Quick summary of HSE models for forecasting mesothelioma incidence
Internal document

Timothy Wyant

8/30/13

Contents

List of Figures

1 Introduction

As I mentioned in an email, I would like you to undertake a project involving estimating a form (or several forms) of the so-called HSE model of mesothelioma incidence in the U.S. I would then like you to forecast future annual mesotheliomas in the U.S. using one or more of these models, along with some error bounds. Finally, we will use these national forecasts to make forecasts for individual asbestos trusts. Typically, I do this by (1) calculating the ratios of recent mesothelioma claims counts against a trust to national counts, by birth cohohrt, and (2) applying these ratios to future predicted national incidence counts from your model, again by birth cohort.

The HSE model is basically a poisson regression of mesothelioma incidence as a function of past exposure. The poisson mean for any birth cohort in a given year is somewhat complex nonlinear function of cumulative exposure, and some biomedical parameters related to latency, intensity of exposure, and the manner in which risk increases as a function of time since exposuree.

This note provides some background on the HSE models, and other models that have been used to forecast asbestos claims. This note also outlines issues that have arisen regarding design and fitting of mesothelioma incidence models, and the treatment and use of demographic, historic incidence, and exposure data that provide the independent varialbes for the fitting.

Along with this note, I am providing a zipped directory that contains:

1. PDFs of relevant references,
2. A summary xls that provides a master list of all the references by category,
3. Demographic and mesothelioma incidence data,
4. Summaries of my previous fits of the HSE and other models,
5. An Rstudio project that includes the Rnw program that produces this document, along with some R programs that process demographic and incidence data.

For purposes of this note, I have created my own idiosyncratic bibliographic reference system. The summary xls mentioned above crosswalks the references to the pdf files in the zipped directory.

This background note does not pretend to be an exhaustive discussion of mesothelioma model fitting and forecasting issues, but hopefully it will give you a good start.

2 Projecting mesothelioma incidence

Mesothelioma has a long latency, so even though U.S. occupational exposure ended (in theory, and disregarding some more recent exposure for brake mechanics) around 1980-1982, we expect to see mesothelioma claims against these trusts until around 2060. [1] Generally, the annual number of mesothelioma claims against one of our trusts is assumed to be a fixed fraction of the U.S. and Canadian mesotheliomas. Canadian claims are generally treated as U.S. domestic claims, in terms of trust evaluation rules and level of compensation. (There are many fewer Canadian mesotheliomas than there are U.S. mesotheliomas, or course, just because of the difference in population size.) Now, "fixed" should be in quotes because sometimes (especially with a small trust) additional exposure worksites are identified for a given defendant. And when

---

[1] You'll see 2050 as the furthest year in a lot of projections, including some of mine. This is because it has been relatively recently that reevaluation of the demographics of exposed workers has suggested that we need to move the outer boundaries further into the future. In truth, it doesn't matter very much, because there are not many claims projected to occur in the 2050s. Other uncertainties in the forecasting process are generally an order of magnitude greater than the uncertainty regarding the scale of continued incidence after 2050. But we should gradually modify our forecasts to go out to 2060.

a new trust comes online after a long time in bankruptcy (when new claim filings are stayed by the bankruptcy process) it may be difficult to know exactly what the fixed fraction is. But caveats aside, we usually assume a fixed fraction. This fraction in the asbestos expert witness vernacular – aka "the lingo" in this document – is often referred to as the "propensity to sue".

Mesothelioma is almost entirely caused by asbestos exposure, and most of that exposure is occupational. In addition, there are reasonably good models of how risk increases as a function of time since first exposure, and of level of exposure. This summary note is full of qualifiers like "generally" and "reasonably good" because, as you might imagine, given the disputatious environment in which we find ourselves, almost every element of a set of calculations has been or will be a cause for disagreement. Nonetheless, the forest is still a lot bigger than the individual trees, so I will try in this note to focus on the former with periodic asides about the latter. Mesothelioma forecasting is central to:

1. Determine how many mesothelioma are likely to be filed in the future against a trust, and when. (The "when" is important because a trust can earn money on its investments, so timing as well as quantity are elements of the cash flow models.) We do this by first making a national projection of mesothelioma incidence, and then applying a propensity to sue. The declining curves of expected future incidence have different monikers in the lingo, such as "future decline curves" or "runoff curves."

2. Assess "how bad it can get" – many of the large trusts have mesothelioma filing rates that are close to the national incidence rates. So the national incidence rates provide an upper bound in worst-case scenarios. As you've seen, I superimpose a national incidence curve on the bar charts in our quarterly reports.

3. Because mesothelioma is so closely tied to asbestos exposure, and has a developed science, projections of other diseases and conditions (lung cancer, colorectal and other digestive tract cancers, asbestosis) are often derived as some function of meosthelioma incidence. This is done by either applying some ratio to meso incidence, or "inverting" meso incidence to calculate the demographic profile of exposed workers likely to file claims against a given trust, and then applying some disease-specific exposure-risk model to this estimated exposed worker population.

3 Mesothelioma projection methods

Okay, so what is the HSE model? HSE stands for "Health Safety Executive," a UK government agency that is sort of a hybrid of OSHA and the CDC. Among other things, HSE tracks and report mesothelioma incidence (a lot better than we do). And they have been active in recent years in creating and using new models for mesothelioma forecasting. Note that I say "models". Not only have they fit different models, but every time they fit "the same" model they tinker with it, so there are models 1a, 1b, 1c, 2, 3, etc. On top of that, the basic models have been adopted by actuaries and public health folks in the UK and in Australia, and with some of us here in the asbestos trust world. All of us secondary adopters have done our own tinkering as well. So "HSE model" is really more of a genus than a species.

Here is a brief (and not exhaustive) overview of projection methods that are, or have been in use:

1. **Nicholson method.** From William Nicholson, in 1983. This is a nice piece of work, but arguably dated. He took airborne asbestos fiber concentration data from several industrial hygiene studies, BLS data on employment by industry, risk as a function of time since exposure from a longitudingal study of insulators, and NHIS death rates, and forecast future mesothelioma and lung cancer incidence.

