FINAL REPORT

Alpha Foundation for the improvement of Mine Safety and Health, Inc.

Grant Number and Title:

AFC6l8-53, "Innovations in applied decision theory for mine surveillance and health and safety efforts"

Organization Name:

University of North Carolina of Chapel Hill

Principal investigator:

David B. Richardson

Contact Information:

Email: david.richardson@unc.edu

Phone: 919 966 2675

Fax: 919 966 2089

Period of Performance:

August 1,2017 to February 28, 2019

Acknowledgement/Disclaimer

This study was sponsored by the Alpha Foundation for the Improvement of Mine Safety and Health, Inc. (ALPHA FOUNDATION). The views, opinions and recommendations expressed herein are solely those of the authors and do not imply any endorsement by the ALPHA FOUNDATION, its Directors and staff.

2.0 Executive Summary:

2.1. Problem

This project addresses two key areas of interest to the Foundation: 1) Development and demonstration of surveillance methods for health outcomes that may have widespread application; and, 2) Development and demonstration of tools, methods, or strategies for the identification and prevention of work-related health effects in a well-defined working environment.

Decisions about interventions to improve miners' health and safety require reasonable estimates of the intervention's impact. Health and safety professionals require a clear idea of where impact can be made, and how big of an impact an intervention may have in terms of disease, disability, and mortality. This project is focused on improving methods to estimate the impact of health and safety interventions for the prevention of work-related disease and mortality.

2.2. Research Approach

In the project's first aim, we addressed the need for valid information to accurately rank-order excess disease or death, by category of disease. We developed and illustrated methods for calculation and ranking of cause-specific excess disease in a contemporary framework for valid decision making; this aim included development of tools for communication in graphical as well as tabular formats of cases of disease and disease-free life. This may inform specific interventions or serve as a basis for framing future intervention efforts.

In the project's second aim, we extended the framework to offer a simple solution to interpretation of competing risks. We developed models for competing diseases potentially affected by the work environment. Here we target estimation of quantities to inform a policy maker about potential impacts on a range of disease endpoints of interventions that effect occupational exposure. We illustrated results in a miner cohort with a focus on models for diseases of the heart and lung.

In our third aim, we considered how intervening on exposure to one agent may have spillover effects. Policy evaluation, however, emphasizes the need for identification of where the greatest impact of a policy occur.

2.3. Accomplishments

First, statistical methods were developed to address each of the project specific aims.

Second, we have communicated the findings. We have presented the findings at international scientific conferences and prepared two scientific manuscripts to describe the results: one a development of the statistical framework used in these aims; and, a second, a manuscript that aims at broadening the uptake of these methods by practicing occupational epidemiologists.

Third, we have facilitated dissemination of the methods. We have developed SAS statistical code for implementation of these methods.

Fourth, we have applied the methods to real data for a large cohort miners. We conducted our empirical illustration of the methods using data files created through the Foundation-supported

project. Specifically, we created an analytical database needed for the proposed work that encompasses nearly a million person-years of observation for almost 30,000 miners who had work experience in uranium mines, and had detailed work history, exposure, and mortality follow-up information. To do this, we established and demonstrated the feasibility of external investigators working with these data through the creation of 'dummy' pilot data to develop and de-bug code, then running final code on the OCRC servers in collaboration with staff there. These data were successfully used for the statistical analyses proposed for the project aims.

2.4. Expected Impact

The proposed work develops innovative methods for leveraging applied decision theory to improve mine health and safety. We develop and illustrate these methods using data for a large cohort of Ontario miners. We expect the approach to become increasingly used by practicing epidemiologists due to its analytical clarity and the usefulness of the approach.

3.0 Problem statement and Objectives

This project addresses two key areas of interest to the Foundation: 1) Development or demonstration of surveillance methods for exposures and/or health outcomes that may have widespread applications; and, 2) Development or demonstration of tools, methods, or strategies for the identification/prevention of work related health effects in one or more well defined working environment. Decisions about possible interventions to improve miners' health and safety require reasonable estimates of the intervention's impact. Health and safety professionals require a clear idea of where impact can be made, and how big of an impact an intervention may have in terms of disease, disability, and mortality. The goal of the proposed work is to develop innovative methods for leveraging applied decision theory to improve mine health and safety.

3.1 Aim 1. Methods to rank order the occupationally-associated health problems of miners. Valid information that allows health and safety professionals to accurately rank-order excess disease or death, by category of disease, may inform specific interventions or serve as a basis for framing future intervention efforts. We develop and illustrate methods for calculation and ranking of cause-specific excess disease in a contemporary framework for valid decision making; this aim includes development of tools for communication in graphical as well as tabular formats of cases of disease and disease-free life. The goal of Aim 1 was to identify the leading categories of excess disease may begin to inform thinking about specific intervention efforts. We established methods for calculation and ranking of cause-specific excess disease in a contemporary framework for valid decision making.

We developed and illustrated these methods using data for a large cohort of Ontario miners. We undertook the statistical analyses to address Aim 1 using data for 28,546 miners employed in uranium mining in Ontario between 1954 and 1996. We analyzed information regarding vital status and cause of death information through 2007.

3.2. Aim 2. Methods to improve decision making about hazards in the mine environment that may affect multiple diseases. We extended the framework to estimate joint models for diseases potentially affected by the work environment. Here we target estimation of quantities to inform a policy maker about potential impacts on a range of disease endpoints of interventions that effect occupational exposure. This approach offers a simple solution to interpretation of competing risks. We focus on joint models for diseases of the heart and lung.

3.3 Aim 3. Methods to improve decision making in a setting with multiple hazards. We considered how intervening on exposure to one agent may have spillover effects. Standard analyses often do a poor job because the focus in on estimation of the independent effect of each agent. Policy evaluation, however, emphasizes the fact that interactions (e.g., departures from additivity of effects) are central to decision making and identification of where the greatest impact of policy choices occur. We will extend the use of Markov chain Monte Carlo methods in a Bayesian analysis to estimate joint models for disease affected by multiple exposures. This approach offers a framework for addressing uncertainty in decision analysis while leveraging external information.

4.0 Research Approach

4.1. Methods for Aim 1.

Consider a study in which deaths have been ascertained without loss-to-follow-up for a closed occupational cohort of *n* men. Define study entry as time 0 and potential study follow-up of *T* years. Define *D* as the time of death (possibly occurring after time *T* in which case *D* is unobserved). We use subscript *i* to denote the values of variables for cohort member *i*.

