The 'Bio' in Biostatistics. SSC 2016 Impact Award Address. SSC Annual Meeting, Montreal, 2018.06.05.

I thank Erica and Robert for nominating me, and the committee for selecting me. I will use the early ROC work, and three extra-mural consultations, as a way to share some of the excitement and satisfaction (and embarrassments) I have had in my life (MY bio) in Biostatistics, and offer advice to young people starting out. I did a mini version of this a decade ago with this piece, which I put on my

website. This mentor (Fred Mosteller) and this colleague (Steve Lagakos) figured big in that piece. 89 / 89

I was lucky at all stages of my career to have outstanding mentors. Another of them is the co-author of this paper. I met Barbara in the Fall of 1977, soon after we moved from Buffalo to Boston. CT scanners were new, and scarce, and she collected information for the first 2000 requisitions for head CT scans in her Harvard hospital. We wrote several papers from these data, but the one that led to the ROC work started out very innocently, and independently of me.

How much more accurate are radiologists when they read images in conjunction with the patient history, rather than without it. Like statisticians, radiologists often complain they are treated as technicians, and asked to analyze imaging information without all the patient facts. And Barbara got 4 radiology residents to read 109 images (51 were of patients who she was able to independently establish had serious brain lesions at the time of the scan, and 58 who went a year post scan with no problems). They each read each image each did so twice, some weeks apart, once with-

out the patient history, and once with. This is the distribution of readings from ONE of the radiologists under ONE of those conditions, on a 5-point rating scale. 124 / 298

Barbara's research assistant was having trouble with the FORTRAN software that implemented this 1969 method to fit 17 / 315

the bi-normal model in the lower right corner of this graph, 11 / 326

The method fits 6 parameters (4 cutpoints, plus a difference in location and a difference in the spread) and there are 8 degrees of freedom in the data frequencies.

Barbara asked me to help get it to work. Now this was my very first encounter with this type of data, or ROC curves, and I thought the best way to learn about it was to ask for a photocopy of the manual – it was about an inch thick. (She also gave me a tutorial paper in 1978 Nuclear Medicine journal, and one in Science in 1973). I brought the material home with me and began to read it the night of August 27, 1980. It was 4 days before I was to stop working at the Sidney Farber Cancer Institute, and to start my McGill position. That night my wife Ann Marie went into labour

and I took her to Framingham Union Hospital to have our second son. We moved to Montreal and the McGill job 10 days later, and I taught my first intro statistics course to 7 students in our epidemiology graduate program. The course grew in popularity and I started recording whether there were duplicate birthdays. My reaching advice: if you can't convince students by algebra, maybe you can by data.

That Fall, I wasn't yet in demand as a collaborator, so I had time to read and think a lot about the ROC curve and the area under it, and I was able to spot the connection with the Prob(Y>X) in the Wilcoxon test. Of course the test is limited to the null of 50%, so the standard error is very simple. But what happens to the SE as one moves away from the null? Empirical work was needed but I had no funds to use the mainframe (I didn't apply to NSERC until Fall 1983). Fortunately, each student in the first and second stats courses got a free computer allowance, and

I requested more accounts than there were students, and used the extra ones to do these SE computations. They confirmed that the SE based on 2 overlapping negative exponential distributions provided a very tight conservative bound. And better than that, this SE has a very nice closed form. At the time, I didn't fully understand the structure and just took it for what it was, a formula. Armed with these insights, I and Barbara planned to write it up for the Journal of Mathematical Psychology, and I went to the library that summer of 1981 to read the instructions for

authors. I came away devastated: while thumbing through the journal, I came across this paper by Bamber. It didn't have ONE practical result, but it had all the insights and theoretical connections I thought I had been the first to discover. 253 / 794

So we went back to the drawing board. Barbara told me there was a big unmet need in the radiology research community, and why not write up a very practical piece for Radiology. So we did. The math formula and the theta hats drove the production people crazy. 48 / 842

We set out the connections, mentioned Bamber, and then got to the practical stuff. 14 / 856

