The purpose of this document is to describe one of the models used by ARPC to forecast future mesothelioma incidences in the United States. The approach described in this document is based on Hodgson et al. (2005), referred henceforth as "HSE".[1]

The HSE model has epidemiological foundations. Similar to earlier models developed by Dr. Peto and others, it uses a "backcasting" technique.[2] The model uses data on historical incidences and an epidemiological model to identify the exposed population, i.e., the number of individuals that were exposed to asbestos in each year throughout the relevant time frame. This technique is referred to as backcasting, because the (observed) number of incidences in a particular year is used to infer the (unobserved) number of individuals exposed in the years prior to the year of incidence. Unlike occupational-based studies that use information on the size of the exposed population, this model infers the exposed population from incidence data and an appropriate epidemiological model.[3] Using the estimate of the exposed population and standard mortality rates, the model can estimate the number of exposed individuals alive in each future year. This, together with the epidemiological model yields an estimate of the incidences in each future year.

1 Data

The primary data inputs used in this model are historical mesothelioma incidences and historical and projected population counts. This section describes the construction of these input data sets.

1.1 Mesothelioma Incidence Data

Information on incidences was collected from two separate sources: from the National Cancer Institute (NCI) for the years 1973-1998 and from the Centers for Disease Control and Prevention (CDC) for the years 1999-2007. The former data has limited geographical coverage of the United States but it goes back further in time. The latter data has virtually full coverage of the United States but it only covers the years from 1999 onward.

The National Cancer Institute administers a data collection program called Surveillance, Epidemiology and End Results (SEER) and publishes historical data on its website. This SEER data contains information on incidence, prevalence, and survival of all types of cancer, including mesothelioma. Data collection began in 1973, with a limited geographical coverage that grew over time. The initial set of 9 regional data centers ("registries") grew to 13 in 1992 and to 17 in 2000. These sets of registries represent about 9.5%, 13.8%, and 26.2% of the US population.[4]

---

[1] Hodgson, J., McElvenny, D., Darnton, A., Price, M., and Peto, J. (2005) "The Expected Burden of Mesothelioma Mortality in Great Britain from 2002 to 2050," *British Journal of Cancer* 4, 587-593
Also explained in Tan, E., and Warren, N. (2009), "Projection of Mesothelioma Mortality in Great Britain," prepared by Health and Safety Laboratory for the Health and Safety Executive 2009.

[2] Peto, J., Henderson, B. E., and Pike, M. C. (1981) "Trends in Mesothelioma Incidence in the United States and the Forecast Epidemic Due to Asbestos Exposure during World War II"

[3] A seminal study that uses information on the size of the exposed population to estimate mortality from asbestos exposure is Nicholson W. J., Perkel, G., and Selikoff, I. J. (1982) "Occupational Exposure to Asbestos: Population at Risk and Projected Mortality – 1980-2030," *American Journal of Industrial Medicine* 3, 259-311

[4] http://seer.cancer.gov/registries/data.html

One of the CDC's divisions, the National Program of Cancer Registries (NPCR) publishes statistics on cancer incidence and mortality. This dataset is commonly referred to as United States Cancer Statistics (USCS). The population coverage is virtually 100 percent.[5]

The remainder of this section explains the data processing steps undertaken to obtain a single, nationally representative incidence dataset.

The first problem to tackle was that the model required incidence data by single years of age, which was not consistently available in the input data. For example, all SEER datasets aggregate the 85 years old and older population into a single category. USCS reports incidences for age categories roughly in 5 year increments: for age 0, for age 1 to 4, for age 5 to 9, for age 10 to 14, etc., and for age 85+. To impute data for single years of ages, a technique called cubic spline interpolation was used. This technique "distributes" the figure for a five-year age group among the five single years of ages, in a way that gives rise to a smooth age distribution. We implemented this method separately, for each arising year (i.e., the year in which the incidences were observed) and gender.

