Epidemiologic Study Designs

Lucia Hindorff
Epidemiologist
Office of Population Genomics

Outline

- Learning objectives
- Study designs
  - Overview
  - Case-control studies
  - Cohort studies
  - Randomized/experimental designs
- The road to GWA studies
  - Overview
  - Family studies
  - Candidate genes
  - Genome-wide association (GWA) studies

Outline

- Learning objectives
- Study designs
  - Overview
  - Case-control studies
  - Cohort studies
  - Randomized/experimental designs
- The road to GWA studies
  - Overview
  - Family studies
  - Candidate genes
  - Genome-wide association (GWA) studies

Learning objectives

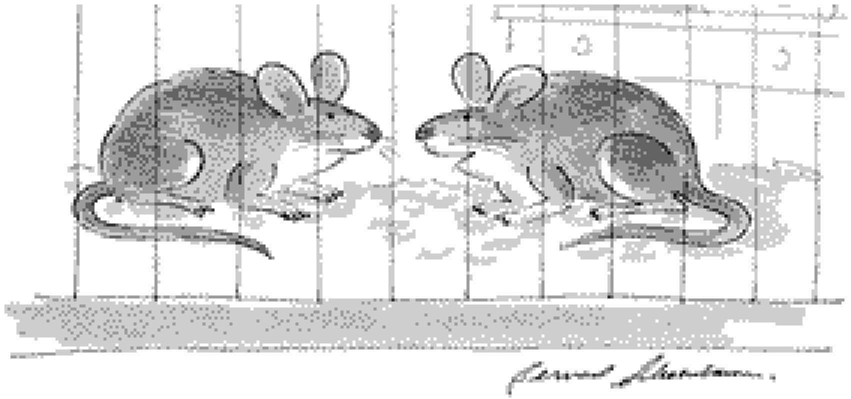
- *Course objective #4:* To know the various study designs, their assumptions, advantages, and disadvantages that could be applied to identify associations between phenotypes and genomic variants
- *Course objective #8:* To appreciate use of epidemiologic study designs for a variety of applications of potential practical importance
- To read a GWA study and be familiar with data presentations unique to GWA studies

Outline - overview

- Learning objectives
- Study designs
  - Overview
  - Case-control studies
  - Cohort studies
  - Randomized/experimental designs
- The road to GWA studies
  - Overview
  - Family studies
  - Candidate genes
  - Genome-wide association (GWA) studies

Who you study is as important as *what* you study

- Need to measure genotype and phenotype *in the appropriate participants* for the question you want to answer



"He's nice, but all his best qualities were bred out of him."

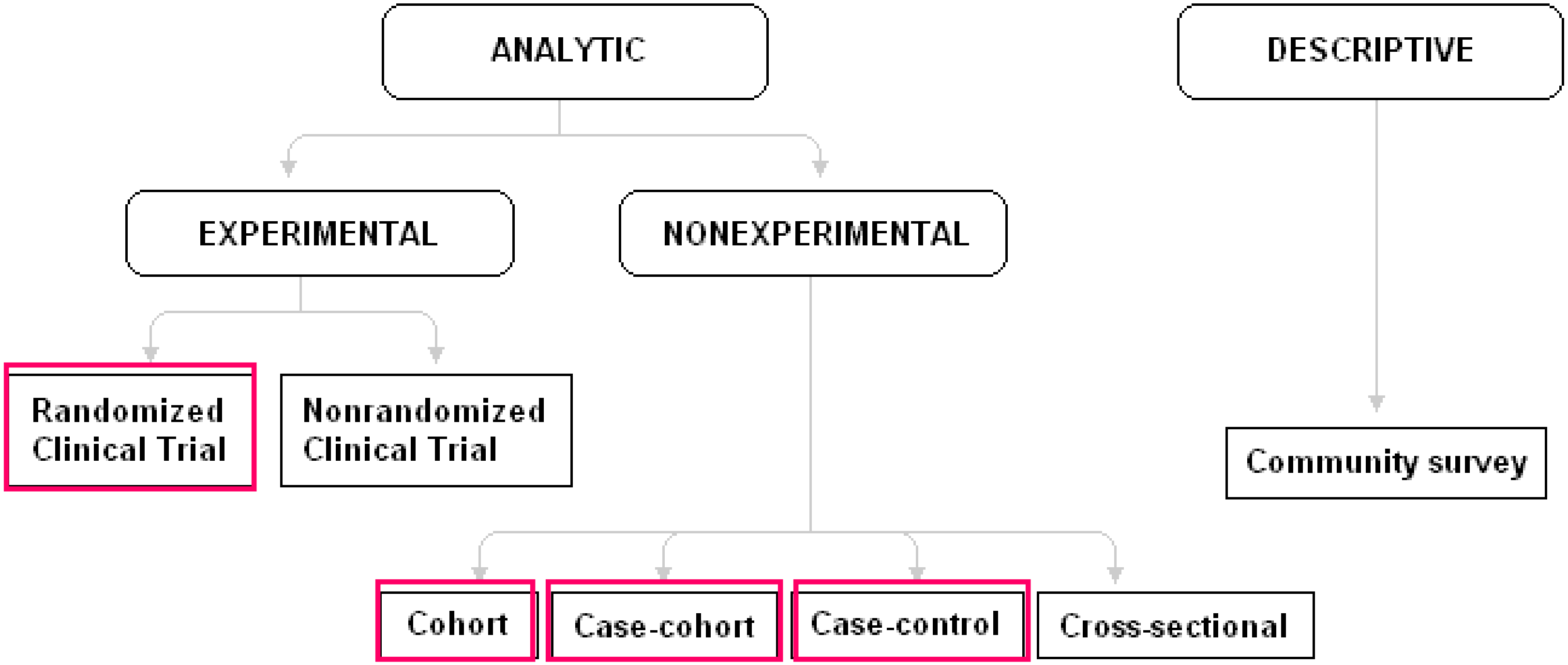
Which study design?

- Purpose of the study
  - Hypothesis-testing versus hypothesis generating
  - Finding signal versus quantifying the signal
- Available resources
- Need for data collection
- Choice of outcome
- Ability to draw valid causal inference

Population-based designs

- Relevant to any study design
- Can you define the *source population* from which the study sample is drawn?
- Ability to define the population
  - Challenge for convenience, volunteer samples
- Why is population-based design important?
  - Validity
  - Generalizeability

Types of epidemiologic study designs



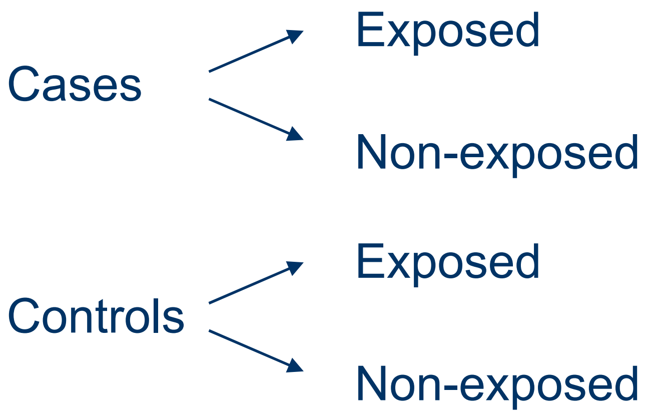
From Wikipedia

Outline

- Learning objectives
- Study designs
  - Overview
  - Case-control studies
  - Cohort studies
  - Randomized/experimental designs
- The road to GWA studies
  - Overview
  - Family studies
  - Candidate genes
  - Genome-wide association (GWA) studies

Case-control studies: design

- Design: identify participants based on their disease/outcome status, compare presence of risk factor



Cases

Exposed

Non-exposed

Controls

Exposed

Non-exposed

Assumptions

- Cases representative of all cases of disease
- Controls drawn from the same population as cases (and at risk for the outcome)
- Exposure data collected similarly in cases and controls

Case selection

- Cases are identified on the basis of their disease/phenotype, representative of all individuals who develop disease
- Distinguishing incident from prevalent or recurrent cases important
- High participant rates important

Control selection

- “Compared to whom?”
  - Controls are representative of the general population who do not develop the disease
  - Selected from population at risk to become case
  - Families, population registries, neighborhood
- Who is the population at risk?
- How do you know they don’t have the disease?