We gave some example raw data 6 / 862

and the detailed steps to calculate the standard error. I think this step by step recipe, not just for the SE but also for the W statistic (auc) itself, was what would make this paper the hit it was. 39 / 901

Others things helped too. At that time this first spreadsheet was the killer app: people were buying a 2,000 dollar Apple II to run this 100 dollar software. 28 / 929

And this family medicine researcher was interested in distinguishing streph. throat from other sore throats, and wrote a Visicalc program that would complete the auc and SE calculations in my table in less than a minute, He extolled the virtues of the non-parametric AUC, saying that the Dorfman and Alf binormal assumptions may not be accurate, and the FORTRAN program is not easily transferred to microcomputers. 66 / 995

We showed the tight bounds for the standard error, and how it is somewhat like the binomial SE. 18 / 1013

And we showed power and sample size calculations – all new for the time. 14 / 1027

The paired images case (the new aspect) became our second paper. 11 / 1038

Our correlation estimates was clunky and inelegant, and heavy hitters Sam Wieand and Mitchell Gail were quickly able to improve on it –in Biometrika no less – but we were the first to capitalize on the pairing. 37 / 1075

DeLong and DeLong later came out, in Biometrics, with a much more elegant set of non-parametric variance calculations for the paired situation, but to me it was still very much a black box. 33 / 1108

So our last contribution, in this paper, 7 / 1115

we demystified the U statistics, and made the variance structure very clear. 12 / 1127

I am very proud of this elegant and minimalist but far less cited paper. I also proud of the fact that Margaret Pepe took my idea of 'placement values', and ran with it.

Margaret told me that our papers in Radiology and Medical Decision Making were not 'real biostatistics' and that we should have put them in Biometrics or Statistics in Medicine where real biostatisticians read them. I don't have a counterfactual, but you be the judge. In 1994 these guys looked into how quickly statistical techniques transfer. 55 / 1215

These are the ones they looked at 7 / 1222

Here, arranged by the decade the were published, is how fast they spread 13 / 1235

Of course, we need different scales for these classics. But its interesting to go back to 16 / 1251

the four from this decade, and ask where they have gone since. What would you guess? 16 / 1267

The top one is a book, and the next 2 were published outside statistics. George Styan told me my citations would stop once textbooks came out, such as these two that appeared in 2002 and 2003, 36 / 1303

but they don't seem to have. 6 / 1309

Mind you I think levels of scholarship are falling. We no longer have the situation, as I had with an old-time researcher I collaborated with, who insisted we had a hard copy of every paper he referenced. Today, with cut and paste, researchers don't even have to type in the name (or remember the gender) of the person they are citing, let alone have looked at the actual article. They remind me of what this man said: 77 / 1386

I will now address three extra-mural consultations, finishing with the earliest one that led to a 20+ year research topic of my own. 23 / 1409

My involvement in this one started with a call from a lawyer who asked me if I knew this article, and I said yes– another colleague had used it as a teaching example in a course we co-taught. The lawyer's client, Canwest, was asking for laxer Canadian rules about DTCA, and the Government lawyers were using the study to defend the status quo. He wondered what I thought of the soundness of the study, and if I would help him evaluate it. I told him we liked the study, and that I also liked the advertising rules they way they were. I also said that I had had a very nasty

experience with a consultation for a small company that was developing a diagnostic test for Alzheimers Disease, so I wasn't willing to help him. He then said, "I know we are the skunk in the corner, but can you suggest any good Canadian statistician who might be interested in helping us out?" I gave him the names of the ones I considered the best in Canada, but I told him I didn't know their politics, or how likely they were to help. 193 / 1602

A year later, a Government of Canada lawyer called to say that a statistician had written an affidavit that was very critical of the article, and asked if I would advise her. It was by one of the persons I has suggested to Canwest. 44 / 1646

I was quite surprised at the over the top criticisms, and so I prepared a counter-affidavit. Canwest filed for bankruptcy protection, so the case was called off and never went to the judge. Both affidavits are available online, so you can judge for yourself. 44 / 1690

Here I am picking out this one generic issue that goes well beyond the specifics of the case. It concerned this so-called 'rule of thumb', that one needs 10 events per variable' for logistic regression. This is what the statistician hired by Canwest wrote. 44 / 1734

