First Conditional Logistic Regression Biostatistics Seminar Series 2024.11.27

- Lyrics to accompany the slides.

Thank you. Good afternoon and welcome. The LI in the title means it' the 51st article in the history section. 28 / 28

ONLINE The 50th article was 12 years ago, and the editor who initially rejected my piece wasn't aware of the series until I asked him if I could take the title with me to another journal. He then asked me to resubmit it, and one reviewer said 'it was a privilege to review this paper, which felt like the statistical equivalent of finding and opening an unknown Pharoah's tomb.'

- Conditional logistic regression is important in epidemiology because it goes hand in hand with the individually-matched (and possibly nested) case-control study, a study design widely used in this school and worldwide. For orientation I will describe one easier to follow non-epidemiological context in which conditional logistic regression is used today, and just LIST several modern epidemiological ones I have annotated ONLINE. They all have the same statistical structure as the non-epidemiological example. I will then go back to earlier versions of the case-control study, and the gradual refinements that gave it the respectability it has today. 97 / 194

- Through their strategic sampling of cohorts and open populations, McGill epidemiologists and biostatisticians played a very large role in its coming of age. They also connected it with a major development in survival analysis.

- I will say a good bit about the 1970s. That is the decade when what is still referred to as the ‘case control’ study came of age, and gained respectability. I was starting out at the time, and was fortunate enough to see both the statistical and the design aspects come together and to have known and to have had as colleagues several of those who brought the two together. The story is one that McGill should be very proud of. Since I have already told the Penrose version of conditional logistic regression in writing, I am not going to spend much time on it today, but I do go over it in the full set of slides and lyrics that I have already put on my website. 129 / 358

ONLINE [In 1958 McGill gave Penrose an honorary doctorate for his work in genetics. It did not have a department of epidemiology and biostatistics at that time. But if it did, it could have pointed out that he and RA Fisher's 1934 publication was probably the very first application of a no-name brand of conditional logistic regression]

- Here are some reasons why I am telling this story. It combines several areas; is a chance to celebrate, and to be proud, but also to reflect and to learn.

- For orientation I will show an easy to follow, non-epi, example that is used in a statistical competition that Andrea Benedetti runs every year. The method, developed in the early 1970s in the economics world, also helped Daniel McFadden win a Nobel prize in the year 2000. It was only then that biostatisticians realized that THEY had independently developed the same method at about the same time.

- Pardoe updates his dataset and analysis each year, but when he first began looking at them, these were the 5 important predictors. The last 3 predictors come from competitions that doesn’t go all the way back to when the Oscars started. Here are some reasons why ‘regular’ (unconditional) logistic regression is inappropriate

- You can see them better here if you focus first on the items in black, i.e. on the dataset structure. For the personal awards (Director/actor/Actress) there tend to be 5 nominees each year, but for the best picture nowadays there are often more than 5. The set of Xs (what I call the profile) are the predictors, and you will see that some of them don't go all the way back to the beginning. In each year of the competition, there is 1 and only 1 winner. So you see why the regular logistic regression won't work (You could get it to run, but it doesn't align with the data structure).

- Now let's look at the PARAMETER MODEL and the FITTING, all shown in red. Take the 2024 competition. If we didn't know who was who or what their profiles we, each would 1 chance of 5 of winning. But suppose the K weights (betas, parameters with unknown values) for the K elements in the profiles produced relative probabilities of omega 1 to omega 5. Then the fitted probabilities would be the 5 P's shown. The 2024 winner (which one has Y=1) is now known, so we can write down the log likelihood contribution from 2024. It will be a function of the betas. Doing the same for each earlier year and adding them up gives us a LogLik function which we can maximize with respect to the betas.

- And these ‘best weights’ applied to the profiles of the 2025 nominees will give us the predicted probabilities. Once all the 2025 profiles are known, Andrea will announce the McGill version of the competition, and I see that she has already scheduled a seminar on March 19 to hear about your prediction performances.

