Analysis of sequencing data with survival outcomes in family-based study using a correlated frailty model.

Yun-Hee CHOI

Dept. of Epidemiology and Biostatistics, University of Western Ontario

Joint work with
Agnieszka KRÓL[1], Shelley BULL[1], Razvan ROMANESCU[1],
Virginie RONDEAU[2], Laurent BRIOLLAIS[1]

[1] Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Canada
[2] INSERM U1219, Biostatistics team, University of Bordeaux, Bordeaux, France

August 9, 2018

Objective

- Few methods exist for the analysis of sets of genetic variants associated with time-to-event data in family setting
- To integrate genetic information into the analysis of time-to-event data arising from family studies to evaluate the genetic association.
- To quantify unexplained heterogeneity within families
- To detect rare susceptible variants associated with disease risks

Outline of the talk

- Correlated shared frailty model with IBD and kinship matrices
- Tests for association using the variance components
- Applications :
  - Lynch Syndrome and Familial Colorectal Cancer Type X families
  - Breast Cancer affected sib pairs study

Notations

Clustered data from n families.
For individual i, $i = 1, \ldots, n_f$, from family f, $f = 1, \ldots, n$
We observe (T_{fi}, δ_{fi})

$$T_{fi} = \min(T^*_{fi}, C_{fi})$$

$$\delta_{fi} = I_{\{T_{fi} = T^*_{fi}\}}$$

- T^*_{fi} : time of the event (e.g. disease onset)
- C_{fi} : time of censoring
- $\boldsymbol{X}_{fi}$: non-genetic covariates

Correlated shared frailty model

The hazard function for individual i from family f

$$r_{fi}(t|b_{fi}) = r_0(t) \exp\left(b_{fi} + \boldsymbol{X}_{fi}^\top \boldsymbol{\beta}\right)$$

- Correlated random effects :
  $\boldsymbol{b}_f = (b_{f1}, \ldots, b_{fn_f})^\top \sim \mathcal{MVN}(0, \boldsymbol{\Sigma}(\sigma))$
- The covariance matrix $\boldsymbol{\Sigma}(\sigma)$ represents the within-familial correlation explained by
  - kinship coefficients matrix $\boldsymbol{K}$: $\boldsymbol{\Sigma}(\sigma) = \sigma^2\, 2\, \boldsymbol{K}$
  - Identity-by-descent (IBD) probabilities matrix $\boldsymbol{D}$: $\boldsymbol{\Sigma}(\sigma) = \sigma^2\, \boldsymbol{D}$
  - both : $\boldsymbol{\Sigma}(\sigma) = \sigma_1^2\, 2\, \boldsymbol{K} + \sigma_2^2\, \boldsymbol{D}$

Inference

Parameters to estimate $\boldsymbol{\Theta} = (r_0(\cdot), \boldsymbol{\sigma}, \boldsymbol{\beta})^\top$.
The marginal likelihood :

$$
\begin{aligned}
L_f(\boldsymbol{\Theta}) &= \int_{\boldsymbol{b}_f} \prod_{i=1}^{n_f} f_{T_{fi}|\boldsymbol{b}_f}(T_{fi}, \delta_{fi}|\boldsymbol{b}_f; \boldsymbol{\Theta}) f_{\boldsymbol{b}_f}(\boldsymbol{b}_f; \boldsymbol{\Theta}) d\boldsymbol{b}_f, \\
&= \int_{b_{f1}} \cdots \int_{b_{fn_f}} \prod_{i=1}^{n_f} \left[\left\{ r_0(T_{fi}) e^{b_{fi} + \boldsymbol{X}_{fi}^\top \boldsymbol{\beta}} \right\}^{\delta_{fi}} e^{-R_0(T_{fi}) e^{b_{fi} + \boldsymbol{X}_{fi}^\top \boldsymbol{\beta}}} \right] \\
&\times \frac{\boldsymbol{b}_f' \boldsymbol{\Sigma}(\sigma)^{-1} \boldsymbol{b}_f / 2}{(2\pi)^{n_f/2} |\boldsymbol{\Sigma}(\sigma)|^{1/2}} d\boldsymbol{b}_f \\
L(\boldsymbol{\Theta}) &= \prod_{f=1}^{n} L_f(\boldsymbol{\Theta}) \qquad (1)
\end{aligned}
$$

where $\boldsymbol{b}_f = \{b_{fi}, i = 1, \ldots, n_f\}$ and $r_0(\cdot)$ and $R_0(\cdot)$ are baseline hazard and cumulative hazard functions, respectively.

