- Read, and be prepared to discuss on <u>*Nov 5*</u>, sections 1, 2, 3 and 6 of the 1976 article by Miettinen. Use the Questions* and Notes to guide your reading – and make a note of the items that are still unclear and that you wish to have discussed at the session. *If* you run short of time, start with the numerical example in section 6.

- Complete (and *email* by Nov 12) exercises 1-4 on Aird & on Woolf.

- If you have time, read exercises 5-8 and run the R code – to reinforce the efficiency message in the Autism-MMR simulations.

Exercises on 1954 article by Aird et al. and 1955 article by Woolf

1. Read the "Selection of Controls" section in Aird et al. (BMJ, Aug 7, 1954) and summarize (in point form) their arguments for/against the "two ways" they considered. List any concern(s), not discussed by the authors, that *you* have about the method they selected.

2. The article by Woolf (1955) does not have an abstract. As a way to summarize the contents and messages of the article, write a structured 'abstract', of 200 words or fewer, paraphrasing the article. Devote a good portion of your abstract to the "Background/Rationale" section, summarizing the first two paragraphs of his paper.†

3. How 'modern' is Woolf's view of the case-"control" study?‡ Is it more or less modern than the earlier Cornfield (1951) article that Woolf seemingly missed (the website has a pdf version JH found on the Web)? Give your reason(s).

4. In an investigation of the role of blood group in the etiology of ovarian cancer, would a denominator-series consisting of males be acceptable? Must those in the denominator-series be at non-zero risk of ovarian cancer? Explain.

Exercises on 1973 article by Mantel

Focus on the portion *in italics* (added by JH) on the following excerpt.

> If we chose π_1 as 1 and π_2 as 0.15, we would have all the cases and 3.5 negatives per case. By the reasoning that $n_1 n_2/(n_1 + n_2)$ measures the relative information§ in a comparison of two averages based on sample sizes of n_1 and n_2 respectively, we might expect by analogy, which would of course not be exact in the present case, that this approach would result in only a moderate loss of information. *(The practicing statistician is generally aware of this kind of thing. There is little to be gained by letting the size of the control group, n_2, become arbitrarily large if the size of the experimental group, n_1, must remain fixed.)*
>
> But the reduction in computer time would permit much more effective analyses. Ostensibly we would be meeting the additional conditions assumed for validity of the retrospective study approach; that is the retained individuals would be a random sample of the cases and disease-free individuals arising in the prospective study.... p.481.

5. Use the data from the London portion of the data in Woolf's Table 1, and the R code provided on the Website, to numerically illustrate how "little (is) to be gained" by having the denominator series become arbitrarily large: Calculate what the variance (and its reciprocal, the amount of '*information*') would have been if the denominator series were (i) the same (ii) twice (iv) four times (x) 10 times (c) 100 times and (m) 1000 times the size (as)of the numerator series. ¶ Using the amount of information in the'1000 times as large' scenario as the *maximum* information I_{max}, calculate and plot the ratio of each of the

[*] You are not expected to hand in answers to these.

[†] Table III and the Appendix in Aird 1954 are examples of where case-control studies 'were at' in 1954: clearly, news of Cornfield's estimator has not yet reached the authors or the BMJ reviewers.

[‡] As many of today's epidemiologists still do, does he regard a case-"control" study as a comparison of the exposure frequency (or exposure odds) in the 'case' group with that in the 'control' group? Or, does he take the modern view? It holds that 'in better families, *one never compares cases vs. controls*'. Even though the exposure data were collected after the fact, 'epidemiologists are students of *(event) rates*' and so we compare (the event rate in) the *exposed experience* with (that in) the *unexposed experience*, even if it means having to *estimate* the (relative) magnitudes of the denominators involved in these event rates.

[§] Note that $n_1 n_2/(n_1 + n_2)$ is algebraicly equivalent to the reciprocal of $1/n_1 + 1/n_2$. The product of $(1/n_1 + 1/n_2)$ and the square of the within-group SD is the square of the SE of the difference of two averages.