Here are two further excerpts. 5 / 1739

The first bullet did not cite a source, but is similar to the messages from this paper. But is this criticism relevant or justified in this case.? 27 / 1766

This paper, which had appeared just before the Canwest affidavit was deposited, is much more focussed, and relevant. 18 / 1784

How many of you saw the 1967 movie Cook Hand Luke and the phrase "What we've got here is failure to communicate" . This was a theme I addressed in my piece on my early mis-communications and what they didn't teach us in graduate school. What we have in the 1996 paper, and the affidavit, represent an even more fundamental failure, a failure to DISTINGUISH. The same regression equation can have different uses in different contexts, such as 1. to make a particular comparison fairer (and when Y is quantitative, also sharper), 2, for prediction, and 3. this much

more demanding task. 102 / 1886

This very recent article advocates a 'two subjects per variable' 'rule of thumb' for multiple regression. It was derived from a simulation, although mathematical statistics had long since provided an exact and comprehensive answer (provided the question was posed correctly!) 40 / 1926

I used this inappropriate 'rule' as a pretext to bring out (in the same J. Clinical Epidemiology) some long-established and intuitive sample size considerations – all based on closed-form formulae – for simple and multiple linear regression. Nowadays, these considerations are less well understood, and there is less intuition, because variances are seldom calculated manually. My next topic deals with much trickier regression models. 64 / 1990

Paediatricians and parents worry about growth charts. All doctors rely on clinical chemistry labs and their reference values. 18 / 2008

People who establish these values are more worried about the extremes (and percentiles) than the mean, and homoscedasticity is never the norm. 22 / 2030

The WADA project also concerns extremes. It started out innocently enough when my next door neighbour knocked on my door one evening in early 2013. I knew he worked for WADA, but he didn't know what I did, until he went on the website earlier that day looking for a Montreal statistician, and saw my picture on my website. His first words were we need your help with our reference values. The court for Arbitration in Sport had ruled against WADA, and for an Estonian skier, because of statistical issues with the detection limits values WADA was using. I still remembered

the unpleasant aspects of the Canwest case, and if it hadn't been that he was my neighbour, I would have refused. I said that if I could enlist some colleagues to share the stress, I would consider it. He said time was of the essence, but I told him that no matter what, we wouldn't be able to do anything until the semester was over. We spent the summer on it, and delivered our final report in August, the week of JSM. I appeared briefly at a court case, involving a German cyclist, in Lausanne at the end of August, and we submitted the manuscript in January 2014. The court

ruled against the cyclist at the end of February, and WADA began using the new limits later that year. 230 / 2260

The manuscript and the report have the details, so I will just pick out a few points. As the abstract says, the limit is based on a ratio of a value from an assay that tends to pick up more of the artificial hGH and another assay that picks up more of the hGH produced by the body. To complicate matters, the distribution of the ratio is also a function of the concentration involved. Another paediatrician I had worked with told me about the LMS method, developed by British statistician Tim Cole, she had used to make growth charts. 99 / 2359

His x axis is age, ours is concentration.

If the concentration is low (they are low more often in men), the ratio is unstable, and not used against an athlete. Our first decision involved what we meant by low. We used the Geometric Mean of the 2 concentrations, which you will see on the x axis. We have the ratio on the Y axis, on a log scale. But if we used their way of doing it, we would have excluded this green region of the x-y space, and the x-specific distributions of the logRatio would violate the usual independence between the epsilon and the mu. 99 / 2466

So we harmonized it so that the boundary for low (green) was vertical. 13 / 2479

We needed to set the detection limit at the 99.99 percent points, and the quartiles are already nasty, and even after Box-Cox transforms that helped induce Gaussian variation, we have the huge issue of the shapes of 3 (LMS) functions . We didn't want the fitted function to have problems at the left and right extremes, where there are fewer data points. So instead of modelling L, M and S as functions of X, we modelled them as a function of the rank of X, so that there would be equal amounts of data everywhere, and then back-transformed. And we used the LMS on the already-logged

ratios. 107 / 2586

I skipped over one additional complication. There are two separate kits, as a double check. So we had to choose limits that took this into account. Here is the final fit for the first kit, for females. The thicker dotted line is the fitted 99.99 percentile, and the thinner one is the upper 95Here are the limits for both kits, and both genders. 10 / 2750