Case-control studies: examples

- Aspirin and Reye’s syndrome in children
- Oral contraceptives and reduced risk of ovarian/endometrial cancer
- *LOXL1* and exfoliation glaucoma
- *TCF7L2* and type 2 diabetes

Advantages of a case-control study

- Suitable for rare outcomes
- Suitable for outcomes with long induction period
- Cheaper
- Need fewer people in some cases
- Readily evaluate multiple exposures
- Convenient
- If assumptions are met, valid estimates of relative risk

Disadvantages of a case-control study

- Doesn’t estimate risk directly
- Special considerations (more later)
  - Exposure-related
    - Recall bias: Disease status may influence reporting
    - Etiologic time period
  - Outcome-related
    - Are studying survivors of the disease
- Difficult to study rare exposures

Case-control study designs: variations on a theme

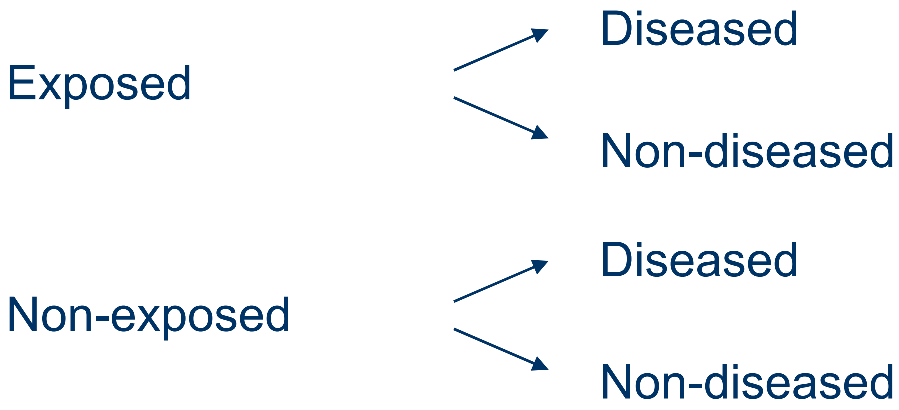
- Nested case-control
  - Within a cohort study, compares all cases to a subset of persons who did not develop disease
- Case-cohort
  - Within a cohort study, compares all cases to a random subsample of the cohort
  - Subcohort can be used for multiple case groups
- Super-cases and super-controls
  - Extremes of the phenotypes
  - Maximizes opportunity to detect signal

Outline

- Learning objectives
- Study designs
  - Overview
  - Case-control studies
  - Cohort studies
  - Randomized/experimental designs
- The road to GWA studies
  - Overview
  - Family studies
  - Candidate genes
  - Genome-wide association (GWA) studies

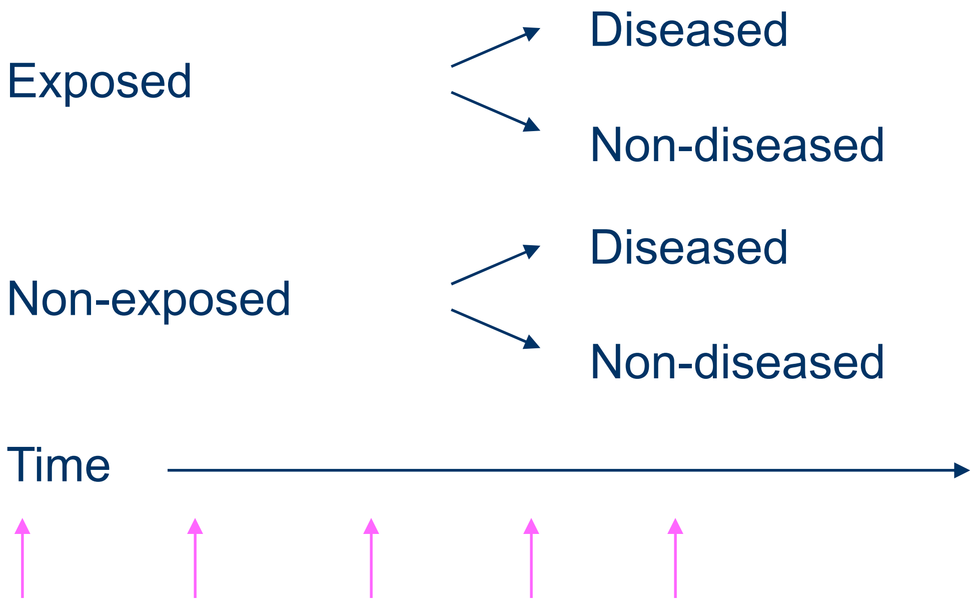
Cohort studies

- Identify individuals based on their exposure status, follow forward to ascertain disease/outcome status



Cohort studies

- Longitudinal: multiple measurements over time



Exposed

Diseased

Non-diseased

Non-exposed

Diseased

Non-diseased

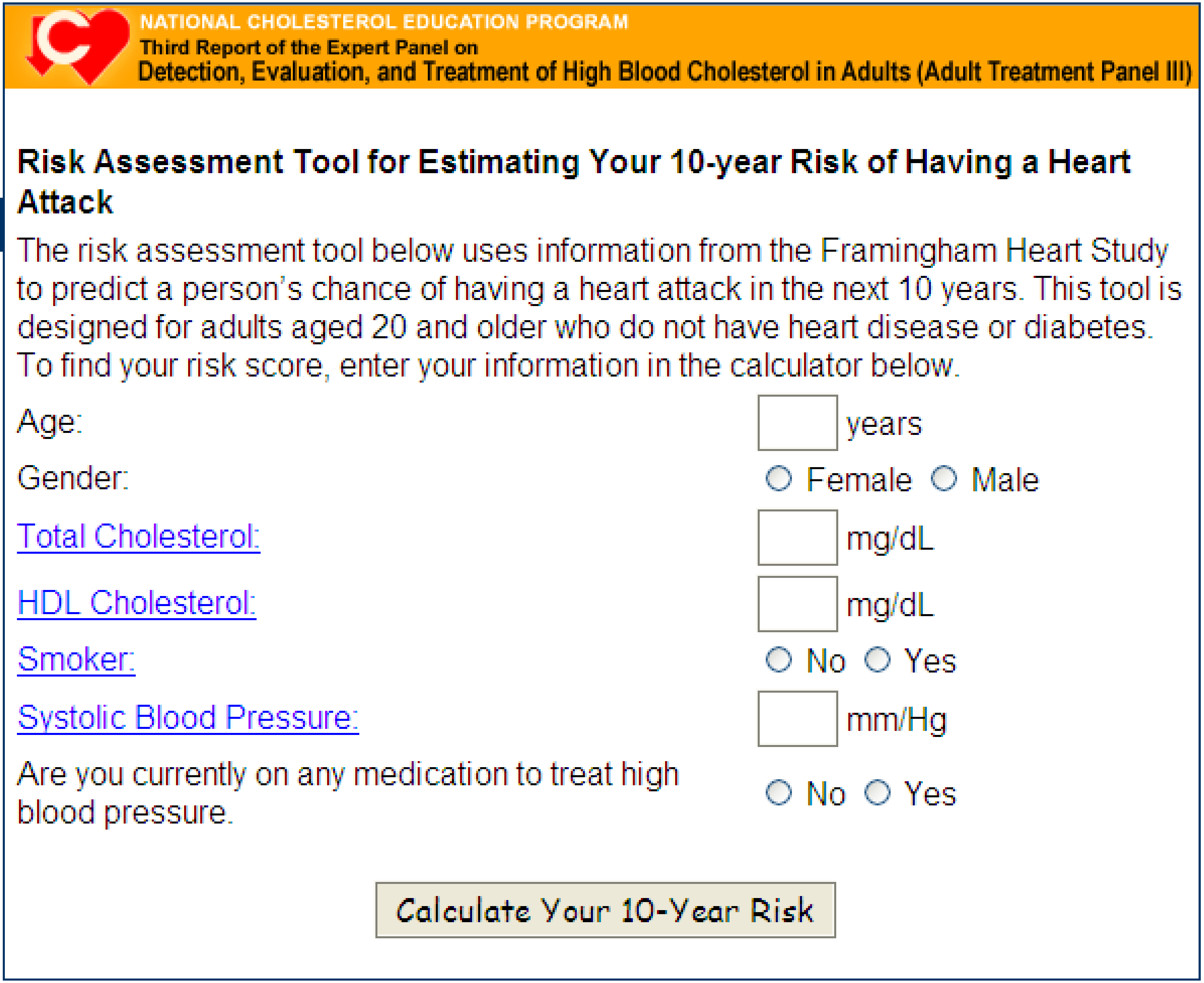
Time

Assumptions

- Exposed and non-exposed groups are representative of a well-defined general population
- Absence of exposure well defined
- Outcome assessment comparable between exposed and non-exposed

Example: Framingham Heart Study

- Original cohort: 5,209 residents of Framingham, MA (1948)
- Offspring cohort: 5,124 children + spouses (1971)
- Framingham III: 3,500 grandchildren (ongoing)
- Identification of major risk factors for heart disease



NATIONAL CHOLESTEROL EDUCATION PROGRAM
Third Report of the Expert Panel on
Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III)

Risk Assessment Tool for Estimating Your 10-year Risk of Having a Heart Attack

The risk assessment tool below uses information from the Framingham Heart Study to predict a person's chance of having a heart attack in the next 10 years. This tool is designed for adults aged 20 and older who do not have heart disease or diabetes. To find your risk score, enter your information in the calculator below.

