uwo.png



# EpiBio Research Day

## Correlated Shared Frailty Model Incorporating Ascertainment Correction with Missing Covariates in Family-Based Studies

Jiaqi Bi

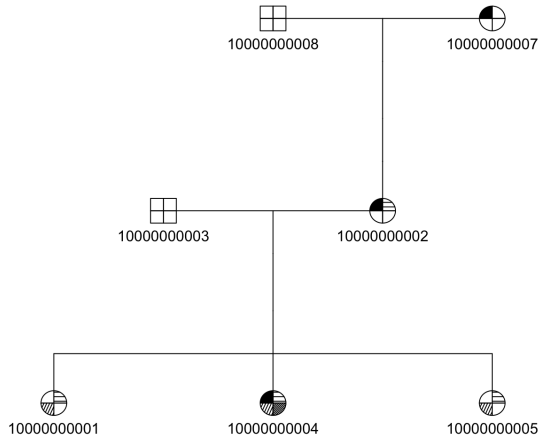
Apr. 9, 2024

# Background

## Breast Cancer

- There were estimated 25,200 new cases of breast cancer in Canada in 2015, and approximately 5,100 deaths, making it the second leading cause of cancer-related death among women [1].
- Hereditary breast-ovarian cancer (HBOC) is an autosomal dominant disease characterized by germline pathogenic mutations in the BRCA1/2 genes [2].
- Some genetic studies based on the family have been conducted to investigate the hereditary breast cancer and ovarian cancer due to the mutation genes of BRCA1/2 [3].
- Time-To-Cancer as an outcome, mutation gene status & PRS are predictors - Problems: There are missing data!

# Pedigree Tree



# Background

## Family-Clustered Frailty Model

- Many different frailty models have been proposed for the analysis of BRCA1/2 families by Choi et al. [4], Chen et al. [5]
- When the missing data occurs in the frailty model due to other mechanisms than the Missing Completely at Random (MCAR), one may use the Monte Carlo Maximization-Expectation (MCEM) [6, 7, 8, 9] or Multiple Imputation (MI) [10] methods to make the inference

## Missing Data

- The issue of the missing data was firstly brought by Rubin [11] in 1976.
- Three missing mechanisms: MCAR, Missing At Random (MAR), Missing Not At Random (MNAR)

# Missing Mechanisms

Denote $Y$ as the complete data matrix, and $M$ as the missing data indicator matrix. Define $y_{ij}$ and $m_{ij}$ as $i$-th row (observation) and $j$-th column (variable) for the matrix $Y$ and $M$. The conditional distribution of the missingness for MCAR is said to be

$$f(m_j|y_{ij}, \phi) = f(m_{ij}|\phi) \tag{1}$$

The MAR is defined as

$$f(m_{ij}|y_{ij}, \phi) = f(m_{ij}|y_{i,obs}, \phi) \tag{2}$$

The MNAR is defined as

$$f(m_{ij}|y_{ij}, \phi) = f(m_{ij}|y_{i,mis}, y_{i,obs}, \phi) \tag{3}$$

# Parametric Survival Analysis

Without loss of generality, everything on the current research will be Weibull baseline hazard.

## Weibull Parametric Survival Analysis

The hazard function is defined as

$$h_{ij}(t_{ij}|\mathbf{x}_{ij}, z_j) = h_0(t_{ij}) \exp(\beta \mathbf{x}_i) z_j \tag{4}$$

In our case, for the simplicity,

$$h_{ij}(t_{ij}|z_j) = h_0(t_{ij}) \exp(\beta_1 x_{1,ij} + \beta_2 x_{2,ij}) z_j \tag{5}$$

In Weibull baseline hazard, $\lambda$ is the shape parameter, $\alpha$ is the scale parameter

$$h_0(t_{ij}) = \alpha \lambda t_{ij}^{\lambda - 1} \tag{6}$$

# Complete Likelihood

Models are meant to be evaluated on the optimized parameters!