Obviously, how well we did in producing Gaussian residuals was important. We included one prefect panel as well. From where you sit, without looking at the legends, which is it? 30 / 2780

WADA didn't just take our word for it. They had commissioned 2 independent reports, and brought the 3 Montreal authors and the 1 French author together in Montreal that Fall to come up with the paper for publication. In addition to what I have just shown you, you might be interested in the boxplots we showed for results from people in different sports categories. When I introduced Dick Pound to open the Inyernational Biometrics Conference in Montreal in 2006, he laughed when I told our 800 attendees he was going to test all the speakers for performance-enhancing

drugs, but was not amused when I said he was the only statistician I knew who could estimate what proportion of ice-hockey players were doping without examining individual players, using the Dick Pound Estimator. Don Cherry's estimator gave a very different estimate. I kept this third consultation/campaign for last, because it is the only message I want you to remember. 157 / 2937

I am dedicating this last section to three exceptional mentors. First, my teacher at Waterloo, my first boss, and – if that were not enough – a lifelong contributor to cancer screening. 32 / 2969

Second to this cardiologist. After serving as Dean of Medicine both at McGill and at Witswatersrand, in 1994 he asked me to help his Quebec Health Technology Council to advise the Health Ministry. Should it be paying for PSA tests to screen for prostate cancer? We told them the harms were large and the benefits uncertain (but probably small).

And third to this colleague. In this seminar paper on mammography screening in 2002, he pointed out the obvious – that cancer cures that are achieved by earlier detection and treatment don't show up as 'mortality deficits' until many years later, and that the hazard ratio (comparing screening with no screening) must have this bath-tub shape. He once told me that when a hole is too small, people have trouble seeing it, but also when a hole is too large people also have trouble seeing it. 86 / 3114

A single ratio is appropriate if the reduction in hazard rates is IMMEDIATE, and SUSTAINED. For example, adult circumcision continues to protect against getting HIV; a vaccine gives decades of protection. Its is also appropriate if we STOP COUNTING EVENTS as soon as the agent stops working – e.g. soon after people stop taking a blood thinner or beta-blocker. 59 / 3173

Cancer screening generates a different again hazard ratio time pattern. The reductions appear after a VERY LONG DELAY in PROSTATE CANCER SCREENING 22 / 3195

In this RCT, after an average of almost 9 years, the 'AVERAGE' hazard ratio was 0.8, i.e. the 'AVERAGE' reduction was 20%. But the hazards only begin to diverge after about 7 years. 33 / 3228

The shape of the Hazard Ratio function becomes clearer when we calculate YEAR-SPECIFIC hazard ratios. 15 / 3243

At about 7 years the hazard ratios begin to show the impact of the FIRST screens, but there is insufficient follow-up to see when the effect of the LAST SCREEN WEARS OFF. The HR of 0.8 is an average of 7 years of 0That trial took a long time to enrol men, so many of the deaths are towards the front end and they weight a single hazard ratio away from the nadir or asymptote 32 / 3328

Here is a graphical version Olli Saarela and some of us are trying to popularize. [If there is time at the end I can tell you the statistical espionage we used to get all these data from just those two Nelson-Aalen plots]. You see again how little info we have about the rate ratio in the time window where you would expect to see the biggest deficits. 67 / 3395

Here is a trial that DOES have ENOUGH follow-up to see when the effect of the last screen wears off. 20 / 3415

They screened for colon cancer once every year or every 2 years. 12 / 3427

They reported mortality reductions of 22 and 32 percent, and claimed that the effect persists after 30 years. 18 / 3445

BUT these curves conceal a lot.