Age:	□ years
Gender:	○ Female ○ Male
Total Cholesterol:	□ mg/dL
HDL Cholesterol:	□ mg/dL
Smoker:	○ No ○ Yes
Systolic Blood Pressure:	□ mm/Hg
Are you currently on any medication to treat high blood pressure.	○ No ○ Yes

Advantages of a cohort study

- Able to directly estimate risk
- Optimal for short induction periods
- Can look at multiple outcomes
- Potential to investigate natural history of disease
- Amenable to both quantitative and binary outcomes
- Risk factors ascertained prior to disease

Disadvantages of a cohort study

- Not suitable for rare exposures or rare outcomes
- Requires large populations
- May be more expensive, time consuming

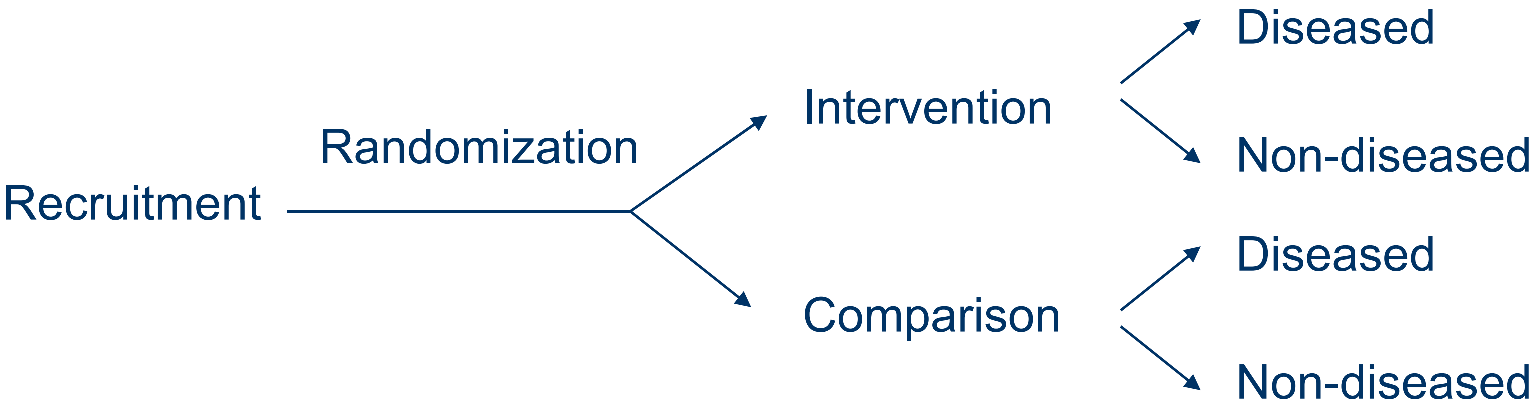
Outline

- Learning objectives
- Study designs
  - Overview
  - Case-control studies
  - Cohort studies
  - Randomized/experimental designs
- The road to GWA studies
  - Overview
  - Family studies
  - Candidate genes
  - Genome-wide association (GWA) studies

Randomized designs

- Definition: a comparative study in which study subjects are assigned by a formal chance mechanism between two or more intervention strategies
- Gold standard for inferring causality
- Also called “randomized controlled trials, randomized clinical trials, experimental studies”

Randomized trials



Recruitment

Randomization

Intervention

Comparison

Diseased

Non-diseased

Diseased

Non-diseased

Randomized designs

- Hallmark: participant assigned to intervention group by a formal chance mechanism
- Assumptions
  - Exposure must be potentially modifiable
  - Primary outcomes are relatively common, occur relatively soon

Randomized designs

- Methods of randomization
  - Several choices, from “flipping a coin” to stratified randomization
- Blinding/masking
  - Participant, study investigator (and anybody else involved in follow-up)
  - Ideally, double-blinded
- Analysis: intention-to-treat

Randomized designs: examples

- Women’s Health Initiative
  - Clinical trial component: 68,131 postmenopausal women
  - Multiple interventions: Dietary, hormone therapy, calcium/vitamin D
- Physician’s Health Study
  - PHS-1
    - 22,071 male physicians
    - Assess benefits and risks of aspirin and beta carotene
  - PHS-2:
    - 14,642 male physicians
    - Multiple interventions: vitamin C, vitamin E, beta carotene, multivitamin

Advantages of randomized designs

- Similar distribution of baseline characteristics in comparison groups
- Protection against confounders, both known and unknown
- Able to directly estimate risk
- Allows comparison of multiple outcomes

Disadvantages of randomized designs

- Limitations on types of interventions
- Costly
- Not suitable for rare outcomes
- Not suitable for outcomes requiring long or extensive follow-up
- Potential challenges to the generalizeability of findings
  - Eligibility: strict inclusion/exclusion
  - Adherence/withdrawal issues

Summary of epi study designs

Design	Well suited for
Case-control	Rare outcomes, long induction periods Multiple exposures
Cohort	Common outcomes Multiple outcomes
Randomized trials	Short induction periods Multiple outcomes Exposures prone to confounding

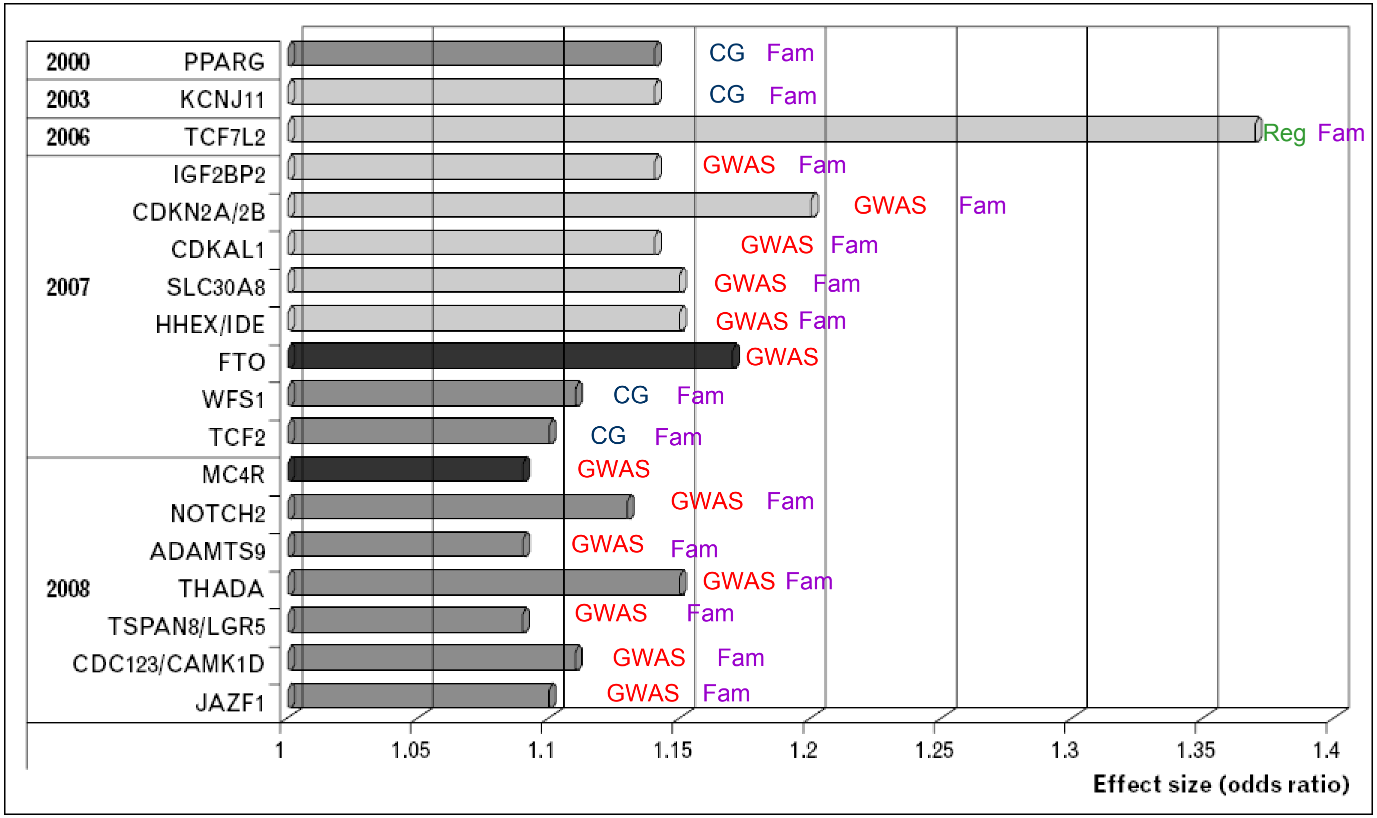
Outline

- Learning objectives
- Study designs
  - Overview
  - Case-control studies
  - Cohort studies
  - Randomized/experimental designs
- The road to GWA studies
  - Overview
  - Family studies
  - Candidate genes
  - Genome-wide association (GWA) studies

Progression of genetic epidemiology

- Twin studies, family studies → candidate SNPs → candidate genes → genome-wide association
- Intersection of developments in biology, technology and statistical methods
- Emphasis shifting from hypothesis-driven to agnostic study designs
- Expanding focus from single gene disorders to common, multigenic diseases

Identification of T2D loci



Perry and Frayling, *Curr Opin Clin Nutr Metab Care,* 2008

Outline

- Learning objectives
- Study designs
  - Overview
  - Case-control studies
  - Cohort studies
  - Randomized/experimental designs
- The road to GWA studies
  - Overview
  - Family studies
  - Candidate genes
  - Genome-wide association (GWA) studies

Why family studies?