Ascertainment correction

- Population-based designs : proband is affected mutation-carrier *Choi et al., 2008*
- Prospective log-likelihood :

$$pl_f(\boldsymbol{\Theta}) = \log L_f(\boldsymbol{\Theta}) - \log P(T < a_{fp} | G_{fp}, \boldsymbol{X}_{fp})$$

$$= \log L_f(\boldsymbol{\Theta}) - \log \left\{ 1 - \int_{b_{fp}} e^{-\Lambda(a_{fp}|G_{fp},\boldsymbol{X}_{fp})e^{b_{fp}}} f_{b_{fp}}(b_{fp}) db_{fp} \right\},$$

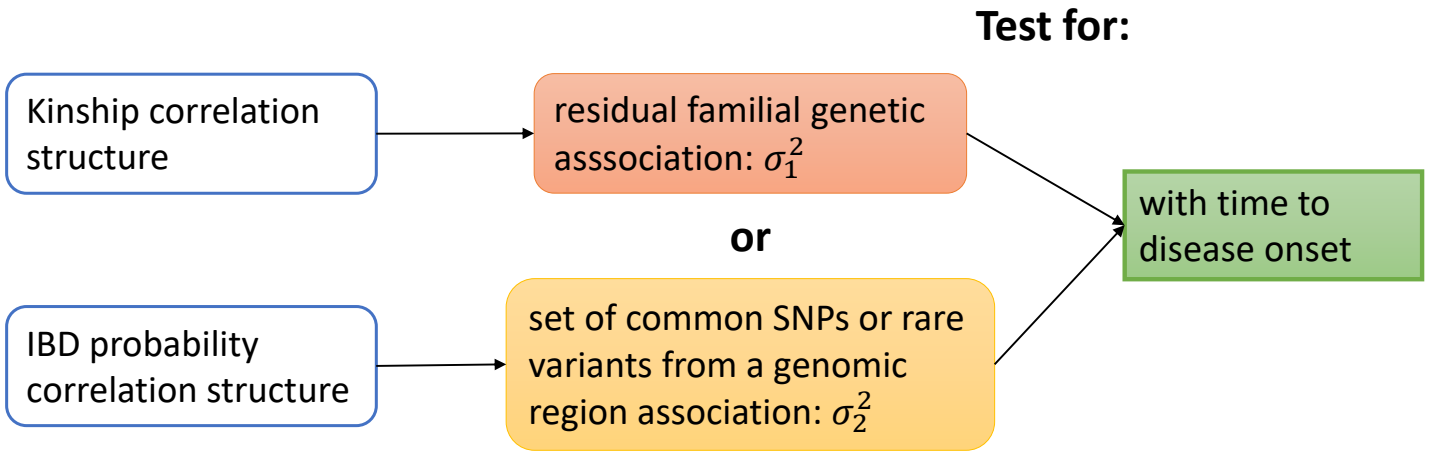
$$pl(\boldsymbol{\Theta}) = \sum_{f=1}^{n} pl_f(\boldsymbol{\Theta})$$

where a_{fp} is the current age of the proband p from family f, G_{fp} denotes the genotype and $\Lambda(\cdot)$ is the cumulative hazard function.

