Suggestions regarding the analysis of the UKCTOCS data

James Hanley

2019-10-14

Dear Professors Parmar and Menon

I am happy to respond to your invitation of 24 September, and to provide my views to the UKCTOCS Trial Steering Committee. Section 1 is devoted to bringing out the guiding principles that arise quite naturally if we reflect on the intent of cancer screening and what the time-graph showing the yearly 'returns on investment' would look like if screening (a) helped nobody, or (b) helped some people. Guided by these principles, section 2 suggests various approaches to the data analysis. The Appendix contains my comments on some of the points/proposals (A B C) made in the material you sent me.

Section 1: Principles

Lessons from oncology trials

I spent the first seven years of my career as a clinical trials statistician for the Eastern Cooperative Oncology Group and the Radiation Therapy Oncology Group, with Zelen, Pocock, Schoenfeld, Begg, Lagakos et al. In these trials, at the time of randomization already, every patient had been diagnosed with a cancer, and the aim was to confront it **there and then**. We got to see quite quickly, within a matter of at most a few years, if the newer treatments were any better than the older ones at changing its course. A trial that took more than 5 years was considered very long. Hazard functions were usually proportional (*HR==1*) **after** 5 years, when only the cured remained. Before 5 years, since one was often dealing with an unknown mix of aggressive and less aggressive tumors, the hazard functions themselves had sharp early peaks. Fig. 3 on page 17 of this work http://www.biostat.mcgill.ca/hanley/Reprints/HanleyMiettinenIntJBiostat2009.pdf shows the fitted hazard function for 1 arm. A hazard ratio of 0.8 (an average over years 0-5) might be interpreted as telling us that the newer treatment cured 20% of the ones that the older treatment could not cure. The extra cures were **achieved up front**, using the radical therapy at time 0, but the biggest portion of the **deficit in mortality** (the return on that newer treatment) **showed up** at the time the hazard rate peaked in the control arm.

Cancer screening

I then moved to McGill University, and switched to teaching, diagnostic test evaluation, and collaborative epidemiological research. Even though I had not been involved in any screening trials, I began to take a serious interest in this topic 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.

After reading Miettinen's paradigm-shifting 2002 Lancet article on the Malmo mammography trial http://www.biostat.mcgill.ca/hanley/screening/Miettinen2002Screening2articles.pdf, I not only revised how I taught the analysis of screening data to our medical and graduate students, but also began to revisit past screening trials. I began with the Minnesota FOBT trial and in a 2005 piece http://www.biostat.mcgill.ca/hanley/Reprints/hanley_screening_epidemiology2005.pdf I re-iterated some important first principles (**all to do with timing**) that go back to Alan Morrison's textbook in 1985/1992. [Fig 1 in that 2005 piece suggests a mathematical basis for the hazard ratio curve in Miettienen's piece] In 2011, having published in the 2010 J of Medical Screening http://www.biostat.mcgill.ca/hanley/Reprints/Hanley-JMS-2010.PDF a re-analysis of the ERSPC data (scraped from the 2008 publication) and having reviewed several screening trials in an article in Epidemiological Reviews, 2011 http://www.biostat.mcgill.ca/hanley/screening/EpiReviews2011.pdf I obtained a Canadian Institutes of Health Research grant to address the quite consequential deficiencies that I had noted.

The dedicated website http://www.biostat.mcgill.ca/hanley/screening/ contains not only the grant application itself, but also annotated output from that grant, along with several relevant publications by others and by us, as well as several conference presentations and posters. You will also see that I include the 2015 report on UKCTOCS, and a rough re-analysis of the data, shown in this graphic http://www.biostat.mcgill.ca/hanley/screening/UKCTOCShrFunction.pdf . The bottom half of this figure appeared in a poster Sisse Njor and I presented at the International Cancer Screening Network (ICSN) conference in Washington DC in the summer of 2017 (see below).