Suppose we denote the hazard rate as h^a, where a=1 indicates the occupational cohort of interest and a=0 indicates a reference mortality rate (e.g., from a region or nation). We use superscript 1 and 0 to denote exposed and (reference) unexposed, respectively, throughout the paper. We allow that the hazard may vary with baseline characteristics, and vary over time on study. Let $\boldsymbol{W}$ denote a vector of baseline characteristics such as race, sex, age at entry, and calendar year of entry. Denote the mortality rate at time t in the occupational cohort by $h^1(t|\boldsymbol{W})$; and, denote the reference hazard rate function (e.g., race, sex, age, and calendar period-specific death rates for a region or nation) by $h^0(t|\boldsymbol{W})$. Furthermore, suppose that we allow that potential follow-up time to vary between cohort members, as occurs when there is staggered entry into the study and the administrative end of study follow-up is a single calendar date. Therefore, we again define study entry as time 0 and now denote potential study follow-up as T_i years for person *i*.

Suppose that follow-up time has been grouped into discrete time intervals, where $L(u)$ is the duration of follow-up over the u^{th} time period. Let $S^a(u|\boldsymbol{W}_i)$ denote the probability of surviving through time u or one minus the probability of being dead by time u, i.e.,

$$S^a(u|\boldsymbol{W}_i) = 1 - \sum_{v=0}^{u} h^a(v|\boldsymbol{W}_i)S^a(v-1|\boldsymbol{W}_i)L(v),$$

where we define $S^a(-1|\boldsymbol{W}_i) = 1$ and assume the rate is suitably small for this approximation; namely, our estimands are conditional on survival until entrance into the study.

The expected value of Y_i^a for a person who experienced the hazard rate, $h^a(t|\boldsymbol{W_i})$, over the span of T_i years is

$$\sum_{u=0}^{T_i} h^a(u|\boldsymbol{W}_i)\, S^a(u-1|\boldsymbol{W}_i)L(u).$$

Letting *O* denote the total number of deaths in a cohort of size *n* given occupational exposure (a=1), the expected value of *O* is

$$\sum_{i=1}^{n}\sum_{u=0}^{T_i} h^1(u|\boldsymbol{W}_i)\, S^1(u-1|\boldsymbol{W}_i)L(u).$$

Letting *E* denote the overall number of deaths in a cohort of size *n* in the absence of occupational exposure (a=0), the expected value of *E* is

$$\sum_{i=1}^{n}\sum_{u=0}^{T_i} h^0(u|\boldsymbol{W}_i)\, S^0(u-1|\boldsymbol{W}_i)L(u).$$

An estimator for person-time, P^a, in the cohort of interest (a=0 or a=1) is,

$\sum_{i=1}^{n}\sum_{u=0}^{T_i} S^a(u|\boldsymbol{W}_i)\, L(u)$.

This formula is attained since the average time contributed to the study by a cohort member, *i*, is the area under the survival curve, estimated by $\sum_{u=0}^{T_i} S^a(u|\boldsymbol{W}_i)L(u)$.

Using these quantities we can define the difference in the expected number of deaths under policy options, the difference in person-time in the cohort under policy options, and difference in rates of death. 95% confidence intervals were calculated using Byar's method.

4.2. Methods for Aim 2.

Suppose we are interested in cause-specific mortality rather than all-cause mortality, and allow for competing causes of death. We now allow that the probability of survival depends upon two categories of cause of death: A and B, where B denotes death due to all causes other than A. Allowing $h_A^0(t|\boldsymbol{W})$ to denote the discrete time hazard rate of outcome A in the absence of exposure, and $h_B^0(t|\boldsymbol{W})$ to denote the hazard of outcome B in the absence of exposure, the expected numbers of deaths due to A, denoted E_A is calculated as,

$\sum_{i=1}^{n}\sum_{u=0}^{T_i} h_A^0(u|\boldsymbol{W}_i)S^0(u-1|\boldsymbol{W}_i)L(u)$,

where $S^0(u|\boldsymbol{W}_i)$ is defined as the overall probability of survival up to time *u,* given as

$S^0(u|\boldsymbol{W}_i)=1-\{\sum_{v=0}^{u} h_A^0(v,\boldsymbol{W}_i)S^0(v-1,\boldsymbol{W}_i)L(v)+\sum_{v=0}^{u} h_B^0(v,\boldsymbol{W}_i)S^0(v-1,\boldsymbol{W}_i)L(v)\}$.

4.3. Extension of methods for Aims 1 and 2 to a Quantitiative Exposure Metric

Now, consider a cohort mortality study where X_i denotes a binary point exposure of primary interest, and $\boldsymbol{Z}_i$ denote baseline covariates (such as age at study entry, race, and sex), noting that *i* indexes cohort members and we use subscript *i* to denote the values of variables for cohort member *i.* Suppose the cohort information is recorded in discrete time, meaning that continuous time has been divided into a sequence of contiguous time periods of equal duration (e.g., person-years). Define study entry as time 0 and administrative censoring at end of study, τ. Denote by T_i person *i*'s failure time, C_i person *i*'s censoring time due to loss-to-follow-up (possibly occurring after τ, in which case the person was observed through the end of study), and let d_i denote an indicator of failure during the study period (d=1), or censoring (d=0). Denote the time of last observation for person *i*, T_i*, equal to T_i, C_i, or τ, whichever occurs first. A person-oriented data structure for such cohort data may include one row of data per person in the cohort study (Figure 1A), which records *i*, X_i, $\boldsymbol{Z}_i$, τ, T_i, C_i, and d_i.

An alternative data structure is a person-period data structure in which each person has multiple records in the data file. Suppose person *i* contributes τ rows of data, where τ corresponds to the number of time periods for person *i* from study entry until administrative end of study follow-up (Figure 1B). Let *j* index discrete time from study entry (j=0), and subscript *j* denote the values of variables at time period *j*. Denote by Y_{ij} a binary time-varying indicator of the outcome status associated with person *i* at period *j* that takes a value of '0' except at time T_i, when Y_{ij} is assigned

the value of the binary indicator of failure status for person i, d_i. In addition, for each record in the data structure, we define a time-varying variable, q_{ij} that equals 1 for periods, $j \leq T_i^*$, and equals 0 for periods $T_i^* < j \leq \tau$.

By partitioning the study period into narrow time intervals, a data analyst may be able to estimate the incidence proportion over each of these sub-periods, with little or no censoring occurring during any given interval; and, over a narrow interval of observation, the epidemiological cohort data may closely approximate a closed cohort. Using the term 'risk set' to refer to the group of people observed at the start of an interval who are at risk of the event, we let P_{ij} denote the probability that person i in risk set j will experience the target event during the unit time interval j, conditional on their event-free survival up to the start of time interval j (Singer and Willett 2003).