- Good route for gene discovery in Mendelian disorders
  - Strong familial clustering suggests genetic basis
  - Sentinel families good for studying specific phenotypes
  - Less susceptible to population stratification
- Estimation of special parameters
  - Familial relative risk
  - Risk penetrance

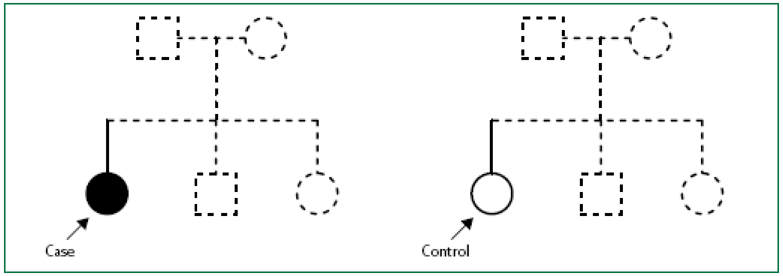
Early family study designs

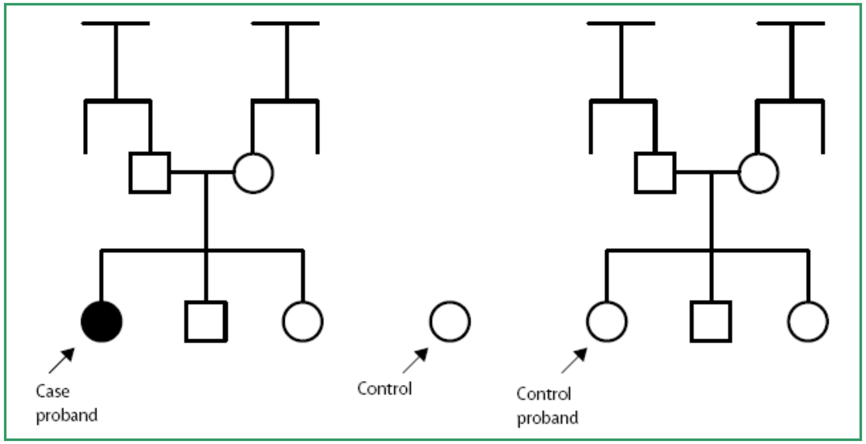
- The original agnostic approach
- Heritability analysis
  - Objective: quantify the fraction of total phenotypic variance attributed to genetic differences
- Linkage analysis
  - Objective: identify genomic regions where genes associated with the phenotype might lie
- At best, identify large chromosomal regions, not specific genes
- Further fine mapping of causal locus required

Family-based association studies

- A twist on a familiar theme: cases + their relatives
  - Family history, e.g., first-degree relative
  - Parent-child trios: compare observed to expected transmission of alleles
  - Extension to siblings, nuclear families, extended pedigrees,

Family studies: example

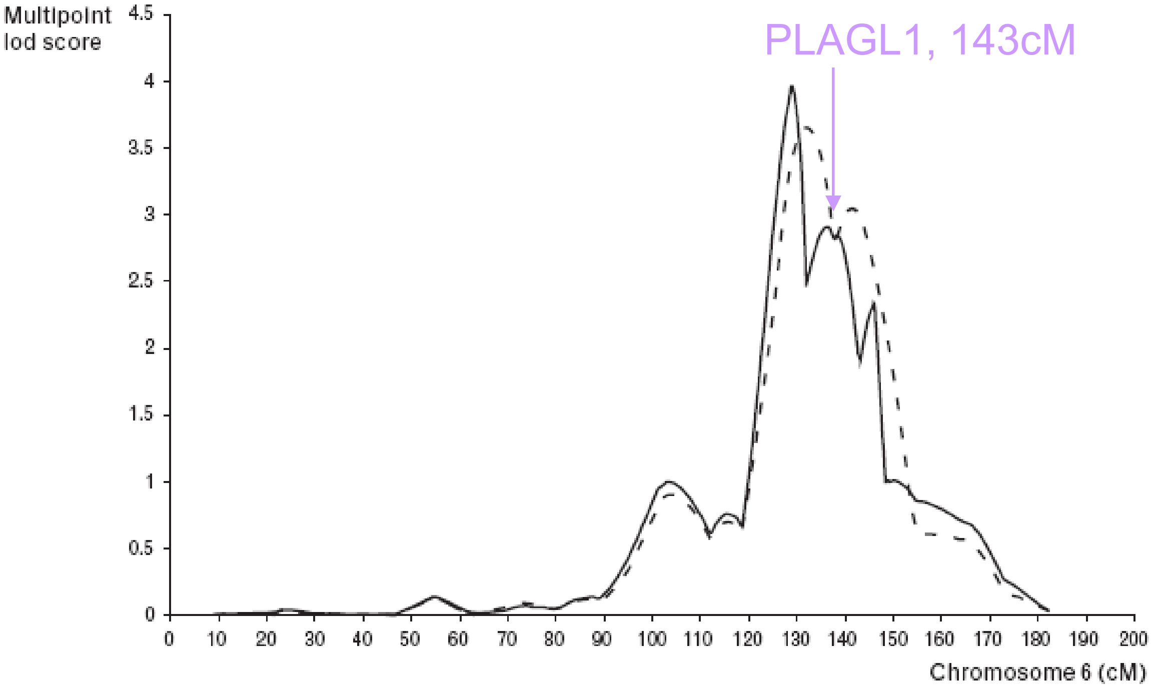




Hopper, et al., Lancet, 2005

Family studies: example

- Linkage and association data: HDL_3C



Cupples, *Curr Opin Lipidol*, 2008

Transmission disequilibrium test (TDT)

- Null hypothesis: If neither linkage nor association is present between marker and disease locus, then alleles from heterozygous parents will be randomly transmitted to affected offspring

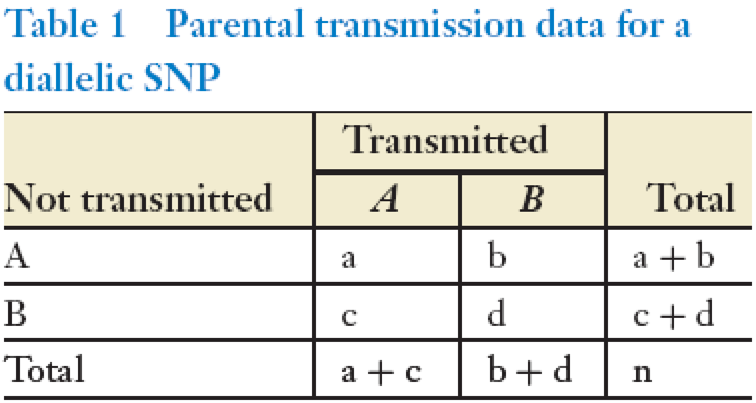


Table 1 Parental transmission data for a diallelic SNP

Not transmitted	Transmitted *A*	Transmitted *B*	Total
A	a	b	a + b
B	c	d	c + d
Total	a + c	b + d	n

Elston, et al. *Annu Rev Genom Hum Genet, 2007*

Advantages of family studies

- Less prone to population stratification
- Rich context for evaluating shared genetic and environmental influences

Disadvantages of family studies

- Difficult to separate shared environmental from genetic influences
- Reduced power due to exclusion of uninformative families
- Challenging for outcomes of older age
- Estimates may not apply to general population

Outline

- Learning objectives
- Study designs
  - Overview
  - Case-control studies
  - Cohort studies
  - Randomized/experimental designs
- The road to GWA studies
  - Overview
  - Family studies
  - Candidate genes
  - Genome-wide association (GWA) studies

Candidate gene studies - biology

- Driven by current state of knowledge
- Assumptions about genes, SNPs
- Common disease, common variant hypothesis
- One or a few common (≥5%) SNPs in one or a few genes, associated with outcome