Beyond these, of particular relevance to this submission are

- The 2014 thesis http://www.biostat.mcgill.ca/hanley/screening/LiuMeasuringMortalityReductionsDueToCancerScreening pdf by Amy Liu, which develops a statistical model motivated by cancer screening principles, and applies it to data from colon and lung cancer screening,
- The 2014 International Statistical Review paper http://www.biostat.mcgill.ca/hanley/screening/LiulIntStatReview2014.pdf (a shorter version of the thesis),
- The 2017 poster, with Sisse Njor, presented at the International Cancer Screening Network http://www.biostat.mcgill.ca/hanley/Reprints/HanleyNjorEJE2018.pdf begins with guiding principles, reviews past studies, and (before turning the population data from a screening and a non-screening region of Denmark) **re-analyzes the data scraped from the UKCTOCS Lancet article**.
- The 2018 EJE paper with Sisse Njor http://www.biostat.mcgill.ca/hanley/Reprints/HanleyNjorEJE2018.pdf (where we apply our model to population-based data on mammography from two regions of Denmark, and are a good bit less technical than in the ISI Review article),
- The commentary http://www.biostat.mcgill.ca/hanley/Reprints/CommentaryEJE.pdf from a senior Dutch academic who has long been involved in cancer-screening. Paragraph 1 of his piece nicely summarizes the core issues (the same ones I expound on here) and the first sentence of the second paragraph is a partial answer to your query. I return to this in option 4 of section 2.

However, I will first suggest a simpler starting point for the analysis of the UKCTOCS data. The poster presentation at the 2017 ICSN contains, at the bottom of column 2, an empirical hazard ratio function I derived from the data I scraped from the 2015 UKCTOCS report.

And among my assignments for our biostatistics graduate students http://www.biostat.mcgill.ca/hanley/bios601/intensity-model-inference-plan-2019.pdf, you will see (Q15, page 22) an assignment of sample size for the Mayo Clinic Chest X-ray lung cancer screening trial that began in the early 70s – and that has a 25 year follow-up in the year 2000 – 20 years after the last chest X-ray. It still has lessons for us, even today.

If I were to contrast my 7 years in oncology with my 25 years studying screening, I might summarize it by saying that cancer screening works by an entirely different mechanism than the traditional therapeutic 'intervention' aimed at persons all of whom are already known to have the cancer in question. It also induces far more complex time relationships. The 25:7 time ratio is pertinent: considerable patience and 'waiting around' are required to see the fruits of cancer screening. This should not be surprising, especially if one asks how can/does screening achieve its intended goals.

Different mechanisms and more complex time relationships: patience required

The 2017 ICSN poster http://www.biostat.mcgill.ca/hanley/screening/HanleyNjorICSN2017.pdf contains many real instances of the 'delayed effect' that the material you send me refers to, and also shows several cancer-screening-principles-based approaches to measuring the mortality reduction**s**. Note the use of the **plural** in 'reductions'. Of all the applications of the proportional hazards model that one could envisage, cancer screening is the least suitable – unless of course we have a completely useless screening test (or useless treatment), in which case a single-number hazard ratio (HR) would be ideal: $HR = 1$ for each and every year after randomization.

Examples of **immediate and sustained reductions** in mortality or infection or other event rates are very rare, and limited to contexts such as vaccines, adult circumcision, abdominal aortic aneurysm (AAA) screening, screening for heart abnormalities in young athletes, etc.

[A passing comment: the report on the UK Multicentre Aneurysm Screening Study (MASS) - a randomized trial of one-off abdominal aortic aneurysm screening by S G Thompson et al. – http://www.biostat.mcgill.ca/hanley/Reprints/LonelinessOfLongDistanceTrialist.pdf highlights an important point that goes to the core of the issue we are concerned with. The hazard ratio is about 0.6 (the reduction in mortality is about 40%) right from the get-go, and stays so for many years thereafter. The reason for this constant-over-time HR is that (ultimately fatal) AAAs do not relentlessly progress the way (ultimately fatal) cancers do. One might think of them more like bulging tire-tubes that rupture if they strike a random nail on the road, and that nail is equally likely to appear next month or next year as it is 5 or 10 years from now. Finding and repairing AAAs today will bring dividends every year, just like adult circumcision carried out today is a gift that keeps on paying dividends (reducing the rate of acquiring HIV), **right from the start**. Vaccines do the same, but some may need a 'booster' 20 or 30 years from now. We cannot expect the same immediate and sustained effect of PSA screening or ovarian cancer screening: the biology is different and the mortality-reduction tools are different. One 'round' does not give immediate and sustained returns; moreover, those whose latent cancers were not detected at round one (and therefore were not helped by this round) have to wait for the next round ('boost') where the process starts all over again.]