## Assuming missing data & frailties are observed

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{J} \prod_{i=1}^{n_j} h(t_{ij}|\mathbf{x}_{ij}, z_j)^{\delta_{ij}} \exp(-H(t_{ij}|\mathbf{x}_{ij}, z_j)) \tag{7}$$

## Ascertainment Correction

In genetic epidemiology studies, families with multiple affected individuals are more likely to be studied than those with only one or no affected individuals. Consider $A$ as the event of being ascertained, we then have $P(D, A|\theta) = P(A|D, \theta)P(D|\theta)$. Thus, we know $A$ is included in $D$, from Baye's rule

$$P(D|\theta) = \frac{P(D, A|\theta)}{P(A|D, \theta)} \propto \frac{L(\theta|D)}{P(A|D, \theta)} \tag{8}$$

# Complete Likelihood

## Assuming missing data & frailties are observed

Denote $A(\boldsymbol{\theta})$ be the ascertainment, and $p_j$ be the proband in family $j$, we have

$$A(\boldsymbol{\theta}) = 1 - S_{p_j}(a_{p_j}|\mathbf{x}_{p_j}) \tag{9}$$

Then the complete likelihood becomes

$$L_C(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta})}{A(\boldsymbol{\theta})} \tag{10}$$

# Complete Log-Likelihood

## Simply take the log

$$\ell_C(\boldsymbol{\theta}) = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \delta_{ij} \log h(t_{ij}|\mathbf{x}_{ij}, z_j) - H(t_{ij}|\mathbf{x}_{ij}, z_j) \tag{11}$$

$$- \sum_{j=1}^{J} \log(1 - S_{p_j}(a_{p_j}|\mathbf{x}_{p_j}, z_j)) \tag{12}$$

$$= \sum_{j=1}^{J} \sum_{i=1}^{n_j} \delta_{ij} \log h(t_{ij}|\mathbf{x}_{ij}) z_j - H(t_{ij}|\mathbf{x}_{ij}) z_j \tag{13}$$

$$- \sum_{j=1}^{n_j} \log(1 - \exp(z_j H_{p_j}(a_{p_j}|\mathbf{x}_{p_j}))) \tag{14}$$

## Recap on $h(\cdot)$ and $H(\cdot)$

Denote $\xi_{ij} = \exp(\boldsymbol{\beta}^\top \mathbf{x}_{ij})$. Note that we can derive

$$h_{ij}(t_{ij}|\mathbf{x}_{ij}, z_j) = \alpha\lambda t_{ij}^{\lambda-1}\xi_{ij}z_j = h(t_{ij}|\mathbf{x}_{ij})z_j \tag{15}$$

With one function in the survival analysis, you can derive the rest! So,

$$H(t_{ij}|\mathbf{x}_{ij}, z_j) = \int_0^t h_{ij}(u|\mathbf{x}_{ij}, z_j)du \tag{16}$$

$$= \alpha\xi_{ij}z_j\lambda \int_0^t u^{\lambda-1}du \tag{17}$$

$$= \alpha\xi_{ij}z_j\lambda \cdot \frac{1}{\lambda}t_{ij}^{\lambda} = \alpha\xi_{ij}z_j t_{ij}^{\lambda} = H(t_{ij}|\mathbf{x}_{ij})z_j \tag{18}$$

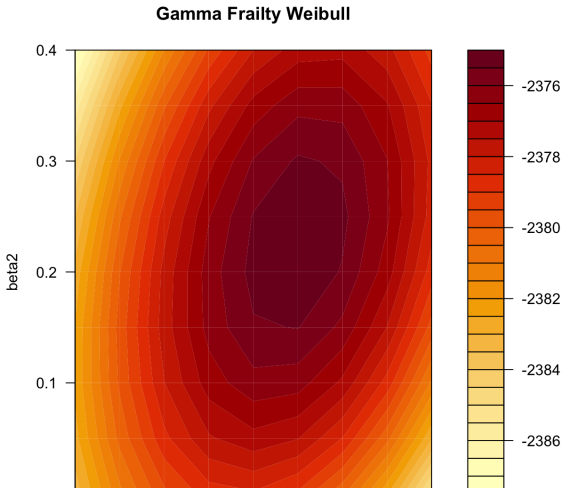
# Frailty Term and Missing Data

**MCAR**

If, by any case, one can verify their missing data are MCAR. A complete case analysis (CCA) is enough by MCAR definition. Unfortunately, our data was not this case.