4.3.1 Using occupational cohort data to estimate a baseline disease risk function
Using just the records for person-periods observed among the unexposed (q_{ij}=1 and X_i=0), we can estimate the baseline discrete time hazard of the outcome (i.e., in the absence of exposure), which we refer to as the disease risk score. Since this is bounded by 0 and 1 over unit discrete time intervals, the disease risk score can be modeled as having a logistic dependence on a set of predictors by fitting a pooled logistic model to the discrete-time data for the unexposed, X=0, of the form

$$log\left(\frac{P_{ij}|X=0}{1-(P_{ij}|X=0)}\right) = \boldsymbol{\alpha}_j + \boldsymbol{Z}_i\boldsymbol{\beta},$$

where the vector of parameters, $\boldsymbol{\alpha}_j$, describe the baseline logit hazard function and $\boldsymbol{\beta}$ is a vector of parameters associated with covariates $\boldsymbol{Z}$. The logistic transform implies that the predictor variables are linearly associated with the logistic transform of the hazard. Note that product terms between the time-scale and covariates can of course be included in the model as well, if appropriate. Also note that estimation of a vector of j parameters associated with discrete time intervals of the baseline logit hazard function, $\boldsymbol{\alpha}_j$, may be inefficient; and, often a smooth parametric function of time, j, might be specified. Using the estimated coefficients from the fitted model, the disease risk scores (Figure 1C) may be calculated for all members of the cohort, i, and all periods j as $g(j, \boldsymbol{Z}) = expit(\widehat{\boldsymbol{\alpha}}_j + \boldsymbol{Z}_i\widehat{\boldsymbol{\beta}})$.

4.3.2 Estimating standardized rate ratios
Consider a standardized rate ratio contrasting X=1 to X=0 where the target population is those exposed at X=1. The numerator of this discrete time hazard ratio is the observed rate of the outcome among those exposed (X=1). The denominator of this ratio is the expected rate of the outcome that would have been observed if, contrary to fact, the exposed (X=1) had been unexposed (i.e., set to X=0).

Suppose that we expand the person-period data set to include 2 rows for each person-period: the first corresponds to the information needed for estimation of the numerator of the standardized rate ratio; and, the second corresponds to information for estimation of the denominator of the standardized rate ratio (Figure 1D). Letting k index these two rows of information for each person-period. The expanded data set includes rows indexed by person i, period j, and stratum k.

In this expanded data set, k=1 corresponds to the information needed for estimation of the numerator of the rate ratio (Table 1). Let $m_{ijk=1}$ and $n_{ijk=1}$ correspond to the events and person-time, respectively, contributed by person i at time j. Therefore, $m_{ijk=1}$ is a binary indicator of case status for person i at period j, Y_{ij}; and, $n_{ijk=1}$ is a binary indicator of observed person-time for person i at period j, q_{ij}.

In the expanded data set, k=0 corresponds to the denominator of the rate ratio (Table 1). Let $m_{ijk=0}$ and $n_{ijk=0}$ correspond to the expected failures and person-time, respectively, contributed by person i at period j if exposure had been set to X=0. We calculate $m_{ijk=0}$ and $n_{ijk=0}$ using the disease risk score, $g(j,\boldsymbol{Z})$, and the survival function $S(j,\boldsymbol{Z})$. Therefore, $m_{ijk=0}= g(j|\boldsymbol{Z}_i)S(j|\boldsymbol{Z}_i)$; and, $n_{ijk=0} = S(j|\boldsymbol{Z}_i) = 1 - \sum_{t=0}^{j} g(t|\boldsymbol{Z}_i)S(t-1|\boldsymbol{Z}_i)$, where we define $S(-1|\boldsymbol{Z}_i) = 1$.

A regression model fit to the expanded data structure may take the form,

$$\log\big(E[m_{ijk}]\big) = \delta_1 k + \log\,(n_{ijk}),$$

where, using just the records with X=1, the antilog of estimated parameter δ_1 estimates the standardized rate ratio reflecting the observed rate at X=1 to the expected rate among those with X=1 had exposure been set to X=0. Estimation of robust confidence intervals are recommended given the two stage regression (first estimation of disease risk score and second fitting the marginal structural model) (Huber 1967).

In the Appendix to this Alpha project report we provide illustrative SAS code to obtain discrete time hazard ratios and associated robust confidence intervals.

4.4 Methods for Aim 3 and extensions to assess impacts of occupational policy options

4.4.1 Policy options (what we can do)

Suppose that a company owner is interested in the impact of a policy that would lead to a change in workers' exposure to an agent. For example, suppose that policy A is to leave conditions unaffected and as they were in the past. Policy B is to invest in personal respiratory protection just for workers in areas of a worksite where dust levels exceeded a specified limit, m. Any number of proposed policies could be evaluated, but only two will be discussed for simplicity. The first step in applied decision theory is to clearly state our options or choices about what we can do. We wish to consider the question, "What would be the impact on the occurrence of outcome Y of a policy that affected exposure X?"

4.4.2. What we know.

Applied decision theory draws upon what we know; this may involve information drawn from external sources (e.g., prior findings) as well as empirical data from an occupational cohort. Suppose that we have an occupational cohort of n individuals. Let i = 1, 2, …, n, index cohort members, and let u index time on study. Let X denote the exposure that will be affected by policy B, allowing that it may be time-varying so that we let $X(u)$ denote the exposure level at time u. Let $\boldsymbol{W}$ denote baseline covariates (such as age at hire and sex), and let Y denote a binary indicator of the outcome of interest. Define study entry as time u=0, and denote the time from

study entry until administrative end of study follow-up as T_i for person i. Let $\acute{T}_i \leq T_i$ denote time to last observation for person i, which will end at the time of death, loss to follow-up, or administrative end of follow-up, whichever occurs first.

A discrete time representation of the person-time at risk in the cohort study divides time on study into a sequence of contiguous time intervals of uniform duration. In the resulting data structure each cohort member, i, contributes one row of data for each time period of observation, letting u index discrete time intervals, from entry time 0 to $\acute{T}_i$ (Table 1a). The binary indicator of the outcome of interest takes a value of '0' for each time period of observation for which $u<\acute{T}_i$; at the last period observation for each person i, a value of '1' is assigned if the outcome of interest was observed while a value of '0' is assigned to censored observations (Richardson 2010). This data structure is used in standard approaches to epidemiological cohort analyses and allows for estimation of the effects of exposure on health outcomes using empirical data.

4.4.3. A classical regression modeling approach

Let $P(Y_{iu}=1)$ denote the probability that person i will experience the outcome event of interest during time interval u, conditional on their event-free survival up to the start of time interval u. By partitioning the study period into a sequence of narrow discrete time intervals, we are able to estimate the probability of the event, over each of these sub-periods. Since $P(Y_{iu}=1)$ is bounded by 0 and 1, it can be modeled using the data structure in Table 1a as having a dependence on a set of predictors as,

$$log\left(\frac{P(Y_{iu}=1)}{1-P(Y_{iu}=1)}\right) = \alpha_u + \beta x_{iu} + \gamma_1 w_{i1} + \gamma_2 w_{i2} \quad \text{.... Equation 1,}$$

where x_{iu} is the exposure variable of interest, and w_{i1}-w_{i2} are covariates.

