Special Topics in (Genetic) Epidemiology

Li Hsu
lih@fredhutch.org
Outline

- Course overview
- Course main topics
- Basics
  - Common epidemiologic study designs
  - Basic genetics
Acknowledgements

- Bhramar Mukherjee (U Michigan)
- Shawn Lee (U Michigan)
- James Dai
- Chongzhi Di
- Charles Kooperberg
- Ross Prentice
- Yingye Zheng
Overview

- This course will review many (but selected) current developments in statistical methods for (genetic) epidemiologic studies
  - Cover some basic models and standard analysis techniques but is not meant to be comprehensive
  - Focus on opportunities and challenges
- Things that are not covered:
  - Data preprocessing and QC steps
  - Software packages
Main topics

- Various topics that are related to observational studies
- Why?
  - A lot of data (genetics, environmental risk factor, biomarkers, various molecular data on the phenotypes)
  - Lots of interesting questions
  - Messy and complex
Topics: Genetic association studies

- This course will review many statistical methods in genetic association studies
  - Single variant analysis and confounding; effect size estimation and winner’s curse
  - Meta-analysis
  - Set-based (genetic) association analysis
- No biology prerequisites
Topics: Environmental risk factors

- Measurement error (Ross Prentice)
- Functional data analysis and applications to high-resolution bio-signal data from wearable devices (Chongzhi Di)
Topics: Genetics and Environment

- Gene-environment interaction, genome-wide search strategies and machine learning approaches (Charles Kooperberg)
- Mendelian randomization and instrumental variable analysis (James Dai)
Topics: Study design and risk estimation

- Current developments of sub-sampling study designs for epidemiologic and biomarker studies (Yingye Zheng)
- Absolute risk estimation from case-control and cohort studies
Evaluation

- Credit/No Credit course; audit is welcome
- No homework, but there will be recommended readings
- Final will be a 20-30 minute presentation of a self-chosen topic
  - An in-depth reading of a particular topic covered in this course or related to the research interests of (guest) lecturers
Complex Diseases

- Many complex diseases are contributed by both genes and environment
- Twin studies suggest that about 30% of colorectal cancer risk is due to genetic factors

**Colorectal Cancer: Etiology**

- Environmental Exposures
- Genetic Susceptibility
- Colorectal Cancer
Observational Studies

- Powerful tools for studying disease etiology (risk factors or disease causation)
- Investigator has no control over exposure
- Two common study designs
  - Case-control studies
  - Cohort studies
Case-control studies

- Identify risk factors for a disease/outcome

Example: Diet, Activity, and Lifestyle Study (DALS)

- **DALS** is a population-based case-control study of colon cancer.

- **Case-definition** Participants were recruited from three locations: the Kaiser Permanente Medical Care Program (KPMCP) of California, Utah, and Minnesota. Eligibility criteria included age at diagnosis of primary colorectal cancer between 30 and 79 years in 1991-1994. Individuals with familial adenomatous polyposis, Crohn's disease, or ulcerative colitis were excluded.

- **Control-definition** Controls from KPMCP were randomly selected from membership lists. In Utah and Minnesota, controls were randomly selected through random-digit dialing and driver license lists.

- **Matching** Cases and controls were matched by age (5-year) and sex.
Case-Control Data Analysis

- $Y$: Disease status
- $Z$: Covariates
- Logistic regression model

$$
Pr(Y = 1|Z) = \frac{\exp(\beta_0 + \beta'Z)}{1 + \exp(\beta_0 + \beta'Z)}
$$

(1)
Bayes’ rule and logistic regression model (1) imply that

\[
\frac{\Pr(Z|Y = 1)}{\Pr(Z|Y = 0)} = \frac{\Pr(Z, Y = 1)/\Pr(Y = 1)}{\Pr(Z, Y = 0)/\Pr(Y = 0)} = \frac{\Pr(Y = 1|Z)\Pr(Y = 0)}{\Pr(Y = 0|Z)\Pr(Y = 1)} = \frac{\Pr(Y = 0)}{\Pr(Y = 1)} \exp(\beta_0 + \beta'Z) \quad (2)
\]
Case-Control Data Analysis

