Statistical Issues in Studies of Genetic Susceptibility to Disease

Gerarda Darlington

April 3, 2002

Studies of disease etiology

- epidemiology
  - environmental risk factors
  - population based
  - family history reported

- genetic epidemiology
  - genetic risk factors
  - environmental variables?
  - family ascertainment

Some genetic concepts

- chromosomes contain genes
- locus is position on chromosome
- alleles are specific genes at locus

Some genetic concepts

Haplotype

- alleles found on group of loci on chromosome

$$\begin{array}{ll} A_1\text{-} & \text{-}A_1 \\ B_1\text{-} & \text{-}B_2 \\ C_3\text{-} & \text{-}C_5 \end{array}$$

3 loci: A, B, C

A locus: A_1, A_2, A_3

B locus: B_1, B_2

C locus: C_1, C_2, C_3, C_4, C_5

Some genetic concepts

A-   -A   a-   -a

homozygous

A-   -a

heterozygous

phenotype: observable characteristic

Some genetic concepts

- Mendelian (particulate) inheritance
- parental haplotypes
- equal probabilities

Genetic role in disease

- unknown genes
  - linkage studies
  - limited families
  - extreme outcomes (e.g. early onset)
- candidate genes
  - known biological function
  - information from other species

Genetic linkage

- specific "marker" locus or loci

- unknown disease susceptibility locus

- are marker and disease loci linked?

Genetic linkage

- affected sibling studies

- distribution of parental haplotypes

- haplotype sharing

Affected sibling studies

- unaffected siblings not included
- not susceptible
- susceptible but
  - late onset
  - never get disease

Affected sibling pairs

What data do we have?

Parent 1	Parent 2
ab	cd

possible offspring:
ac ad bc bd

observe 2 siblings:

e.g. 1: ac, ac

e.g. 2: ac, ad

e.g. 3: ac, bd

Affected sibling pairs

- share 0, 1, or 2 haplotypes identical by descent (IBD)

- under H_0: Mendelian inheritance
  observe:
  0 with probability 0.25
  1 with probability 0.5
  2 with probability 0.25

- chi-squared test

- other methods

Studies of candidate gene

A- -A A- -a a- -a

- 1 locus
- 2 possible alleles: A, a
- potential susceptibility allele: A
- allele frequency: p =Pr(A)

dominant indicator:
1 if AA or Aa; 0 if aa

recessive indicator:
1 if AA; 0 if Aa or aa

Candidate gene study designs

- association studies
- common disease
  - cross-sectional or cohort study
- rare disease
  - case-control study
  - family controls
- can include environmental factors
- generalized linear models

Challenges of data collection

- epidemiology
  - random sample
  - often data from questionnaire
  - non invasive

- genetic epidemiology
  - random sample?
  - need blood samples
  - family ascertainment

Candidate gene studies for common disease

- select individual from population (proband)

- data from proband and relatives

- correlated observations

- unbiased effect estimates

- standard error estimates incorrect

Candidate gene studies for common disease

- size of impact on standard error estimates
- study design; sample size calculations
- design effect

Design effect

Scott and Holt (1982) JASA;
Donner (1984) J Chron Dis

- 2 stage sampling
- regression context (OLS)
- identify “design effect” or inflation factor
- effect of omitting correlation

OLS design effect

- $E(Y) = X'\beta$
- $\widehat{\beta} = (X'X)^{-1}X'Y$
- $var_c = \sigma^2(X'X)^{-1}D$
- $D = (X'RX)(X'X)^{-1}$
- expression for single coefficient estimate?