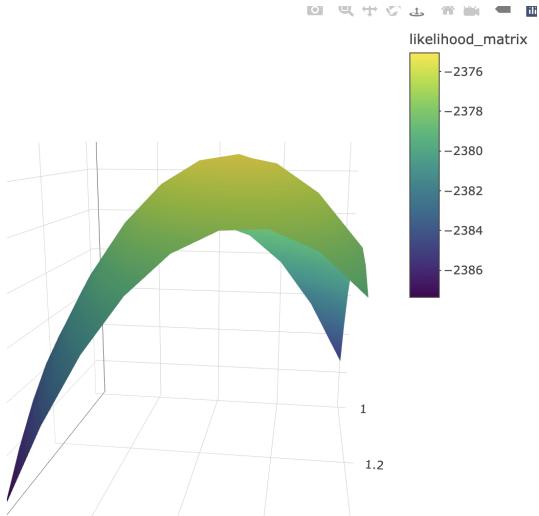
# Frailty Distributions

If we assume $z_j \sim \text{Gamma}(\upsilon, \upsilon)$, as shape and rate parameters. then the likelihood will look like these



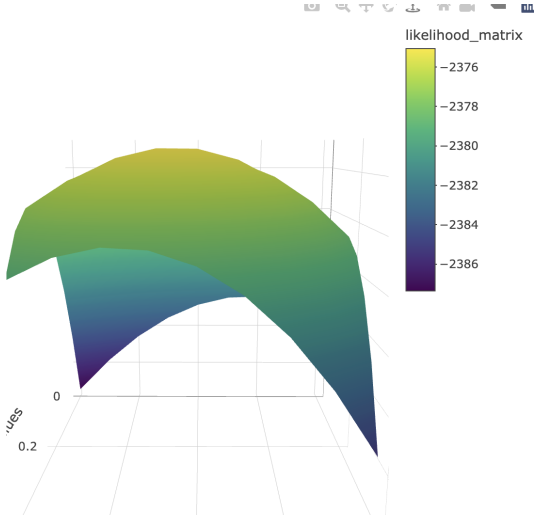
# Frailty Distributions

If we assume $z_j \sim \text{Gamma}(\upsilon, \upsilon)$, as shape and rate parameters. then the likelihood will look like these



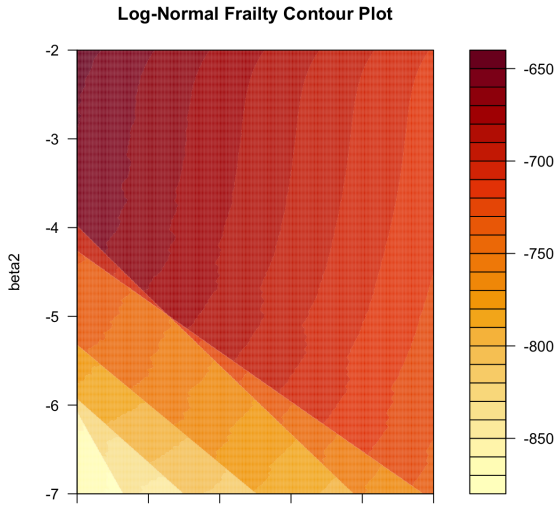
# Frailty Distributions

If we assume $z_j \sim \text{Gamma}(\upsilon, \upsilon)$, as shape and rate parameters. then the likelihood will look like these