Estimation

- Maximization of log-likelihood by Marquardt algorithm *(Marquardt, 1963)*
  - Combining the Newton-Raphson and the steepest descent algorithms
  - Stable behavior in complex problems while preserving fast convergence
  - Three conditions for convergence : parameters, likelihood and gradient of the likelihood
  - Variances of the parameter estimates directly obtained from the inverse of the Hessian matrix
- Baseline hazard function estimated using parametric Weibull distribution
- Multivariate integrals approximated with a non-adaptive procedure for the Gauss-Hermite quadrature proposed by *Genz and Keister, 1996*

Tests for the genetic association



$$H_0 : \sigma^2 = 0 \text{ versus } H_1 : \sigma^2 > 0$$

- Likelihood Ratio Test statistic : $T_{LR} = 2(pl(\widehat{\sigma^2}) - pl(0))$
- Wald test statistic : $T_W = (\widehat{\sigma^2})^2 / \widehat{Var}(\widehat{\sigma^2})$

$$T \sim 0.5\chi_0^2 + 0.5\chi_1^2, \quad T = \{T_{LR}, T_W\}$$

Molenberghs and Verbeke, 2007

Tests for the genetic association : score test

- Score test statistic for a likelihood of observations $\mathcal{L}(b_1, \cdots, b_N)$:

$$T_S = U^\top W U - \text{tr}(\hat{I} W),$$

- $U_i = \frac{\partial(\log \mathcal{L})}{\partial b_i}(0) \rightarrow$ martingale residuals at T_i under the null
- $\hat{I}_{ij} = -\frac{\partial^2(\log \mathcal{L})}{\partial b_i \partial b_j}(0)$
- $W = \mathbf{\Sigma} - I = (w_{ij})_{N \times N}$
- $T_S \sim \mathcal{N}(0, Var(T_S))$, with $Var(T_S) = 2 \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}^2 v_i v_j$, where v_i is the cumulative baseline hazard at T_i under the null
- $T_S / \sqrt{Var(T_S)} \sim 0.5\chi_0^2 + 0.5\chi_1^2$

Commenges and Jacqmin-Gadda, 1997

Simulation study

- Evaluation of the likelihood ratio genetic association test : power and type I error
- Data generation steps :
  1. Generation of pedigrees (three generations with max 10 members)
  2. Generation of genotypes
  3. Generation of survival data using the correlated frailty model
- Model estimation and p-values calculation using modified R package `frailtypack` *(Rondeau et al., 2012, Krol et al., 2017)*

Generation of genotypes

- R package `sim1000G` $\rightarrow$ simulates genomic regions for unrelated and related individuals *(Dimitromanolakis et al., in preparation)*
- Genotypes generation using regions from 1000 genomes Phase III sequencing data VCF files and a genetic map GRCh37 from chr 4 (online github database)
- Unrelated individuals' genotypes are generated using two simulated haplotypes from package `hapsim` *(Montana, 2005)*
- Minor allele frequency (MAF) range (0.02%, 0.1%)
- Number of families : $N = 100$, 200 and 500

Generation of survival data

- R package `FamEvent` (modified `simfam` function) $\rightarrow$ simulates age-at-onset traits in family data obtained from population-based designs *(Choi et al., 2017)*
- Survival times simulated using the correlated frailty model with mean IBD correlation structure

$$\boldsymbol{mD} = \boldsymbol{D}_1 + \boldsymbol{D}_2/2,$$

  $\boldsymbol{D}_i$ is the IBD matrix of sharing i alleles
- Survival times generated with Weibull baseline, one non-genetic variable ($\beta = 0.5$) and 5 genetic variants to modify the disease risk.
- The effects of the 5 genetic variants are fixed at $\beta_1 = \beta_2 = \cdots = \beta_5$ = 0, 0.5 or 1.
- In order to evaluate the association test for the assumed genotypes, the frailty variance parameter, $\sigma^2 = 0$.

Simulation Results

TABLE: Estimated type I error and power over 500 simulations for the association LRT.