The most common method that experts have used in asbestos bankruptcy proceedings over the past decadesd has been to calculate a "propensity to sue" for a particular debtor, and apply it to Nicholson forecasts to get projecte claims. However, the industries that generate the majority of today's claims are not industries that Nicholson included. In particular, the new industry have much younger workforces. Nicholson did not model women, or family members of exposed workers. (It takes very little exposure to elevate the risk of mesothelioma. Whenever an asbestos worker came home from work and bounced a kid on his or her knee, or handed a spouse some overalls to was, he was creating a poential mesothelioma case.) Nicholson looked only at U.S. incidence. He attenuated his worker cohorts in the future by applying death rates that were valid in the 1970s, not today's lower death rates. His risk-as-a-function-of-exposure models have been modified and updated in the epidemiology literature.

This method is still used by some experts, most notably by the experts for the TACs – each trust has a Trust Advisory Committee made of lawyers representing current claims. These experts have principled reasons for espousing the Nicholson method, but I would be remiss if I did not note in passing that this method yields lower future meso counts – by as much as 30-35% – than all other methods currently in use.

2. **KPMG Nicholson** Along about 2000-2005, KPMG did an updated version of the Nicholson method that essentially replicated his work, but updating the industeries (and their workforce age profiles) included in the calculation.

3. **Revised KPMG Nicholson** In 2010 or thereabouts, ARPC (a firm that for while was a subsidiary of KPMG, and is now both the claims administrator and the forecasting expert for many trusts), at my urging, updated KPMG Nicholson with current and projected U.S. death rates.

4. **Peto method** In 1990, ARPC asked me to come up with a metod for forecasting mesothelioma claims for the UNR trust. The Nicholson method was not widely in use at the time. I came up with what is called an inversion method – I took meso claims, calculated the inverse of the risk-as-function-of-time-since-exposure formula, estimated the size and age profile of the exposed workforce for this trust, and moved this workforce forward in time, applying death rates and probabilities of disease in each futue year. Unbeknownst to me, the Brit epidemiologist Julian Peto had done something similar, so ARPC at some subsequent date rechristined this approach as "the Peto method". ARPC also refined my original methods for calclating future lung and other cancers, and nonmalignants, as a function of mesoteliomas.

Somewhere in the late 1990s, ARPC also started using a simple model averaging for their forecasts – they took the median forecast in each future year of the Nicholson, the KPMG Nicholson, and the Peto methods.

5. **Revised Peto method** In 2010 or thereabouts, experts for the TAC in the BW Trust challenged the Peto method. In fact, this challenge had some merit as the method was currently being implemented. In 1990, there were very few meso claims and many nonmalignant claims, so I had jury-rigged a method to supplement the meso claims with nonmalignant claims. In consultation with me, ARPC revised their Peto method to rely only on mesothelioma claims for "backcasting" the exposed population, improved the smoothing methods they used for year of initial exposure crossed with duration of exposure, and while they were at it starte using four models in their averaging – Nicholson, revised KPMG Nicholson, revised Peto, and the HSE model.

In current payment percentage reports like the ARPC report for BW that I gave you, ARPC does not use the word "revised", but in fact their KPMG and Peto methods are now the revised ones that I have just described.

6. **Age-period-cohort methods** These are standard sort of "main effects" models in epi-

demiology. A lot of public health types and academics have published quick forecasts of mesothelioma incidence in various countries, states, and provinces using this method. Unfortunately, it doesn't work very well. Fits and predictive power have been crummy. The basic problem is that asbestos exposure blossomed in the 1920s, dipped down during the depression, exploded during WWII (think liberty ship construction), grew further in te 50s and 60s boom times, and then precipitously declined in the 1970s because of OSHA requirements. There is basically a complex and nonignorable interaction between age at first exposure duration of exposure that is not captured by the vanilla epidemiologic models.

THe elimination of asbestos exposure in U.S. workplaces was a triumph of safety regulation. Many European countries adopted similar workplace standards much later than the U.S. did, and they will continue to experience many more asbestos-caused deaths per capita than occur in the U.S. (In some of these countries, there may also be cancer risks higher than those in the U.S. because more of the asbestos used in manufacturing was amphibole asbestos (crocidolite, amosite, tremolite). U.S. companies mainly used chrysotile asbestos, which is thought to be a less potent carcinogen.)

The main reference you will find to the use of these age-period-cohort models for projecting U.S. incidence is in the papers by Price and Ware. These guys also testify in asbestos bankruptcy about the likely number of future mesotheliomas.

7. **Stallard and Manton methods** These are two demographers from Duke who were hired by the Manville Trust in the 1990s to do a mesothelioma forecast. They developed and inversion method somewhat like the Peto method, but with more bells and whistles. They have published a book about their method, which I have – but in electronic form. It may be worth your picking up either an electronic or hardcopy at some point – you can bill me for it. They work as experts for two trusts right now, and have updated their model to reflect current mesothelioma incidence and claim filing, and current and projected death rates. They're a little more guarded in releasing their updated model, and I don't know of any experts who have tried to reproduce their models exactly. However, I have seen the output of their model for one of our trusts, and it appears to give results very similar to those of the HSE model – which is reassuring. It yields future mesotelioma counts aobut 30% higher than the Nicholson method would suggest.

8. **HSE method(s)** More on this method in later on in this note. It basically fixes the age-period-cohort method by incorporating a more complex interaction between age and period.

9. **Latency methods** This is a poor term, but I wasn't consulted. One could model level of occupational exposure in a country for any past year by looking at the number of tons of asbestos either mined or imported in each year, perhaps with a lag of a year or two to reflect time for the asbestos to move through the manufacturing and distribution system. This seems like like a way to add useful information. However, level of exposure in past years can be inferred from the HSE method for each past year, based on the shape of the period effect component of the model. The HSE-inferred patterns of exposure over time do not match up well with the tonnage imported patterns. And the HSE models yield substantially better fits to the mesothelioma. This is not necessarily surprising, since these models use more degrees of freedom, but it is unclear why they differ so much. Differences in manufacturing patterns and processes that are not captured by sheer tonnage? The thing is, the HSE inferred exposure patterns make intuitive sense. Exposure levels go up in the 20s, down in the 30s, up and up in the 50s and 60s, and way down in the 70s. It's just the slopes and ultimate heights of the exposure curve over time that differs from raw tonnage.