- The word logistic comes into it because the relative probabilities for any two nominees have a familiar loglinear form. What makes it CONDITIONAL is that each competition has exactly 1 winner, no matter how many nominees. There has ALREADY been a winner, and you are wondering why the person who won (rather than someone else in the same competition) was the winner. The reasoning has an after-the-fact ring to it. This retrospection is even more obvious when (at your leisure) you look ONLINE at my list of recent epidemiologic applications of conditional logistic regression.

ONLINE But then, you could say that causality assessment is always an after the fact exercise. In the lyrics I revisit the two in red in these 6 examples.

The one from 1953 shows what it would be like if every journal insisted on putting the study design/methods in the title! 53 / 1010

ONLINE What distinguishes these from the more classical case-control studies I will show you later is that they don't compare rates in the exposed person-time with rates in the unexposed person time contributed by OTHER persons. Each person's time is divided into exposed and not.

ONLINE The first and very famous example was done at a time when fewer people had cell phones, and before there were restrictions on their use while driving. The Toronto researchers identified 699 drivers who were involved in motor vehicle collisions resulting in substantial property damage but no personal injury, carried a cell phone in the car, and were willing to have their phone bill records examined. The primary analysis focused on use of the cell phone at any point in the 10 minutes before the collision.

ONLINE Here is the breakdown from the records they received from the cell phone providers. Almost 1/4 of the drivers had been on the phone at some point in the 10 minutes prior. 33 / 1175

ONLINE This statistic by itself doesn't tell us any more than the older statistic that 95% of all driving accidents occur within 10 Km of home. So should we avoid driving close to home? Obviously, we need to know the denominators, i.e., how much driving by cell phone owners is done on and off the phone? Instead of studying OTHER cell phone owners who had not been in accident to learn how often they use it when driving, the investigators studied whether these SAME 699 drivers had been driving on the phone at the same time the day before. So, now you have 4 possible configurations. So, how much more dangerous is it to drive on the phone? 118 / 1293

ONLINE and these are the frequencies. Can the epidemiologists in the audience tell us what they would make of them?

ONLINE That was almost 30 years ago. Today investigators who carry out self matched comparisons use more extensive sampling to more precisely estimate the exposure distributions. In this study, the exposures are environmental, not personal, and not alterable by the person, So in addition to the date of ER visits, they use the other 3 of 4 days in the same month as candidate days, not just 1 previous day.

ONLINE As a toy example, here us another small, but real, Ontario dataset bearing on the relationship between an event of concern and the daily temperature. We start with the 10 tornadoes, and then find the temperatures on the dates of the 10 tornadoes (shown in bold) and the 3/4 days on the same weekday of the same month in the same year, That's all the fancy 'time stratified' means. 70 / 1453

ONLINE The beta (or betas if more than one X) in the model are varied until the (joint) probability of the 10 events happening on the days they did is as large as possible. It is the same thinking in the Bridge at San Luis Rey novel (*) about a bridge falling down and 15 persons dying: why did it happen to THOSE particular 15? and the dataset has exactly the same structure as the Oscars one. And it would not be possible to study this PROSPECTIVELY in real time. It has to be after the fact, just like when you try to figure out what it was that gave you an upset stomach or headache.

(*) See the piece by David B. Thomas (U. Washington) in the link to case control studies link in course c609. He asks “Does not the following passage from Thorton Wilder’s 1927 book, The Bridge of San Luis Rey although fictional, constitute an independent, original description of the case-control method of epidemiologic enquiry?”

ONLINE The next set of examples are the mix: some involve transient exposures where the risk from a medication disappears soon after one takes it (so one can do WITHIN-person comparisons); in the last one, we are talking long term, and are forced to do BETWEEN-person comparisons. But the data structure and the basis for the likelihood (why me? or why her?) are still the same as in the Oscars example.

ONLINE We might call it a NESTED case control study, because the controls are sampled from a known base. Here you have a combination of matching and regression (I call it matching + modelling: modelling the long list of variables after the asterisk). The cell-phone investigators could have included unmatched but possible relevant, variables, such as whether it was raining or not, as modelled variables in the cell phone study.