- Now imagine cases and controls are members of a second, hypothetical population of individuals whose disease probability is \( \pi \), the proportion of sampled subjects are cases, but the covariate distribution \( \Pr(Z|Y = 1) \) and \( \Pr(Z|Y = 0) \) still satisfy (2).
- In this hypothetical population, from Bayes’ rule

\[
P_Z \equiv \Pr(Y = 1|Z) = \frac{\pi \Pr(Z|Y = 1)}{\pi \Pr(Z|Y = 1) + (1 - \pi) \Pr(Z|Y = 0)}
\]

\[
= \frac{\pi/(1 - \pi) \Pr(Z|Y = 1)/\Pr(Z|Y = 0)}{1 + \pi/(1 - \pi) \Pr(Z|Y = 1)/\Pr(Z|Y = 0)}
\]

\[
= \exp(\beta^* + \beta Z)
\]

\[
= \frac{1 + \exp(\beta^* + \beta Z)}{1 + \exp(\beta^* + \beta Z)}
\]

- \( \beta^* = \beta_0 + \log\{\pi/(1 - \pi)\} - \log\{\Pr(Y = 1)/\Pr(Y = 0)\} \)
- \( \beta_0 \): Baseline disease probability is not identifiable
Case-Control Data Analysis

- Constraint
  \[
  \int P_Z f(Z) dZ = \pi
  \]

- It happens that the constraints are satisfied when maximizing the likelihood function based on \( P_Z \). The score equation corresponding to \( \beta_0^* \) solves the same constraint. Suppose the data consist of \((Y_i, Z_i), i = 1, \ldots, n\)

  \[
  \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} P_{Z_i} = 0
  \]

- Despite case-control studies are retrospective in nature, the data can be analyzed as if they were prospectively collected using a logistic model and the odds ratio approximates the relative risk if the disease prevalence is low.


Confounding

- Exposure of interest may be confounded by a third factor that is associated with exposure and the disease.
Control for confounding

- At the design phase:
  - Sampling cases and controls from the same targeted population
  - Matching controls to cases on factors that are potentially important for disease (e.g., age, sex)
  - If these factors are fixed to be the same in the cases and controls then they can not confound the association

- At the analysis phase:
  - Multivariable adjustment or stratification
Case-Control Designs

- Advantages:
  - Quick and cheap (relatively)
  - Best for rare diseases
  - Evaluation of multiple exposures

- Concerns:
  - Selection bias: Cases and controls selected on criteria related to the exposure
  - Recall bias: Presence of disease may affect ability to recall or report the exposure
  - Beware of reverse causation: the disease results in a change in behavior (exposure)
(Prospective) Cohort Design

- Cohort is a group of people with something in common, usually an exposure or a defined population group, identified before occurrence of disease under investigation.
- The study population is followed over a period of time to determine the frequency of disease.
- Study the association of exposures with diseases in this group of people.
The Women’s Health Initiative (WHI) initiated in 1991 consists of three clinical trials (hormone therapy, dietary modification and calcium/vitamin D) and an observational study.

Investigate major health issues causing morbidity and mortality (e.g., cardiovascular disease, cancer, and osteoporosis) in postmenopausal women.

WHI enrolled more than 160,000 postmenopausal women aged 50 - 79 years (at time of study enrollment) over 15 years.
Cohort Data Analysis

- Survival analysis with time to event outcome
- Suppose $T$ is time to the event of interests and $Z$ are covariates
  \[ \lambda(t|Z) = \lambda_0(t) \exp(\beta'Z) \]
- $\lambda(t|Z) = \lim_{\Delta t \rightarrow 0} \Pr(t \leq T < t + \Delta t | T \geq t, Z)$
- Longitudinal analysis with repeated (or longitudinal measurements)
- Suppose $Y_{ij}$ is the measurement (e.g., blood pressure) for the $i$th subject at $t_{ij}$th time
  \[ \mathbb{E}(Y_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2' Z_{ij} \]
Subsampling Design

- Case-cohort and nested case-control study designs.
- Commonalities:
  - Sampling from a prospective cohort where disease outcomes and some baseline information are known for all the individuals
  - Include all individuals who develop the disease during follow-up (cases)
- Differences:
  - In case-cohort studies, controls come from a subcohort sampled from the entire cohort at baseline, while in nested case-control designs, controls are sampled from individuals at risk at the times when cases are identified.
- Dr. Yingye Zheng will cover (newest) methods development for subsampling data
Example: WHI