To make matters more confusing, HSE folks in one of their publications[2] look at what

_____

[2] HSE Resarch Report #876 (2011)

happens when you use a couple of different models, one of them being the latency model. But they don't call it the "latency model", so it sort of ends up being "an" HSE model but not "the" HSE model. (HSE refers to it as the "revised risk" model. But they don't discard their original model, and assert that the best estimate of future the future incidence peak is still what they estimated from "the" HSE model, and that the upper bound is still the upper bound that they calculated as one end of the 95% posterior density interval for that model's forecast.

10. **Two-stage clonal expansion methods** This the other of the two alternative models from the HSE research report cited just above. I don't know much about it, and I have not seen anybody else use it to make actual incidence forecasts.

4   The basic "risk as function of exposure" models

The basic biomedical models for risk of mesothelioma and lung cancer as a function of (1) time since first exposure and (2) intensity of exposure come from an EPA review of available models conducted in the early 1980s and published in the Congressional Record in 1986.[3] A good recent article about these models, which concludes (with a bunch of ifs and buts) that the 1986 EPA models are still appropriate.[4]

---

[3]EPA 1986 Assessment (1986)
[4]Berman and Crump (2008)

Mesothelioma Model

The absolute risk model adopted for mesothelioma in the EPA 1986 update is a particular adaptation of the multistage model for carcinogenesis, which predicts that incidence is independent of age at first exposure and increases as a power of time since first exposure (Armitage and Doll, 1961; Pike, 1966; Cook et al., 1969; Peto et al., 1982). The model adopted in the EPA update was based on evidence of the time course of the mortality rate in several cohorts. This model can be derived by assuming that the mortality rate at time t after the beginning of exposure is the sum of the contributions from exposure at each increment of time, du, in the past. The contribution to the mortality rate at time t from exposure to $E(u)$ fibers (f)/ml (as measured by PCM) at time u is assumed to be proportional to the product of the exposure rate, $E(u)$, and $(t-u-10)^2$, the square of the elapsed time minus a lag of 10 years. Thus, as with the lung cancer model [Eq. (2)], the mesothelioma model assumes a 10-year lag before exposure has any effect upon risk. With the additional assumption that the background rate of mesothelioma is zero, the mesothelioma mortality rate at time t after the beginning of exposure is given by

$$I_M(t) = 3 \times K_M \times \int_0^{t-10} E(u) \times (t - u - 10)^2 \, du \quad [3]$$

where t and u are in years, and $I_M(t)$ is the mortality rate per year at year t after the beginning of exposure. (The factor of "3" is needed to retain the same meaning of the mesothelioma potency factor, K_M, as defined in the EPA 1986 update.)

If exposure is at a constant intensity, E, for a fixed duration, D, beginning at $t = 0$, this model can be written in the form presented in the EPA 1986 update:

$$I_M(t) = \begin{cases} 0 & 0 \leq t < 10 \\ K_M \times E \times (t - 10)^3 & 10 \leq t < 10 + D \\ K_M \times E \times [(t-10)^3 - (t-10-D)^3] & \\ & 10 + D \leq t \end{cases} \quad [4]$$

This model predicts that the mesothelioma mortality rate varies linearly with exposure intensity E (for fixed duration, D, and time since first exposure, t). Upon rewriting the model in the

Figure 1: The 1986 EPA mesothelioma model, from Berman and Crump.

5   Previous vesions of the HSE model

6   Historic incidence data for model fitting

6.1   U.S. data

Recent (1999-2009) U.S. mesothelioma incidence data come from the USCS (United States Cancer Statistics) system. This is a collection of registries that covers 98% of the U.S. population. Navigating the USCS website poses some challenges. As of 9/03/13, the full downloadable data (all cancers, not just mesothelioma) live here. You can produce interactive tables of incidence

by age and sex here. These tables are useful for checking computer programs for extractiong mesothelioma data from the full downloads. Also, if you need a quick download, you can download the interactively-produced age by incidence tables. You can create and download corresponding U.S. population tables here as well.

Assembling mesothelioma incidence data for the U.S. is more of a pain than it is for other countries. For years that precede the establishment of the USCS, the only (approximately) comprehensive national tumor registry is the SEER system. You may already be familiar with this system, in which case you can skip to the next paragraph. SEER-9 goes back the furthest. It contains data from nine tumor registries that reporesent a substantial, but nonrandom, portion of the U.S. population. Nonetheless, many researchers just apply incidence rates from SEER-9 to U.S. population counts (by sex and age group) to get estimated national incidence counts. In the micro-world of asbestos disease estimation, there have been papers about "coastal bias". These articles describe attempts to adjust for the fact that mesothelioma cases come disproportionately from coastal areas, due to shipbuilding exposure. The SEER-9 data (it is argued) provide a skewed view of the U.S. because they don't adequately represent coastal regions. I mention this issue just an historical aside – it doesn't really affect what we're doing.

Additional registries have been added to SEER over time, so as the years advance we start seeing, in addition to the SEER-9 series (which continues to be provided), SEER=11 and SEER=13.

For the U.S., the gold standard is the USCS. So the analytic challenge in building a complete historic record of U.S. incidence is to impute what USCS would have looked like had it been in existence as long as SEER-9 has.

ARPC and I have both done this kind of imputation, but the methods have been pretty crude. For initial fits of the HSE model, I suggest that you use UK and Australian meso incidence data. These data are provided in easily used formats, and tend to be more up-to-date than the U.S. data. In addition, since most of the development and use of the HSE model has been in these two countries, we can compare our initial fits and predictions to fits and predictions that have been published. (I'm not looking for identical results – our data are likely to be of slightly different vintage, and there are a variety of minor tweaks that have been made to the HSE model by different analysts.)