Delayed effects of screening

Delayed but transient effects are more common, for example with the use of alcohol (on driving ability), blood thinners, statins, and beta blockers, but even then we need to be careful to be guided by the biology when defining our estimands, and not to hope that the effects of a last drink/beta-blocker/round of screening some 24 hours/ 90 days/ 20 years ago will still be evident now, when we analyze the data.

Pharmacologists use **time-curves** to show the full pattern of repeated drinks or drugs and screeners would do well to think and analyze data the way they do. Screening has a lot in common with taking statins. As one statistician told me once, to keep people's cholesterol down, one must keep pumping statins into people; to keep a population's cancer mortality rates down, one must keep pumping screening into it. Indeed, in the second picture in the first column of that 2017 poster, I show time curves from early trials of statins in humans. These curves should encourage us to reflect on the form of the hazard ratio curve we should expect from repeated 'administrations' of our early detection tests (coupled, one should add, with effective treatment).

The principles that should guide analyses of data from screening trials might be practiced more if we replaced the word 'screening' by the term used by the early biostatisticians (such as Marvin Zelen) who worked in this area 50 years ago, namely **early detection**, or better still, the phrase **earlier detection and treatment**. If we were to use, and think about the meaning and implications of, this phrase, we would immediately see that it makes little sense to look for immediate reductions in mortality rates. A bit like with law, the material you sent me seems to rely a lot on data instances/precedents to justify some of the approaches. The extensive knowledge we already have about cancer, together with the foundations on which cancer screening is built, could be given higher priority. We might want to modify Deming's maxim "In God we trust; all others must bring data" to 'we start with logic and test it using data'.
The delayed dip in the HR function should be the default model – unless of course the screening tools and treatment are so poor that there are no mortality deficits anywhere along the time-scale.

If the early detection is **today** (at the first screen) then the 'normal' (unaided) diagnosis occurs at some time in the future, well after randomization to the first screen. Moreover, that otherwise-fatal cancer won't prove fatal until some time after that again. And, if some early detections (and earlier treatments) are at/after the 7-th round of screening, then the deficits in mortality they generate are maybe at year 13.

Even if there were just 1 screen (today), any mortality deficits it generates have to be well into the future. Miettinen sees this delay as the sweet spot in the 'detectability-curability trade off'. Not-yet-diagnosed cancers that – in the absence of screening – will prove fatal only a few years from now might be readily detected today, but would probably be already incurable. Not-yet-diagnosed cancers that will prove fatal 20 years from now would be less easily detected today, but would presumably be more curable if they were detected today. The 'sweet spot' for the timing of this one screen (or not-too-early and not-too-late, as Goldilocks would say) can be thought about by working back from the year/time (t) when the cancer proved fatal in the absence of screening, to the ideal screening time t-x. Miettinen told me that when the US-based TV journalist Peter Jennings https://en.wikipedia.org/wiki/Peter_Jennings died of lung cancer in 2005, Miettinen's academic chest physician colleagues asked themselves: if Jennings could have had (just) one low-dose screen x years before, what would have been the best x?

When there are multiple rounds, it is a bit more complicated, but we should and can still think about it biologically. Analyzing the data using a **convenient** model that does not reflect the principles from cancer screening is likely to lead to results that are not realistic.

The model that Liu and our group developed (see URL above) contains that x as one of its parameters, and we estimate it (the 'when'), along with the 'how much good' each one round does, directly from the data.

A further consideration peculiar to screening: slow/fast, late/early

Starting back with Zelen and Morrison, the early thinking about the detection process and the biology and the early writings on screening highlight an important consideration, namely length biased sampling. Naturally, it is insidious in a one-arm study with no comparison arm, but it is also very relevant in an RCT, particularly if some of the cancers in question remain silent until they are quite far along. Screening preferentially detects slower-progressing cancers. The deaths in the control arm in the earlier years of a trial would be produced by faster progressing ones, and so one should not expect their counter factual (screening arm) course to be much better than that. If there is some (latent) mix of progression rates, then one would expect to see any effects of earlier detection and earlier treatment in the slower progressing ones – i.e., in the ones that (otherwise, i.e., in the absence of screening) would prove fatal in the **later** years of the trial.

Why adopt a purely null-hypothesis testing approach? Why not pursue a relevant screening-principles-based estimand?