[¶] Scale the observed frequencies in the denominator series accordingly – in practice, because of sampling variation, they would not scale exactly.

other amounts of information, I_{lesser}/I_{max} as a function of the ratio of the size of the denominator series to the size of the numerator series. *The computations can be carried out using the R code provided.*

6. Can you see from your plot why a 'control:case' ratio of 4 has come to be regarded as a compromise? Can you think of situations where this magic number doesn't always work? As one possible situation, *simulate* various size denominator series sampled from the 40,046 homes known to have been supplied by the Southwalk and Vauxhall Water Company and the 26,107 known to have been supplied by the Lambeth Company (in the 300 homes in the numerator ('case') series, the split was 286:14). In this example, why is there so little extra gained even in going from 1:1 to 2:1? *Hint: find the 'weakest link' in the Woolf variance for the log(idr) estimate.*

7. Suppose the denominators of 40,046 and 26,107 homes (and thus known person-time denominators) are *known and without error*. Calculate the variance of the log(idr) and use it to superimpose a 95% CI for the IDR on the graph produced in Question 6.

What would the Woolf variance have been if Snow were forced to *estimate* the ratio of the number of homes supplied by the Southwalk and Vauxhall Water Company to the number supplied by the Lambeth Company, using a 'denominator-series' of a total of just 100 homes, and an observed split of 65:35?

Treat the ratio of the denominators (O and A, called H and K by Woolf) in Woolf's example as *random-error-containing* estimates of the true ratio of the two person-time denominators that gave rise to the numerators o and a (called h and k by Woolf).

What is the formula for the variance of the log(idr)? What is the role of the $(1/O + 1/A)$ term in this variance formula? the role of the $(1/o + 1/a)$ term?

8. Use the comparison of the variance formula in which the denominator ratio has to be estimated with the variance formula in which the denominator ratio is treated as known, as a motivation for how to explain the main data-analysis difference between so called 'case-control' and 'cohort' studies:

> "When estimating an incidence density ratio, the main data-analysis difference between doing so in a so-called 'case-"control" ' and in a'cohort' study is that in the
>
> ________ study the person-time denominators are _______
>
> whereas in the
>
> ________ study the person-time denominators are _______.

Sections 1, 2, 3 and 6 of Miettinen's 1976 article

Questions and Notes:

Section 2. The nature of the study (...) Thus, with reference to *populations* it is desired to learn about *rates* of occurrence of the illness in relation to the exposure (possibly in causal terms), within categories of age and other characteristics; and for *individuals* the concern is with *risks* (for various time periods) of the development of the illness in relation to the exposure, conditional on age and other characteristics.

Q: Where, in the 19th century writings covered earlier in this course, did we encounter this same distinction between *rates* and *risks*?

"**Section 3.1** ***The parameters*****. Incidence density ("force of morbidity" or "force of mortality")** - perhaps the most fundamental measure of the occurrence of illness - is the number of new cases divided by the population-time (person-years of observation) in which they occur."

Note the distinction: population-time (*concept*) vs. person-years of observation (*units of measurement*).

Q: Who was the first writer to use the term *force of mortality*?

Note: The population time ‖ $C(t''-t')$ in equation 1 and in the Figure is a *product* of C persons × $(t''-t')$ time-units.

Q: Could we use Figure 1 to depict the I.D. of accidents in those operating an automobile over a short period (say between 8:35 am and 8:40 am) of a particular day, with 'exposed' and 'nonexposed' defined as 'on' and 'not on' a cell phone, respectively?

"**Section 3.2.** ***Estimability of ratio***. (...) C/D is estimable from the reference series (as c/d, the ratio of the sample numbers of exposed and nonexposed referents from the total pool of candidates for the illness)."

Note: C/D [or $C(t''-t')/D(t''-t')$] is the **unknown** '*denominator ratio*', whereas c/d [or $c(t''-t')/d(t''-t')$] is an **estimate, i.e. the empirical version of,** this '*denominator ratio*'.

"If the reference series (of size n) is drawn from the total source populations, i.e., without excluding prevalent cases in the ascertainment, then it is always feasible to estimate..."

Q: Any comments?

Formulae (6) and (7), expressing the exposure-category-specific ID_1 and ID_0 in terms of the overall ID, the IDR and the etiologic fraction EF:

Q: Where do these relations come from?