6.2   OECD data

Most countries report their data cancer incidence to the OECD in a reasonably timely fashion. So the OECD data, which are publicly available online, represent a source from which we can get incidence data in a standard format. If we ever feel like it, once we get the mdel fitting where we like it, we could write an article or build a website in which we report meso forecasts by country. So work on your shiny and markdown chops. The U.S. reports only the USCS data to the OECD. It is usually more laggard than other developed counteries about doing this in a timely fashion.

6.3   Use of mesothelioma incidence data in fitting HSE and releated models

Incidence data for HSE (and similar) model fitting pose the same challenges as demographic data do for HSE fitting: (1) going from grouped data to "by single year of age", and (2) extending the oldest age interval to single years of age, or to some finer grouping levels. See the section below on demographic data. Usually, parallel methods are used for the demographic and incidence data.

Ad hoc methods are often necessary for breaking the oldest age interval into finer categories. For example, Australian data usually go through age 85+, and UK datq through age 90+. For

most purposes, it would be adequate to more finely bin the Australian date by making use of the ratio (UK 90+ count) / (UK 85+ count).

Another issue with regard to use of incidence data in our context is that at different times we deal with one or more of three different dates associated with a case of mesothelioma:

- Date of diagnosis (i.e., incidence)

- Date of death

- Date a claim is filed against a particular trust

This issue doesn't matter much for current purposes, as these three dates tend to be within a year or two of each other. Mesothelioma is almost always fatal within a fairly short period of time, and U.S. attorneys are very efficient at identifying and signing up claims. (The most expensive search term on Google for buying paid ads is "mesothelioma".) In some expert analyses, the distinctions across these dates are ignored.

However, it is also reasonable to assume something like a 1-year or a 2-year lag between incidence and filing, particularly forecasting. In forecasting, it might make sense to assume that the decline (aka "runoff") curves for claims lag the national incidence decline curves by a couple of years.

Some other issues in terms of processing data on the claims end are:

1. Claims during a bankruptcy proceeding are stayed – i.e., firms are prohibited from filing them even they would have been filed against the debtor had the debtor not filed for Chapter 11 bankruptcy. So at the beginning of a trust, there is a catch-up period (usually up to an initial statute of limitations (SOL) date that is typically three years after the initial claims received date) when a lot of the claims would have been filed earlier. You will see the term "date arising" or "year arising" in some of the claims data that we receive, which translates to "year would have been filed."

2. Some mesothelioma claims are filed in the tort system against solvent defendants. If there is a judgment in favor of the plaintiff, usually there is a damage award that is apportioned among the codefendants. If the plaintiff has settled a claims with one of the codefendants, then that settlement amount may be subrtracted from the total award. So sometimes a law firm will hold off on filing a claim with any trusts, because then the claim has not been settled with regard to the defendants associated with the trusts, and there is no offset to the tort system award. Only when the claim is "finished" in the tort system is it filed with the trusts. This can create a longer gap between diagnosis and filing than the 1-2 years I cited above.

3. Sometimes there can be a delayed diagnosis. For example, recently a lot of lung cancer claims were filed because law firms started reviewion x-rays of workers who died of lung cancer, and who had some exposure to asbestos in the workplace. Many such claims are not filed against asbestos trusts, because the likely cause of the lung cancer is smoking. However, if there is x-ray evidence of asbestosis in addition to lung cancer, the claim is compensable according to most trust TDPs. Here again is a situations in which the diagnosis dates in our claims data may not correspond to diagnosis dates that would be counted in public health systems.

In R programs provided in the Rstudio project associated with this note, I have collected some UK and Australia incidence data and converted them into R data frames. I provide some details below.

6.4   Australia data

Australia mesothelioma data through 2009 were downloaded from a spreadsheet found here. Government summaries can be found here. These links may change when the data are updated.

I found that googling the terms AIHW, mesotheloma, and data got me to the desired sites.

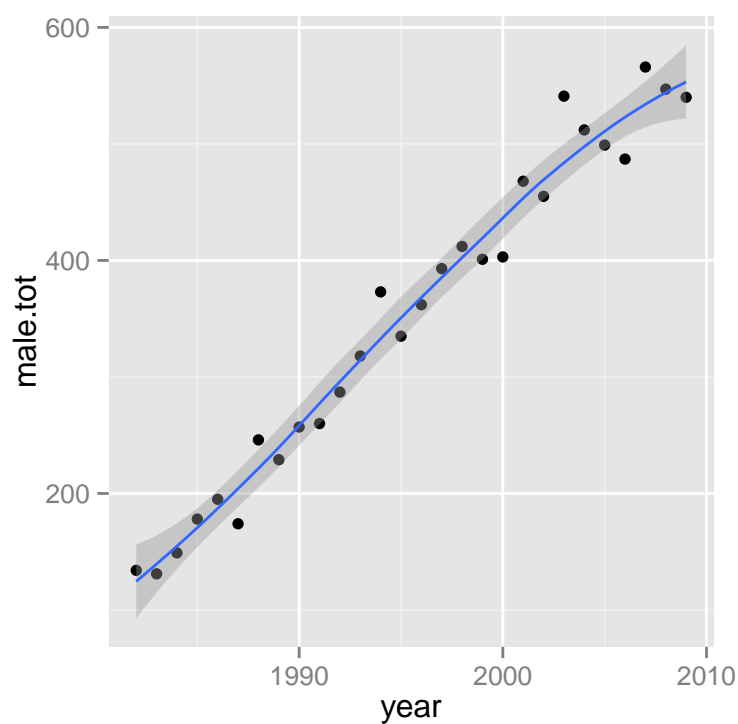Figure 2: Mesothelioma incidence - Australia males