	N=100	N=200	N=500
$\beta_1 = \beta_2 = \cdots = \beta_5 = 0$	5.6	4.2	5.0
$\beta_1 = \beta_2 = \cdots = \beta_5 = 0.5$	35.7	53.2	60.0
$\beta_1 = \beta_2 = \cdots = \beta_5 = 1.0$	96.8	99.4	100.0

- Results for the Wald test were similar to LRT
- Score test was more conservative but had higher power under n=500 and intermediate effects ($\beta_s = 0.5, s = 1, \ldots, 5$)

Application 1 : colorectal cancer families

- Lynch syndrome (LS) : most common hereditary syndrome for colorectal cancers (CRCs) due to mutations in DNA mismatch repair (MMR) genes. Family members experience higher risks of developing CRC with reported lifetime risk among gene carriers
- Familial Colorectal Cancer Type X (FCCTX) : family members meet Amsterdam Criteria, but whose tumours are DNA-mismatch-repair-proficient, unlike LS
- 781 LS and 168 FCCTX families of 3 generations identified from Colon Cancer Family Registries,

Residual familial correlation in LS and FCCTX families

- Comparison of genetic variability via kinship matrix
- More elevated familial unexplained heterogeneity in LS families ($\hat{\sigma}^2 = 1.56$, SE $= 0.07$) than in FCCTX families ($\hat{\sigma}^2 = 0.92$, SE $= 0.15$)
- Simulation study $\rightarrow$ about 6-7 high-risk variants transmitted under a dominant model (MAF range (0.5%, 2%) and HR=7.0) or alternatively about 15 to 20 intermediate-risk variants (MAF(1%,10%) and HR=1.5) would be necessary to explain the CRC familial aggregation in FCCTX families
- Further efforts are therefore needed to identify and characterize these predisposing genetic variants $\rightarrow$ more targeted surveillance programs and treatment among FCCTX families.

Application 2 : breast cancer sib pairs study

- Sister pairs diagnosed with breast cancer recruited from Ontario Breast Cancer Family Registry (OBCFR)
- Probands' disease onset was at 45 or younger and they had one or more sisters also diagnosed with breast cancer
- All the individuals were noncarriers of BRCA1 or BRCA2 mutations, and CHEK2*1100delC were excluded.
- Whole exome sequencing was carried out for all 42 women from 21 families.
- We excluded 4 single probands and analyzed 17 families (3 families with three sisters and 14 families with two sisters), 37 individuals in total

Genetic associations in sister pairs study

Frailty model with two correlation structures :

$$r_{fi}(t|\boldsymbol{b}) = r_0(t)\,\exp\left(b_{fi}\right),$$

$$\boldsymbol{b} = \{b_{fi}, i = 1, \ldots, 3, f = 1, \ldots, 21\} \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Sigma}(\sigma_1^2, \sigma_2^2)),$$

$$\boldsymbol{\Sigma}(\sigma_1^2, \sigma_2^2) = \sigma_1^2\, 2\,\boldsymbol{K} + \sigma_2^2\,\boldsymbol{D}$$

- for identifiability : one of the parameters σ is fixed and the other is estimated :
  - A model using only $\boldsymbol{K}$: $\hat{\sigma_1}^2 = 0.321$ (SE=0.528) used in the models with the kinship and IBD matrices
- Three sets of 6900 models, one for each region of genes, considering $\boldsymbol{D_1}$, $\boldsymbol{D_2}$ and $\boldsymbol{mD}$
- Tests (LRT and Score test) :
  - null model $\rightarrow$ only the kinship matrix
  - model under alternative $\rightarrow$ full model with both, IBD and kinship matrices

Results

- More significant regions found using the score test but with more conservative p-values
- The most significant regions :
  - $\boldsymbol{D_1}$: chr 13, 2 and 5
  - $\boldsymbol{D_2}$: chr 1, 5 and 10
  - $\boldsymbol{mD}$: chr 4, 5 and 10
- Further investigation using Epstein's method on the significant regions confirmed the results found on chr 13 using $\boldsymbol{D_1}$ *(Epstein, 2015)*
- Perspectives : Use segregation-based approach (e.g. FamEvent) of rare variants in the significant regions to find causal variants

