Main : 2008-07-18 NIH Epidemiology in Genomic Studies Seminars

This page last changed on Jul 23, 2008 by wfitzhugh.

Introduction

NIH held a day-long series of lectures on the epidemiology of genetic/genomic studies. It was organized by the Office of Population Genomics within NHGRI, and all the speakers are from that office. All the slides are available here and the video is available here(well, as of now it is not there, but hopefully will be). The office's web site shoul d also have the slides and video but has not been updated yet either. If you want to limit your commitment then I recommend just looking at the talks by Lucia Hindorff (Lecture 4) and Jeff Struewing (Lecture 7).

Introductory Presentation

Thomas Pearson

Basically a motivational talk about how whole genome association studies are feasible and becoming commonplace, but that issues of reproducibility of associations are also common. Epidemiology, being the study of how diseases are distributed in a population, deals heavily in the biases that can affect these kinds of studies. It also deals with discerning the difference between a factor that is associated with a disease and a factor that is causative of the disease.

Measuring Phenotype

Erin Ramos

What is a phenotype? An observable expression of a individuals genotype (their DNA). Phenotypes can be discrete (diabetic or not) or quantitative (height or blood cholesterol level). Simple diseases are primarily caused by one genetic difference and complex diseases are caused by many differences as well as environmental factors. Simple diseases generally have a limited range of severity while complex diseases can have a variety of severities that arise at various times.

Measuring a phenotype can also have issues. Any measurement has an associated error. Some of this data is gathered through interviews and medical examinations while some is obtained through laboratory tests. She made the distinction between reliability (how consistent repeated measurements are) and validity (how close to the true value a measurement is). Both are desirable, obviously, but you can have a reliable result that is not valid (if your scale is always reading 10 pound too heavy, for instance) and a valid result that is not reliable (if repeated measurements give numbers averaging the right value but with a large variance).

Mentioned a project called Phenx which has a goal of producing standards so that measurements of phenotypes will be consistent across studies.

Measures of Assocation and Risk

Emily Harris

How do you measure association between a genotype and a disease? Keep in mind that genetic factors are just part of the story. The physical and social environments are also relevant. There are different kinds of measures of risk, too. There is the risk that one can get a disease at a particular point in time (incidence), the risk that one can have an existing case of a disease (prevalence) or the risk of developing a disease over a period of time (cumulative risk). Because age is often a factor in getting disease, the incidence rate at different ages can vary.

There are several measures of association between genotype and phenotype. One is relative risk, which is the ratio between the incidence rates for two different populations. So if the incidence rate if you have allele A of a particular polymorphism is 0.1 and the incidence rate if you have allele B of that SNP is 0.2 then the relative risk for having allele B is 0.2/0.1 = 2.

Odds ratio is similar. It is the ratio between the odds of getting a disease for two different populations. So for the previous example, the odds for people with allele A are 1-9 and odds for people with allele B are 2-8 (or 1-4). So the odds ratio for allele B is (1/4)/(1/9) = 2.25.

If a disease is relatively rare then relative risk and odds ratio are virtually the same.

Population attributable risk (PAR) is the proportion of disease risk attributable to a certain factor. More clearly, it is the proportion of cases of a disease which would not occur is a certain factor was removed. This is useful because for complex diseases there are often multiple factors and it's useful to know how much risk each one contributes to the overall risk.

All the calculations can be done for single genetic factors or for combinations of genetic factors or for combinations of genetic factors and environmental factors. So allele B of a polymorphism might give a relative risk of 2 but the relative risk for people who had allele B and who smoke might be 3.

She also touched on nature versus nurture issues. How do you determine whether a disease is caused by genetic or environmental factors or both? There are several different ways:

- Ecologic studies compare incidence rates across countries or between ethnic groups in the same country.
- Migrant stuides compare incidence rates between ethnic populations in their home country and people from that same population who migrated to a different country.
- Adoption studies look at incidence rates for adopted children compared to their non-adopted siblings. Presumably this means that environmental and social factors are similar but genetic factors are not.
- Twin studies look at the differences between identical twins and fraternal twins. Since identical twins are genetically identical and fraternal twins share half their DNA, but both are presumably raised identically, these studies can be used to tease apart genetic and environmental factors.

Heritability is the proportion of the total phenotypic variation that is attributable to genetic variation. So if a twin studies showed that if a disease is always found in both identical twins, it has a heritability of 1.

Epidemiologic Study Designs

Lucia Hindorff

She described the difference between case-control studies and cohort studies. Case/control studies are done when it is already known who has a disease and who doesn't. So you identify a population of people with a disease (cases) and people without (controls) and genotype both populations. Then you can look for genotypes that occur more frequently in the cases than controls.

Cohort studies are ones in which a population is genotyped before they are have any disease. You then follow them until some of the get a disease you want to study. An example is the Framingham Heart Study which identified risk factors for heart disease. Advantages are that you can potentially study many different phenotypes with the same study and that presumably there's less bias between the cases and controls because you didn't know which was which when you started. The disadvantages are that you need to study a lot more people to get a good number of cases and that you have to wait a while for the disease to develop.

Randomized designs involve splitting populations into groups randomly, typically used for comparing two or more treatments to a disease. An example if the Women's Health Initiative which found that hormone therapy was not useful for treating post-menopausal women.