ONLINE It still has the same REGRESSION STRUCTURE as the Oscars dataset. There the object was PREDICTION, but the purpose here is different: you are just interested in the magnitude of the effect of the exposure of interest, but not in the other variables. These authors have not always been able to convince reviewers and editors that what they are estimating is a hazard ratio or incidence density ratio or just incidence ratio. But I would try to avoid the term odds ratio, especially if I were speaking to a journalist or to a lay audience, and I suspect they would too.

102 / 1866

ONLINE I like this blurb, which tries to educate the reader and, I suspect, the reviewers. 16 / 1882

- The educational effort you just saw in the previous slide is a good example of what the eminent epidemiology teacher Ken Rothman was referring to when he wrote these words. I call it ”the coming of age” of the etiologic study. “The sophisticated use and understanding of case-control studies is the most outstanding methodologic development of modern epidemiology.” I have made a timeline of some of the design and statistical developments that helped case-control studies become more rigorous and respectable.

- Lets work up the left column of the timeline first. It is personal selection of some etiology TOPICS. You are all familiar with some of these classics, and maybe even the names of the authors, and the types of control groups they used. I'm curious how you would study the role of parents' ages and birth order in Down's syndrome (1934) and what controls you would use to look at the role of the ABO Blood Groups in cancer and peptic ulcers and cancer in the mid 1950s.

- The Surgeon General's report included 29 case-control studies of lung cancer, 2/3 of them without a clear population base. Matching (individual or pre- or post hoc strata) was the usual (and often the only) way to control for confounding.

ONLINE [But sometimes exquisite matching, as there was in the study on abortion and secondary infertility, does not make it a good study. The first author later headed the department of epidemiology at Harvard, but I think that the data from this study should be removed from the datasets package in R. Can you see any issues with the study?]

- Now, as for the STATISTICAL developments: statisticians might recognize the author of this 1935 work that fixes (and CONDITIONS ON) all four margins of the 2×2 table, and epidemiologists might recognize the author of this 1951 insight that helped them communicate their case-control results, and that greatly annoyed the tobacco industry. Until then, epidemiologists could only say that *the affected were more likely to have been exposed that the unaffected*, but now they could warn that *exposed people were more likely to be affected than the non exposed people.*

- I will now show you a neglected piece that uses plain language and a great deal of common sense and wisdom. Samy Suissa teaches that the Woolf formula for the log of a crossproduct ratio (the $1/a + ...1/d$) goes to the core of study design choices in epidemiology. What I like about it is that doesn't even mention the term odds or exposure odds or odds ratio, or 'talk backwards.' The word odds is not to be found in the article: it uses INCIDENCE ratio. What I also like is his lower and upper case notation borrowed from genetics: lower case letters for the exposed and unexposed in the case series, and upper case from the numbers in the corresponding risk categories in the control series, or what Miettinen would call the denominator series.

- Even more notable is that the two series are entirely INDEPENDENT of each other. The denominator series is presumably anonymous, and there is no mention of checking if any of the cases happen to also belong in the control series, and excluding them so we would be comparing cases with controls. Indeed, the London denominator series is not from the blood bank, but from pregnant women. But it could still be used if one was say looking at the incidence of prostate cancer in relation to blood group. As Miettinen argued later on, it is enough that the denominator series be a FAIR SAMPLE of the BASE from which the cases emerged (one does not have to be eligible to become a case).

ONLINE ? But now look at the way the incidence ratio ends up as a CROSSPRODUCT ratio. The sampling fraction drops out, and we have is h/H divided by k/K. The ONLY reason it LOOKS like and ODDS ratio is that when people had to do calculations by hand, they minimized the number of operations, and it is simpler to do 2 two multiplications and 1 division, than it is to do 3 divisions. If we do it as h/k divided by H/K, we get that crossproduct ratio. Arithmetically, if we do it as h/k divided by H/K, we get the same crossproduct ratio: Sadly some editors and reviewers hang on to thinking that because it looks like an odds ratio, the ad/bc in this example is an estimate of an odds ratio. And we will continue to have trouble convincing them that we are in the business of comparing disease rates, not exposure rates if we don't speak up. As Miettinen often said, ‘we are students of rates’. Woolf said the SAME thing in 1955, and objected strongly to comparing the exposures of cases versus controls. Sander Greenland included Woolf's paper in his ‘classic articles’ book.