Such a model is a discrete time approach to estimation of a model similar to a Cox proportional hazards models for continuous time. The model in Equation 1 includes parameters α_u that describe temporal variation in the hazard function; in practice, temporal variation in the baseline hazard is often modeled more parsimoniously as a polynomial or spline function of the time scale. The output from fitting model Equation 1, is often used to communicate with decision makers. For communication with experts, it is standard to report the estimated coefficient, β, which corresponds to the change in log relative hazard per unit of exposure of interest, or the antilog of β which is the covariate-adjusted discrete time hazard ratio.

To derive simpler summary statistics that are often desired in presentations and communication with stakeholders, the estimated coefficients from model Equation 1 sometimes will be used to calculate other quantities to express the impact of X on Y in the cohort. For example, one summary quantity is the number of cases that can be attributed to exposure. This often is calculated based on the predicted values for the outcome that are derived using the estimated parameters for the fitted model and the covariate pattern associated with each record in the person-period file (Appendix 1). The summation of fitted values over person-periods is reported as the predicted number of events; and, by using the estimated parameter values, multiplied by the appropriate covariate patterns, one can also estimate the background cases as the predicted

values for the outcome *Y* in the absence of exposure (i.e., at *X*=0), the excess cases (i.e., the fitted values minus background cases), and the attributable fraction (excess cases over fitted values).

4.4.4. A counterfactual approach

An exposure that affects mortality will affect the distribution of person-time in the study cohort; therefore, a policy that affects exposure (e.g., reducing *X*) will affect the time of onset of *Y* (Richardson, Keil et al. 2017.) For example, a policy that reduced hazardous exposure would lengthen the average time to death among a cohort of workers and increase the amount of person-time accrued in a long term follow-up of those individuals. Consequently, the classical approach described above to calculation of excess cases and attributable fractions, although performed in the epidemiological literature, is liable to be misleading for decision-makers (Greenland and Robins 1988, Greenland and Drescher 1993, Richardson, Keil et al. 2017) because the calculations are based on the observed records in the person-period file (which reflect the person-time accrued under the observed exposure conditions, rather than what would have been observed under a policy that affected exposure (e.g., reduced exposure).

We propose an alternative approach that frames the estimate of the impact of a policy that leads to a change in exposure *X* on outcome *Y* in a counterfactual framework. The following steps are taken to make the necessary calculation.

First, an extended data structure (Table 1b) is constructed in which we add to the discrete time data structure one record for each person period from $\acute{T}_i$ until T_i such that the extended data structure includes records for each person period from u=0 through the administrative end of follow-up, T_i. In addition, for each record in the data structure, we define two indicator variables: i) a time-varying binary variable, *c*, that equals 1 for observed person periods, $u \leq T'$, else 0; and, ii) a time-varying binary variable, *f*, that equals 1 for observed person periods, *u*, in which person *i* conformed to the exposure level of a given policy, else 0 (e.g., for a policy capping exposure below a specified level *m*, we would assign *f*=1 at all times *t*<=*u* if $X_i(t) \leq m$, else 0).

Second, we fit a weighted logistic regression model with terms to describe temporal variation in the baseline hazard and effects of covariates, where we define the weight as *cf* denoting the product of *c* and *f*. We use the fitted logistic model to generated predicated values of the estimated probability (hazard) of the outcome for each person-period in the data set. Standard statistical packages allow the investigator to output a person-period data structure that now includes the predicted value of the outcome, which is the estimated hazard of the outcome for each record in the person-period file (Appendix 2). This predicted value is based on the estimated parameters for the weighted regression model where the weighting leads to estimates based solely on the observed data for those workers who conformed to a given policy.

Finally, we use the predicted value of the hazard of the outcome during time interval *u* to calculate the probability of surviving through that time interval, for each person *i* and time interval *u*. Given the extended data structure, we can calculate the survival probability of the outcome to any time or age. The survival probability is calculated as one minus the product of the hazard of the outcome during that time interval and the probability of surviving up to the start

of that time interval (Appendix 2). We define the probability of surviving up to the first time interval as 1, because our estimates are conditional on survival until entrance into the study. Using this information, we can readily calculate the expected risk of death under a proposed policy (such as policy *B*) for a worker. The extended data structure in Table 1b allows for the expected risk of death to be calculated to any given attained age or time on study.

This approach readily extends to analyses of cause-specific mortality in which we model two (or more) competing risks for mortality (Appendix 3).

4.4.5. Graphical and tabular communication

The observed and counterfactual person-time and numbers of cause-specific deaths can be used as the basis for graphs, charts, figures, and tables that illustrate comparative survival, absolute numbers of events, years of life lost, and risk difference in the study population under alternative policies, standardized to the baseline covariate distribution of the cohort. These values can be summed over all records to yield estimates of the overall number of deaths and years of life in the cohort through the administrative end of follow-up under contrasting policies *A* and *B*. An important aspect of decision analysis is representing uncertainty in outcomes under differing policy alternatives. Projections regarding future events, such as the expected numbers of events if the cohort were followed to a specified date in the future, are readily generated by extending the data structure to include additional records for person-periods that span the desired time period.

4.5 Implementation in this Alpha Foundation project

During this project we established data access agreements and developed a process for working together on analyses. We identified 28,546 miners employed in uranium mining in Ontario for at least 1 week between 1954 and 1996, obtained work history, exposure, and mortality follow-up information for these workers. We created an analytical file of person-time of observation from date of entry into the analysis until end of follow-up or administrative censoring of workers alive at age 90 years. As proposed, all of the aims were addressed using this analytical data structure. As proposed, Dr. Richardson led the development of statistical methods, pilot testing development of statistical software code for proposed work, and fit models by collaborating with partners at OCRC. Analyses were run on computers at OCRC under the direction of Drs. Demers and Arrandale, with Colin Berrault performing the computer runs. Summary output was shared with Dr. Richardson.

Using information on estimates of annual radon progeny exposure in this cohort, we considered a policy B that caps annual exposure to radon using the time-varying exposure compliance variable *f*. Using the approach described above, we fit a model with terms for year of birth and attained age, and estimate survival times and deaths under the policy cap. We use the resultant values as the basis for graphs, charts, figures, and tables that illustrate comparative survival, absolute numbers of events, years of life lost, and risk difference in the study population under alternative policies. We further extend the calculations to examine cause specific mortality and further illustrate comparative cause-specific survival, rank ordering of leading causes of death under different policies, and survival.

5. Results, Summary of Accomplishments, Conclusions and Impact Assessment

We illustrated Aim 1 methods using data for a cohort of men who entered follow-up of miners study in 1954 and were followed through 2007 to ascertain deaths. The expected number of deaths due to all causes were calculated. Extending this to a competing risks framework in Aim 2, we calculated expected numbers of deaths due to all causes and due to lung cancer.