Family studies look for linkage between genomic regions and disease in a set of related people. Because genetic recombination does not occur that frequently these kinds of studies have limited resolution. But they eliminate biases between case and control populations.

Candidate gene studies pick a small number of genes to look at rather the entire genome. Less worry of false positives and multiple hypothesis correction not generally needed. APOE and Alzheimer's disovered this way. Another problem is that hard to pick the right SNPs. Not all disease-associated SNPs are 'functional' in a classical way.

Genome-wide Association Studies look at 10's of thousand or 100's of thousands of SNPs. False positives are an issue. Linkage disequilibrium needs to be taken into account. Identifying associated vs. causative SNPs is an issue. Needs independent replication studies to confirm results.

Why are Findings Hard to Reproduce?

Teri Manolio

Most published association findings don't replicate in other studies from different labs. Replication is difficult because different labs use different platforms and genotype different sets of SNPs. In some examples a gene is replicated but the SNPs and/or haplotypes associated are different.

'Winner's Curse' is that the first to discover an association often finds a stronger association than it ends up being across other studies. Because they found it just because it was a stronger than normal in their sample set.

Beware results that show up in only some analytical methods, or those that depend on phenotype definition. Propose randomizing case/control labels and looking at number of 'significant' associations to get baseline.

Need to know genotyping quality measures and use a second genotyping technology to confirm. Also do QC of genotyping such as gender checks, Hardy-Weinberg checks, and use of forensic markers.

Criteria have been developed to define successful (or unsuccessful) replication and a proposal has been published in Nature.

Low minor allele frequency SNPs are a problem and some platforms have problems calling some SNPs. In general need to check quality of SNPs and samples.

Bias in Human Genome Research

Teri Manolio

She referred to a paper on why most research findings are false, although the accuracy of that paper is, ahem, under scrutiny.

From this talk it sounds like an epidemiologist's job is identify and catalog ways that a study can be wrong, presumably to eventually craft the perfect, bias-free study which illluminates all. Here are the types of biases she described.

Selection Bias: systematic differences between those who are selected for study and those who are not. Are cases in case/control study typical of all of those with disease? Are those selected in some group that is not typical (blood donors, army recruits, etc.). Are cases in different studies defined identically? Are controls like cases in all ways except their disease status?

Information Bias: systematic differences in data collection. Families may remember something only if member has disease, or questioning may be different for cases and controls. DNA collected and handled differently?

Confounder: another factor that is associated with disease. So if obesity is risk factor for diabetes then you need to make sure that your case/control study has equal numbers of obese and non-obese people in each category, or at least that you know about the factor so you can control for it. Unfortunately not all confounding factors might be known in advance.

Population Stratification: differences in populations compared with differences between cases and controls. So the case population might have been enrolled at one hospital and the controls at another in a neighborhood with a large population of people of Scandinavian descent.

Genetic Screening and Diagnosis

Jeff Struewing

Bayes' Theorem: how do probabilities relate to relationships between events? P(A|B) is posterior probability of A given B, etc. So A could be breast cancer and B could a be a abnormal mammogram result. Before mammogram, it would be prior probability. After mammogram, it is posterior probability.

EGAPP: Evaluation of Genomic Applications in Practice and Prevention

Sensitivity and specificity are not dependent on frequency of disease. Positive predictive value (PPV) and negative predictive value (NPV), however, are. Rarer diseases cause problems because high sensitivity still can lead to relatively low PPV's.

Clinical Utility: depends on the PPV/NPV and on what the next step would be. What test, treatment and/or intervention would be done based on a positive test? What is its cost, what is its effectiveness and what are its side effects?

Talked about BRCA1/BRCA2 alleles and the fact that you can now get those results from DeCODE or 23andMe. But what would you do if you were positive for those alleles? He quoted a study recommended screening starting at age 25-35 using mammography (PPV = 2.5%) or MRI (PPV = 8.3%). Also, there are other breast cancer alleles with varying odds ratios and allele frequencies. Said 7 SNPs together give risk of 4.2% to 23% (avg is 9.4%) so screening could be tailored to where you were on that distribution. Most of this described in article in New England Journal of Medicine.

Practical Applications of Epidemiological Methods

Thomas Pearson

How can we determine whether an association is causative? It's very hard. Temporal relationship is easy as genotype obviously precedes phenotype. But does expression (or other functional thing) occur prior to phenotype? Strength of association is important, obviously, but single SNP vs. multi-SNP results are an issue. Does dose-response relationship (copy number associated with phenotype) occur all the time? Not really in all cases.

Replication is important. Biological plausibility. How would you do this? Can you find a story for anything? The idea of what is a functional variant has changed over time. Linkage disequilibrium also complicates matters.

In general the Surgeon General's Criteria for Causation seems to be hopelessly outdated when applied to the results of genetic association studies (my editorial contribution, not his...).

Explores desire to know the mechanism of how the genotype is truly related to the phenotype. Is the gene expressed and produces a protein and alters physiology which causes disease? Maybe, if you believe in the Easter Bunny (sorry, editorializing again...). But in general he says there could be intermediate phenotypes which capture the steps in the chain relating genotype to phenotype.

Discusses experiments for showing these steps, including siRNA knockdown, mouse knockouts, etc.

Pharmacogenetics: genetic factors affecting drug response or adverse reactions. Drug metabolism is known to have genetic factors, for instance.