6.5   UK data

UK mesothelioma data through 2010 can were downloaded from a spreadsheet found here. These UK data reflect date of death, not date of diagnosis.
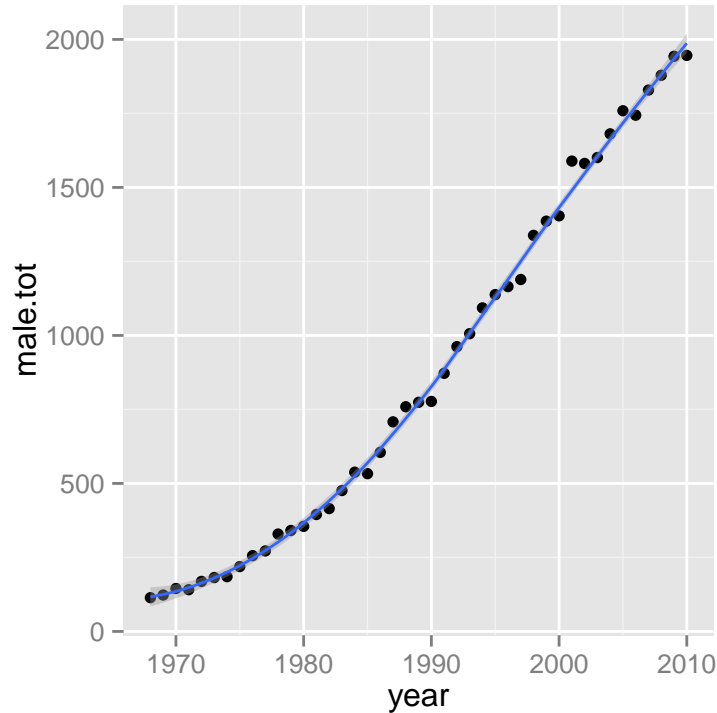
Figure 3: Mesothelioma incidence - UK males

7 Demographic data for model fitting and forecasting

Population counts are used as an exposure variable in the Poisson regression, and in forecasts made from it. Historic population counts and death rates can be obtained from the Human Mortality Database. See the program "Get populations and death rates" in the Rstudio project that accompanies this note. The program downloads population and death rates for Australia, UK, and US, by gender. The counts and rates are 1x1 – age is not grouped, and there counts and rates for each individual year – but can obviously be grouped if grouping is desired for consistency with meso incidence data.

This R program also uses the demography package to generate Lee-Carter projections of the death rates. These projected death rates can be used to translate the most recent population counts into future counts, for use in the HSE model projections. Once we (well, you) have fit a satisfactory HSE model, we should probably do one projection just assuming that the most recent historic death rates will remain constant, and one using the projected death rates that continue to decline.

In my past projections of U.S. mesothelioma incidence using an HSE model, I relied on the historic and projected death rates that the Social Security Administration uses. See `http://www.ssa.gov/oact/NOTES/as120/LifeTables_Body.html`, authored by the wonderfully named Felicitie Bell (Who wouldn't trust a forecast from her?). I had a number of conversations with her back when I was projecting future tobacco mortality and health care costs for the U.S. DoJ civil rico case against the tobacco companies.

11

I have all of those SSA rates in R formats from my previous work, both with tobacco and with my HSE modeling. For now, I suggest we use the Lee-Carter projections for all three of our countries. We can plug the SSA stuff in later on. I have read that the SSA uses Lee-Carter projections, although that doesn't really seem to be the case based on Felicitie's description. I'm not a demographer, so keep that in mind. I think there are a number of newer methods of death rate projections that do better than Lee Carter by some measures, but Lee-Carter still seems to be the most commonly applied method.
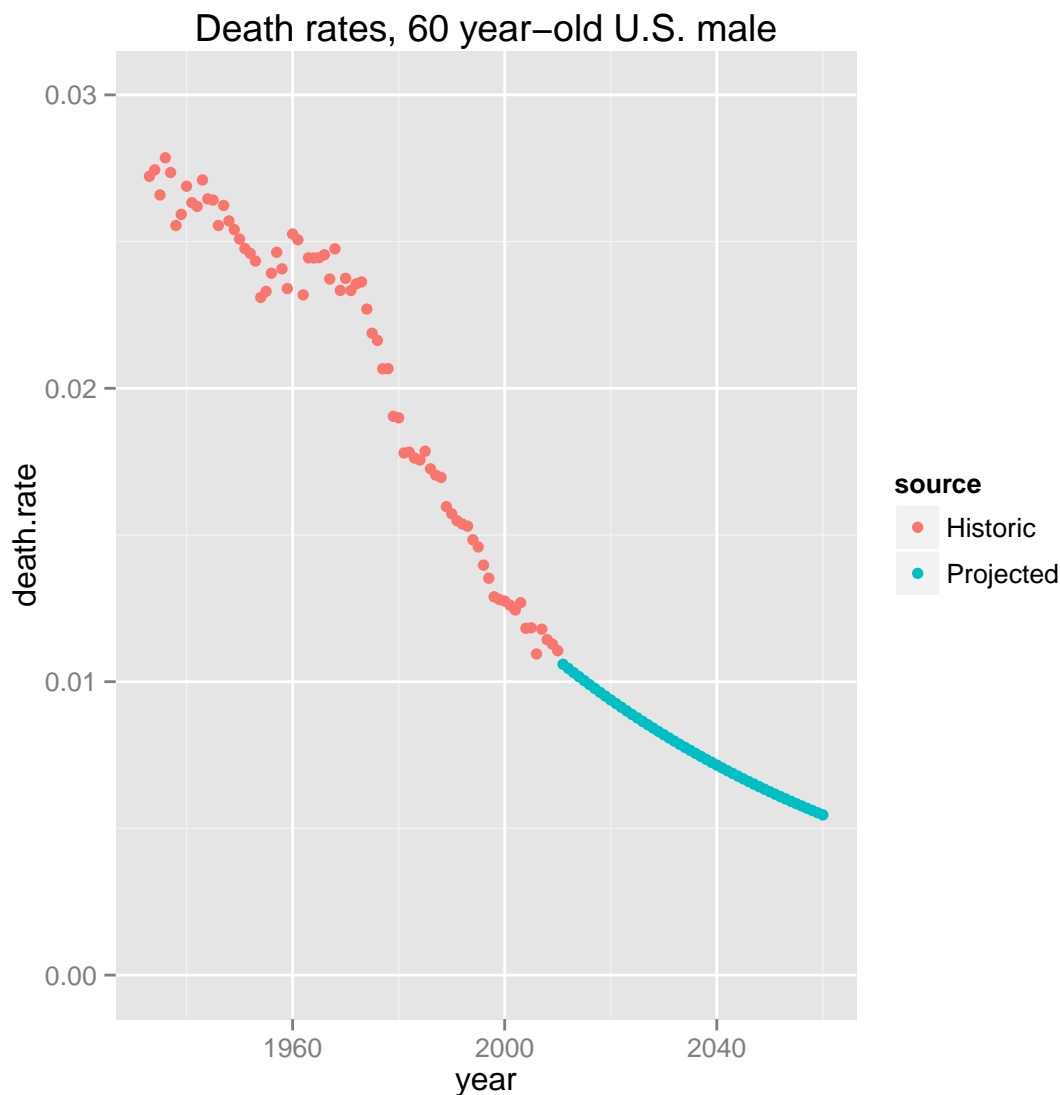
Figure 4: U.S. historic and projected death rates by year, 60 year-old men.