The null-hypothesis testing approach dates back to the HIP mammography trial and the Mayo Lung trial, where the focus, and the sample size considerations have always been on a test of the null. But surely the question is not whether (a particular, and particularistic) regimen of screening does **some** good, versus **no** good. Just about everyone would agree that earlier treatment prompted by a proper science-based screening regimen (or even just one round) will avert cancer deaths in **some** (possibly a few, but definitely non-zero) persons. So, why formally test the null, when we are pretty sure the absolute null is not true? Surely, the question is **how much** good it would do if applied for a specified duration (e.g., starting at age 50, until age 69 say and frequency (e.g., every year, every two years, . . .) and what the 'costs' (downsides) are. To describe the good, the benefits (the yearly dividends, i.e., the mortality-reductions) need to be documented each year for as long as these yearly dividends are non-zero – or modelled for as many years as they would be non-zero in a program where everyone is invited starting at (say) age 50, and invitations stop at say age 74.

The likely form of this **yearly-dividends-curve** was set out by Miettinen in the Lancet in 2002, and earlier (a bit more loosely) by Morrison in 1985 and 1992.

As we say in our 2017 (Hanley-Njor) poster, the usual single-number hazard ratio is an **average**, not only over all the follow-up time **there happens to be** as of the time the data are analyzed, but also over women who (because they were near the upper end of the screening ages when the study/trial began) had just **1 or 2** screens or invitations, and other women (at the lower end of the screening ages) who had **many more**. See the diagram (in the 2017 ICSN and 2018 EJE material cited) showing the varying screening/invitation histories in the Lexis space relevant to the women in Funen and the 'Rest' of Denmark.

Sadly, just about all of the meta-analyses of the mammography trials – still the basis for Task Force recommendations – produce a single number hazard ratio, one that is averaged over whatever number of **rounds** each trial used, however many invitations the **different** birth cohorts in the trial received, however many years of **follow-up** there were, and however long it was **since the last screen**. It makes little sense to look for effects early on, but it also makes little sense to average over the person time in the 20-th year after the last screen.The key concept in cancer screening itself is time-sensitivity; being **time-aware** is just as important in the **data-analysis**.

Averaging of mortality reductions over all available follow-up time is a bit like summing/averaging over all of the fatal cancers in the entire GI tract even if the scope only reaches part of it, or averaging over all persons and arriving at a mean of 1 ovary and 1 testicle, or, as Francis Galton said in his 1889 book Natural Inheritance

"*It is difficult to understand why statisticians commonly limit their inquiries to Averages, and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once*"

The overall 20% reduction in the first report on the ERSPC is a vivid example of this: an average of 7 years of (much more precisely measured) null reductions and five years of sizable (but very imprecisely measured) ones. In addition, because of the very long accrual period (with some countries joining in as many as 10 years after the first ones) the man-years, and thus the numbers of deaths, were heavily front-loaded (see the population-time plot in the right hand side of Figure 7.3, page 140 in Amy Liu's 2014 thesis). UKCTOCS and NLST were much speedier in their accrual, and so the population-time plot (No. at risk plotted versus years since randomization) is much more "rectangular', and closer to what it would be if all women started at age 50 – as one might want to consider when trying to project the impact if the trial regimen were to be put into practice. (Moreover, there is no averaging over different screening frequencies, as there was in the ERSPC)

A further downside of the prevailing practice of announcing results as soon as the p-value dips below 0.05 is that it makes the single-number hazard ratio quite arbitrary. It is merely the (over all follow-up time to date) HR that happens to prevail at the time then the p-value dipped below 0.05. In my 2011.06.23 presentation http://www.biostat.mcgill.ca/hanley/Reprints/HANDOUT-EpiCongress-June23-2011.pdf to the North American Congress of Epidemiology Symposium under the topic Randomized Trials of Cancer Screening: How Useful are They?, I asked "Randomized Trials of Cancer Screening: How Big Are The Mortality Reductions Produced by Cancer Screening? Why do so many trials say 20%?". One of my headings was "How to stop a screening RCT at a 20% mortality reduction? [Theorem]". It seemed too co-incidental that no matter the cancer, or the screening tool/strategy, several recent reports (FOBT, ERSPC, NLST) were all stopping when the HR was 0.8.