The second problem to tackle was the disparate levels of population coverage encountered in the various datasets. Due to the fact that the different SEER datasets and the USCS dataset had differing population coverage, some scaling was necessary so that these data could be appended together in an effort to create a single dataset. First, for each dataset, we calculated the total population it represents, by arising year and gender. Dividing these figures by the total United States population yields the scaling ratios that we used to estimate national incidence figures, for each dataset, by arising year and gender. However, this scaling does not account for potential sampling bias in the SEER data given the nonrandom location of registries. For example, if registries are typically located in geographic areas where cancer incidences are more likely, this method overestimates national incidences. To correct for this bias, we used the fact that the USCS data had no bias, given that its coverage is almost 100 percent. Next, we calculated the ratios of total incidences in the various SEER datasets to the total incidences in the USCS dataset. These ratios were calculated for the period for which all datasets had information, from 2000 to 2007. We used these ratios to correct the SEER data so that they match the USCS data for the overlap period. Doing so is likely to eliminate the location bias in the SEER data. Finally, a unique incidence series was generated, which was equal to the scaled SEER 9 data for 1973-1991, the scaled SEER 13 data for 1992-1998, and to the USCS data for 1999-2007.

1.2 Population Data

In addition to historical incidence data, historical population data are used to estimate the unknown parameters of the epidemiological model. Also, to forecast future incidences, projected population data are needed.

The source for historical population counts was the SEER website.[6] These data were combined with estimated mortality rates to obtain an estimate of future population counts. The source of the mortality rates was the Social Security Administration.[7]

[5] http://www.cdc.gov/cancer/npcr/uscs/data/00_pop_coverage.htm
[6] The SEER dataset includes population figures covering the entire United States.

The Social Security Administration's mortality tables report historical and estimated death rates by age (from 0 years to 119 years of age) and gender, for every decennial year between 1900 and 2100. These mortality tables are processed in the following way. First, since our model distinguishes between single years of ages only up to 99 years of age, we calculate an aggregate mortality rate for individuals at or above 100 years of age. Second, note that our model requires population figures in each and every year through 2050, not just in the decennial years. Therefore, we interpolate the decennial data to obtain mortality rates by age and gender, for every single year between 1900 and 2100.

The SEER data reports national population figures for single years of age, by gender, for each year from 1969 to 2009. Since mortality rates are significantly different for males and females, it was necessary to distinguish population figures by gender, and forecast population figures separately for the two genders.

The SEER data reports population figures by single years of ages from age 0 to 84, but reports all individuals at or above 85 years of age in a single category. To infer the population figures for single years of ages between 85 and 99, we used the mortality tables. We started by imputing the number of 85 years old individuals. Using the number of 84 years old individuals and their survival rates in the prior year allows for an estimation of the 85 years old individuals in the current year. This calculation was implemented by gender for every year between 1970 and 2009. Then we imputed the number of 86 years old individuals by using the number of 85 years old individuals imputed in the previous round. We did this for all years of ages up to 99. The remainder of the 85+ population (which was reported by SEER), was assumed to be 100 years old or older. Since this represented only a very small fraction of the 85+ population, the procedure appears to yield sensible results.

Forecasting the population from 2010 onwards was based on a similar procedure. For each year, the estimate of the number of people of a particular age was equal to the product of the number of people one year younger and their survival rate, both taken from the prior year. In addition, for each year the number of newborns was estimated too. This was assumed to be equal to the product of a constant birth rate and the total population in the prior year. The constant birth rate was assumed to be equal to the birth rate in 2009 (number of people of age 0 in 2009 divided by the total population in 2008), which turned out to be approximately 1.4%.[8]

2 Model

The model that we estimated follows closely the HSE approach, as described by Tan and Warren (2009).[9]

[7] Bell, F. C., and Miller, M. L. "Life Tables for the US Social Security Area 1900-2100," available at http://www.ssa.gov/OACT/NOTES/s2000s.html

[8] One might note that people born after 2009 are likely to be irrelevant for the calculation of asbestos-related mesothelioma incidences, and therefore, they do not need to be taken into consideration. However, the model predicts background incidences too, not just asbestos-related incidences. While the forecasted incidence figure for people born after 2009 is small, it is not zero.