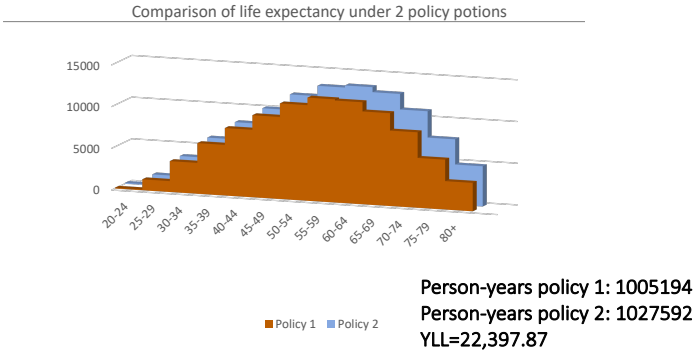
Table1. Male miners followed until 2007.

	Proposed calculations	
Outcome	Observed (*Policy 1*)	Expected (*Policy 2*)
All causes	8572	9280.8
Lung cancer	1246	1407.3
Cardiovascular disease	2788	4468.1

The person-time observed in the cohort was 1,005,194 person-years; reducing occupational exposure would extend working life in the cohort yielding a standardized estimate of 1,027,592 years.

Figure 1. number of years of life lost in the cohort; and,

Values: Life expectancy versus number of deaths
Shape/Frame: Knowledge. Options. Values



Person-years policy 1: 1005194
Person-years policy 2: 1027592
YLL=22,397.87

Extending this to policies affecting a quantitative occupational exposure estimate, as in Aim 3, we first consider a standard regression analysis of these data; this regression model yields an age and birth cohort conditional estimate of the association between the natural log of cumulative exposure and all cause mortality, with the estimated coefficient 0.09 (se=0.02) (Table 2). If we use the model coefficients, we calculate 216 excess deaths among those who were exposed, yielding an estimate of 3% attributable fraction (table 2).

Next, consider policy *B* that removes exposure (i.e., 0 WLM per year). Table 3 reports estimates of the impact of such a policy. The values in table 3 account for the fact that a reduction in exposure would not prevent the inevitability of death, but rather delay it. Under this policy 8176 deaths would be expected, as compared to the 8346 deaths observed (table 3) by ag 85 years. This suggests 170 excess deaths, in contrast to the 216 excess deaths suggested by the calculation in Table 2. A calculation of the excess deaths divided by the total observed deaths suggests an attributable fraction of deaths of about 2%, smaller than the value of 3% reported in table 2. However, while eliminating radon progeny exposure leads to a small change in the number of deaths in the cohort by the end of study follow-up, it leads to greater longevity among the workers. As also shown in table 3, the expected number of person-years observed under policy *B* (limiting exposure) is 13,319 person-years greater than the number observed in the cohort under policy *A* . Under policy *B*, deaths tend to occur at older ages and there is an increase in years of life expected overall in the cohort (approximately 6 additional months of life per worker in the cohort).

Figure 1 illustrates a comparative survival curve, which shows that by attained age 85 years, the percentage of workers dead under either policy is about equal. However, the percentage of the workforce dead at age 65, 70, and 75 is lower under the policy that limits exposure. Most of the years of life gained under policy *B* occur among people in the 60s and 70s (Figure 2).

Table 2 also shows cause-specific mortality for lung cancer and other causes, under policies *A* and *B*. The number of lung cancer deaths is substantially reduced under policy *B* compared to policy *A* (262 fewer lung cancer deaths); in contrast, the number of deaths due to other natural causes is increased. Figure 3 illustrates that policy *B* shifts the distribution of deaths between categories of cause of death. Removing a lung carcinogen may reduce one (or more) cause of death – but in the long run the competing causes of death will lead to an increase in other causes (and perhaps no long term reduction in cumulative mortality). The cause specific cumulative mortality functions illustrate that under policy B the cumulative incidence of lung cancer does not increase as steeply (red line) but other causes of death contribute to causes of death in its stead. Table 2 may help a decision maker consider the role of values/goals in choosing between policy options. The rank order of causes of death may be anticipated to shift substantially when we use counterfactuals (table 2). We must therefore be clear about which outcomes we prefer, reflecting a set of values inherent in decision making.

Table 2. Parameter estimates, fitted values, background, excess, and attributable fraction of deaths obtained by fitting a logistic regression model to the discrete time cohort data.

	Parameters		Fitted values
Parameter	Estimate	Standard Error	
Intercept	-5.62	0.03	
age *	0.074	0.00	
age^2	0.0001	0.00	
age^3	-0.00004	0.00	
born ≤ 1925	0.31	0.03	
1925 < born ≤ 1935	0.102	0.03	
cumulative exposure /100 †	0.0934	0.0231	
Background cases (*BK*)			8130
Excess cases (*EX*)			216
Attributable Fraction (*EX*/[*EX*+*BK*])			0.03

* Attained age up to 85 years.

† Cumulative WLM under a 10 year lag.

Table 3. Expected deaths in the absence of exposure, excess deaths, and attributable fraction of deaths based on the counterfactual failure times. Cohort of underground miners.

Observed (policy *A*)			Expected under policy *B*			Difference (*B*-*A*)		
Person-years	Deaths		Person-years	Deaths		Person-years	Deaths	
998,510	8,346		1,011,829	8,176		13,319	-170	
	Circulatory disease	2,711		Circulatory disease	2,781		Circulatory disease	70
	Cancer excl lung	1,508		Cancer excl lung	1,554		Cancer excl lung	46
	Lung cancer	1,230		Lung cancer	968		Lung cancer	-262
	COPD	334		COPD	348		COPD	14

Figure 1. Cumulative mortality by attained age under policy *A* (red line) and *B* (blue line).

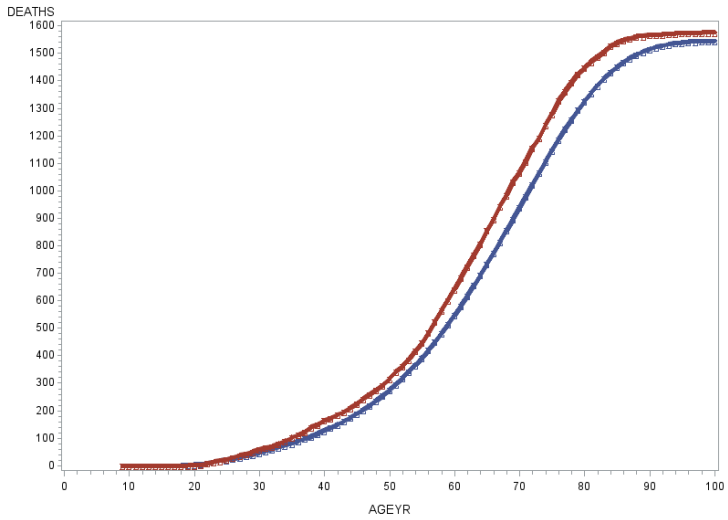


Figure 2. Causes of death shifting. Distribution of deaths under policy *A* and under policy *B* by cause of death.