- I can't skip over this all time classic method for confounder control. Woolf's method is great for strata with lots of information in them, but breaks down as the strata get thin. Mantel and Haenszel's method is very stable, all the way down to have each matched pair as a separate stratum.

ONLINE [Incidentally the numerators in their example are provided by a case series of 64 patients with lung cancer. The two controls for each patient with a diagnosis of cancer were women in the same hospital and service at the time of the interview; one being the next older woman and the other the next younger. About 1/2 had non-respiratory cancers, and 1/4 had cardiovascular-renal disease or diseases of the respiratory system .. not an ideal proxy for a real population-based study base, but it means the adjusted 'relative risk' of almost 11 is an underestimate.]

- Regression had taken over from matching so this is not so important today. But what is still important is Breslow's comment about how modern they were in anticipating the sampling of cohorts, and this statement (that the BMJ still has not taken on board) that the only CONCEPTUAL difference between a cohort study and a modern case control study is that the case-control study used sampling to arrive at ESTIMATED denominators. This is seen directly if we re-phrase, as Alex Walker did for me in the 1970s, and as I did in 2018, what Woolf was saying 70 years ago.

- Few case-control studies up the then had more than 2 controls. This refinement was a step ahead, but was limited to TESTING a null hypothesis, and to a binary exposure (presumably the matching took care of all the confounders). The worked example had tight matching, and tested whether the frequency of positive histories of induced abortion was significantly higher among the cases. Surprisingly, even though he could have used the Mantel-Haenszel estimator, with each row as a stratum, he did not calculate an empirical incidence ratio. 87 / 3041

- This all changed with his 1970 article. Now he thinks of a case-control study as one where, conceptually at least, one compared risk in the exposed versus in the unexposed. More importantly, by fitting the ratio by the maximum likelihood criterion, he specified a fully parametric statistical model, so that variances and confidence intervals came as a byproduct of the fitting, via the information function. He made it very explicit that he was using CONDITIONAL likelihood contributions. Without naming it, he was fitting a 1 parameter conditional logistic regression, i.e., without any confounding variables other than those that were already matched. 102 / 3143

- You might have noticed that one of his references was Cox's 1958 paper on the Regression Analysis of Binary Sequences which introduced the *logit* (the log odds). Logistic regression came into epidemiology in 1962 when Cornfield came up with a risk score from the Framingham Heart Study. He converted a discriminant analysis (which compared profiles of the MI cases with the non-cases) into a logistic regression (which compared exposure profiles). The direct fitting by Maximum Likelihood was formalized in Biometrika in 1967. So, it is not surprising that logistic regression was centre stage in Cox's 1970 textbook on the Analysis of Binary data, and it quickly became a way for for epidemiologists to not have to match on every factor. They could now use regression as a 'poor person's matching' but it was not suitable for individually-matched data.

- Now we come to another Cox classic, this time on survival analysis. At first glance it seems to nothing to do with individually matched or nested case control studies. [When I came to McGill in 1980, I was coming from the experimental RCT world in oncology, and survival analysis, and had not even dealt with non-experimental studies, or case-controlling. Jack and Duncan and Jean-François may remember the faux pas I made at a seminar when Norm Breslow came to talk to us about matched case-control studies. I can still remember where exactly in the seminar room I was sitting and looking around to see if there was hole in the floor I could disappear through.] 116 / 3397

- But if show you how long the 42 patients in his example stayed in remission from their leukemias as the lengths of the 42 horizontal lines (from shortest to longest), can you see the connection? In each vertical set, you have the patients who were being followed that week, and the solid circles are the ones in that set who did fail that week. Cox called each set a 'riskset', but you can also call it a matched set, matched on follow up time. The only part of the patient's profile (Cox calls it z) Cox used in this worked example is which treatment they were assigned to, but it could also have covariates, or confounding factors, just like in the matched sets I showed you from non-experimental studies.