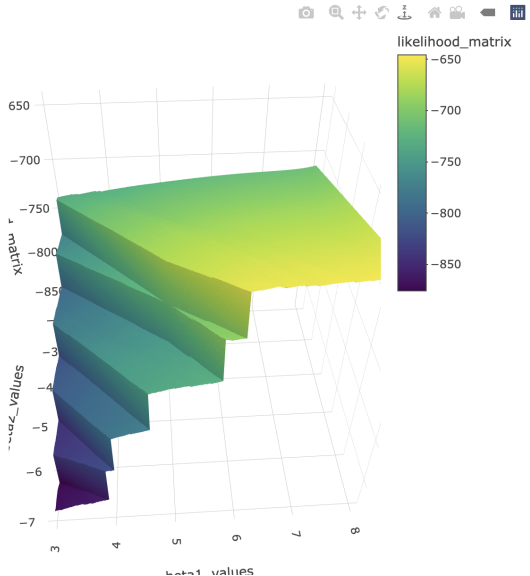


# Frailty Distributions

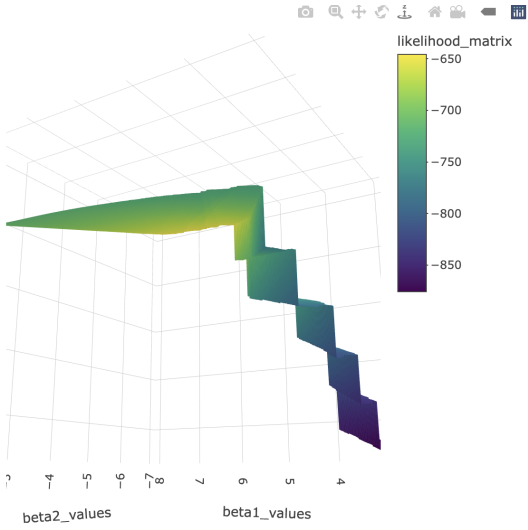
If we assume $z_j \sim \text{logN}(0, \upsilon^2)$, then the likelihood will look like these



# Frailty Distributions



# Frailty Distributions




frailty 3D 2.jpeg

# Frailty Term and Missing Data

## MAR

However, the MCAR is a very strong assumption, and usually not verified. By definition, we need to take the expectation with respect to the frailty term $z_j$ and the missing data $\mathbf{x}_{ij,mis}$. Assume the frailty distribution is chosen to be $f(z_j|\upsilon)$, and one might take the missing PRS $x_{ij,1,mis} \sim N(\psi_0 + \psi_1 x_{ij,2,obs}, \tilde{\psi}^2)$ to sample from.

$$E(\ell_C(\boldsymbol{\theta})|\boldsymbol{\theta}^{(r)}) = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\int_{\mathbf{x}_{mis}}\int_{z_j}\Big(\delta_{ij}\log h(t_{ij}|\mathbf{x}_{ij},z_j) - H(t_{ij}|\mathbf{x}_{ij},z_j)\Big) \tag{19}$$

$$\times f(\mathbf{x}_{ij,mis}|\mathbf{x}_{obs,ij},\psi^{(r)})f(z_j|\upsilon^{(r)})d\mathbf{x}_{ij,mis}dz_j \tag{20}$$

$$-\sum_{j=1}^{J}\int_{\mathbf{x}_{mis}}\int_{z_j}\log(1-\exp(z_j H_{j_p}(a_{j_p}|\mathbf{x}_{j_p}))) \tag{21}$$

$$\times f(\mathbf{x}_{ij,mis}|\mathbf{x}_{obs,ij},\psi^{(r)})f(z_j|\upsilon^{(r)})d\mathbf{x}_{ij,mis}dz_j \tag{22}$$

