgen M, Matsumoto H, Nowosad I, Sliwerska E, Burmeister M. Association between val66met brain-derived neurotrophic factor (bdnf) gene polymorphism and post-treatment relapse in alcohol dependence. *Alcoholism: Clinical and Experimental Research*. 2009; 33(4):693–702.
- Xing J, Myers RE, He X, Qu F, Zhou F, Ma X, Hyslop T, Bao G, Wan S, Yang H, Chen Z. Gwas-identified colorectal cancer susceptibility locus associates with disease prognosis. *Eur J Cancer*. Jul; 2011 47(11):1699–707. [PubMed: 21402474]
- Zheng J, Li Y, Abecasis GR, Scheet P. A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genet Epidemiol*. Feb; 2011 35(2):102–110. [PubMed: 21254217]

Figure 1.
Hazard ratio according to minor allele frequency and uncertainty (R^2).

Figure 2.

Power according to minor allele frequency and uncertainty (R^2), taking GWA significance level, $\alpha = 2 \cdot 10^{-8}$.

Figure 3.

Meta-analysis of 10 studies, simulating censored Weibull response, a Hazard Ratio of 2.5, uncertainty of R-squared = 0.3, and a MAF = 0.5.

Figure 4.

False positive rate (p-value > 0.05) of a meta-analysis taking 5 + 5 studies simulated from a censored Weibull response.

Figure 5. Manhattan plot representing minus log p-values of each analyzed imputed SNP from the Framingham real data.

Figure 6.
p-values Q-Q plot of each analyzed imputed SNP from the Framingham real data.

Table 1

Distribution of imputed SNPs according to combination of R^2 and MAF from the Framingham cohort.

R^2 (%)	MAF (%)			
	1 to 5	5 to 15	25 to 35	45 to 50
5 to 15	9,322	3,862	1,513	600
25 to 35	13,279	9,708	3,893	1,537
45 to 55	13,258	14,892	6,797	2,795
65 to T5	17,687	26,508	14,362	5,773
85 to 95	56,875	93,339	68,742	29,429