Formula (8), expressing the etiologic fraction EF as a function of the exposure rate among incident cases, and the etiologic fraction in the exposed $(IDR-1)/IDR$:

Note: This is the less well understood version of the formula for the overall EF and was the subject of an earlier seminar session.

Section 4. Cumulative incidence-rate and risk. This was the topic of two draft articles by JH (see "From incidence function to cumulative-incidence-rate / risk" under the Rates/Risks tab in the website.). The probability (risk) in fundamental relationship formula 12 can be thought of as the complement of the Poisson probability of no events (transitions) if a *dynamic* population, *always of size 1*, is subjected to the hazard or ID function ID(a) over the age period $a=a'$ to $a=a''$. The integral is the *expected number* μ of *transitions*; the cumulative incidence or risk formula can be thought of as providing the Poisson probability of ≥ 1 transitions, i.e., as $1-\exp[-\mu]$.

Section 6. Worked Example.

The ideas become much more concrete as one works through the example in Table 1.

"The samples of cases and noncases in each category of age allow the estimation of the corresponding incidence density ratio (without any rarity-assumption). For example, for the 50- to 54-years age category, the **incidence density ratio** (IDR) – i.e., the incidence density for smokers (ID_1) divided by that for nonsmokers (ID_0) – is estimated to be $\widehat{IDR}=(24/1)/(22/4)=4.36$ (formula 3)."

Note: The $\widehat{IDR}=(24/1)\div(22/4)=4.36$ arises as

$$\widehat{IDR}=\widehat{ID_1}/\widehat{ID_0}=(a/\hat{C})\div(b/\hat{D})=(a/b)\div(\hat{C}/\hat{D}).$$

The **statistical model** behind Woolf's **variance formula** consists of 2 separate binomial random variables:

(i) conditional on the total $a+b$ of 2 independent Poisson random variables a and b, it is known that

$$a\,|\,(a+b)\ \sim Binomial(n=a+b\,;\,\pi=(IDR\times PT_1)/(IDR\times PT_1+1\times PT_0).$$

(ii) Of the $c+d$ sampled person moments, the (r.v.) c of them that correspond to an exposed person is

$$c\,|\,(c+d)\ \sim Binomial(n=c+d\,;\,\pi=PT_1/(PT_1+PT_0).$$

So, asymptotically, we can estimate the variances of the separate $\log(odds)$:

$$Var[\log(a/b)]=1/a+1/b\,;\,Var[\log(c/d)]=1/c+1/d.$$

so that

$$Var[\log(\{a/d\}\div\{c/d\})]=\underbrace{1/a+1/b}_{}+\underbrace{1/c+1/d}_{}.$$

The $1/a+1/b$ is determined solely the number of cases. If PT_1 and PT_0 were *known*, there would be *no second portion* $1/c+1/d$. However, since smoking status is not asked on the census (not even the long version!) the $PT_1:PT_0$ ratio must be estimated using the 'denominator' series. The $1/c+1/d$ contribution to the variance is the *price for having to estimate this ratio*.

Q: Why has this form of denominator ('control') sampling become known as *incidence density* sampling?

The "**smoking-related fraction of cases**" (0.74 in the 50- to 54-years age category)

"Thus, for the 50 to 54 years category age, for which $IDR(=4.36)>1$, the etiologic fraction is estimated as follows: $\widehat{EF}=[(4.36-1)/4.36]\times(24/25)=0.74$ (formula 8)."

Note: We will discuss this formula later; but the logic (*so* simple, it is not well understood!) is as follows:

- **Q**: What is the *maximum* possible fraction of smoking-related cases? **A**: If *all* 24 cases among smokers were 'due' to smoking, then then 24 of the overall 25, i.e. 96%, are smoking-related.

- **Q**: What fraction of the 24 is smoking-related? **A**: For every 4.36 cases *among smokers*, 1 is a 'background' case; 3.36/4.36 or 77% are smoking-related.

‖ In section 6, JH will refer to the unexposed and unexposed population time as PT_1 and PT_0.

- **Q**: What fraction of the 1 case in the no-smoker is smoking-related? **A**: 0% (unless person lied).

- **Q**: Among *all* cases, what fraction is smoking-related? **A**: 77% of 96%, i.e. 74% or 0.74. [**PS:** what would you expecte the EF to be if the study were repeated today, i.e., 40 years later?]