One would not do this if one wanted to know how much statins reduce a person's serum cholesterol. If we were to measure my serum cholesterol every week after I began taking it, we might be able to show that the reduction was 'statistically significant' by week 12, when it was just 10% down, but it could well be that its has not reached its steady state – maybe a 30% reduction. The only way is to let my serum cholesterol take its full course, and to let the full time-curve emerge.

So far, I have gone on at some length about principles of cancer screening that few could disagree with. Nevertheless, these self-evident but inconvenient truths/principles are not routinely employed when measuring the benefits. Continuing to ignore them by using an arbitrary single-number HR and a p-value will give a very blurred, incomplete and misleading picture of how much/little good screening did for the 100,000 participants, or of how much future women might expect from a screening regimen based on these screening tools.

One needs to pursue an estimand comprising the full HR curve, or as much of it as we can measure.

Pursue a Hazard Ratio Function, not a HR scalar

Fortunately the HR function will be smooth in time, and even parametrizable by a few **interpretable** parameters that can be fitted to the data. I suggest some such parametric forms below. My options 2 and 3 do not take any account of the numbers of screens/invitations; option 4 does.

The effective number of data points is quite small, so spend the precious degrees of freedom wisely.

Even though one has 200,000 subjects and 2-3 million women-years of follow-up, and by now several hundred deaths for ovarian cancer, the estimand is a HR **function** that is a function of years since entry. As Liu et al. have demonstrated, one gains very little by working with individual data and continuous time. One does quite well by using rates (hazards) and rate ratios (hazard ratios) based on aggregated data in 1/2 year or 1 year bins. So if one has 18 1-year bins, the sufficient statistics are the yearly numbers of deaths and person years in each arm. For a good example that shows what is needed, see Figure 1 in the 2010 J Med Screening article http://www.biostat.mcgill.ca/hanley/Reprints/Hanley-JMS-2010.PDF or Figures 3 and 4 in the 2014 ISI Review article http://www.biostat.mcgill.ca/hanley/screening/LiulIntStatReview2014.pdf.

So in all, with 1-year bins, and 3 arms, and say 19 years of UKCTOCS follow-up, there are at most 57 degrees of freedom, or at most 38 if we combine the 2 active arms. Mind you, some of the early years have few deaths (and even if they did have more, there is little information in them on how big the 'returns' on the early screens must have been) so the real degrees of freedom are fewer. In this case, rather than estimating 2 or 3 hazard curves, one per arm, why not estimate just 1 (or 2) **hazard ratio curves**? The baseline hazard function for the control arm is not of intrinsic interest, but rather a nuisance curve that costs precious degrees of freedom to fit.

For the data from other cancer screening trials and from populations, we found that a **conditional** approach, where (with 2 arms) we modelled the year-specific binomial splits of the deaths between the two arms, gave us directly the HR function we were pursuing. So we spent all our degrees of freedom just fitting the HR function. Moreover, because the numbers of woman years in your case are just about 1/3 : 1/3 : 1/3 in every follow-up year, the log of their ratio is a known offset (0) – it wasn't quite zero in the ERSPC data, but it was in the FOBT Minnesota trial and the NLST. As we explain in Fig 1 of the 2018 EJE article http://www.biostat.mcgill.ca/hanley/Reprints/HanleyNjorEJE2018.pdf, the split in each follow-up year is binomial with 'n' = the total deaths that followup year, and the proportion parameter is HR/(1+HR). The HR is then what you parametrize as a function of follow-up time. The Danish data are 2-D (Lexis cells), because there is age and follow-up year. You could keep your HR function as 1-D (follow-up year) or expand it to 2-D if you want to bring in different HR curves for women in different birth cohorts, who got different numbers of invitations. I go into that below, after I first consider some simpler HR-function estimates.

Section 2: Suggested approaches

1. Model-free HR trace, following Miettinen 2002

This merely involves some degree of smoothing of the HRs for different follow-up years. When re-analyzing the data from the Malmo mammography trial – the only trial that screened for more than 4 rounds – Miettinen http://www.biostat.mcgill.ca/hanley/screening/Miettinen2002Screening2articles.pdf faced single-digit numbers of deaths each year (see Table 1 page 2 of the technical article), so he used a 3-year moving window. I used this smoothing as the 'first cut' of the ERSPC data, and for the 2005 Epidemiology article dealing with colon cancer.

