Cancer screening: principles, programs, performance. McGill Epidemiology Seminar, 2019.10.21.

8 years ago I convinced CIHR to fund the neglected question of how best to measure the mortality reductions produced by cancer screening. This website contains the still-continuing output from that grant, and has the slides I am about the show (and many more I will skip). It also has the lyrics and I will add the vocals. I had begun to take a serious interest in cancer screening in the early 1990s, when

– before we had results of any trials – the Quebec Ministry of Health asked Dr McGregor, myself, and several others to advise it on what would be the upsides/downsides of paying for PSA tests to screen for prostate cancer. I knew about lead-time bias and length-based sampling and how comparing mortality rates in the 2 arms of an RCT avoids such artifacts. But it wasn't until I read Miettinen's paradigm-shifting 2002 Lancet article on the data from a mammography trial that I saw the big problem with how such mortality comparisons in trials were being carried out. I took a

second look at the reports from the Minnesota FOBT trial that we used to teach the medical students about screening, and wrote about it in a 2005 piece in Epidemiology. I reminded readers about the important first principles (**all to do with timing**) that go back to Alan Morrison's textbook in 1985/1992 and produced a Fig 1 that suggests a mathematical basis for the theoretical hazard ratio curve in Miettinen's piece. I was very disappointed at the way the 2008 results of the European RCT on prostate cancer screening were reported and in the 2010 J of Medical

Screening I published a re-analysis that emphasized how time-sensitive one needed to be. I then looked at screening trials for several other cancers and continued my campaign with a piece in Epidemiological Reviews, in 2011, the same years I convinced CIHR that we needed better, more principled ways to measure mortality reductions . 326 / 326

My aim today is to introduce you to 1. some principles of screening, and cancer screening in particular. 2. the technologies involved in the various screening activities or programs and 3. better ways to measure the good. I will try to dissuade you from following the crowd and uncritically buying into the PH model. 54 / 380

It don't have time to go into these. I list them to emphasize that screening is a big and costly public health activity. You all had several screening encounters already and the women among you continue to have them. Most of the men probably have no idea about any of these and won't until they reach 50 or so. 59 / 439

I will be making a case later for replacing the word screening with the phrase early detection, or the pursuit of earlier diagnosis and treatment. 25 / 464

This is the classic textbook, from a generation ago. 9 /

This newer one is written by authors involved in the design of some UK screening programs 16 / 489

This is probably the most important principle pertaining to screening programs 11 / 500

This tension – a common one in public health – goes to the core of the problem 17 / 517

Their first chapter gives us some of the history of screen-

These guides have become for screening what Hill's guides became for judging causality. 13 / 541

I will now zoom in on cancer screening. It too has an interesting history, going back more than a century. Women's cancers (uterus and breast) were an early target. In the USA, women, and the Women's Army in the forerunner of the American Cancer Society, spread the message of early detection. Thanks to Papanicolaou and others in New York, and Ayre and others at McGill, the technology for cervical cancer screening was already in place by the mid 1940s, but it took until the early 1960s for it to be widely used. What was it that allowed this to take off then?

Today, screening technologies are available for all of the major cancers. In the case of cervix, breast and colon cancer, they are being used in organized early detection programs that target the population. The programs are typically invitation-based, and run by public health authorities, rather than as a part of clinical medicine. They are quite costly, and complex, and so it is natural to ask how much they have – or would have – reduced the mortality from these cancers? In an extra slide at the end, I give links to 4 articles on Neuroblastoma screening. National neurob-

lastoma screening began in Japan in 1984. In 2003, based on data from 2 major (non-randomized) trials in Quebec and Germany, a Japanese scientific committee concluded that there was sufficient evidence that the current method of screening led to overdiagnosis of neuroblastoma and that there was insufficient evidence that the program reduced the rate of death from the disease. Japan's Health Ministry stopped the program in 2004 after it had been running for 20 years. The impact of cervical cancer screening was never tested in a trial. In a slide at the end, I give links to some of

the population-based evidence from Canada and the Nordic countries. Today, I will address just the 3 I show in red. For colon and prostate, I focus on data from trials. For breast, I focus on non-experimental data from populations.