# Gamma Frailty Term and Missing Data

We can integrate $z_j$ in Gamma frailty via Laplace transform, assuming $z_j \sim \text{Gamma}(k, k)$, which will yield a closed-form likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{j=1}^{J} \left[ \sum_{i=1}^{n_j} (\delta_{ij} \log h(t_{ij}|\mathbf{x}_{ij})) + \log \left( \frac{(k + d_j - 1)!}{k! k^{d_j - 1}} (1 + \frac{\sum_{i=1}^{n_j} (H(t_{ij}|\mathbf{x}_{ij}))}{k})^{-k-d_j} \right) \right] \tag{23}$$

The ascertainment term

$$\begin{aligned} A_j(\boldsymbol{\theta}) &= 1 - S_{p_j}(a_{p_j}|\mathbf{x}_{p_j}) && (24) \\ &= 1 - \int_0^\infty S_{p_j}(a_{p_j}|\mathbf{x}_{p_j}, z_j) f(z_j) dz_j && (25) \\ &= 1 - \int_0^\infty \exp(-z_j \cdot H_{p_j}(a_{p_j}|\mathbf{x}_{p_j})) f(z_j) dz_j && (26) \\ &= 1 - (1 + \frac{H_{p_j}(a_{p_j}|\mathbf{x}_{p_j})}{k})^{-k} && (27) \end{aligned}$$

# Log-Normal Frailty Term and Missing Data

We can integrate $z_j$ in Log-Normal frailty via Gauss-Hermite Quadrature, which will yield a closed-form likelihood

**Definition**

In numerical analysis, the method can be applied in the following form:

$$\int_{-\infty}^{\infty} \exp(-x^2) f(x) dx \approx \sum_{i=1}^{n} \omega_i f(x_i) \tag{28}$$

where $n$ is number of sample points used, and $x_i$ is the roots of Hermite polnomial $H_n(x)$ such that $i = 1, ..., n$, and the weights $\omega_i$ is

$$\omega_i = \frac{2^{n-1} n! \sqrt{n}}{n^2 [H_{n-1}(x_i)]^2} \tag{29}$$

# Log-Normal Frailty Term and Missing Data

$q$ denotes the $q$-th element of Gauss Hermite Quadrature, i.e., $\omega_q$ denotes the $q$-th weight, $y_q$ denotes the $q$-th node, and $N_q$ denotes the total number of quadratures. Thus, substituting into the log-likelihood:

$$\ell_j(\boldsymbol{\theta}) = \sum_{i=1}^{n_j} \delta_{ij} \log(h(t_{ij}|\mathbf{x}_{ij})) + \log\left(\frac{1}{\sqrt{\pi}} \sum_{q=1}^{N_q} \left[\omega_q \exp(\sqrt{2}\sigma y_q)^{d_j} \exp\left(-\sum_{i=1}^{n_j} H(t_{ij}|\mathbf{x}_{ij}) \exp(\sqrt{2}\sigma y_q)\right)\right] \right. \tag{30}$$

Similarly, the ascertainment correction in the log-normal frailty can be written as

$$A_j(\boldsymbol{\theta}) = 1 - \int_{-\infty}^{\infty} \exp(-zH(a_{j_p}|\mathbf{x}_{j_p}))f(z)dz \tag{31}$$

$$= 1 - \sum_{q=1}^{N_q} \omega_q \exp\left(-(\sum_{i=1}^{n_j} H(a_{j_p}|\mathbf{x}_{j_p})) \exp(\sqrt{2}\sigma y_{q_p})\right) \tag{32}$$

# Frailty Term and Missing Data

$$\ell_{C_j} = \ell_j(\boldsymbol{\theta}) - \log A_j(\boldsymbol{\theta}) \tag{33}$$