- Cases included colorectal cancer cases from 2009 database
- A control was selected for each case from the risk set at the time of the case’s diagnosis
- Additional matching variables: age, race, trial arm/observation, and center
- Exclusion criteria: a prior history of colorectal cancer at baseline, IRB approval not available for data submission into dbGaP, and not sufficient DNA available.
Cohort Studies

- **Advantages:**
  - Temporality can be established
  - Several outcomes related to exposure can be studied simultaneously
  - Best for relatively common diseases or rare exposures

- **Concerns:**
  - Large population is needed; time consuming and expensive
  - Study itself may alter people’s behavior
  - Cohort subjects may differ from the general population because of eligibility criteria and characteristics related to their self-selection
The central dogma describes the flow of genetic information in cells from DNA to messenger RNA (mRNA) to protein.

Central Dogma of Gene Expression.

Through the production of mRNA (transcription) and the synthesis of proteins (translation), the information contained in DNA is expressed.
Human Genome

- 3 billion base pairs
- 22 autosomes and 2 sex chromosomes
- Approximately 20,000 protein-coding genes
- Protein-coding sequences account for only a very small fraction of the genome (\( \sim 1.5\% \))
- Other types of genes
  - Non-coding RNA: \( \sim 20,000 \)
  - Pseudogenes: \( \sim 14,000 \)
Human Chromosomes

![Bar chart showing total genes and base pairs for each chromosome in humans.](chart.png)
DNA Mutations

- DNA mutations can occur because
  - DNA damage from environmental agents
  - Mistakes occur when a cell copies its DNA in preparation for cell division
- Mutations are essential to evolution; they are the raw material of genetic variation
DNA Mutations

- **Substitution**
  - CTGGAG
  - CTGGTG

- **Insertion**
  - CTGGGG
  - CTGGCTGG

- **Deletion**
  - CTGGCCCTGG
  - CTGGGG

- **Frameshift**
  - ATG TCG AAT
  - TGT CGA AT CGA
Single Nucleotide Polymorphism (SNP)

▶ SNP is the most common form of polymorphisms
SNP

- **Locus**: Specific location of variant on a chromosome.
- **Allele**: One of a number of alternative forms of the same gene (variant) in a specific locus.
- **Previous example**: A vs G
- **Suppose G is a major allele and A is a minor allele**
- **Homozygote**: individuals with identical pairs of alleles
  - GG: major allele homozygote
  - AA: minor allele homozygote
- **Heterozygote**: individuals with two different alleles
  - AG
SNP

- Number of dbSNP > 60 millions
- The density plot of the minor allele frequencies
Inheritance

- Inheritance mechanism makes the following characteristics
  - One locus: Hardy-Weinberg Equilibrium
  - Multiple loci: Linkage Disequilibrium
Hardy-Weinberg Law

- Hardy-Weinberg equilibrium (HWE) means the frequency of a diploid genotype is the product of the frequencies of its constituent alleles.

- Suppose a locus $A$ has two alleles, $A_1$ and $A_2$.

- $p$ is the frequency of allele $A_1$ and $1 - p$ is the frequency of allele $A_2$.

- Genotype frequency?

\[ \Pr(A_1 A_1), \quad \Pr(A_1 A_2), \quad \Pr(A_2, A_2) \]

- At HWE

\[ p^2 \] is the frequency of $A_1 A_1$ homozygotes.

\[ 2p(1 - p) \] is the frequency of $A_1 A_2$ heterozygotes.

\[ (1 - p)^2 \] is the frequency of $A_2 A_2$ homozygotes.
Properties of HWE

- HWE occurs when the two alleles of an individual are random draws from the population. One generation of random mating produces HWE.