Projected mesothelioma cases (and claims) increase if future death rates decrease, because when people live longer they are more likely to contract cancers like mesothelioma. One issue that nobody but me seems to have thought about in making projections for estimating ultimate trust liabilities is that our exposed workforce may have lower median income than the U.S. as

a whole. (Or maybe not so much – insulators (for example) don't usually become wealthy, but they were at least employed, frequently in well-paying union jobs.) I suggest that we ignore this issue for now. Down the line, we might want to think about it. I provide some references in the accompanying material.[5] In my previous HSE forecast, I adjusted death rates upwards using the Donkin articles.

8 Previous fits of the HSE model

8.1 Initial HSE model(s)

The HSE model was first described in an HSE working paper[6], which subsequently became an article in BJM.[7] I relied on these articles for my own HSE fit. The analysts used a lot of degrees of freedom to describe exposure as a function of age and of calendar year. They used some kind of ad hoc methods that seem a bit antiquated, like what one would expect from some clever statisticians with roots in Bletchely Park who were reared to think of computer runs as an unseemly expense, to be minimized whenever possible.

Nonetheless, I followed their model pretty faithfully. I fit it using optim(), with an option that allowed me to put bounds on the estimates for each parameter. These were needed because, with all the degrees of freedom, the marginal likelihood for some parameters was dead flat, and iterations would wander off into never-never land. Fitting was also very slow, because the optimizer called my likelihood function many times, and each time there were a lot of inefficient summations performed to construct the likelihood with current parameter values. I figured out how to do the summations via matrix multiplication, which should be much faster, but never had time to return to work on this project. Some charts of goodness of fit of my application to U.S. data occur in the latter part of a slide presentation in the accompanying material.[8] (The initial slides in this document relate to a revision of the Peto model – see the previous section on different meso forecasting methods.)

Several other HSW tweaking issues are:

1. The model does not yet include an estimate of background risk.

2. The model at this point is only for males.

3. The age range is somewhat narrow.

4. A lot of emphasis is put on the "diagnostic trend" term – accounting for possible missed cases in earlier years. This element is given less emphasis – or none at all – in subsequent modeling efforts. In Hogsdon (2005), the authors note that presence or absence of this term has essentially no impact on forecasts.

5. The original model in the HSE working paper assumed a long clearance half life of asbestos fibers from the lung. Effectively, "no clearance" was assumed. The assumption of nonclearance is not always made in subsequent models. In the followup Hogsdon (2005) journal article, the authors fit both a clearance and a nonclearance model. The two models gave very similar fits and forecasts.[9]

[5] Donkin (2002), Singh (2006), Waldron (2007)

[6] HSE (2003)

[7] Hodgson (2005)

[8] Wyant (2011)

[9] Two of the model parameters, k (the power of time defining the increase in risk after exposure) and H (the half-life for clearance of asbestos from the lung), are closely correlated and cannot be independently estimated. The effect of reducing the half-life is to increase the best-fit value of k, but the fit of the model is affected only slightly. Since fitting the model with both H and k is unstable, we used two versions of the model, one with (effectively) no clearance (H1000 years) and the other with a clearance half-life of 15 years a value suggested from the modelling of mortality of the Wittenoom workforce (Berry, 1991). Page 587.

The power of time since first exposure, k, is hard to estimate independently of the clearance half life – they are negatively correlated. The nonclearance model gives k within the 2-3 range, as was expected, and also yielded an exposure index over time that more closely matched the pattern of asbestos imports over time. For these reasons, the authors expressed a (mild) preference for the nonclearance model.

6. The model fits the power of time since first exposure, and gets a vale between 2 and 3 – consistent with what one wouuld expect from the EPA 1986 models discussed above. Subsequent fits of the the HSE model by these and other analysts sometimes just treat the power as fixed, and some (notably ARPC) calculate a fitted value that is less than 2 (something like 1.8, if I recall).

7. The predicted peak number of mesothelioma deaths in the UK is about 2,100.

8.2 Bayesian refitting of the original HSE model(s) with MCMC

HSE analysts refit the HSE model using a Bayesian approach.[10] As usual, there were tweaks to the model. Many of the tweaks were just minor fiddling with the binning of the indices of exposure over time, and exposure by age. One substantive addition was that of a background exposure rate parameter. Background exposure was assumed to be related to age.[11]

The clearance half-life was set at 1,000,000. This is really a nonclearance model! And, of course, the Bayesian framework allowed for the calculation of prediction intervals. Basically, though, the fit and the projections from the model are by and large similar to what was produced with earlier classical optimization routines. Note that we have several years more incidence data with which to fit our HSE model than was used in any of the model fits I have so far described.

8.3 Adoption of the HSE model(s) by actuarial societies and public health analysts

Actuarial societies in the UK[12] and Australia[13] did public, systematic reviews of available mesothelioma forecasting models and selected the HSE model as the best. The UK insurers face about $18 billion in estimated liability for such claims. Some public health analysts in Australia started using HSE models as well.[14]

8.4 Initial use of the HSE model(s) in the U.S. asbestos trust world

I fit a version of the HSE model to U.S. data in 2010, as I have mentioned in a couple of places.[15] When ARPC saw my stuff, they decided to both revise their Peto model, and do their own fit of the HSE model.[16] I consulted with them on both. ARPC now uses their mode, but they do a simple model averaging of four models. The HSE model gives the highest forecasts, but because they use the median of their four forecasts, the fact that the HSE models gives substantially higher values than the other three is effectively ignored.