We looked at the HR time-patterns, and I am going to ask you to look at them as well. But before I do, let me ask you to to play radiologist for a minute. I showed these slides when I spoke to radiologists about mammography screening. September 14 was the first birthday picked for the lottery for the 1970 US draft of soldiers to fight in Vietnam. 67 / 3518

Draft numbers ranged from 1 to 366. Here are the 12 boxplots of the draft numbers for each month of birthdays. Does it look random? 25 / 3543

How about if I show you them as a scatterplot? 10 / 3553

Here are are again by month, by now with the months arranged in order, rather than alphabetically. Very different. The only one who noticed it from the scatterplot was a radiologist in one of my summer school statistics courses: he immediately said he saw a 'defect/deficit in the upper right quadrant.' 51 / 3604

Now, here is your chance to play radiologist with the time-specific rate ratios in the colon screening trial. We can divide the 30 years into thirty 1-year bins, fifteen 2 year bins, all the way to one 30 year bin as they did. Do you see any patterns within any of these six different resolutions?

Those were non-overlapping bins. What if we use moving bins? The W shape is a bit clearer here. One reason for it could be biological: colonoscopies have 2 benefits. 29 / 3688

But the bigger reason is that there was a 4-5 year gap in funding and in screening. In these simple Microsoft-smoothed curves, you see the 2 sets of mortality deficits, the lagged responses to the two phases of screening. After a delay of some years, mortality reductions reached a nadir of around 40% before reverting to what they would be in the absence of screening; this pattern is repeated when screening is resumed. So, part of the W-shaped HR curve is artificial, not biological. 84 / 3772

Without the (funding related) hiatus, how large would the reductions have been? A has the HRs fitted to the ACTUAL schedule, and B couples the model parameters with the INTENDED schedule. It With a last screen at year 15, the Hazard Ratio would return to 1 by year 30. The effect of screening will not last for 30 years, unless you screen for 20 years. 65 / 3837

What model did we use to fit this bathtub shaped hazard-ratio function? 12 / 3849

Let's begin at age 50, but NO screening. Each year after, there would a number of deaths (the graph would be entirely grey). To be concrete, focus on the deaths at age 56 or so. 35 / 3884

Now start with just 1 (ONE) round of screening, at age 50. The HR at age 56 is the proportion that would still die at age 56 DESPITE the screening, and 1 minus the HR is the proportional reduction, i.e., the proportion averted, the white part. That one round could impact cancers that are NOT SO FAR ADVANCED and NOT SO SMALL that screening can't detect them; it would have less impact on ones that kill at age 51 or age 71! The delay until the maximum impact is one model parameter, and the maximum reduction is the other. 99 / 3983

What if they had several screens, say every 2 years until age 69? The reductions at any age are the amalgam of the STAGGERED contributions of all the rounds up to then. Here are the delayed contributions from round 5, here from round 10. Amy used this for the colon screening trial, and the lung cancer screening trial, and a Danish biostatistician and I are now using it for a region of Denmark that started breast cancer screening well ahead of most of the rest of Denmark. I am also working with 2 biostatisticians in Ireland on the screening program they introduced there – 1/2

the country started in 2000 and the other 1/2 in 2008. But, given the long lag-times involved, this headstart may not have been enough to use this natural experiment to see how big a benefit screening has had, or will have. 146 / 4129

You see here that the issue is all about TIME. In cancer screening, a proportional hazards model is NOT appropriate for sample size planning or for data analysis. Our model assumed each screen has same impact, but the first screen is different from subsequent ones. With enough follow-data we could fit separate parameters for the impact of 1st and subsequent rounds. And – from a career planning perspective – there is a certain loneliness to being a statistician involved in cancer screening . Be prepared for a very delayed, rather than immediate, gratification, and be prepared, as I

will have to at some point, to pass the work on to the next generation. 113 / 4242

What can I say about my life (so far) in biostatistics? Only that biostatistics has been good to me, and that I hope, that through biostatistics, I have been helped in the public's health and happiness. I highly recommend a career in biostatistics, but please, put the bio part up front. Here are links to our program, my home page, and my funding over the years. I never repaid this travel scholarship of 100-Irish pounds, but I hope that in the big scheme of things, there is only one central final report. 92 / 4334