If the first principle of screening had to do with harm and good, the first principle in measuring these has to do with the TIMING of these. This maxim goes back 2000 years. The opposite seems to be the case for cancer screening, but few of those who measure the benefit follow this 'delay' principle, even though it is follows directly from the very concept of early detection. This blindness is one reason why estimates of the mortality reductions produced by cancer screening are all over the map, and confuse the public. 92

Here are some contexts where it is OK to ignore time, and to treat all of the person-years of followup as interchangeable. The reductions are immediate and sustained, and so a 1-number hazard ratio or percent reductions is reasonable. This has huge implications for study design. We would have the same statistical precision whether we follow 3 million persons for 1 year, or 1/2 a million for 6 years, or 1/4 million for 12.

Here is an example of a nearly-immediate and sustained reduction of 40% or so in the rate of death from ruptured abdominal aortic aneurysms. 24 / 1068

Here are some contexts where the reduced rates persist as long as the agent is present in the body; when the agent is withdrawn, the benefit disappears. This is an important principle in pharmacoepidemiology. 34 / 1102

How soon after the first round of screening might we see the mortality reductions produced by cancer screening? This trial randomized almost 180,000 European men to PSA screening or not, and followed them up for an average of 8 years. 40 / 1142

Here is an ‘AVERAGE’ hazard ratio of 0.8, or an ‘AVERAGE’ reduction of 20%. But the hazards only begin to diverge after about 7 years. 25 / 1167

There is delay of about 7 years before the hazard ratio begins to show the impact of the FIRST screens, but there isn't enough follow-up to see when the effect of the LAST SCREEN WEARS OFF. 36 / 1203

Here is an example where the follow-up WAS long enough to see when the effect of the LAST SCREEN WEAR OFF. It had a no-screening arm and two screening arms, screening every year or every 2 years. 37 / 1240

Here are the results through 30 years of follow-up. 9 / 1249

Does this type of graph help here? Samy Suissa teaches that it seldom does in pharmacoepidemiology. 16 / 1265

I will explain this title soon. What is the prize in this important lottery, whose 50th anniversary is coming up soon? The men born on Sept 14 in the years 1944-1950 were the first to be called up for service in Vietnam. The drum contained 366 birthdays, so your draft number could be anywhere from 1 to 366. 58 / 1323

The draft numbers seemed to be pretty random when plotted as these 12 boxplots, arranged alphabetically. 16 / 1339

How about when the draft numbers are plotted against the days of the year, with Jan 1 on the left and Dec 31 on the right? Every year I showed this scatterplot in 607, students though the distribution was pretty random, until the year I had a radiologist, who instantly noticed "a defect in the upper right quadrant." 58 / 1397

This defect becomes a lot more obvious if we smooth the data by using one month windows: on average, those born later in the year has lower draft numbers. 29 / 1426

The NY Times article had this headline and this graphic, and a statement from a knowledgeable White House official that, "discussions that the lottery was not random are purely speculative." Recently de-classified memos from inside the While house several weeks earlier say otherwise. See Hanley "Lest we forget: US selective service lotteries, 1917 - 2019", the American Statistician, in press. 60 / 1486

Back to that 30-year follow-up. 5 / 1491

I'm going to ask you to play radiologist with the pattern of the rate ratios. Some epidemiologists are 'lumpers' who like 1 number answers, many are splitters who break down results. Indeed, one definition of an epidemiologist is a physician broken down by age and sex. So, here I have broken down the 2 rate ratios (the 0.68 and 0.78 in the last panel) by windows of follow-up time. 3 non-overlapping 10-year windows, 6 5-year bins, all the way to 30 1 year bins. What do you make of them? 90 / 1581

I do the same here, but the bins are 'rolling' or 'moving' bins, like I did with the data from the prostate trial. Any clearer? 25 / 1606

Here is the explanation: there was about a 5-year gap in screening when they lost their funding and had to re-apply. And here are the year-specific rate ratios of hazard ratios, converted to year-specific reductions in mortality. In the bottom panel, Amy was able to use her model to fill in the missing rounds of screening. Time and dose both matter! I will come back to the parameters of her her model below.