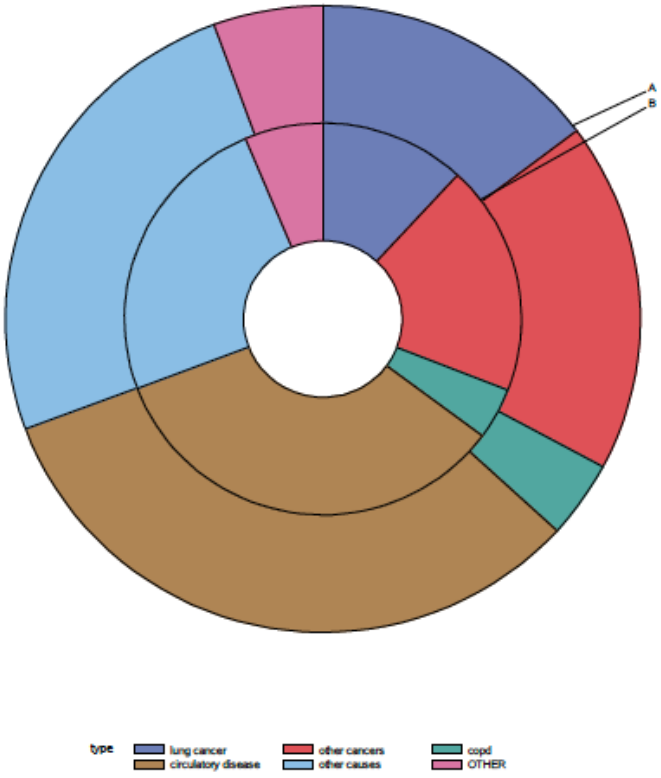
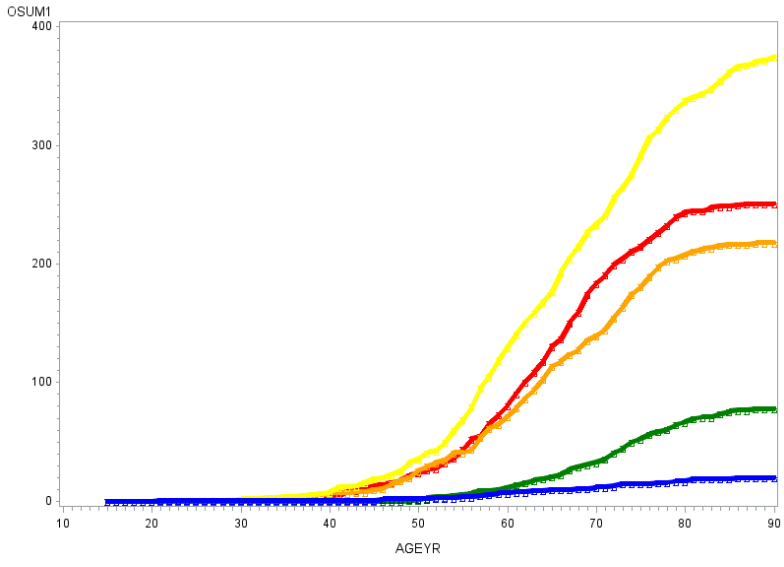
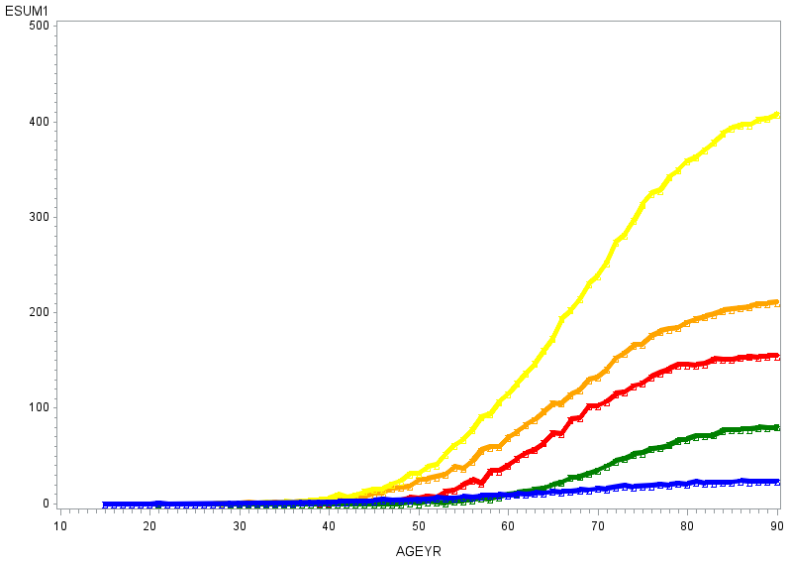


Figure 4. Cause-specific cumulative incidence curves by cause. Red=lung, Orange=other cancer, Yellow=circulatory disease, Green=COPD, Blue=pneumonia



Policy A (observed)



Policy B (expected)

6. References

Greenland, S. and K. Drescher (1993). "Maximum likelihood estimation of the attributable fraction from logistic models." Biometrics **49**(3): 865-872.

Greenland, S. and J. M. Robins (1988). "Conceptual problems in the definition and interpretation of attributable fractions." Am J Epidemiol **128**(6): 1185-1197.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, University of California Press.
Richardson, D. B. (2010). "Discrete time hazards models for occupational and environmental cohort analyses." Occup Environ Med **67**(1): 67-71.
Richardson, D. B., A. P. Keil, S. R. Cole and R. F. MacLehose (2017). "Observed and Expected Mortality in Cohort Studies." Am J Epidemiol: 1-8.
Singer, J. D. and J. B. Willett (2003). Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. New York, Oxford University Press.

7.0 Appendices

7.1 Aim 1.

The expected number of deaths in a cohort of male workers can be obtained by multiplying the appropriate age- and calendar period-specific reference rate by the units of person-time contributed by each person over the period of study. These products are summed over all person-periods that were contributed by all individuals to obtain the expected number of deaths. The program below calculates the expected number of deaths due to all causes in two ways: first using the g-method; and, second using the classical SMR method.
To do these calculations, a person-period file is constructed with one record for each person-period from study entry until the administrative end of follow-up. For example, if the administrative end of follow-up is 31December 2007 then we have a record for each person-period until that date (regardless of whether a given person survives until the end of follow-up). The attained age and calendar period associated with each person-period is determined by the time on study and the person's age at entry and calendar date of entry into the study. The SAS program shown below is used to calculate the expected numbers of deaths in a single pass through the data.