[9] Tan, E., and Warren, N. (2009), "Projection of Mesothelioma Mortality in Great Britain," prepared by Health and Safety Laboratory for the Health and Safety Executive 2009.

It is assumed that the number of incidences for a specific arising year and age combination is a random variable that follows a Poisson distribution.[10] For a Poisson distribution, the set of possible values is the set of non-negative integers, and the distribution has a single free parameter: its mean (λ). This parameter, the mean incidence rate is assumed to vary across different combinations of arising year and age. Furthermore, the mean incidence rate is assumed to have two components: a background rate (assumed to be constant across combinations of arising year and age) and the exposure-related incidence rate (assumed to be varying across combinations of arising year and age). This latter is determined by an epidemiological model described below.

In particular, the exposure-related mean incidence rate is determined by a theoretical model of cumulative asbestos exposure. The rate is assumed to be zero in the first L years of a person's life. Parameter L corresponds to the latency period that it typically takes from first exposure to diagnosis. Following Tan and Warren (2009), we assumed a latency period of 10 years.

Each individual in the population is assumed to accumulate exposure throughout his entire life. The exposure is assumed to be the product of two components: an overall exposure-level (D_t) that varies through the years but does not depend on age, and an age-specific exposure potential (W_a) that varies by age but does not depend on the calendar year. Intuitively, the first component corresponds to the general use of asbestos in industrial applications. An intuitive justification for the second component is the fact that at any given point in time, people of different age have a different likelihood to be exposed to asbestos. In particular, working-age people have a higher likelihood to be exposed to asbestos than children or retired people. The specifications of D_t and W_a varied across the models we estimated. In our preferred model, W_a was a beta-function (with two free parameters to be estimated), whereas D_t was a non-parametric function (with 10 free parameters to be estimated). In particular, the 10 free parameters were the function values at specific grid-points (1910, 1920, 1925, 1930, 1940, 1950, 1960, 1970, 1980, and 1990), and the function values for all other years were obtained by piecewise cubic interpolation.

The exposure of a person of age a in year t is thus equal to D_t times W_a. One approach would be to let the mean incidence rate to be the sum of exposures across the years of a person's life. However, this approach would not account for the possibility that a more remote exposure could trigger an incidence at a different rate than a more recent exposure. In particular, some dose-response studies suggest that the more time has passed since exposure to asbestos, the more likely the incidence of mesothelioma. To account for this relationship, each exposure was weighted by a term that was increasing in time since exposure.[11] In particular, this term was assumed to be a power function with coefficient k, and with the additional restriction that the function is zero if the exposure was in the first L years of the person's life. Finally, these weighted exposures were added up across the years of the person's life.

[10] Count data (such as number of incidences) are often modeled in the literature by a random variable following a Poisson distribution.

[11] In incidence models published by the Occupational Safety and Health Administration (OSHA), the mesothelioma incidence rate was found to be a power function of time since first exposure minus a latency factor. The estimated power term was typically larger than one, indicating that the incidence rate grew more than proportionally with time since first exposure.

To summarize the model formally, the mean incidence rate of the epidemiological model was determined as follows:

Equation 1

$$\lambda_{A,T} = \left(r_{A,T} + b\right) * Pop_{A,T}$$

Equation 2

$$r_{A,T} = \sum_{t=0}^{A} D_{T-t} * W_{A-t} * max(t - L, 0)^k$$

$\lambda_{A,T}$	Mean incidence rate of a person of age A in year T
$r_{A,T}$	Mean exposure-related incidence rate of a person of age A in year T
B	Background incidence rate
$Pop_{A,T}$	Population of age A in year T
D_t	Exposure level (time-specific)
W_a	Exposure potential (age-specific)
L	Latency (fixed, not estimated)
K	Power term of time since exposure

The model's parameters $\{D_t, W_a, k, b\}$ were estimated by fitting the model's predictions to the data. To be precise, the likelihood function was formulated in the following way. Let $Y_{A,T}$ denote the number of incidences in year T by individuals of age A. Since, by assumption, $Y_{A,T}$ is a random variable following a Poisson distribution with mean $\lambda_{A,T}$, one can write the probability distribution of $Y_{A,T}$ as follows.