- The likelihood contribution has the same parametric form, and even the reasoning (*why did the failure happen to the person it happened to?*) as the conditional, after-the-fact '*why me*?' reasoning that we saw in the etiologic studies. The ONLY aspects that look different are that there can be more than one failure (case) in the riskset if the time scale is fairly coarse, and that the matched sets are subsets of each other. In the Penrose example, the matched set is all the children in a family, and more than 1 child can be affected by Down's syndrome. If you set the scores in (14) to zero you have the same estimating equations as those Fisher set up for Penrose when he fit a 7-parameter conditional regression model for the effects of mother's age.

- You know that an efficiency of the case control method had to do with the smaller amount of labour spent on gathering detailed exposure and confounder data, but here is another that you would not even think of nowadays. Because of its prospective data collection, Mantel had data on all the risk factors on all of those in the Framingham Heart study, started in 1948, but when he tried to apply a logistic regression when 165 cases of heart disease had been diagnosed, the computing was taking forever. So, he strategically sampled. In the next two slides, which you can read later, I've included his comments about the intuition one builds up from being a 'PRACTISING statistician'.

- Mantel's article is often cited as the first case control study 'nested' within a cohort, but an even larger one that involving matching was underway at McGill at the same time. The analysis of another one, carried out in the UK between 1959 and 1962, was only published as a thesis, so it is less well known.

- Corbett McDonald came to McGill in 1964 and founded the Department of Epidemiology and Health. This was one of many epidemiology projects carried out in the early days of the department. A register was compiled in the personnel department of each asbestos mining company in the Eastern Townships region of Quebec listing all persons currently or previously employed, as of Nov 1, 1966. One of the first reports, submitted in 1970, involved a cohort of over 10,000 persons. Mesothelioma and other respiratory cancers were a particular worry.

- This broader article in 1974 had a Table 3 (shown on the top right) that compared rates of lung cancer mortality. Please look first at this traditional analysis in Table 3. Now look at the blurb on the left. and then look at Table 4 which shows their relative risks from this ‘nested’ case control analysis. As other epidemiologists did at the time, they called it a ’retrospective’ analysis. You can see that the written description is still couched in case-control language, where one COMPARES the EXPOSURES of the CASES with the EXPOSURES of the CONTROLS. Also the vertical format for the exposures makes Table 4 take up much more space than if they had maintained the same horizontal (dose-response, x→y) format and mentality they had in Table 3. 132 / 4058

ONLINE When I interviewed Miettinen is 2011, he was still kicking himself for putting so much material into that 1976 paper of his. [His worked example involved an open dynamic) population, where it is much easier to see why the rare disease assumption is unnecessary] He regretted that the message that we should compare rates, not exposures, and that there is only one type of etiologic study, had been lost or ignored and still has not been adopted. I don't have time now to go into that insightful paper, which should have put the rare disease assumption out of its misery, but

I do include it on the website, and make a comment, via my 2018 paper, in the lyrics. I encourage you to look at the video of the interview, which I have on my webpage at https://jhanley.biostat.mcgill.ca/Reprints/ . The exchange about the 1976 paper starts at minute 50. 150 / 4208

- This is the paper, again with lung cancer death as the endpoint, that brought the nested case-control study to the attention of statisticians. [Looking back at it now, the terms a priori (cohort) and a posteriori (case-control) seem a bit odd. As Miettinen reiterated, after he had joined McGill, conceptually, the contrast is always between exposed and unexposed; how and when we assemble the data for this is a secondary issue, even if the various ways still confuse reviewers and editors.] Method (a) using standardized rates, was the traditional one at the time. Today, we could use Poisson regression with the person-years as well. 106 / 4314

- Method (b) [sometimes called incidence density sampling, after Miettinen's 1976 paper that introduced lots of new ideas] is the same as the one you saw described in the 1974 paper, and the economy benefits were obvious.