I also used this smoothing to obtain the 'empirical' HR function http://www.biostat.mcgill.ca/hanley/screening/UKCTOCShrFunction.pdf for the data I scraped from your 2015 Lancet article (and put in the 2017 ICSN poster), but I made fancier CI's, using bootstrapping. Note that I did not even need the woman-years in each age bin, since they were essentially equal, and all I was pursuing was a (follow-up-time-specific) HR function. I **matched** on time, **rather than** spend degrees of freedom **modelling** it. Here is a link to the R code that generated this HR function and associated boostrap CI: http://www.biostat.mcgill.ca/hanley/screening/UKCTOCSreAnalysis.R.txt

The benefit of this approach is that it lets the data speak.

2. Model-based HR-function, following Hanley 2010.

The approach is explained in section 2 of the supplementary information for that 2010 article that re-analyzed the ERSPC data. The key here is to be guided by a (smooth) HR-function form such as that in the Figure entitled 'Follow-up experience in a randomised controlled trial comparing screening for cancer with no screening in respect to cause-specific mortality: interrelations of parameters' on page 2 of Miettinen's 2002 Lancet article http://www.biostat.mcgill.ca/hanley/screening/Miettinen2002Screening2articles.pdf. How fancy you want the curve to be depends on how many parameters you can reliably fit, but I expect that even with the amount of data you have, the number cannot exceed 4 or 5 (assuming that you only have fewer than 40 df.)

3. Model-based H functions, 1 per arm, as per Approach C.

I like the idea of 3 splines, as long as they yield a HR function that looks plausible. I worry a bit about spending many more degrees of freedom, and of constraining each H function. I think it would be better to force a form on the HR function itself, and if that could be done with a sensible (constrained) spline, that would be fine. We always have the empirical curve from approach 1 as a way to judge if we have succeeded.

4. Fit a model which contains parameters that describe what (each) round of screening accomplishes.

This is the more desirable model, because it yields parameters that are independent of, and not influenced by, how many rounds of screening each woman had, or how long the follow-up was, or what the stopping rule was. Its fitted parameters can be used to project what screening that starts at say 50 and ends at say 74 might produce over the age-span 50-85 — or whenever the effect of the last screen fades. In the FOBT data from Minnesota, we used the model to 'fill in' some missing years of screening. Because the numbers of deaths in the early years were fewer than had been planned for (I think the UKCTOCS had the same issue), the Minnesota investigators ran out of funding, and screening was interrupted for several years until they secured more. The screening hiatus is quite apparent – albeit with an instructive delay – in the raw data even (see Figure 2. p788 in my 2005 article, and Figure 5.4 page 98 in Amy Liu's thesis, http://www.biostat.mcgill.ca/hanley/screening/LiuMeasuringMortalityReductionsDueToCancerScreening.pdf or the middle column in the 2017 poster.) Sometimes, our mistakes are the best way to grasp the correct principles.

One downside is that the (empirical) HR curve may not have started to revert towards 1, i.e. it may still be at a steady-state nadir. This would make it difficult to fit the model.

Section 3. Comments on proposed approaches A B C

APPROACH A: To continue with a Cox model fitted to all accumulated data, i.e., analyse all the data generated from 2001 to 2020 in the same manner as the original analysis, using a Cox/logrank test which is most powerful under proportional hazards.

The claimed benefits are continuity, appearances, simplicity, and avoiding fitting a hazard function for control arm. But it is also conceded that it is less than optimal.

To me, the first two benefits are more philosophical, and a matter of one's outlook, and hard to quantify.

Proposed option 1 in section 2 does not fit the nuisance function either, but has the advantage of being more cancer-screening-principles based, and also data-based. The worry I have about Approach A is that the answer is not invariant to the amount of experience, and so any average HR is arbitrary. Indeed it is guaranteed to be null if it involves just the first few years, but it is also guaranteed to head back to the null the more follow time is included – just like the newest report (NLST Research team. Journal of Thoracic Oncology, 2019. Lung Cancer Incidence and Mortality with Extended Follow-up in the National Lung Screening Trial) from the NLST. The 20% reduction over 6-7 years has now become 8% over 12 years. The same dilution happened with the Minnesota FOBT trial.