# Frailty Term and Missing Data - Monte Carlo Expectation Maximization (MCEM)

For efficient sampling on missing data,

$$f(z_j, \mathbf{x}_{mis,ij}|\mathbf{x}_{obs,ij}, \boldsymbol{\theta}^{(r)}) \propto f(t_{ij}, \delta_{ij}|\mathbf{x}_{mis,ij}, \mathbf{x}_{obs,ij}, z_j, a_{j_p}, \boldsymbol{\beta}^{(r)}) \tag{34}$$

$$\times f(\mathbf{x}_{mis,ij}|\mathbf{x}_{obs,ij}, \psi^{(r)}) f(z_j|\upsilon^{(r)}) \tag{35}$$

Clearly, we know $f(t_{ij}, \delta_{ij}|\mathbf{x}_{mis,ij}, \mathbf{x}_{obs,ij}, z_j, \boldsymbol{\beta}^{(r)})$ is the likelihood of one single observation $j$ in family $j$, also we know the distribution of $f(\mathbf{x}_{ij}|\psi)$, as well as the frailty distribution $f(z_j|\upsilon)$. Therefore, in our case, we can write

$$f(z_j, \mathbf{x}_{mis,ij}|\mathbf{x}_{obs,ij}, \boldsymbol{\theta}^{(r)}) \propto f(z_j|\upsilon^{(r)}) \Big[ \prod_{i=1}^{n_j} f(\mathbf{x}_{mis,ij}|\mathbf{x}_{obs,ij}, \psi^{(r)}) \tag{36}$$

$$\times h^{(r)}(t_{ij}|\mathbf{x}_{ij}, z_j)^{\delta_{ij}} \exp(-H^{(r)}(t_{ij}|\mathbf{x}_{ij}, z_j)) \Big] \tag{37}$$

# Frailty Term and Missing Data - MCEM

In general, without the specification of the frailty distribution, the E-step in MCEM can be written as

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \sum_{j=1}^{J} \frac{1}{M_j} \sum_{m=1}^{M_j} \sum_{i=1}^{n_j} \left( \delta_{ij} \log h(t_{ij}|\mathbf{x}_{ij}^{(m)}, z_j^{(m)}) - H(t_{ij}|\mathbf{x}_{ij}^{(m)}, z_j^{(m)}) \right) \tag{38}$$

$$+ \sum_{j=1}^{J} \frac{1}{M_j} \sum_{m=1}^{M_j} \log(1 - \exp(z_j H_{j_p}(a_{j_p}|\mathbf{x}_{j_p}))) \tag{39}$$

$$+ \sum_{j=1}^{J} \frac{1}{M_j} \sum_{m=1}^{M_j} \sum_{i=1}^{n_j} \log f(\mathbf{x}_{mis,ij}^{(m)}|\mathbf{x}_{obs,ij}, \psi) + \sum_{j=1}^{J} \frac{1}{M_j} \sum_{m=1}^{M_j} \sum_{i=1}^{n_j} \log f(z_j^{(m)}|\upsilon) \tag{40}$$

Note that we take $M_j$ samples of the missing data and calculate the mean.

# Kinship Matrix

Remember, we made an assumption of the distribution of the missing PRS:

$$x_{ij,1,mis} \sim N(\psi_0 + \psi_1 x_{ij,2,obs}, \tilde{\psi}^2) \tag{41}$$

Is this an adequate assumption?

# Kinship Matrix

No! This is a family-wise genetic study! So within-family correlations need to be accounted for!

$$\mathbf{x}_{mis,j,1} \sim MVN(\boldsymbol{\mu}, \tilde{\psi}_g^2 K + \tilde{\psi}_e^2) \tag{42}$$

such that $K$ is the kinship correlation matrix with diagonal of 1, and $\hat{\boldsymbol{\mu}} = \boldsymbol{\psi}_0 + \boldsymbol{\psi}_1 \mathbf{x}_{obs,j,2}$. $\tilde{\psi}_g$ accounts for the genetic standard errors, and $\tilde{\psi}_e$ accounts for the residual. The multivariate normal distribution is what we are sampling the missing PRS on family-wise.

# Frailty Term and Missing Data - MCEM

## M-Step

In the M-step, I will use Nelder-Mead method, because it is gradient free!