Both my revised Peto and my HSE models seem to yield higher forecasts than ARPC's. I don't know why. I would like to revisit this in detail with them at some point, but I want to have

---

[10] HSE RR728 (2009), the HSE working paper, and Tan and Warren (2010), the subsequent journal article.

[11] The background rate is represented by the number of cases per million in the male population. The age distribution of the background cases in each year is assumed to be $(A-L)^k$. Page 431.

[12] Gravelsons (2010).

[13] I couldn't find the reference right away, but I think it's by Lowe.

[14] Clements (2007a) and Clements (2007b)

[15] Wyant (2011), which also discusses revision of the Peto model.

[16] ARPC (2011a); ARPC (2011b).

carefully gotten all my ducks in a row – with your help – before doing so. One thing about their HSE fit – they get a (k) value of 1.86 or thereabouts, outside the ideal range of between two and three.

I would also like to launch a campaign with them to put more weight on the HSW result. One argument is, as I alluded to earlier, is that Stallard and Manton have revised their forecasting model and are now getting forecasts more like our HSE forecasts – about 30% higher than the old Nicholson model.

8.5 More recent HSE models – two-stage clonal and "imports-based"

Recently, HSE analysts published results of additional model fits.[17] The working paper has the regrettable title "... revised and two-stage clonal models." A better summary of what they did is:

1. Fit one alternative "imports-based" HSE model in which historic annual imports of asbestos in the UK are used as an independent variable,

2. Fit a "two-stage clonal" HSE model, an attempt to reflect some recent mathematical models of tumor development,

3. Fit the "old" HSE model to get a model for women, albeit with usual tweaks to deal with the new data, achieve convergence of numerical methods, etc. – hence the quotes around "old",

4. Basically concluded that the "old" HSE model is still the best model.

The tone and style of this report smacks (to me) of either (1) we have all this machinery, let's get a quick publication by fitting a couple of alternative models, or (2) somebody higher than us in the chain of command keeps asking about these other models – let's just fit 'em to get them off our backs so we can get back to doing something useful. Unfortunately, the authors of this report did not overtly say that the "old" HSW model is still the model of choice." This caused someone (after I had given a presentation on our version of the HSE model for the U.S.) to circulate and email to audience members saying that "the HSE has discarded the model that Tim is using."

In fact, what them authors said was:

1. The estimated of 2,100 mesothelioma cases at the peak is still our best estimate [this is the estimate from the old HSE model],

2. The upper bound on future mesothelioma cases is the upper bound calculated for the old HSE model, and

3. The way we will fit the model for women is the old HSE model (with the usual tweaks and upgrades).

This all says to me that they still think the old HSW model is the best practice – they just don't say so in a simple declarative sentence.

The notion of an "imports-based" model makes some intuitive sense. Using imports (or mining output, but in the cases of the UK and the U.S. import almost all the asbestos is imported) should add information. In the UK, it doesn't matter much. The index of asbestos exposure over time produced by fitting the HSE model mirrors pretty closely the import records. (In fact, some of the earlier HSE papers use "closeness to imports" as a guide in choosing between a couple of model tweaks.)

---

[17]HSE RR876 (2011).

In Australia, the actuarial society (or somebody like that) compared different forecasting models, including the imports-based model (in Australia, probably the mining-based model) – which they called the "latency model" (who knows why) – and concluded that the imports-based model did not fit well, and that the HSE model was superior.

My fit of the HSE model yielded an index of exposure over time that does not match the record of U.S. imports. I have no idea why. Poor imports data (e.g., how much was then re-exported?). Some mistake that I made? Change in the mix of manufacturing and construction processes that use asbesto, with resulting changes in intensity of exposure? Change in the type of processes? Change in job safety standards? (Nobody seems to believe the latter.) Who knows. In any event, once you have fit a new HSE model for the U.S., a subsequent step will be to fit an imports-bassed model. The U.S. imports data are in the accompanying materials.[18]

I don't know anything about the two-stage clonal model, but we should look at fitting that one once we have gotten more important things out of the way.

9   Our HSE modelling plan

The word "template" is too strong, but I would like to use something like the model fit by CLements[19] as a starting point. See Figure 5.

Clements, Berry, Shi, et al

APPENDIX

AGE AND CALENDAR YEAR MODEL
Following Hodgson *et al*,[6] we assumed: (1) that the mesothelioma rate for a cohort aged a at time t exposed to asbestos at time u in the past was proportional to the earlier asbestos dose (that is, rate$(a,t|u)$ was proportional to dose$(a-u,t-u)$); (2) that dose was multiplicative by age and time, that is, dose(a,t) was proportional to $W(a)D(t)$, for dose potentials by age ($= W$) and by time ($= D$); (3) that the rate was proportional to time from exposure to malignant conversion raised to some power k together with an effect due to clearance of asbestos fibres from the lung, where the half-life is represented by H and τ is the average latency time from malignant conversion of the cancer to clinical detection; and (4) that the observed number of cases followed a Poisson distribution with mean μ_{at}.[12] To model the mesothelioma rate, we averaged the rates across all of the times since exposure from τ to a. The predicted number of cases μ_{at} was calculated by weighting the mesothelioma rate by the person-years at risk estimated by the population ($= P_{at}$) and scaled by the fraction of correctly diagnosed cases at year t ($= D_x(t)$). Taking β as a constant, the Poisson regression model was:

$$\mu_{at} = \left[\beta \int_{\tau_c}^{a} D(t-u) \, W(a-u)(u-\tau)^k \, 0.5^{(u-\tau)/H} du\right] P_{at} D_x(t)$$

We modelled the rates for the mid-point of the five-year age intervals (for example, for ages 40–44 years, $a = 42.5$). The mean rate μ_{at} was numerical integrated using Simpson's rule, with functional evaluations at single-year increments for

$u = \tau+0.5$ to $u = a$.[20] The dose potential functions $W(a)$ and $D(t)$ were specified as exponentials of natural splines with one internal knot and two parameters to be estimated.[21] For $W(a)$, the spline functions were defined for boundary knots at ages 20 and 65 years and $W(a)$ was defined to be 1 at age 50 years; for $D(t)$, the spline functions were defined with boundary knots at 1945 and 1980 and $D(t)$ was defined to be 1 in 1970; internal knots were assumed to be at the mid-point between the boundary knots.