Equation 3

$$Pr\left(Y_{A,T} = y\right) = \frac{exp\left(-\lambda_{A,T}\right) * \left(\lambda_{A,T}\right)^y}{y!} \qquad for\ y = 0,1,2,\ldots$$

It follows then that the likelihood of the data can be written as follows.

Equation 4

$$L = \prod_{A,T} \frac{exp\left(-\lambda_{A,T}\right)\left(\lambda_{A,T}\right)^{Y_{A,T}}}{Y_{A,T}!}$$

Taking the natural logarithm yields

Equation 5

$$logL = \sum_{A,T} -\lambda_{A,T} + Y_{A,T} log\left(\lambda_{A,T}\right) - log\left(Y_{A,T}!\right)$$

The maximum likelihood estimate is obtained by maximizing Equation 5 with respect to the unknown parameters: $\{D_t, W_a, k, b\}$.

3 Results

We estimated a number of different specifications, but to keep the exposition brief, we only report the results from our preferred model.

In our preferred model, we used males and females in both the estimation and the forecasting phase. The time-varying exposure level function (D_t) was assumed to be non-parametric, with 10 free parameters at specific gridpoints. The age-specific exposure potential function was assumed to be a parametric beta function, and it was also assumed that the two free parameters are the same (symmetric function). The estimated parameters are displayed in the table below.