OLS design effect

- assuming common residual correlation: ρ_y
- $D_{EX} = I + \rho_y(M - I)$
- single covariate
- $DF = 1 + (m-1)\rho_x\rho_y$
- $var_c(\widehat{\beta}) = [var_I(\widehat{\beta})]DF$
- ρ_x: within subject correlation wrt x
- ρ_y: residual correlation

Design effect

Neuhaus and Segal (1993) Stats in Med

- binary outcomes
- several link functions
- simulation studies
- $DF = [1 + (m - 1)\rho_x \rho_y]$ still valid

Design effect for association studies

- $DF = [1 + (m - 1)\rho_x \rho_y]$
- often $\rho_y > 0$
- direction of DF depends on ρ_x
- $\widehat{\rho}_x = 1 - (mXX_W)/[(m - 1)XX_T]$
- multiple covariates
- association studies

ρ_x for genetic studies

association studies

- dominance model:
  $x_{ij} = 1$ if AA or Aa; 0 if aa
  $E(\widehat{\rho}_x) \approx (1 - 0.75p)/(2 - p)$

- recessive model:
  $x_{ij} = 1$ if AA; 0 otherwise
  $E(\widehat{\rho}_x) \approx (0.25 + 0.75p)/(1 + p)$

- range$= (0.25, 0.5)$

Variable family size

- variable family sizes, m_i
- use $\tilde{m} = \sum m_i^2 / \sum m_i$
- $DF = [1 + (\tilde{m} - 1)\rho_x \rho_y]$

Simulation study

Shin (1998) MSc thesis

- non exchangeable correlation
- variable family size
- 100 nuclear families

Simulation study

- generated 2 parents
- generated sibships
- geometric distribution for sibship sizes
- mean sibship size of 2.2
- 2 candidate genes (CG's)
- 1 environmental factor (EF)

Simulation study

- allele frequency $p = 0.4$
- dominant effect assumed for CG1
- additive effect assumed for CG2
- EF normally distributed

Simulation study

- disease status
- $exp(u_{ij} + s_{ij})/[1 + exp(u_{ij} + s_{ij})]$
- $u_{ij} = X'_{ij}\beta$
- s_{ij} additional unmeasured shared factors

Simulation study

- $\beta_0 = -6$; $\beta_{CG1} = 2$; $\beta_{CG2} = 2$; $\beta_{EF} = 0.01$
- disease prevalence of 25%
- 100 families generated
- logistic regression modelling
- repeated 500 times

Simulation study

- variance of coefficient estimates
- average of independence variance estimates
- compute simulation design effect
- compute DF
- $\rho_y = 0.3$
- $\rho_x = 0.5, 0.5, 0.7$ for CG1, CG2, EF

Simulated Versus Approximated Design Effects

Covariate	Simulation Design Effect	DF
CG1	1.69	1.72
CG2	1.61	1.72
EF	1.91	2.01

Continuous trait example

GAW9 (1995) Genet Epidemiol

- 23 extended families
- common complex trait
- A3, age, sex, EF
- all immediate relatives
  - $\rho_y = 0.14$; $\rho_x = 0.5, -0.02, -0.05, 0$
- siblings only
  - $\rho_y = 0.33$; $\rho_x = 0.28, 0.47, 0.08, -0.11$

Observed Versus Approximated Design Effects from GAW9 Data On 23 Extended Families

Covariate	Family Group	Univar. Ratio	Multivar. Ratio	DF
A3	All	1.54	1.80	1.59
	Sibs	1.25	1.56	1.37
Age	All	1.17	1.37	0.98
	Sibs	1.35	1.14	1.39
Sex	All	0.64	0.58	0.94
	Sibs	0.85	0.79	1.10
EF	All	1.02	0.94	1.00
	Sibs	0.77	0.83	0.96

Final remarks

- design must consider correlation
- DF adjustment of independence sample size
- need $\tilde{m}, \rho_x, \rho_y$
- $\rho_x \leq 0.5$ for candidate gene
- range of 0.1 to 0.3 for ρ_y
- efficient designs

Acknowledgements

Collaborators: S.B. Bull,

N.H. Chapman, C.M.T. Greenwood, J. Shin

Funding agencies:

NSERC

MITACS