Results

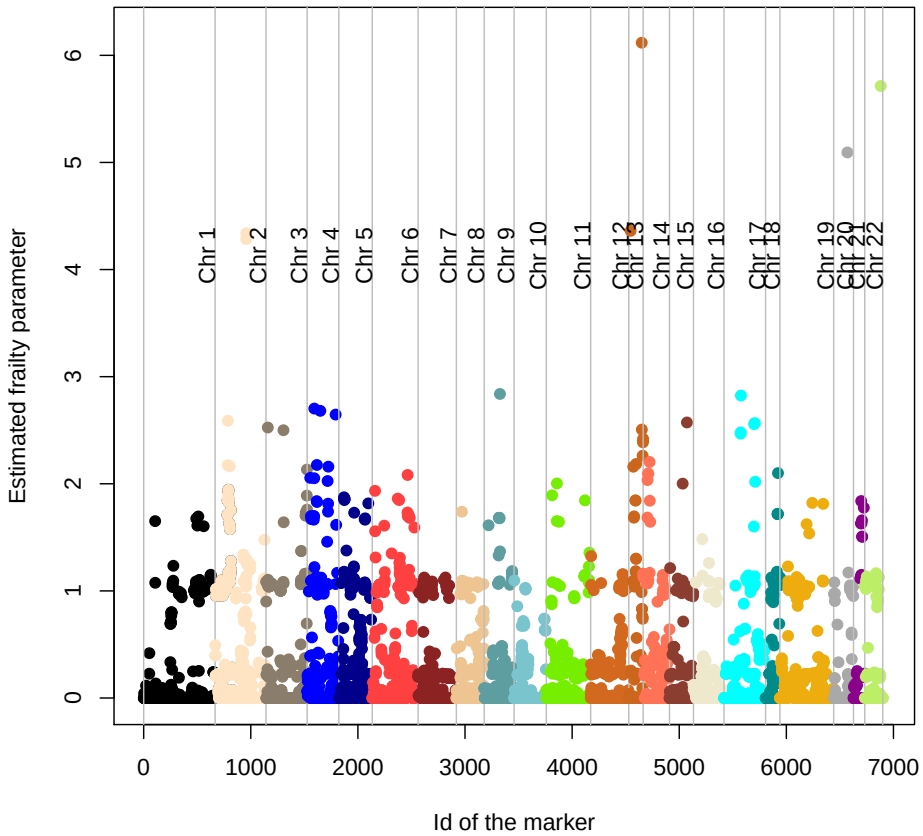


FIGURE: Estimates of σ_2^2 using D_1 in the frailty models.