The five parameters to be estimated included the log of the constant β and four parameters for the spline functions $\log(W)$ and $\log(D)$. We assumed that k was fixed at 3.5. The asbestos half-life H was given a value of 15 years. For incidence data, the latency period τ was assumed to be 5 years. As in Model 1 of Hodgson *et al*,[6] the diagnosis fraction $D_x(t)$ was assumed to be almost complete (98%) in 1997, with a 5% annual percentage trend in the missing cases, such that $D_x(t) = 1-0.02 \times 1.05^{1997-t}$.

The models were fitted using the mle() function in the R statistical package,[22] which uses a quasi-Newton approach to maximise the Poisson log-likelihood. The design matrices for the spline functions were calculated using the ns() function in R. We assessed goodness of fit by performing a likelihood ratio test using the residual deviance with the residual degrees of freedom. For interval estimation for predictions, we used the bootstrap, re-sampling from Pearson's residuals.[23] Following Friedl, we standardised the residuals by dividing by the square root of (1–model degrees of freedom/number of observations) and scaled the residuals to have zero mean.[24]

Figure 5: Clements 2000b technical appendix. Possible starting point for our HSE fit.

Here are my thoughts about working off of this model:

1. I don't like the HSE's baseline HSE model (and the similar one that I originally fit using optim()) because it uses too many degrees of freedom and too ad hac an approach to building indices of exposure over time and exposure as a function of age. I thought splines

---

[18]USGS (1997)
[19]Clements 2007b

would be a better way to go – as did Clements. However, I'm not totally sold on his splines. I do think some assymetry in the indices is good to incorporate, which to me means perhaps three degrees for freedom for each of the two splines, not two. I haven't quite followed the significance of his knot placement. I think I understand roughly how he's forcing the indices to have fixed values at certain points, but my level of understanding is probably lower than I think. Anyway, we're into your territory now, so I leave you to deal with this issue, and (I hope) educate me a bit.

2. I don't know about using numeric integration, rather than discrete summing. Discrete summing can be slow, as I found out, but could presumably be speeded up by matrix multiplication or other cleverness. Plus, fits will go faster simply due to usine fewer and better-defined independent variables. Anyway, I again turn to you. Maybe numeric integration is the bee's knees.

3. As other articles I have cited would indicate, halflife and the k exponent are negatively correlated, and models that try to fit both of them are unstable. I don't know why Clements chose to make both of them fixed, and at extreme values. (They're not extreme in a way – due to the negative correlation you can get the same fit making the exponent big and the halflife small as you get with more moderate values.) Given the discussions in the other articles, I would tend to fit k and ignore half-life – or fix halflife at 1,000,000 years as one paper did. For one thing, there is some biological validation of our fit if the estimated k is between 2 and 3, as it was in my fit of the HSE model.

4. I think beta in the Clements model is basically the background rate. I would like to fit a background rate. As discussed above, assuming that the background rate increases with age makes sense – we might add that. This is not a big issue in a way, as background rates are small. But it again lends some biologic validation to our model if we fit a background rate and it is indeed small.

5. Clements uses stats4::mle(), rather than optim(). Of course, mle() is just a wrapper for optim(). I have not used mle(), but it has seemed to me a good thing to consider where (as here) one is optimizing a likelihoodl. It would seem to clarify how exactly we are using a poisson regression (not clear at all in the early HSE papers), and I think could yield some initial error bounds by numicalluy inverting the second derivative matrix. There is also now bbmle::mle2() to consider. Anyway, I leave all this in your capable hands, hoping again for some education down the line that is simple enought for me to absorb.

6. In U.S. meso incidence data, I have noticed a flattening in the incidence rates somewhere above age 80. This may reflect underdiagnosis. If an oldster in a nursing home with pneumonia dies, I don't think anybody will be requesting an autopsy to see if there is something else going on in the lung region. However, "flattening" has been an issue from time to time in the meso modeling literature. Some analysts find it hard to believe that risk can go up indefinitely as a power of two or three. But in any event, I have found it useful to build in a "flattening age" like at 85 in my modeling. You might consider doing the same, at least as a check.

7. I don't know about fitting to midpoints of age ranges, as Clements does. Certainly it should make initial fitting routines run a little faster. In the past, I have converted to single-year-of-age by applying splines to cumulative incidence by age in each calendar year. It would certainly seem necessary to do something like this in the forecasts, where we need annual forecasts and it is hard to calculate how many people move from one age group to the next in a year without – you know – actually calculating what happens at the single year of age level. But I again leave this kind of pondering to you.

8. I think (I think you would have done this anyway) we should get a standard numerical fit before moving on to Rjags. I also think we should start by fitting to UK and/or Aussie data, as these data are tidy, and the we can also check our fit against the several fits in the literature.

9. You would do this anyway, but just so I have a reminder here for myself – the modeling programs should generate the kinds of diagnostic fit plots that appear in my and ARPC's presentations[20], including plots of the indices overlaid by import data, and also some measures of goodness of fit like log-likelihood, AIC, etc. And we need some error bounds on forecasts and on parameters such as k.

10. Next up in rough priority after doing a new "standard" U.S. male model is to:

   (a) Compare with my and ARPC's HSE fits,

   (b) Do some forecasts for individual trusts – as I say, this has been done by outputting historic and projected U.S. national counts by birth cohort group, identifying a stable claimiing period in the trust history, calculating trust to U.S. ratios by birth cohort, and then applying these ratios to the projected U.S. counts from the HSE model.;

   (c) Do a female model,

   (d) Do a combined forecast of male plus female,

   (e) Fit some kind of imports-based model,

   (f) Take a look at the two-stage clonal model, and

   (g) Look at an Rjags fit.

---

[20]ARPC 2011a, Wyant 2011