Candidate gene studies - methods

- Started by interrogating known functional regions – promoters, exons
- Increasing knowledge about linkage disequilibrium → tagSNPs
- HapMap
- Concern for false positives moderate
- Problems with replication

Candidate gene studies - examples

- APOE and Alzheimer’s Disease
- BRCA and breast cancer
- PPARG and type 2 diabetes

Outline

- Learning objectives
- Study designs
  - Overview
  - Case-control studies
  - Cohort studies
  - Randomized/experimental designs
- The road to GWA studies
  - Overview
  - Family studies
  - Candidate genes
  - Genome-wide association (GWA) studies

GWA studies - biology

- Robust associations not always with functional variants
- Success of candidate gene approach depended on correct specification of genes
- Early GWA studies identified promising regions that were previously unknown
- “Agnostic” approach

GWA studies - methods

- Genotyping platforms developed to look at hundreds of thousands of genes
- Same analysis (and relative risks or odds ratios) as before, but repeated hundreds of thousands of times
- False positive results a major concern
- Statistical adjustment of p-values, replication

GWA studies - overview

- Selection of large number of individuals with the trait of interest, including a suitable comparison group
- DNA isolation, genotyping, data review to ensure high genotyping quality
- Statistical tests for associations
- Replication of associations in independent population(s) or experimental confirmation of function
- Reports of allele frequencies, p-values, association statistics

Adapted from Pearson and Manolio, JAMA, 2008

Anatomy of a GWA study – colorectal cancer

Zanke, et al. Nat Genet 2007

Stage 1: Ontario Familial Colorectal Cancer Registry
1,226 cases / 1,239 controls
99,632 SNPs

Stage 2: Seattle and Newfoundland case-control studies
1,139 cases / 1,055 controls
1,143 SNPs

Stage 3: Scotland case-control study of early onset disease
975 cases / 1,002 controls
76 SNPs

Stage 4: Scotland case-control study of early onset disease
1,910 cases / 1,985 controls
9 SNPs

Anatomy of a GWA study – height

Weedon, et al., *Nat Genet,* 2007

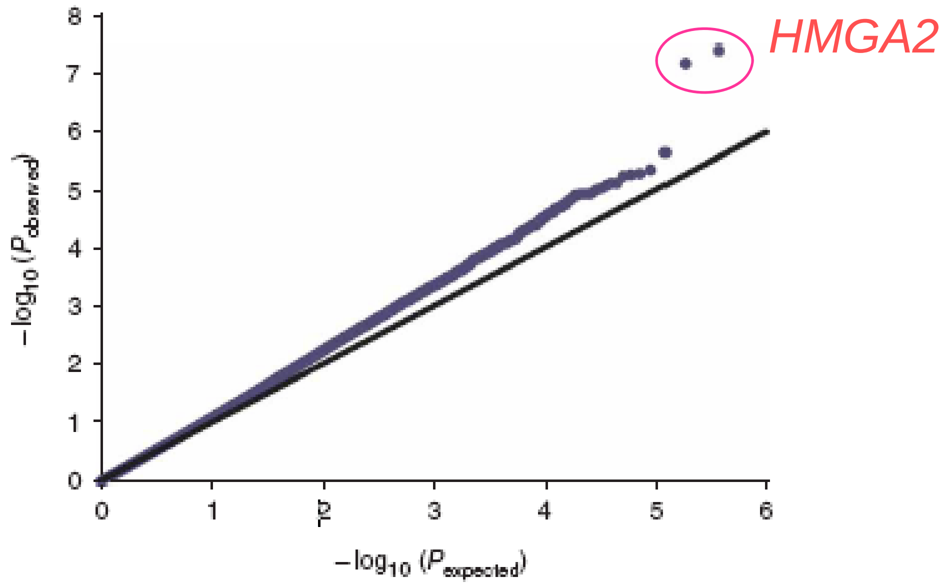
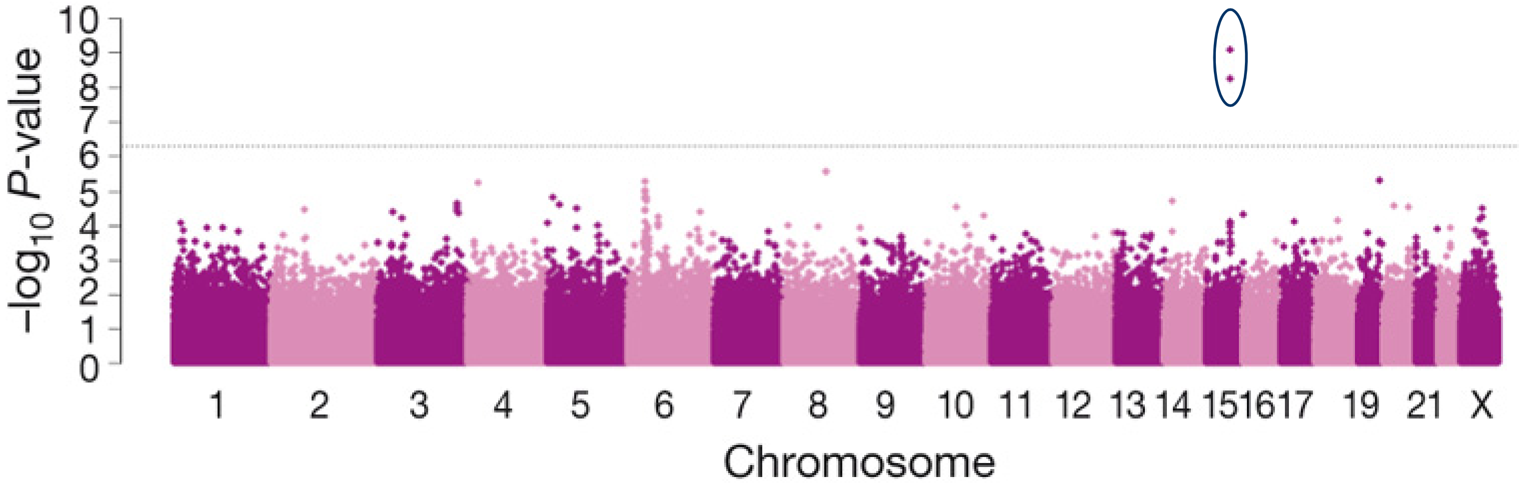


Figure 1 Quantile-quantile plot of 364,301 SNPs from the meta-analysis of DGI and WTCCC genome-wide association statistics. Blue dots represent observed statistics, and black line represents expected statistics.