- Allele frequency of the next generation

\[
\Pr(A_1) = \Pr(A_1A_1) + \Pr(A_1A_2)/2 = p^2 + p(1 - p) = p
\]

- Allele frequency doesn’t change

- A variety of factors can disturb the equilibrium
  - Inbreeding and other forms of non-random mating
  - Subdivision of the population
  - Natural selection and genetic drifts
HWE

- HWE had a profound effect in early genetics
- In genetic association studies, HWE is primarily used to check genotyping quality
  - Association studies assume that samples are unrelated individuals in HWE.
  - Genotype calls are very precise with error rate $\sim 10^{-3}$.
  - Genotyping technology may be affected by sample preparation, DNA quality, lab conditions, and SNP conditions. Badly called SNPs may be out of HWE.
Genotype QC

> Example: Suppose 10% of $A_1A_2$ was mis-called to $A_1A_1$.
> Probabilities of observed genotypes are

\[
\hat{\Pr}(A_1A_1) = p^2 + 2p(1-p) \times 0.1, \quad \hat{\Pr}(A_1A_2) = 2p(1-p) \times 0.9
\]

> Based on the observed genotypes we can calculate the allele frequency

\[
p' = (p^2 + 2p(1-p) \times 0.1) + p(1-p) \times 0.9 = p + 0.1p(1-p)
\]

> By HWE, the probability of $A_1A_1$ should be

\[
p^2 + 2p(1-p) \times 0.1 + 0.1^2 p^2(1-p)^2.
\]
Inferring population substructure

- HWE has also been used for deriving population substructure
- Suppose in the sample there are two sub-populations Pop1: Pop2 = 1:1
- The allele frequency is: 0.7(Pop1) and 0.3(Pop2)

\[
\begin{align*}
\Pr(A_1A_1) &= (0.7 \times 0.7 + 0.3 \times 0.3)/2 = 0.29 \\
\Pr(A_1A_2) &= (2 \times 0.3 \times 0.7 + 2 \times 0.3 \times 0.7)/2 = 0.42
\end{align*}
\]

- The allele frequency is 0.29+0.42/2 = 0.50
- Under HWE, \( \Pr(A_1A_1) = 0.25 \) and \( \Pr(A_1A_2) = 0.50 \)
Test for HWE

- Compare observed genotypes vs expected genotypes from HWE
- Pearson $\chi^2$ test:
  \[ T = \sum_{j}^{3} \frac{(O_j - E_j)^2}{E_j} \]
- Test statistic follows $\chi_1^2$ distribution
- Fisher exact test can be used
Recombination

Crossover

Homologous chromosomes aligned

Chromosome crossover

Recombinant chromosomes

Non-recombinant chromosomes
Recombination

- Introduce genetic diversity.
- Crossovers more likely to occur between genes that are further away; likelihood of a recombination event is proportional to the distance
- Allows for mapping genes
Linkage disequilibrium (LD)

- Two loci $A$ (two alleles $A_1$ and $A_2$), $B$ (two alleles $B_1$ and $B_2$)
- Haplotype (alleles that are located closely together and that tend to be inherited together)
- Frequency of haplotype and allele

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1B_1$</td>
<td>$x_{11}$</td>
</tr>
<tr>
<td>$A_1B_2$</td>
<td>$x_{12}$</td>
</tr>
<tr>
<td>$A_2B_1$</td>
<td>$x_{21}$</td>
</tr>
<tr>
<td>$A_2B_2$</td>
<td>$x_{22}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Allele</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>$p_1 = x_{11} + x_{12}$</td>
</tr>
<tr>
<td>$A_2$</td>
<td>$p_2 = x_{21} + x_{22}$</td>
</tr>
<tr>
<td>$B_1$</td>
<td>$q_1 = x_{11} + x_{21}$</td>
</tr>
<tr>
<td>$B_2$</td>
<td>$q_2 = x_{12} + x_{22}$</td>
</tr>
</tbody>
</table>

- Under no LD
  \[ D = x_{11} - p_1 q_1 = 0 \]
- When only genotypes are available, EM algorithm can be used to estimate haplotype frequencies
Linkage Disequilibrium (LD)

- LD block structure for a particular region.
LD

- LD is an extremely important feature in genetic data.
- It allows to investigate diseases with fewer markers.
LD makes it hard to adjust for the multiple tests and to find causal variants.
Summary

Today we cover

- Course overview
- Epidemiologic study design
- Genome, DNA-mutation
- Hardy-Weinberg equilibrium
- Linkage disequilibrium
Recommended to read