- Method (c) used Cox's model on the full cohort of almost 11,000 persons, with age as the time variable that Cox didn't want to model, and so it is matched out. To align with (b) calendar year of birth was modelled. As the abstract stated, it worked well but took a huge amount of computing time, and so no doubt, they thought of sampling. Fortunately, the sampling that had already been done for the nested case control study in (b) fitted the bill, and (after the paper was submitted) when the paper was read, Duncan Thomas was able to add an Addendum confirming the efficiency: 1/20th of the computing cost of the full analysis, a sizeable saving back then. 120 / 4471

ONLINE [More recently, one of the researchers I collaborated with was looking at whether elderly persons taking certain medication were more likely to fall, and had full riskets that were 1.5 million each, so the computer froze when the Cox model was used. I finally convinced that collaborator to take a sample of 1000 persons each day to represent the 1.5 million elderly in Quebec alive that day. I think that finally convinced my collaborator that (despite the BMJ's snobbishness) doing analyses by nested case-controlling (with just over 1000 in each riskset) was feasible and just as valid as doing it by the full cohort way

] 107 / 4578

- Note also Duncan's way of saying what the conditional likelihood is. It's the same as computing the probability that the Oscar went to the nominee it went to. Instead of 'And the Oscar GOES to', it's 'And the Oscar WENT to'. And so again, it is effectively doing conditional logistic regression. Indeed, if you do logistic regression via the `clogit` function in `R` (not sure in SAS) it simply calls the routine that was used for the Cox model (I say this in the Bka. article). So you should not be surprised that it is inside the `survival` package in `R`.

- A year later Prentice and Breslow gave more theory In Biometrika, and Breslow et al. wrote it up in The American Journal or Epidemiology, applying conditional (or stratified) logistic regression to a matched case control study of esophagus cancer*. [He told me once he liked to follow a theoretical article with a more expository one] There was no physical cohort. Later on, Breslow gave some of the back story. Note the hospital controls. And again he emphasized the CONCEPTUAL nesting.

ONLINE In his 1976 paper, Miettinen shows very nicely that Cole's cc study of bladder cancer is nested in an OPEN (dynamic) New England population.

* "All persons admitted to hospital for investigation of dysphagia and weight loss were identified as potential cases. Four controls were matched to each of these on the basis of age (within five years), sex, and race (Chinese, Malay, Indian or Pakistani, and European). Two controls were drawn from the same ward as the case, while two were drawn from an orthopedic unit."

- He used a different matched case-example in Chapter 7 of the textbook, which came out in 1980. 18 / 4868

ONLINE I put this in so that the old fogeys among us can tell our grandchildren about ’BACK THEN’ or ’BC’ (’Before Computers’, almost!) For the internet generation: this code is in the FORTRAN language and it takes 4 pages in the textbook. 43 / 4911

ONLINE And this is some of the output from fitting different models.

- Before I go back to that UK study from 1959-62, I will fast forward to 2018 (The backstory as to why I wrote it is ONLINE)*. I gave a draft of my manuscript to Miettinen, and he insisted in using the first two paragraphs of it to put his own thinking in there. After thinking hard about it for 50 years, he insisted that there is only ONE study: THE etiologic study, one that sits on a proper base. The denominators of the rates to be compared can be estimated by sampling if need be. What about that, Editors of the BMJ? 103 / 5026

ONLINE * I wanted to give a bit of a history lesson to some Journal of Clinical Epidemiology authors; they had made their own extension of the matched pair crossover design to allow 2 controls per case; they invoked the rare disease assumption, and conveniently overlooked Mantel and Haenszel's 1959 paper, and Miettinen's 1970 paper. I took the chance to revisit the fundamentals in a piece I called "Individually-matched etiologic studies: classical estimators made new again". I didn't think it would be appreciated by the Journal of Clinical Epidemiology, and Miettenen encouraged me to send it to the editors at European Journal of Epidemiology (2018), who were quite receptive.