Anatomy of a GWA study – lung cancer

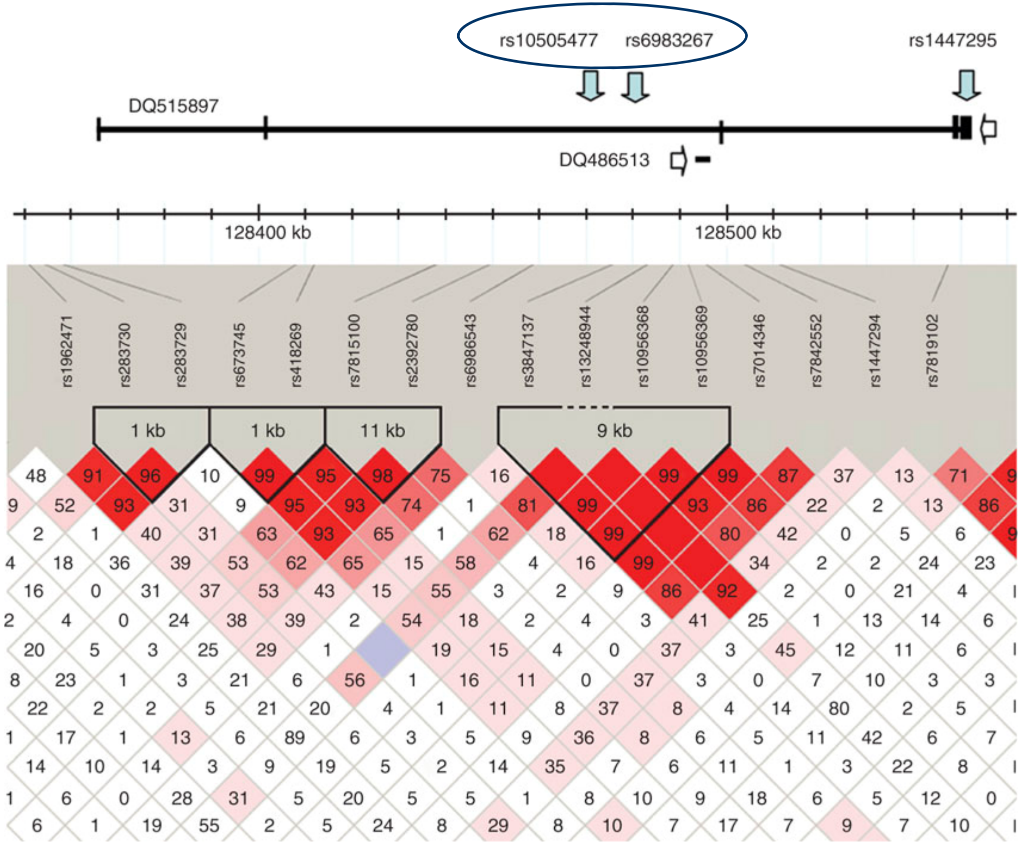
Hung, et al., *Nature*, 2008

CHRNA3,CHRNA5, CHRNB4

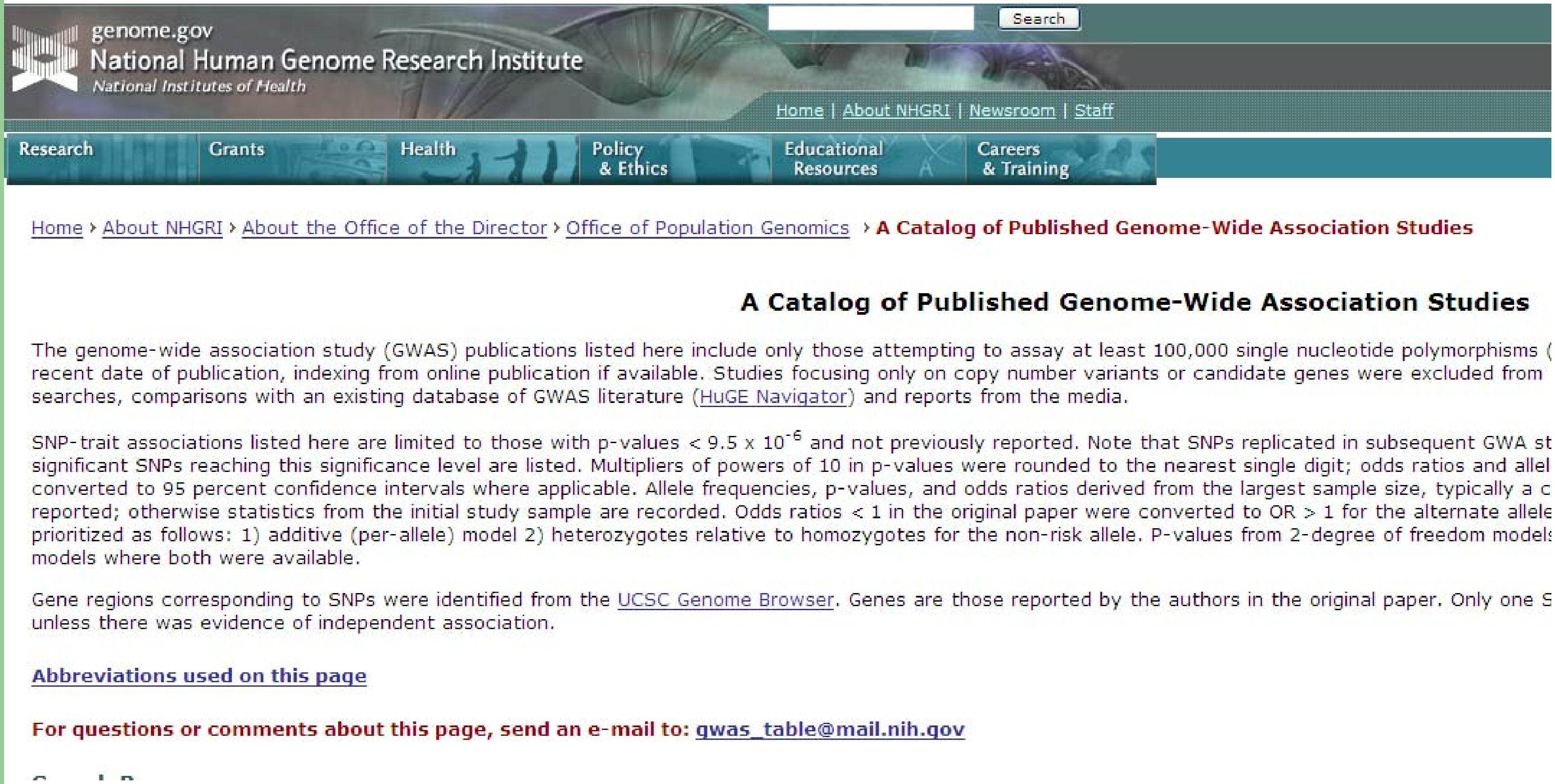


Anatomy of a GWA study – colorectal cancer

Zanke, et al., *Nat Genet*, 2007



NHGRI GWA study catalog
www.genome.gov/gwastudies


genome.gov
National Human Genome Research Institute
National Institutes of Health
Search
Home | About NHGRI | Newsroom | Staff
Research
Grants
Health
Policy & Ethics
Educational Resources
Careers & Training

Home › About NHGRI › About the Office of the Director › Office of Population Genomics › A Catalog of Published Genome-Wide Association Studies

A Catalog of Published Genome-Wide Association Studies

The genome-wide association study (GWAS) publications listed here include only those attempting to assay at least 100,000 single nucleotide polymorphisms (recent date of publication, indexing from online publication if available. Studies focusing only on copy number variants or candidate genes were excluded from searches, comparisons with an existing database of GWAS literature (HuGE Navigator) and reports from the media.

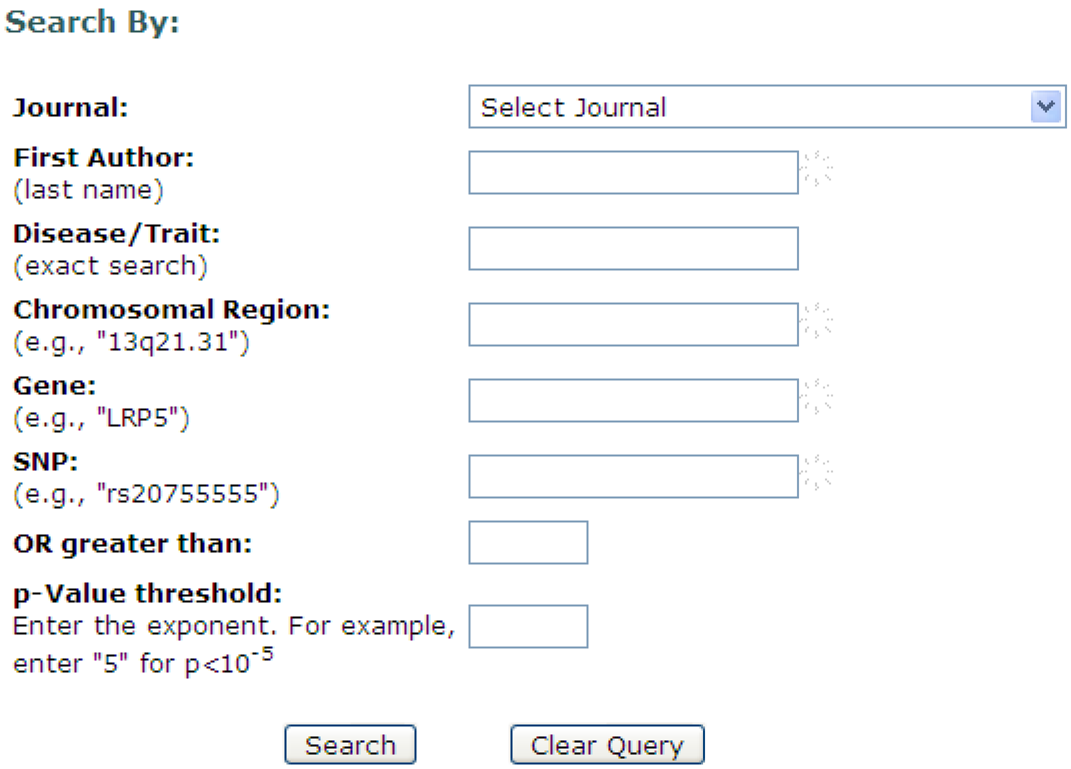
SNP-trait associations listed here are limited to those with p-values $< 9.5 \times 10^{-6}$ and not previously reported. Note that SNPs replicated in subsequent GWA st significant SNPs reaching this significance level are listed. Multipliers of powers of 10 in p-values were rounded to the nearest single digit; odds ratios and allel converted to 95 percent confidence intervals where applicable. Allele frequencies, p-values, and odds ratios derived from the largest sample size, typically a c reported; otherwise statistics from the initial study sample are recorded. Odds ratios < 1 in the original paper were converted to OR > 1 for the alternate allele prioritized as follows: 1) additive (per-allele) model 2) heterozygotes relative to homozygotes for the non-risk allele. P-values from 2-degree of freedom model models where both were available.