7.1.1. SAS code for all cause mortality calculations

Assume the study data consist of a person-period file, named *DS*. Each person is identified by a unique study id, *i*. There is one record per unit of person-time, with time indexed by the variable *u* which takes a value of 0 at study entry and increases monotonically until time T_i, the administrative end of follow-up. Each record of this dataset includes the following:
i, a unique study id for each person;
u, a variable that indexes potential follow-up time, from 0 at study entry to *T* at end of the study follow-up;
age, a variable equal to age-at-entry plus *u*,
period, a variable equal to date-of-entry plus *u*;
c, a binary variable that equals 1 for time prior to date of last observation ($u \leq T'$), else 0
rate, reference death rate, expressed in deaths per person-year at risk (e.g., reference rate for the sex group of subject *i*, and the 5-year age and calendar period intervals associated with follow-up time *u*);
lu, unit of person-time (e.g., 1 if each record represents one person-year).

The data are sorted by *i* and *u*.

The following code calculates *E*, the g-estimate of the expected number of deaths if the cohort had experienced the reference hazard rate. The code also calculates *Q*, the expected number of deaths using the classic SMR method.

```
data MR ;
set DS end=eof ;
 by i u ;
retain S_i expected_i expected_smr_i  E Q;
if _n_=1 then do;   E=0; Q=0; end;
if first.i then do;   S_i=1; expected_i=0; expected_smr_i=0; end;
expected_i=expected_i+ (rate * S_i * lu); expected_smr_i= expected_smr_i + (rate * c * lu);
S_i=1-expected_i ;
if last.i then do ;  E=E+expected_i;  Q=Q+expected_smr_i;  end;
if eof then output MR; run;

proc print data=MR ;  var E Q; run;
```

7.2. Aim 2. Competing risks.

Suppose that the dataset, *DS*, also includes the variables *rateA* and *rateB*, reference death rates for causes A and B, expressed in deaths per person-year at risk (e.g., reference rates for causes A and B for the race and sex group of subject *i*, and the 5-year age and calendar period intervals associated with follow-up time *u*). The data are sorted by *i* and *u*.

The following code calculates *E* and *Q* for a specific cause of death, A, allowing for mortality due to competing causes, B.

```
data MR ;
set DS END=EOF ;
by i u ;
 retain S_i expectedA_i expectedB_i expected_smr_i   E Q ;
if _n_=1 then do;  E=0; Q=0; end;
if first.i then do;  S_i=1; expectedA_i=0; expectedB_i=0; expected_smr_i=0; end;
expectedA_i=expectedA_i+ (rateA * S_i * lu); expectedB_i=expectedB_i+ (rateB * S_i * lu);
expected_smr_i= expected_smr_i  + (rateA * c * lu);
S_i=1-expectedA_i-expectedB_i ;
if last.i then do ;   E=E+expectedA_i;   Q=Q+expected_smr_i; end;
if eof then output MR; run;

proc print data=MR ; var E Q; run;
```

7.3. Calcuation of counterfactual failure times and events with quantitative exposures

A logistic regression model can be fit to the discrete time data. The model is weighted so that the estimated parameters are based on records for observed person-periods during which workers had consistently conformed to the policy. We use the resultant estimates of the discrete time hazard to calculate counter factual failure times and events under the policy of interest for each

cohort member, *i*, under policy *a*. Let $S^a(u|\boldsymbol{W}_i)$ denote the probability of surviving through time *u* (i.e., one minus the probability of being dead by time *u*). We define $S^a(-1|\boldsymbol{W}_i) = 1$, and then can proceed to calculate survival through each person-period for subject *i* as one minus the product of the discrete time hazard in each period, $h^a(t|W)$, and the survival through the preceding period,

$$S^a(u|\boldsymbol{W}_i) = 1 - \sum_{v=0}^{u} h^a(v|\boldsymbol{W}_i)S^a(v-1|\boldsymbol{W}_i)L(v),$$

where $L(v)$ equals 1 in our example, given uniform durations of person-periods in our data structure. We can calculate the probability of being dead by time *u* if policy *a* had been implemented as, $Y^a(u|\boldsymbol{W}_i) = \sum_{v=0}^{u} h^a(v|\boldsymbol{W}_i)\, S^a(v-1|\boldsymbol{W}_i)L(v)$.

```
proc logistic data=Table1b;
 model Y(event='1')= t w1 w2;
 weight cf;
 output out = table1c p = h; run;


data table1d;
  set table1c;
  by i u ;
  retain S_u D_u;
  if first.i then do; S_u=1; D_u=0; end;
    delta_D_u = (h * S_u );
    D_u = D_u + (h * S_u );
  output table1d;
    S_u = 1 - D_u ; run;
```

7.4. Aim 3 methods extended. Calculation of counterfactual failure times and events for cause-specific mortality under altenrative policy options.

Suppose there are 2 categories of cause of death of interest, *Y1* and *Y2*, where category *Y2* denotes death due to all causes other than categories *Y1*. Let $h_I^a(t|\boldsymbol{W})$ denote the discrete time hazard rate of outcome *1* at time *t* under policy *a*, and $h_{II}^a(t|\boldsymbol{W})$ denote the hazard of outcome *2* under policy *a*. Under our proposed approach, each hazard function is estimated by a regression model fit to the empirical data for those who comply with the policy. Under this setting of competing risks, survival is calculated simply by extending the expression to include all categories of cause death,

$S^a(u|\boldsymbol{W}_i)=1 - \{\sum_{v=0}^{u} h_I^a(v,\boldsymbol{W}_i)S^a(v-1,\boldsymbol{W}_i)L(v) + \sum_{v=0}^{u} h_{II}^a(v,\boldsymbol{W}_i)S^a(v-1,\boldsymbol{W}_i)L(v)\}$. The

approach readily extends from two categories to any number of categories (as long as they are mutually exclusive and exhaustive, such that they encompass all causes of death). Below, we illustrate calculations using the SAS package.

```
%let n=2;

%macro models;
%let cnt=0;
```

```
data m0; set table1b;
%do %while (&cnt<&n);
proc logistic data=m&cnt;
%let cnt=%eval(&cnt+1);
model Y&cnt(event='1')=t w1 w2; weight cf; output out = m&cnt p = h&cnt ; run;
%end;
%mend;
%models;

proc sort data=m&n; by i u; run;

data MR MRunit;
set m&n END=EOF ;
by i u ; du=1;
array expi{*} expected_i1-expected_i&n;
array ec{*} ec1-ec&n; array rate{*} h1-h&n;
array hexpi{*} hexpected_i1-hexpected_i&n;
retain S_i expected_i1-expected_i&n  Ec1-Ec&n P_i P;
if _n_=1 then do; do a=1 to &n; ec{a}=0; end; P=0; CP=0; end;
if first.i then do; S_i=1; do a=1 to &n; expi{a}=0; end; P_i=0; end;
do a=1 to &n;
expi{a}=expi{a}+ ( rate{a} * S_i * du); hexpi{a}= ( rate{a} * S_i * du); end;
P_i =P_i + (S_i * du) ;
hP_i = (S_i * du) ;  * hold counterfactual person-period length;
S_i=1-sum(of expected_i1-expected_i&n);
output MRunit;
if last.i then do; do a=1 to &n; ec{a}=ec{a}+expi{a}; end; P=P+P_i; end;
if eof then output MR; run;
```

8. Acknowledgement/Disclaimer

This study was sponsored by the Alpha Foundation for the Improvement of Mine Safety and Health, Inc. (ALPHA FOUNDATION). The views, opinions and recommendations expressed herein are solely those of the authors and do not imply any endorsement by the ALPHA FOUNDATION, its Directors and staff.