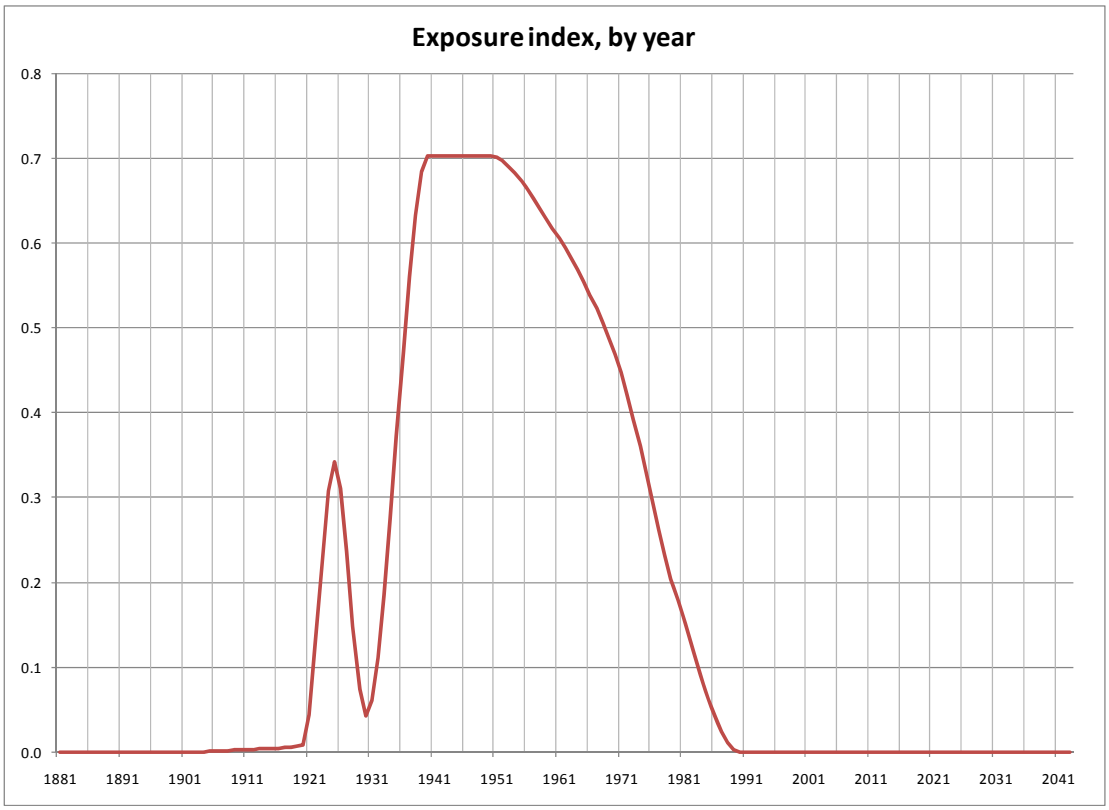
Table 1 Parameter Estimates

Parameter	Coefficient Estimate	95% Confidence Interval	
		Lower Bound	Upper Bound
D_{1910}	0.0029	0.0000	0.0638
D_{1920}	0.0082	0.0000	0.0727
D_{1925}	0.3422	0.0753	0.4358
D_{1930}	0.0428	0.0000	0.3468
D_{1940}	0.7019	0.5030	0.8748
D_{1950}	0.7019	0.5363	0.9008
D_{1960}	0.6171	0.4834	0.8016
D_{1970}	0.4690	0.3521	0.5911
D_{1980}	0.1795	0.1088	0.2607
D_{1990}	0.0000	0.0000	0.0000
$W_{alpha} = W_{beta}$	3.0387	2.7522	3.3391
K	1.8928	1.8434	1.9492
B	0.0566	0.0486	0.0639

The 95% confidence intervals were calculated by a technique called parametric bootstrap. The data was redrawn from the observed dataset randomly 200 times and each time, a new parameter-vector was estimated. The confidence interval was calculated on the basis of these 200 estimated parameter-vectors.

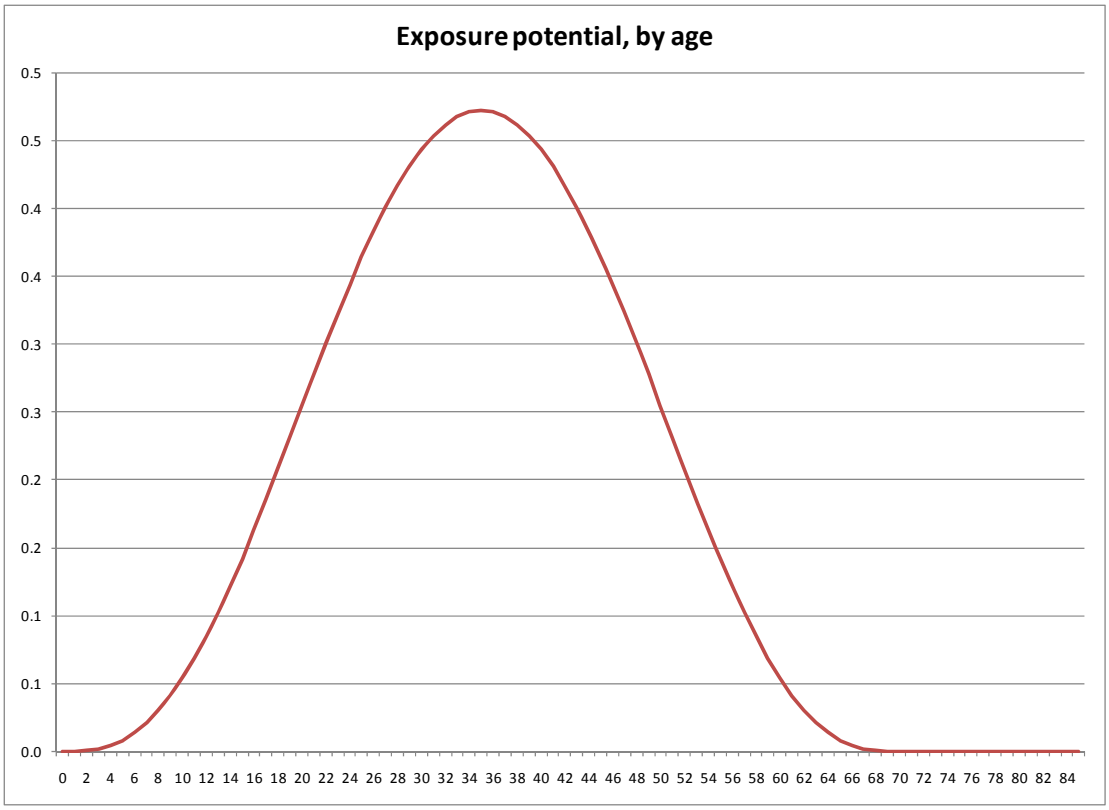
To visualize the results, we display the two main exposure components, the nonparametric exposure-level function and the parametric exposure potential function.