Why we have to rely on population-based data and not on trial data; why we need to be principles-based 19 / 1698

We know a lot about the COSTS of mammography screening programs: the financial outlays and the individual harms. The easiest BENEFIT to measure should be the number of breast cancer deaths averted, but even on this measure analysts cannot agree. One big reason is the arbitrariness of their estimands. There are contemporary population-level data from countries that staggered the introduction of their organized programs. But the Big country-level Data that are easily obtained are not sharp enough, and dilute the reductions. 81 / 1779

The sharper studies use diagnosis dates from the cancer registry to define the women targeted by the screening program. Let's start with Denmark. Copenhagen, here, was the first area to introduce a screening program. 34 / 1813

I will focus on the province of Funen, here, which began in 1993, well before most of the rest Denmark. In 2015, Sisse and colleagues compared the mortality in the relevant woman-years 14 years before and after it started. In case this Lexis diagram is new to you, COHORTS proceed along the diagonal, and become 1 year older in AGE every calendar YEAR; all three critical elements – age, period and cohort – are shown in one diagram. The shaded areas are the woman years that would be impacted if screening was from age 50 to 69. Some of the pre-post difference in

mortality rates might be due to improved management and treatments over time, rather than screening per se, and so they used the pre-post difference in the still-not-screening regions of Denmark to estimate this and calculated a double difference to measure the portion attributable to screening.

Ireland's BreastCheck program began in 2000 in these 11 eastern counties. it was extended to these 3 in these years, and the last 12 at the end of 2007. We focused on the earliest and latest. Different from most programs, screening in BreastCheck used to end at 64 rather than 69 (the extension 69 is being phased in now). 59 / 2020

First, Sisse's 2015 analysis and results for FUNEN and the 8 times bigger non-screening comparison experience. 16 / 2036

The cross-product ratio of breast cancer mortality rates gives an adjusted hazard ratio of 0.78, or a 22% 'reduction' that they (cautiously of course) attributed to the screening program. i.e., there were an estimated 22% fewer breast cancer deaths than there would have been if they hadn't screened for these 14 years. 52 / 2088

Now to Ireland. Recall the basic comparison, between 2 regions that started screening almost 8 years apart. But what if these two regions had different mortality rates even in the absence of screening? The Irish Cancer Registry did not begin until the mid 1990s so we could not use the same type of historical comparison that was used in Denmark. So we opted to stay entirely in the 21st century, and for a 'control', use the experience of women who were already OLDER than the upper screening age of 64 when screening was first introduced in 2000. These woman-years allowed us

to check if there were differences in the background cancer death rates in the 2 regions and to correct for them. ↓ 123 / 2211

Let's look first at these older WY – lived by women born before 1936. As you can, see the death rates are very close, but slightly lower in the first (eastern) region to start screening that the western region where women 50-54 had to wait. So when we compare the rates in the screen-eligible WY, we will have to handicap the east just a tad. 65 / 2276

What happened in the same 14 years in the woman-years targeted by screening? As you can see the 2 rates were 12% lower in the region that started first. 29 / 2305

So when we take the ratio of the hazard ratios so as to handicap the East, we get a corrected HR of 0.91, ie. a 9% difference. We can interpret this as saying that the almost 8 years' more screening led to 9% fewer deaths in East in the 14 years. But what if we asked the more relevant counterfactual comparison: how would the rates in the East have looked relative to those we would have seen there if the program had not been introduced at all? Or if there were a full 14 year gap between the East and the West, like in Denmark. The 9% is merely a lower bound. Because of

the delays before the full results in the East are realized, we can only conjecture as to how much more than 9% it would have been. 140 / 2445

Two of the problems with the meta-analyses of the old trials, and even the better population-based comparisons, is that they largely ignore the delays before mortality reductions show up, ie that hazards are inherently non-proportional, and the variation in numbers of invitations. Amy Liu's thesis took the fundamental parameters to be the effect of 1 round, and used them to built up a bathtup shaped HR function over the trial follow-up time. She only applied her model to trials, but we dont have recent ones in mammography. 87 / 2532