- I promised to say something about that UK study of 1959-62. I had often heard Doug Liddell talk about this, but I hadn't realized (or had forgotten) until I re-read his 1988 review that it had been written up in an MSc thesis in this department in 1974. It dealt with the sequelae of infections in pregnancy. And now that I read it, the McGill nested case controlling of the 1970s is no longer such a surprise. I have crammed a lot into this one summary slide, but let me take you through it bit by bit. 98 / 5233

- The data collection was carried out under the direction of Corbett McDonald. He had joined the Public Health Service in the London suburb of Colindale in 1951 and was head of the epidemiological research laboratory from 1960 to 1964, working on the epidemiology of viral infections, particularly influenza.

- For 21 months in 1959-60, some 31 antenatal clinics in England and Scotland enrolled almost 12,000 pregnant women at their visit at the end of or just after their first trimester. Colindale is number 6 on the map. They gave their reproductive history and a history of any infection they had had in this pregnancy, and they gave a sample of blood. Some of it was used for Rh and blood typing, and the rest was frozen and stored. 80 / 5362

- The pregnancy and 1 year outcomes were recorded, and the number with adverse outcomes was just over 1,100. The bloods of these 1,100 mothers, and of a 1 in 10 random sample of those who had normal outcomes were thawed out and analyzed for 20 different types of antibodies. The analysis of this nested case control study became John Stewart's MSc thesis at McGill in 1974, under the supervision of Doug Liddell.

ONLINE For those who like to follow visually, Here it is as a flow diagram. 15 / 5450

ONLINE In his 2001 review, Doll mentions an early nested case-control study using **biobanking** , published in 1980, of low serum vitamin A and subsequent risk of cancer. In the cohort of 16,000 men, after an average of 3 years of followup, some 86 had developed cancer. The retinol concentrations in their stored blood, and in the blood of 192 mean who had not, were determined and used to produce relative risks for the 5 quintiles of retinol. I noticed in their table 1 that they could not resist a comparison of the mean concentration of the “subjects” (cases) and the “controls.”

But then that was back in the end of the 1970s. 113 / 5563

0 / 5563

- The final two McGill studies I will mention are from post 1980, and again illustrate the strategic use of various forms of sampling from a cohort, again born of necessity. One, from the then-separate department of Occupational Health, has the term nested right in the abstract. It was called a "nested case-control study within a retrospective cohort" by Doll in his 2001 review of 'retrospective' cohorts. It had not one, but TWO sampling aspects, one to make it possible to assemble job histories, and an entirely separate one, done in current workers to estimate job exposures.

/ 5660

- The other study is one I mentioned to another McGill colleague , Jean-François Boivin, last week when I emailed him to invite him to this session, and to ask him for the backstory 34 / 5694

- Here is his reply. When you dig into these papers on the case-cohort study, you will see that there are some attractive features. 24 / 5718

- I put this historical material at the end of these slides and the lyrics, and you can read it ONLINE along with the Biometrika article from earlier this year.

ONLINE The next two slides are from the Wikipedia entry on Down Syndrome

ONLINE The epidemiology on the effects of parents' ages and birth order goes back to the **statistical work of Dr Penrose** and his behind the scenes collaborator Ronald Fisher. Penrose's main work was on genetics of intellectual deficit, but he had wide ranging interests. As a Quaker, he opposed war, and spent the World War II years working in Ontario. After the war, he returned and became the chair of genetics at UCL when Fisher moved to Cambridge University.

ONLINE This was his first paper, in 1933. I spoke about it in a seminar last November. 17 / 5857

ONLINE Every family included was visited personally and, among other things, the ages of the parents at the birth of each child was carefully recorded: miscarriages and all individuals in whom a diagnosis of normality or Down's syndrome could not be made with certainty were excluded. His next disentanglement project presented a much more difficult statistical problem, involving, again, two highly correlated suspects. 63 / 5920

ONLINE The first statistical method he used was a repeat of what he had done with the parents' ages paper, but this time, Fisher insisted that he **show** the full 3-D data .. you have the mothers' ages as rows, and birth orders as columns, the margins as before, and the d's (the numbers of Downs children in the cell) as suffixes. 62 / 5982

ONLINE Fisher had a much more central and hands-on role in this second study, starting with his letter directly to Penrose after he got Penrose's paper to review. 28 / 6010

ONLINE WOW. And even as he says it's a lot of work, his encouragement is admirable. 16 / 6026

ONLINE Below I include Fisher's critical observation. The rest of the letter explained how to get around this, and it meant that Penrose and Fisher had a lot of computing and iterating to do. There are several letters back and forth and Penrose came in to UCL to see Fisher a few times.