Figure 1 Exposure level, by year



As expected, the estimated exposure level reaches its peak in the 1940's. It gradually declines afterwards but does not fully disappear until the early 1990's.

Figure 2 Exposure potential, by age

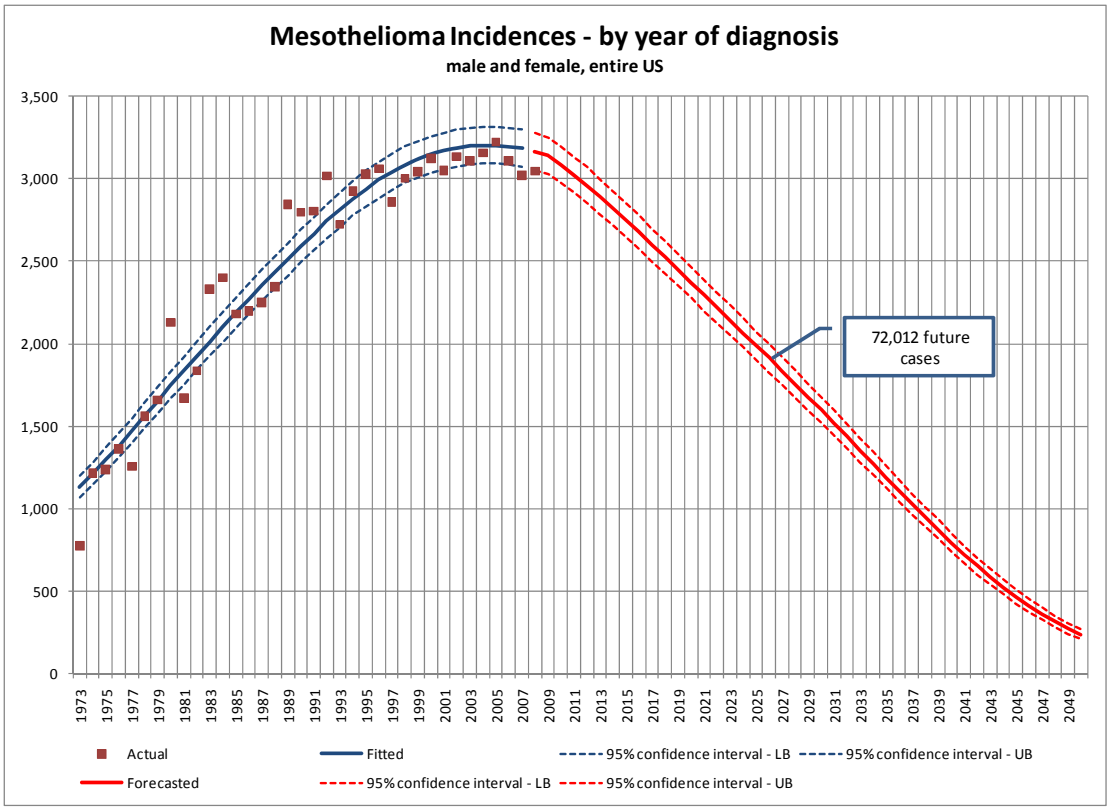


The estimated exposure potential is also sensible. It gradually increases until age 35, and decreases thereafter. Of course, with a simple, one-parameter specification one could not expect the model to generate a more complex pattern of age-specific exposure potential.