Remember the FUNEN data I showed you earlier. 8 / 2540

In 2015, I contacted Sisse and suggested we try Amy's simple parametrization but add the age-dimension to the program-year dimension. 20 / 2560

She was keen, and updated the follow-up to the 22 years shown in this compact Lexis diagram that drops the pre phase. The black dots every 2 years are the invitations to Funen women, stopping at 69. (This right hand wall is when the next part of Denmark started screening). Those aged 69 when the program started got just 1 invitation. Other birth cohorts got many more. The Rest of Denmark is white dots every year. 76 / 2636

The easiest way to understand the 2 fundamental parameters in our model is to think of an unscreened population, and women whose cancers that proved fatal in say 2019. Then ask oneself, if these women could have been offered just one screen, when in the past would have been optimal and what percentage of them would have had these deaths averted because of the earlier detection and treatment? 68 / 2704

In the blue curve in the diagram on the right, the sweet spot (tau) is about 7 years earlier and the maximum percentage (delta) is about 8%. The blue curve is the probability of being helped if the 1st and only screen were x = 0, 1, ..., 22 years before the cancers would(otherwise) have proved fatal. The black curves show the probabilities for 2, 3, 4 .. rounds and can be thought of as convolutions or amalgamations of the benefits of multiple screens. I will leave the details on the left to question time. 95 / 2799

On the left are what the data look like, one row per Lexis cell. The first row is for those aged 87 in 2014. 11 died of breast cancer in the Rest of Denmark, and 1 in Funen, among approx 17,000 and 2,000 respectively. The Funen women had received 2 invitations, 20 and 18 years earlier, when they were 67 and 69. Those in the second row had had 4 and those in the third had had 7. The no. and timing of the invitations are the x's in the HR regression function.

Here are the fitted mortality deficits or % reductions, based on convoluting the fitted parameters and invitation histories. Those in the uppermost diagonal had just 1 invitation, so the top numbers are the fitted blue curve for 1 round of screening, reaching a nadir of 8% at year 7. The lower down ones had more invitations and so the trough is deeper and longer. The overall reduction is about 19%, an average of reductions ranging from 0% to 30%. 79 / 2971

FUNEN's 14 year time gap made it a bit easier to fit the 1-round parameters. When I initially tried the model on the Irish data, with less than an 8 year gap, and treated the west as entirely unscreened, I had trouble, so the PLOS article only had the less-meaningful overall 9% difference. I have since refined the data-analysis to allow for the second startup and am now able to report the 2 fitted parameters and the fitted HR function over the Lexis space. 84 / 3055

Here are the 2 sets of fitted mortality deficits, one for each region. Women in the uppermost diagonal in the FIRST region (where the larger bold numbers are) had just 1 invitation – in 2000; the fitted curve for 1 round of screening reaches a nadir of 6% at year 6. In the lower down ones the troughs are deeper and longer. The largest reduction is about 19%. In the second region, women in the several uppermost diagonals had no invitations, so had 0 benefit; in the most-often invited, the fitted reductions have only reached 10%. The 2 sets of Hazard ratios explain why the

average difference between the regions was only 9%, and it will get smaller as the follow-up is extended. 123 / 3178

we have more detail here [I will repeat this at end] 11 /

3189

So, to summarize... We have been trying to get those who calculate the benefts to use the basic cancer screening principles. Our parametrization is minimalist, and leads to non-PH HR functions that are more realistic and more meaningful than 1-number summaries. It applies both to trial and population data. As for Breastcheck in particular: When they started, this was their stated goal. to reduce breast cancer mortality by 20% in 10 years. For those cohort of women who on the main diagonal, i.e., invited from age 50 onwards, we think that close to a 20Thank you to my

collaborators, to my funder over the last 8 years, as well as the Institute that paid my Air Canada ticket when I went to the University of Waterloo 50 years ago this September.

37 / 3500

EXTRA MATERIAL 2 / 3502

Some references, 2 / 3504

This Big-Data study in a high-impact journal overlooked two important principles. Its clever idea was to look at pairs of jurisdictions where one started screening well before the other. For example, the North of Ireland started 10 years before the first half of the Republic did. 46 / 3550