Gene regions corresponding to SNPs were identified from the UCSC Genome Browser. Genes are those reported by the authors in the original paper. Only one S unless there was evidence of independent association.

Abbreviations used on this page

For questions or comments about this page, send an e-mail to: gwas_table@mail.nih.gov

NHGRI GWA study catalog www.genome.gov/gwastudies



Search By:

Journal: Select Journal

First Author:
(last name)

Disease/Trait:
(exact search)

Chromosomal Region:
(e.g., "13q21.31")

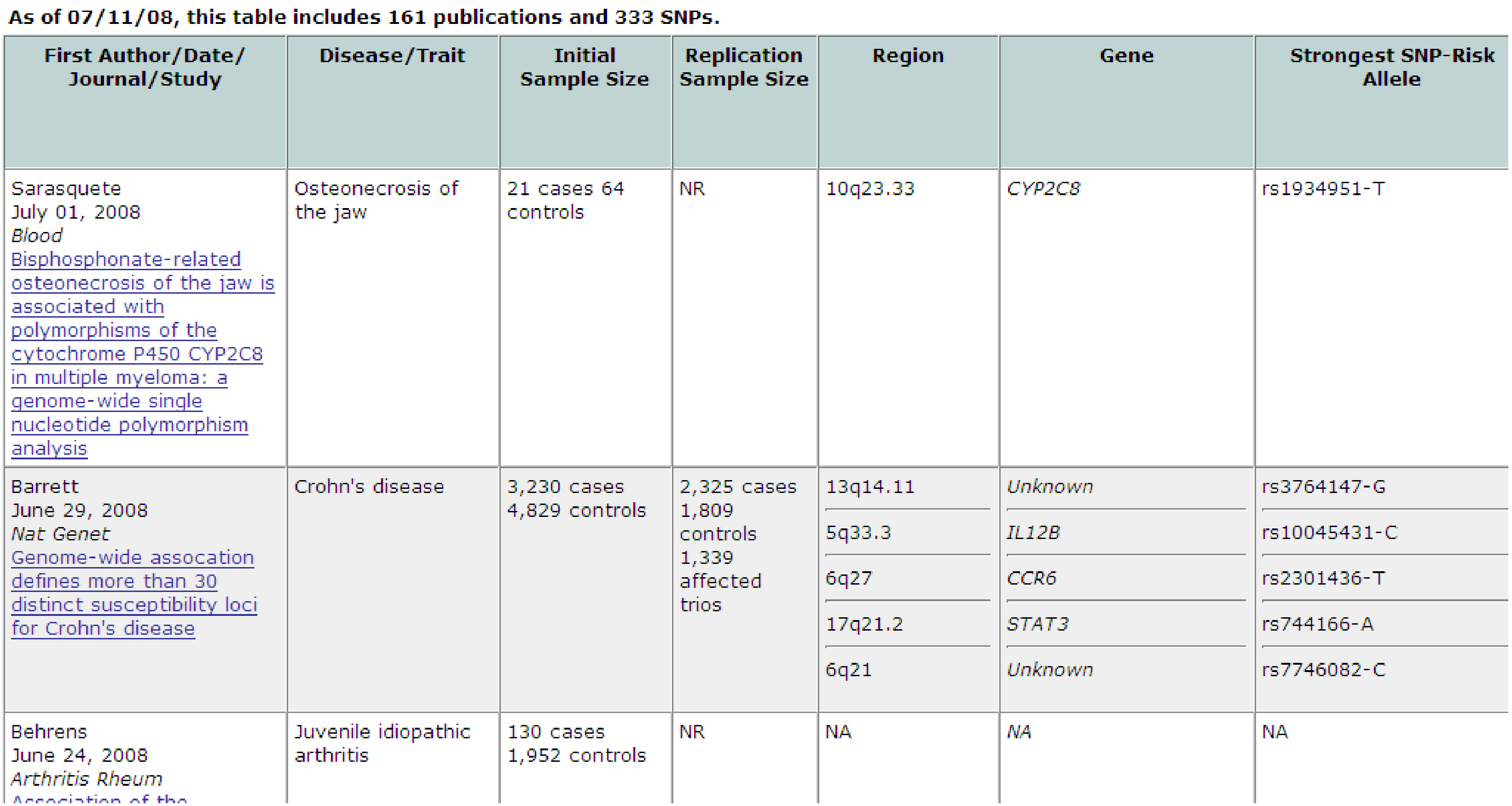
Gene:
(e.g., "LRP5")

SNP:
(e.g., "rs20755555")

OR greater than:

p-Value threshold:
Enter the exponent. For example,
enter "5" for $p<10^{-5}$

NHGRI GWA study catalog www.genome.gov/gwastudies



As of 07/11/08, this table includes 161 publications and 333 SNPs.

First Author/Date/ Journal/Study	Disease/Trait	Initial Sample Size	Replication Sample Size	Region	Gene	Strongest SNP-Risk Allele
Sarasquete July 01, 2008 *Blood* Bisphosphonate-related osteonecrosis of the jaw is associated with polymorphisms of the cytochrome P450 CYP2C8 in multiple myeloma: a genome-wide single nucleotide polymorphism analysis	Osteonecrosis of the jaw	21 cases 64 controls	NR	10q23.33	*CYP2C8*	rs1934951-T
Barrett June 29, 2008 *Nat Genet* Genome-wide assocation defines more than 30 distinct susceptibility loci for Crohn's disease	Crohn's disease	3,230 cases 4,829 controls	2,325 cases 1,809 controls 1,339 affected trios	13q14.11	*Unknown*	rs3764147-G
				5q33.3	*IL12B*	rs10045431-C
				6q27	*CCR6*	rs2301436-T
				17q21.2	*STAT3*	rs744166-A
				6q21	*Unknown*	rs7746082-C
Behrens June 24, 2008 *Arthritis Rheum* Association of the	Juvenile idiopathic arthritis	130 cases 1,952 controls	NR	NA	*NA*	NA

NHGRI GWA study catalog
www.genome.gov/gwastudies

- First author
- Date
- Journal
- Study
- Disease/trait
- Initial sample size
- Replication sample size
- Chromosomal region
- Gene (author)
- Strongest SNP/allele
- Minor allele frequency
- P-value
- OR or beta (95% CI)
- Platform
- Number of SNPs passing QC

Take-home messages

- Design or read each study to make sure assumptions are met
- Incorporate population-based designs whenever possible
- Consider: for which study designs are your scientific questions suitable?
- Appreciate wealth of information available from GWA studies

Which study design(s) are most suitable for investigating the following associations?

- 1) Toxic shock syndrome and tampon use?

  Case control

- 2) Cigarette smoking during pregnancy and low birthweight?

  Cohort

  Randomized trial

- 3) Antidepressants and quality of life?

  Randomized trial

- 4) Genetic variants and celiac disease?

  GWA case control study

QUESTIONS?

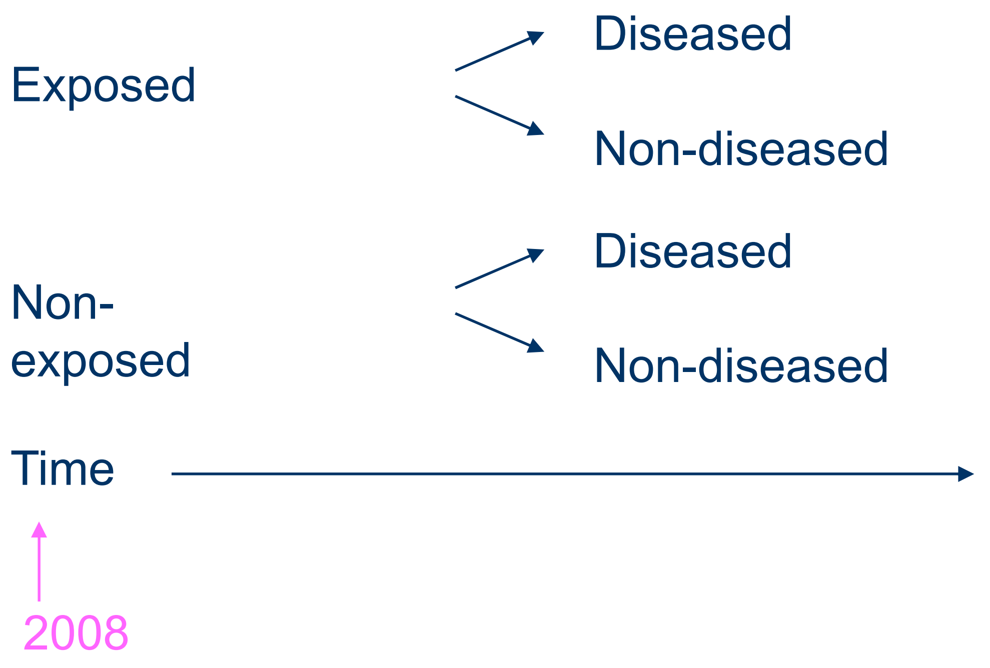
ALL RIGHTS RESERVED
http://www.cartoonbank.com

"According to an article in the upcoming issue of 'The New England Journal of Medicine,' all your fears are well founded."