ADDENDUM TO FINAL REPORT

Alpha Foundation for the improvement of Mine Safety and Health, Inc.

Grant Number and Title:

AFC6l8-53, "Innovations in applied decision theory for mine surveillance and health and safety efforts"

Organization Name:

University of North Carolina of Chapel Hill

Principal investigator:

David B. Richardson

Contact Information:

Email: david.richardson@unc.edu

Phone: 919 966 2675

Fax: 919 966 2089

Period of Performance:

August 1,2017 to February 28, 2019

In this addendum, we discuss communication of findings in graphical and tabular formats.

Decisions about possible interventions to improve miners' health and safety require reasonable estimates of the intervention's impact, an indication of where impact can be made, and how big of an impact an intervention may have in terms of disease, disability, and mortality. In our project, we established quantitative statistical methods for calculation and ranking of cause-specific excess disease in a contemporary framework for valid decision making.

For example, consider a decision-maker who is contemplating a policy change that would reduce hazardous exposure on-the-job. For example, suppose that policy *A* is to leave conditions unaffected and as they were in the past. Policy *B* is to invest in respiratory protection. If an occupational agent is hazardous then reduction of exposure to the agent may lengthen the average time to death among members of a cohort of workers. This will increase the amount of person-time accrued in a long term follow-up of those workers because reduction of hazardous exposures will lead to an increase in longevity. In the long-run regardless of whether these men and women are exposed to the occupational hazard they will die. A reduction in occupational hazards does not prevent mortality. Rather, it may affect when death occurs, and potentially alter what cause a worker will die from.

When assessing the impact of a policy, we propose examining the policy's impact on longevity and the expected distribution of causes of death. Below we illustrate how graphs, charts, and tables of comparative survival, absolute numbers of events, years of life lost, and risk difference in the study population under alternative policies.

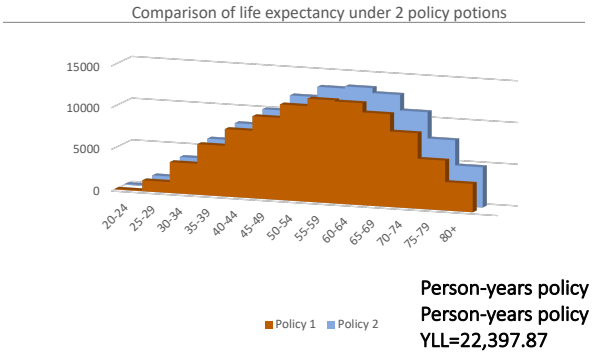
To illustrate, we use information about a cohort of 28,546 miners employed in uranium mining in Ontario for at least 1 week between 1954 and 1996, and were followed through 2007 to ascertain deaths. We considered a policy to effectively remove exposure to radon (i.e., 0 WLM per year). We note that the reported comparisons are 'fair comparisons' with regards to factors such as age, race, and sex because they are standardized to the baseline distribution of these factors in the occupational cohort.

The expected number of deaths due to all causes were calculated. Extending this to a competing risks framework in Aim 2, we calculated expected numbers of deaths due to all causes and due to lung cancer.

The person-time observed in the cohort was 1,005,194 person-years. Reducing occupational exposure would extend life in the cohort yielding an estimate of 1,027,592 person-years. Figure 1 illustrates the gain in life expectancy. The red bars in the bar chart illustrate the number of person-years of life in each category of attained age that were observed in the cohort. The blue bars in the bar chart illustrate the number of person-years of life in each category of attained age that would be expected in the cohort if the policy to cap radon exposure had been in place. The blue bars are as high or higher than the red bars in each category of age, indicating that if the protective occupational policy had been in place there would be more workers alive at each attained age.

Figure 1. number of years of life lost in the cohort; and,

Values: Life expectancy versus number of deaths
Shape/Frame: Knowledge. Options. Values



Person-years policy 1: 1005194
Person-years policy 2: 1027592
YLL=22,397.87

In the occupational cohort 8346 deaths were observed (Table 1). Next, consider the number of deaths under policy *B*. Under this policy 8176 deaths would be expected (table 3) by age 85 years. This suggests 170 excess deaths. The number of lung cancer deaths is substantially reduced under policy *B* compared to policy *A* (262 fewer lung cancer deaths); in contrast, the number of deaths due to other natural causes is increased.

Table 1. Expected deaths in the absence of exposure, excess deaths, and attributable fraction of deaths based on the counterfactual failure times. Cohort of underground miners.

Observed (policy *A*)		Expected under policy *B*		Difference (*B*-*A*)	
Deaths	8,346	Deaths	8,176	Deaths	-170
Circulatory disease	2,711	Circulatory disease	2,781	Circulatory disease	70
Cancer excl lung	1,508	Cancer excl lung	1,554	Cancer excl lung	46
Lung cancer	1,230	Lung cancer	968	Lung cancer	-262
COPD	334	COPD	348	COPD	14

Figure 2 illustrates that policy *B* shifts the distribution of deaths between categories of cause of death. Removing a lung carcinogen may reduce one (or more) cause of death – but since we all die in the long-run – the workers in the cohort will eventually experience an increase in other causes of death.

Figure 2. Causes of death shifting. Distribution of 1576 deaths under policy *A* and 1547 under policy *B* by cause of death.

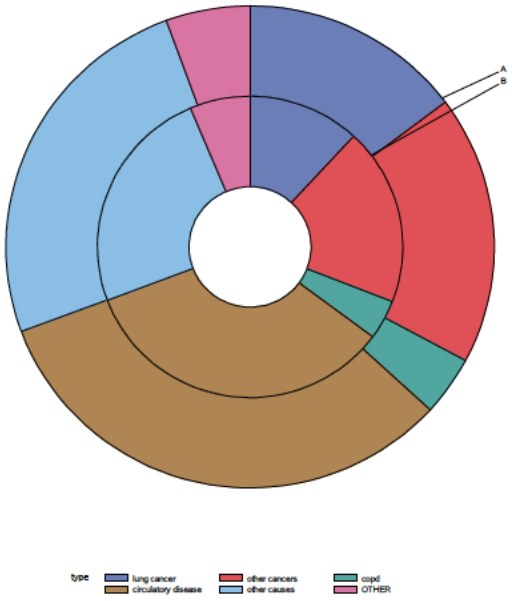


Table 1 and Figure 2 may help a decision maker consider the role of values/goals in choosing between policy options. We must be clear about which outcomes we prefer, reflecting a set of values inherent in decision making.