Results

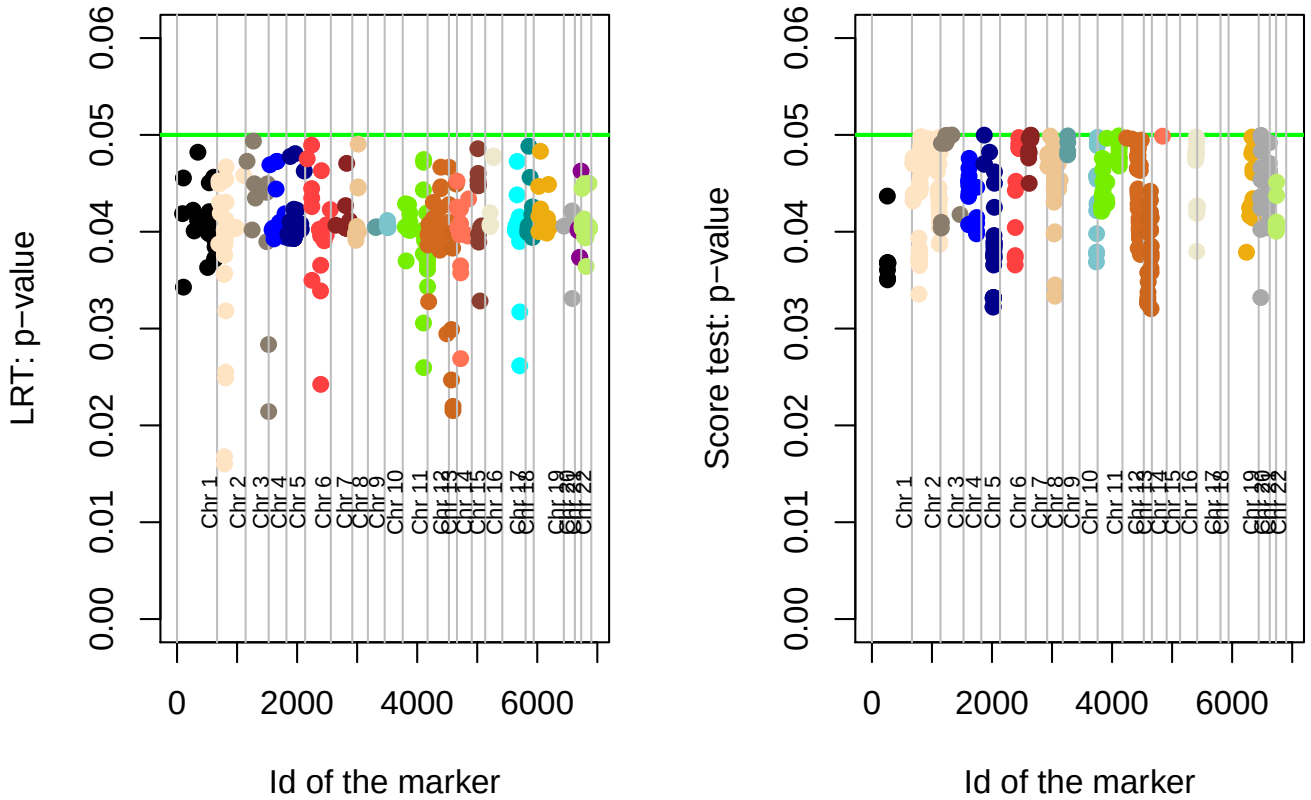


FIGURE: Results of the association test : the most significant p-values using likelihood ratio and score tests (D_1 correlation structure).

Conclusions

- A new approach for correlated frailty models applied to family studies with genetic variants.
- Novel methodology for estimating correlated frailty models and evaluating the genetic familial association with survival times
- Easily applicable to other chronic diseases with variable age at onset and clustering within families.
- Method application to two different cancer studies :
  - Colorectal cancer : Evaluation of the residual familial correlation among FCCTX families to further pursue causal genes/variants
  - Breast cancer : Identification of associated regions with the time to cancer via IBD probabilities matrix to further elucidate causal variants segregating within families

Software : R packages

- `sim1000G`

  https ://cran.r-project.org/web/packages/sim1000G/index.html

- `FamEvent`

  https ://cran.r-project.org/web/packages/FamEvent/index.html

- `frailtypack`

  https ://cran.r-project.org/web/packages/frailtypack/index.html

References

1 Choi, Y.H. et al. (2008). Estimating disease risk associated with mutated genes in family-based designs. *Human heredity* 66(4) 238-251

2 Choi, Y.H. et al. (2017) Modeling of Successive Cancer Risks in Lynch Syndrome Families in the Presence of Competing Risks Using Copulas. *Biometrics* 73(1), 271-282

3 Epstein, M.P. et al. (2015) A Statistical Approach for Rare-Variant Association Testing in Affected Sibships. *Am. J. Hum. Genet* 96, 543-554.

4 Genz, A., & Keister, B. D. (1996). Fully symmetric interpolatory rules for multiple integrals over infinite regions with Gaussian weight. *J Comput Appl Math* 71(2), 299-309.

5 Król, A. et al. (2017). Tutorial in joint modeling and prediction : a statistical software for correlated longitudinal outcomes, recurrent events and a terminal event. *J Stat Softw* 81(3), 1-52.

6 Leclerc, M. et al. (2015). SNP set association testing for survival outcomes in the presence of intrafamilial correlation. *Genetic Epidemiology* 39(6) 406-414.

7 Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics* 11(2), 431-441.

8 Molenberghs, G., & Verbeke, G. (2007). Likelihood ratio, score, and Wald tests in a constrained parameter space. *The American Statistician* 61(1), 22-27.

9 Rondeau, V. et a. (2012). frailtypack : an R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *J Stat Softw* 47(4), 1-28.