END

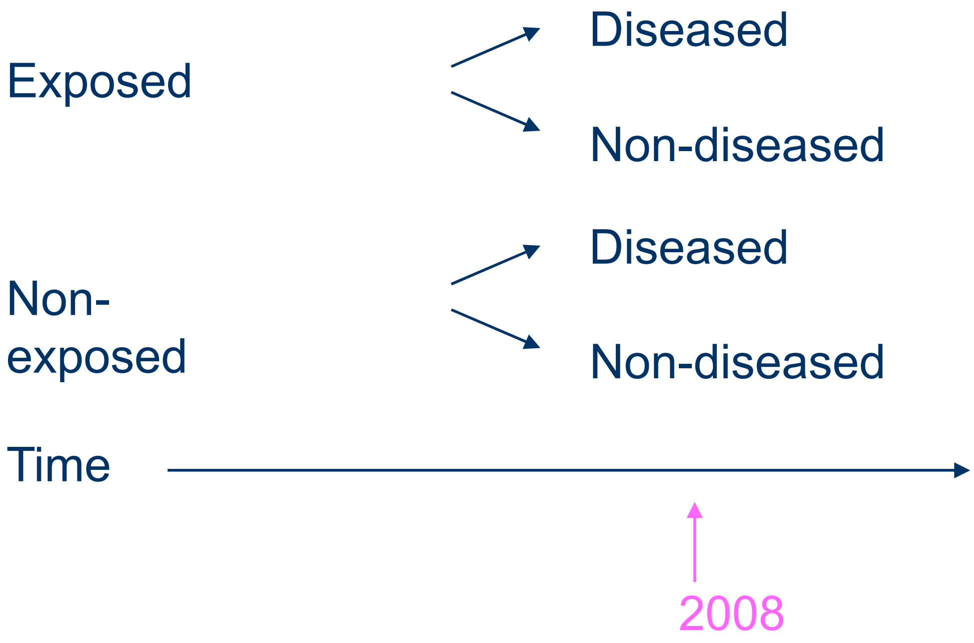
Cohort studies

- Prospective: study initiated before follow-up for outcome occurs



Cohort studies

- Retrospective: study initiated after follow-up for outcome occurs (e.g., atomic bomb survivors)



Exposed

Diseased

Non-diseased

Non-exposed

Diseased

Non-diseased

Time

2008

Example of TDT

- G72/G30 locus on 13q33 associated with bipolar disorder (Hattori, AJHG, 2003)

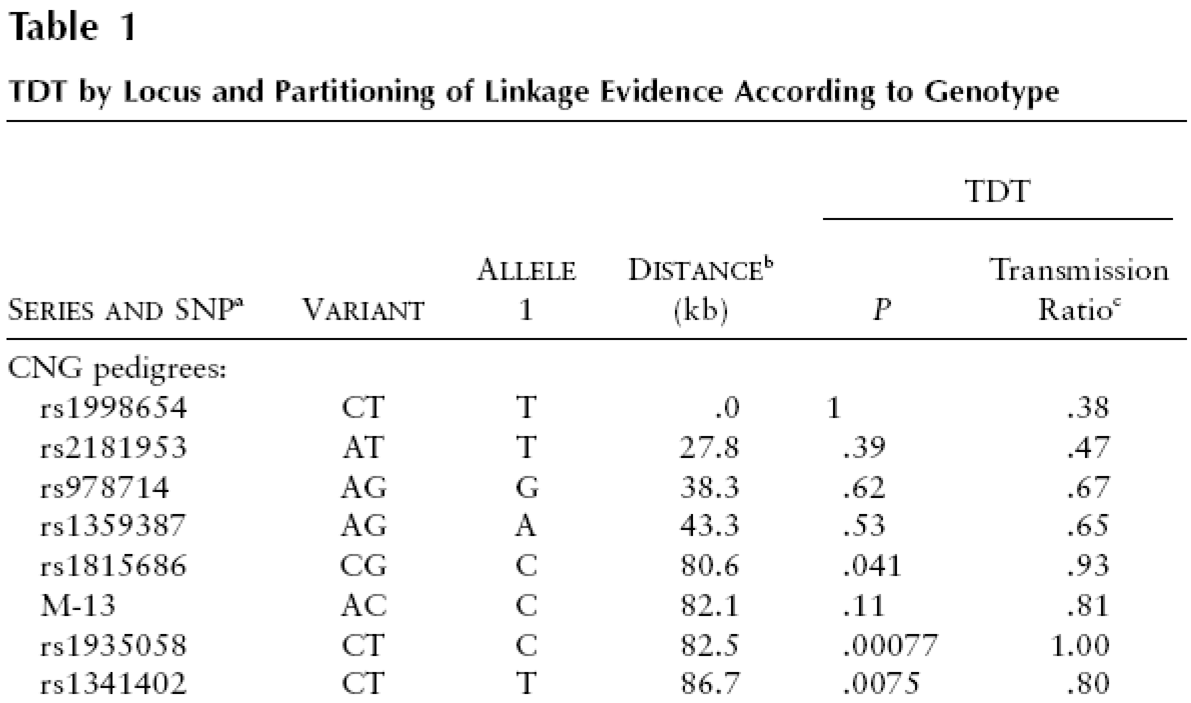


Table 1

TDT by Locus and Partitioning of Linkage Evidence According to Genotype

				TDT	
Series and SNP[a]	Variant	Allele 1	Distance[b] (kb)	*P*	Transmission Ratio[c]
CNG pedigrees:					
rs1998654	CT	T	.0	1	.38
rs2181953	AT	T	27.8	.39	.47
rs978714	AG	G	38.3	.62	.67
rs1359387	AG	A	43.3	.53	.65
rs1815686	CG	C	80.6	.041	.93
M-13	AC	C	82.1	.11	.81
rs1935058	CT	C	82.5	.00077	1.00
rs1341402	CT	T	86.7	.0075	.80

Family studies - examples

- Cystic fibrosis
- Neurofibromatosis
- Bipolar disorder
- Familial hypercholesterolemia

Case-control study: control selection

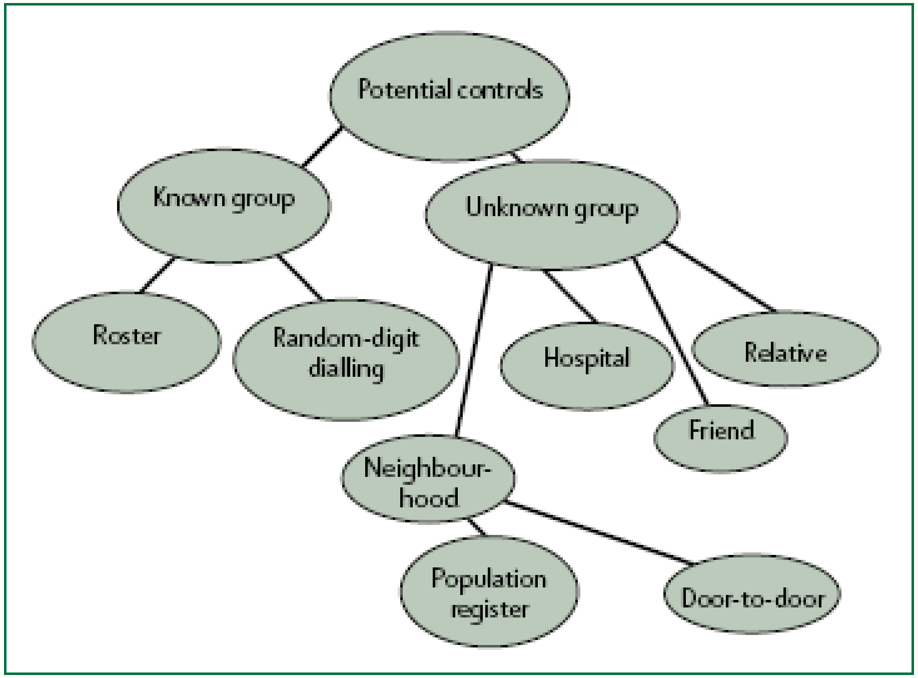


Figure 2: Choosing controls with known and unknown group of study participants

From Grimes and Schulz, Lancet, 2005

© Cartoonbank.com

"And it was so typically brilliant of you to have invited an epidemiologist."