## Convergence Rule

The convergence criterion is

$$(\boldsymbol{\theta}^{(r+20)} - \boldsymbol{\theta}^{(r)})^2 < 10^{-4} \tag{43}$$

# Preliminary Analysis Results

Table: Parameter Estimates on BRCA1 Family (MCEM - Assumed MAR)

| Parameters | Gamma (CCA) | Log-Normal (CCA) | Gamma (MCEM) | Log-Normal (MCEM) |
|---|---|---|---|---|
| $\alpha$ | -4.10 | -10.91 | -4.71 | -19.05 |
| $\lambda$ | 1.06 | 1.41 | 0.84 | 1.12 |
| $\beta_1$ | 1.26 | -5.12 | 2.42 | 3.97 |
| $\beta_2$ | 0.23 | 6.62 | 0.34 | 0.42 |
| $\upsilon$ | 4.35 | 2.73 | 3.71 | 2.89 |

The Log-normal distribution can introduce a wider range of heterogeneity due to its ability to model right-skewed distributions effectively. This wider range means that, once the Log-Normal frailty is accounted for, the remaining baseline hazard needs to adjust significantly to fit the data, hence the more negative scale parameter. A smaller (more negative) scale parameter indicates a more rapid initial occurrence of events, with the rate of increase over time again depending on the shape parameter.

# Next Step

- Due to the computational cost - The MCEM runs 48+ hours, I have not obtained the preliminary result on the case of the MNAR, but I will.
- A simulation study based on the correlated family will be conducted.
- Stay tuned for the final thesis!

# References I


[1] Canadian Cancer Statistics Advisory Committee. Canadian cancer statistics 2023, 2023. URL `https://cancer.ca/en/research/cancer-statistics`.

[2] Colin C Pritchard. New name for breast-cancer syndrome could help to save lives. *Nature*, 571(7763):27–29, 2019.

[3] Yun-Hee Choi, Mary Beth Terry, Mary B Daly, Robert J MacInnis, John L Hopper, Sarah Colonna, Saundra S Buys, Irene L Andrulis, Esther M John, Allison W Kurian, et al. Association of risk-reducing salpingo-oophorectomy with breast cancer risk in women with brca1 and brca2 pathogenic variants. *JAMA oncology*, 7(4):585–592, 2021.

[4] Yun-Hee Choi, Hae Jung, Saundra Buys, Mary Daly, Esther M John, John Hopper, Irene Andrulis, Mary Beth Terry, and Laurent Briollais. A competing risks model with binary time varying covariates for estimation of breast cancer risks in brca1 families. *Statistical Methods in Medical Research*, 30(9):2165–2183, 2021.

# References II


[5] Lu Chen, Li Hsu, and Kathleen Malone. A frailty-model-based approach to estimating the age-dependent penetrance function of candidate genes using population-based case-control study designs: an application to data on the brca1 gene. *Biometrics*, 65(4): 1105–1114, 2009.

[6] Amy H Herring, Joseph G Ibrahim, and Stuart R Lipsitz. Frailty models with missing covariates. *Biometrics*, 58(1):98–109, 2002.

[7] Amy H Herring and Joseph G Ibrahim. Maximum likelihood estimation in random effects cure rate models with nonignorable missing covariates. *Biostatistics*, 3(3):387–405, 2002.

[8] Joseph G Ibrahim and Geert Molenberghs. Missing data methods in longitudinal studies: a review. *Test*, 18(1):1–43, 2009.

[9] Samuli Ripatti, Klaus Larsen, and Juni Palmgren. Maximum likelihood inference for multivariate frailty models using an automated monte carlo em algorithm. *Lifetime Data Analysis*, 8:349–360, 2002.

# References III

[10] Donald B Rubin. Multiple imputation. In *Flexible Imputation of Missing Data, Second Edition*, pages 29–62. Chapman and Hall/CRC, 2018.

[11] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

# Q&A

Question Time!