Slide 1: Some of you are in my lab session of Biostats course, you might be super angry and curious on what this guy was doing when I did not respond to your emails on time. So, I will present my research to beg your forgiveness.

Slide 2: According to the official department, the breast cancer has been a second leading cause of cancer-related disease among females. There are already many studies to understand how to accuse the genetic effect such as genes BRCA1/2. The model on the time-to-cancer based on the gene and the PRS is also investigated, but there are challenges such as missing data!

Slide 3: Here is a pedigree tree showing one family structure of the data…

Slide 4: The problem arises on how we are going to handle the missing data when it is a clustered dataset, especially with genetic links. Although we have many proposed models on the clustered survival analysis, which is called the frailty model, the missing data still remains a problem when making the inference.

Slide 5: If you are a graduate student, you probably have taken the missing data course by Dr. Zou, and there are three missing mechanisms to be considered. When MCAR, the missingness does not depend on anything, when MAR, there may be some observed covariates related to the missingness, but not the covariate contains missing itself.

Slide 6: In the Cox proportional hazard model, which is more frequently used in the epidemiological study, you may have seen the first formula. But in order to attain a higher power of the sample size, a parametric survival analysis is preferred. WLOG, we will assume a Weibull baseline hazard in the parametric setting.

Slide 7: In most of situations, the model is being evaluated from a likelihood to obtain the highest possible model parameters. In the complete data likelihood, that is saying frailties and missing data are completely observed, we can simply follow the procedure. Moreover, due to the nature of the genetic study, we need to correct the ascertainment bias to make conditions on the likelihood, which can be simply derived from the Bayes' Rule.

Slide 8: Which will lead to the following formula after we define the probability of the proband being affected in the study.

Slide 9: If you have taken the calculus, you know that the likelihood using product is hard to be optimized, so we take the log will yield the following equation. As you may notice, the frailty term z is being seperated, but why?

Slide 10: In the survival analysis, if we have one function, such as hazard function, we can easily derive others such as survival function, cumulative hazard function. In the Weibull model, the hazard and the cumulative hazard functions can be written in the form that multiply the frailty term. Which answers our previous slide's question.

Slide 11: According to many publications, when the data is MCAR, we can simply run the analysis without considering the missing data at all. But our data was not this case.

Slide 12-14: I am saying this for a reason, let's delve into the likelihood plot, when we assume the frailty is Gamma distribution with rate and shape parameters being the same, the mean is 1 and variance is 1 over upsilon. The likelihood looks OK.

Slide 15-17: But if we assume the frailty is log-normal distribution, the plot is problematic with a flat top that gives us a hard time to optimize the model parameters.

Slide 18: Therefore, we need to account for the missing data and the frailty term jointly, and take the expectation with respect to both.

Slide 19: Before talking about the missing data, I want to show you how the frailty term is handled if there is no missing data. In the Gamma frailty, the integral can be calculated from the Laplace transform, which will give us a closed form likelihood as well as the ascertainment correction.

Slide 20: But in the log-normal case, due to it's not a power-variance-function, we are not able to get a closed form formula without using the numerical integral. Therefore, applying the Gauss-Hermite quadrature, which you first need to find your equation in the form like this, then estimate the integral using the following equation.

Slide 21-22: Which will give us an estimation of the log-likelihood and the ascertainment term.

Slide 23: Now let's get back to our missing data problem, we need to identify the joint distribution of the missing data and the frailty term since we cannot make a strong assumption of these two are completely independent. So I have obtained the posterior distribution for efficient sampling. We are able to conduct Gibb's sampler to find our target joint distribution by first sampling the missing data, then sampling the frailty iteratively.

Slide 24: Thus, we again can use the sampled M complete datasets do run the MCEM, that the expectation is now without the integral because we have sampled the joint distribution of the latent structures.

Slide 25: In the previous slide, I have made a little assumption that for example, when PRS is missing and we want it to be a normal distribution with the mean being estimated using a linear regression. But, it is not that simple for the genetic study.

Slide 26: Instead of assuming the univariate distribution, we want to make sure we capture the genetic associations between individuals within the family, so we will use a multivariate normal distribution for each family, and we will obtain a kinship matrix of diagonal of 1, because my genetic association with myself is clearly 1. So the sampling for the missing data is conducted on a family-wise.

Slide 27: In the M-step, I will use Nelder-Mead because we are lazy of obtaining the gradient and hessian. Due to the computational cost, the convergence rule is more lenient than usual EM algorithm.

Slide 28: Here is the preliminary result assuming missing at random for the BRCA1 family. As you can see, the log-normal case is largely pulled back from a biased result than complete case analysis.

Slide 29: Thank you very much for your listening, please stay tuned for my final thesis!