Statistical Inference in Gene Association Studies of Cancer Risk with Partially Missing Family History Data

- **Ingo Ruczinski**
  Johns Hopkins Bloomberg School of Public Health, Baltimore, MD.

Timothy J. Jorgensen
Georgetown University, Washington, DC.

Abenaa M. Brewster
University of Texas, M.D. Anderson Cancer Center, Houston, TX.

Kathy J. Helzlsouer
Mercy Medical Center, Baltimore, MD.

Email: ingo@jhu.edu URL: http://biostat.jhsph.edu/~iruczins/

Abstract

Family history is an important cancer risk factor to consider in association studies of putative cancer risk alleles. If an allele is associated with risk, it should also account for some proportion of familial clustering of cancer. A major obstacle to evaluating genetic variants in the context of family history is often the lack of complete family history data.

The most common approach for dealing with missing data is to omit the observations that have missing records in the model's covariates. However, this approach has several shortcomings, including the loss of power, and the potential introduction of biases in the parameter estimates. As an alternative, multiple imputation can be used to draw valid statistical inference from data with missing values when the data are missing at random.

We used a decision tree based multiple imputation strategy in a case-control study of a breast cancer nested within a cohort of women in Washington County, Maryland (CLUE II Cohort). The risk associated with the polymorphisms in XPD (Lys751Gln) and XRCC1 (Arg399Gln and Arg194Trp) DNA repair genes, accounting for breast cancer family history, critically depended on the statistical handling of the missing data. We believe that this effect may account for some of the conflicting results of association studies between genetic polymorphisms and cancer risk, and stresses the importance of proper statistical approaches for handling missing data.

The Study

CLUE II Population: A nested case-control study was conducted using the population-based CLUE II cohort, established in 1989. Individuals residing in Washington County, Maryland, and surrounding regions were invited to donate blood for cancer and heart disease research. The CLUE II cohort consists of 32,892 individuals, including approximately 30% of county residents. Cancers that developed among cohort participants were ascertained through linkage to the Washington County and, since 1992, the Maryland State Cancer Registries. In 1996, and about every two years afterwards, participants were asked to complete a follow-up questionnaire asking about health events, medication use, and cancer risk factors.

Study Population: For this study, 321 cases of breast cancer that occurred after blood donation to CLUE II in 1989 were identified. Cases were excluded if they had a diagnosis of any other cancer except for non-melanoma skin cancer and cervical cancer in situ. Controls were individually matched to cases (1:1) by age at blood donation and menopausal status at blood donation. Selected controls were cancer-free.

Family History: Information on breast cancer risk factors was obtained from several sources, including a questionnaire that was sent in 1995 to a portion of the CLUE II breast cancer cases and controls who were part of a case-control study on organochlorine compounds. In addition, a 1996 follow-up questionnaire that was sent to all CLUE II participants. The questionnaires contained detailed information on family history of breast cancer, reproductive history, medication history, and selective dietary intake.

Throwing out Data

The most common approach for dealing with missing data is to omit the observations that have missing records in the model's covariates. This approach can have several shortcomings, including:

⟶ Loss of power.

⟶ Bias in the parameter estimates.

A good reference on this is Greenland S and Finkle WD (1995).

As an alternative, multiple imputation can be used to draw valid statistical inference from data with missing values when the data are missing at random (Little RJ and Rubin DB 1987, Schafer JL 1997).

In essence, multiple imputation acknowledges the uncertainty due to missing data, instead of simply ignoring it: several complete data sets are generated, and the uncertainty in the model parameter estimates incorporates the standard errors of the parameter estimates as well as the variability between the parameter estimates from the replicate data sets.

Note: While the hypothesis of missing at random cannot formally be tested, it is a lot less stringent than the requirement of missing completely at random, which is the underlying assumption made when observations are omitted.

Statistical Inference (1)

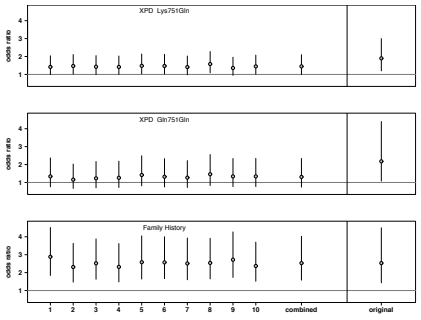
	Number of Pairs	Odds Ratio	Confidence Interval
XPD Lys751Gln			
original data set	202	1.90	(1.20 – 3.00)
multiple imputations	321	1.45	(1.00 – 2.10)
XPD Gln751Gln			
original data set	202	2.18	(1.08 – 4.40)
multiple imputations	321	1.31	(0.74 – 2.34)
Positive Family History			
original data set	202	2.53	(1.43 – 4.50)
multiple imputations	321	2.53	(1.58 – 4.03)

Missing Data Patterns

The polymorphisms of the DNA repair gene XPD (751) for case/control pairs and family history reporting status, in the breast cancer study. This highlights the fact that missing data can not simply be ignored.

	Family History not complete				Family History complete			
	AA	AC	CC	na	AA	AC	CC	na
raw numbers								
case	43	54	5	5	61	121	25	7
control	35	57	12	3	90	102	22	0
percentages								
case	40.2	50.5	4.7	4.7	28.5	56.5	11.7	3.3
control	32.7	53.3	11.2	2.8	42.1	47.7	10.3	0.0

Imputation



The missing data were imputed using decision trees. See Dai et. al. (2006) for details.

Statistical Inference (2)

	Number of Pairs	Odds Ratio	Confidence Interval
XRCC1 Arg399Gln			
original data set	199	1.91	(1.05 – 3.46)
multiple imputations	321	1.64	(1.00 – 2.69)
XRCC1 Gln399Gln			
original data set	199	1.33	(0.69 – 2.57)
multiple imputations	321	1.22	(0.73 – 2.06)
XRCC1 Arg/Trp194Trp			
original data set	204	1.40	(0.70 – 2.80)
multiple imputations	321	1.20	(0.72 – 2.00)

References

- Brewster AM, Jorgensen TJ, Ruczinski I, Huang HY, Hoffman S, Thuita L, Newschaffer C, Lunn RM, Bell D, Helzlsouer KJ (2006).
  Polymorphisms of the DNA Repair Genes XPD (Lys751Gln) and XRCC1 (Arg399Gln and Arg194Trp): Relationship to Breast Cancer Risk and Familial Predisposition to Breast Cancer.
  Breast Cancer Research and Treatment, 95(1): 73-80.
- Dai J, Ruczinski I, LeBlanc M, Kooperberg C (2006).
  A Comparison of Haplotype-based and Tree-based SNP Imputation in Association Studies.
  Under review.
- Greenland S, Finkle WD (1995).
  A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses.
  American Journal of Epidemiology, 142 (12), 1255-1264.
- Little RJ, Rubin DB (1987).
  Statistical Analysis with Missing Data.
  John Wiley Sons, New York.
- Schafer JL (1997).
  Analysis of Incomplete Multivariate Data.
  Chapman & Hall.

Grant Support: The National Cancer Institute Breast Spore P50CA88843, and the National Institute of Aging 5U01AG018033. I.R. was supported, in part, by the Maryland Cigarette Restitution Fund Research Grant to the Johns Hopkins Medical Institutions and NIH grant CA 074841. T.J.J. was funded by a Ruth L. Kirschstein Senior Fellow Award (NCI F33 CA09817-01). A.M.B was supported by Minority Supplement of Clinical Oncology Research Career Development Program (K12-CA01709) from the National Cancer Institute.