ONLINE Each iteration involved computing new values for **each** member of **each** family, and summing these. So they showed the individual-level data in both the Royal Society paper and in the Methods paper that followed in the Annals of Eugenics – a journal that Karl Pearson started and Fisher took over in 1934. 53 / 6132

ONLINE Both articles showed that the birth order did not matter. 11 / 6143

ONLINE The methods paper also showed a fitted **relative** risk curve to show how strong the mother's age effect was, but the conditional nature of the data did not allow them to calculate **absolute** risks. Here is how he went from the model in the population (in red) to the model for the data in his outcome-based sample of families (in blue).

ONLINE Fisher was never one to spell out all the details: it was all so 'evident' and intuitive to him. I show in the Bka. paper and supplement that it was the same ML criterion we use with conditional logistic regression today.

ONLINE Those who have dealt with survival data know that there can be ties in the failure times if the time scale is a bit coarse, and that the computing can get heavy if there are a lot of them and the risk sets are large. A very nice feature in the (companion) Methods paper in 1934 are the worked examples with specific sibships, including a sibship with two affected children. 71 / 6318

ONLINE This is how in 1933 (and even in my time) one included mathematical material in a typed letter. 19 / 6337

ONLINE This is the population model of risks, albeit that they could only be relative to each other. If you want to see some DIRECT population-based risks, have a look at the work of Stark and Mantel: cf. https://jhanley.biostat.mcgill.c

39 / 6376

ONLINE Penrose listed the raw data, sibship by sibship, in both of the 1934 reports. Here are the (head) and (tail) of that file. 24 / 6400

ONLINE This is page one of a more visual representation I amde. I just show the first 44 sibships. The red dots are the cases. You can again see, but at a sibship level, the very large (structural) collinearity between age and birth order. In the Bka. supplement, I address this, and the implications, in a bit more detail. 59 / 6459

ONLINE Fisher was knighted in 1952 and received honorary degrees from several universities. Harvard University (1936), University of Calcutta (1938), University of London (1946), University of Glasgow (1947), University of Chicago (1952), University of Adelaide (1959), University of Leeds (1961), and the Indian Statistical Institute (1962). 46 / 6505

ONLINE Penrose got an honourary doctorate from McGill in 1958, when he attended the 10th International Genetics Conference here (thats about the same time that the 'trisomy' basis of Down Syndrome was established – in France)

ONLINE Starting from the right in this photo, you see principal Cyril James, then the two other recipients, then Penrose. And at the extreme left is Sewall Wright, who chaired the Congress. Remember the 1926 author who showed that it was the age of the mother than mattered for defects in guinea pigs; that was the same metric Penrose started out with in 1933, before he consulted Fisher about standard errors for correlated data! 74 / 6615

- To wrap up, ... 5 / 6620

- I think we could dispense with the old study design labels, and when the BMJ insists on putting a subtitle, we could call our study an ETIOLOGIC study. That would nicely distinguish it from two other genres: the DIAGNOSTIC study and the PROGNOSTIC study. I must mention, in passing, that sampling of the base (follow up experience) allows one to use conventional logistic regression to fit smooth in time incidence density functions and produce profile-specific x-year risks (prognostic probabilities). The Cox model almost treats time as a nuisance, whereas in prognosis, it is central that it be modelled or dealt with in some way. When I showed this 2009 paper to Andrea Benedetti's class one year, two of the students ran with it, and, now with the help of a few more people as well, it is available as the `casebase` package in `R`. 145 / 6765

- I would like to acknowledge two archives, and many many people, going all the way back to Tadgh Carey at UCC in Cork. He awakened my interest in statistical history by (?deliberately) keeping the journals behind locked doors. 39 / 6804