Research Areas in Statistical Genetics

Jaihee Choi

September 13, 2024

Background

- Twin studies have revealed the importance of the role of genetic heritability in many biological outcomes
- Advent of lower-cost sequencing has allowed scientists to study the genome more easily
- Genome-wide Association Studies (GWAS) have discovered thousands of genes that are linked with different outcomes

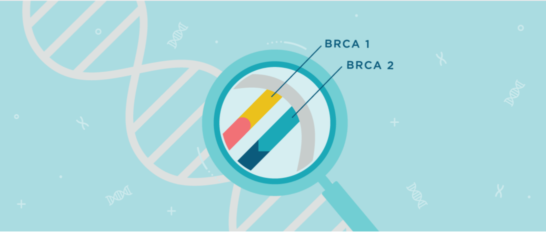


Figure: Source: https://www.rmany.com/blog/understanding-brca

Motivation

- Huge national and international efforts to build massive databases containing genetic and health information
- Information is often collected through periodic questionnaires
- Can be modeled as time-to-event (survival) outcomes

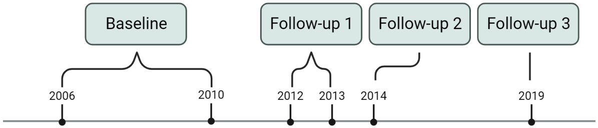
biobank uk

MVP MILLION VETERAN PROGRAM

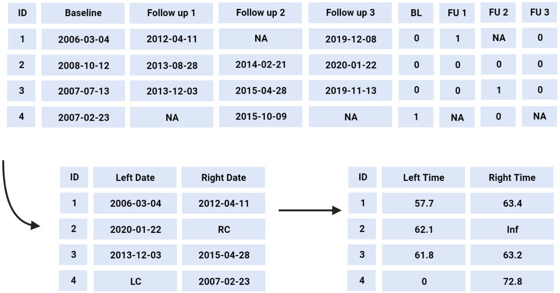
All of Us℠ THE FUTURE OF HEALTH BEGINS WITH YOU | The Precision Medicine Initiative®

Censored Time-to-Event Outcomes

- Much of the data is interval-censored due to the periodic questionnaires and repeat assessments.



Example: Fractured bone



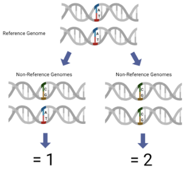
ID	Baseline	Follow up 1	Follow up 2	Follow up 3	BL	FU 1	FU 2	FU 3
1	2006-03-04	2012-04-11	NA	2019-12-08	0	1	NA	0
2	2008-10-12	2013-08-28	2014-02-21	2020-01-22	0	0	0	0
3	2007-07-13	2013-12-03	2015-04-28	2019-11-13	0	0	1	0
4	2007-02-23	NA	2015-10-09	NA	1	NA	0	NA

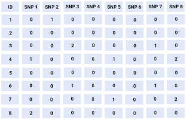
ID	Left Date	Right Date
1	2006-03-04	2012-04-11
2	2020-01-22	RC
3	2013-12-03	2015-04-28
4	LC	2007-02-23

ID	Left Time	Right Time
1	57.7	63.4
2	62.1	Inf
3	61.8	63.2
4	0	72.8

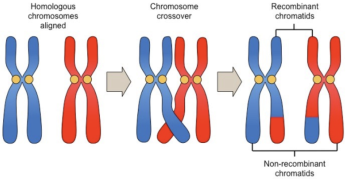
Challenges in this Work

- There currently are only a few methods tailored to interval-censored time-to-event outcomes
- Rare genetic variants are difficult to test for due to their low frequency in the population
- Complex correlation structures in the data due to linkage disequilibrium





ID	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7	SNP 8
1	0	1	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	0	0	2	0	0	0	1	0
4	1	0	0	0	1	0	0	2
5	0	0	0	0	0	0	0	0
6	0	0	1	0	0	0	1	0
7	0	0	0	0	1	0	0	2
8	2	0	0	0	0	0	0	0



Projects

Research focus: Developing robust and scalable tools to extract insights from complex data to better understand biological systems

Previous works:

- Set-based inference for genetic association with multiple interval-censored outcomes
- Interval-censored Bayesian variable selection for genome-wide association studies

Current interests:

- Multi-omic data integration methods with interval-censored outcomes
- Gene-environment (G x E) interaction methods for survival outcomes