Next, we used our estimated model to generate predictions of future incidences. Although the model generates predictions by arising year and age, it is difficult to illustrate these predictions in both dimensions at the same time. Therefore, we aggregated our predictions along each of those dimensions and present the results separately.

First, the figure below represents the total predicted incidences, by arising year.

Figure 3 Mesothelioma Incidences, by arising year



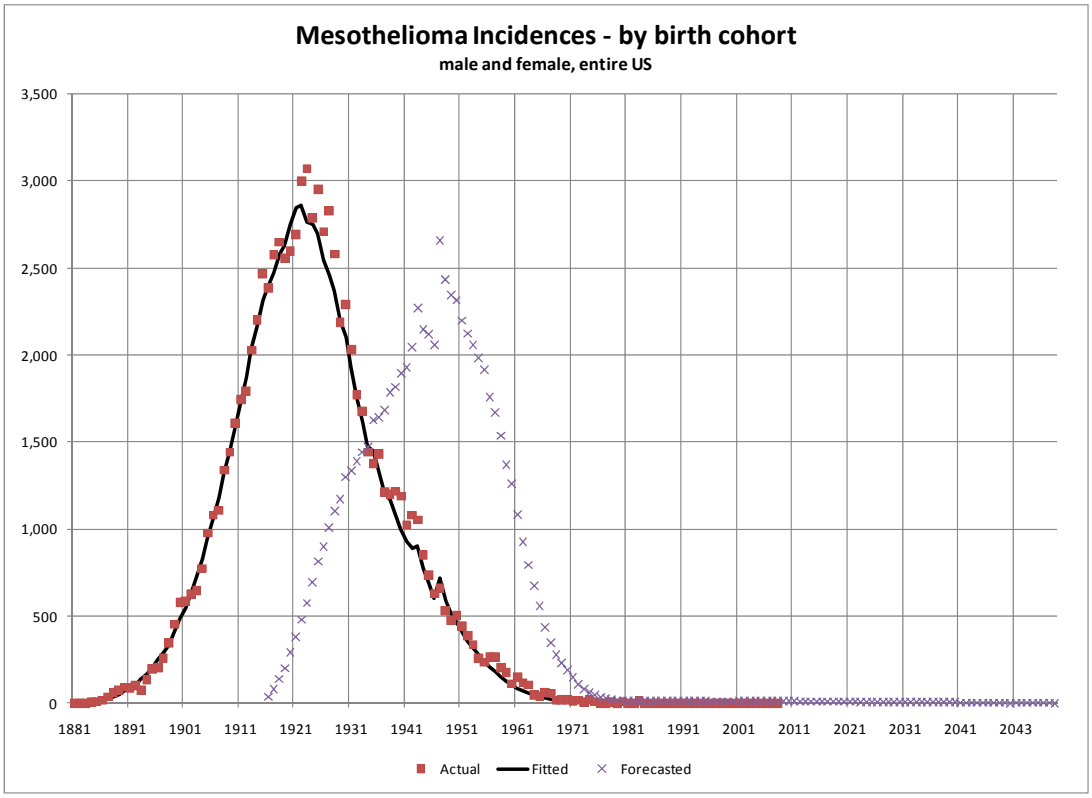
In total, 72,012 future cases are predicted from 2008 to 2050.[12] We also estimated a 95% confidence interval around the predictions, and we note that this bound is reasonably tight.[13]

[12] Although available, the 2008 data point was not used in the estimation of the model.

[13] The confidence interval was estimated by a technique called parametric bootstrap. Holding the estimated parameters (thus the year and age specific means of the Poisson distribution) fixed, we redrew the forecasted data 10,000 times. Note that while this approach is commonly used, it only accounts for sampling error, not for parameter uncertainty.

Finally, the chart below displays predictions by birth cohort.

Figure 4 Mesothelioma Incidences, by birth cohort



The distribution of future cases by birth cohort shows a sensible pattern. We predict the largest number of future cases to be born in the 1940's and 1950's. Very few cases are predicted from the cohorts born after 1980.