One way to judge – a priori, before even embarking on a trial – whether approach A is justifiable is to ask whether the 200,000 x say 15 = 3 million woman-years of follow-up are exchangeable. In other words, can we design the trial to assemble these 3 million years of experience and get the same HR (or other metric) by following 3 million women for 1 year, or 1 million women for 3 years, or 1/2 a million for 6, or 200,000 for 15, or 100,000 for 30 years?

All of these designs would give similar answers if we were studying vaccination-induced reductions in HPV infections in college age students, or HIV acquisitions following adult circumcision. Approach A would not be suitable if the HPV vaccination were at age 5, or the circumcision at age 0. Cancer screening works by a different mechanism where time is of the essence, and the person-years are very far from exchangeable with each other. One has no option but to be patient and let the data play out and to look in the right window.

Once it was pointed out that the observed HR function in the ERSPC made biological sense, the Lancet report did split the results by follow-up time. The sad part is that the first NEJM 1-number HR of 0.80 (20% reduction) got to be the one that was remembered. [There is also the added issue, seldom remarked upon, that the sizable reductions in years 8-12 in the first report were largely driven by the data from Sweden, which was screening every 2 years (everyone else was screening every 4). At that point only the man-years in years 8-12 were mainly contributed by Sweden, one of the earliest to accrue men. In subsequent analyses, many of the man-years in this window came from countries that screened less intensively, and the stronger effect in Sweden was diluted out]

So, in summary, it is very difficult to defend Approach A, all the more so when several defensible biological alternatives exist.

APPROACH B: To model just the 'new' data acquired since censorship of the original analysis (2015 onwards)

This approach would ignore the data that has been previously analysed, and start the analysis point from the end of censorship of the original analysis.

The main motivation seems to be statistical purity, and statistical correctness. But any HR it would produce would be arbitrary in its women-years-composition, and disconnected from the earlier data. As I argue in sections 1 and 2, the full HR function is a more justifiable estimand.

Also, by focusing on an a-priori estimand (such as a presumed non-null HR function), we get away from the focus on p-values and power and null-hypothesis testing in Approaches A and B.

As for the issue of time scale, there are in fact two scales, both relevant, particularly if one were to embark on proposal 4 in section 2. One is time since randomization, and the second is age (one where the HR function is modified by the number of invitations each birth cohort received). See the Hanley-Njor EJE paper of 2018, and the Lexis diagrams in it. These provide a natural approach, and they make the 'before 2015' versus 'after 2015' boundary look artificial. The HR function is a function of both scales.

[Incidentally, the p-value-based arguments invoked in approach B are similar to ones used after the first report of the ERSPC, when one editorial noted that the ERSPC study had 'not fully matured, and it is essential to continue the follow-up in each group'. It added, 'unfortunately, the authors of the ERSPC have already performed 3 interim analyses. The criteria for statistical significance in subsequent analyses have become much more rigorous as the number of interim analyses has increased. The ERSPC has 'eroded its alpha,' meaning it may have difficulty conducting future statistically valid analyses. It may be impossible for future analyses of the ERSPC to have a statistically significant finding that screening is beneficial.' But maybe the alpha was spent looking in the wrong time window, like the person who only searched for lost keys under the lamp post. The same question needs to be raised in the present context.]

APPROACH C: To use a model and test on all accumulated data that allows for a delayed effect

It would "analyze all 19 years of mortality data from 2001 to 2020 with a model that allows for a delayed effect. The Royston-Parmar model is a flexible parametric model estimating the cumulative hazard with cubic splines for each arm of the trial. The test for a mortality effect is the multivariate Wald test for the difference between the coefficients of the two splines - as specified by Royston and Parmar."

It makes good biological sense to adopt the delayed effect as an essential element of the analysis model, and good statistical sense to be somewhat flexible. However, one needs to ensure that the model parameters have a direct interpretation. Moreover, instead of a single p-value as the output, it might be better to produce a multivariate confidence region for the parameters.

Section 2 above also made a case for being flexible but suggested modeling the HR function itself, and possibly taking advantage of the flexibility of splines. There is a saving in degrees of freedom if instead of modelling follow-up time, one matches on it and removes it by conditioning. That way, the parameters are also more directly interpretable, and one has a way to check the fit against the (locally smoothed) HR curve described in option 2 of section 2.

I am happy to elaborate on this if need be.

Respectfully submitted

James Hanley

http://www.biostat.mcgill.ca/hanley/

and

http://www.biostat.mcgill.ca/hanley/screening/