An Application of the G-formula to Asbestos and Lung Cancer

Stephen R. Cole

Epidemiology, UNC Chapel Hill

Slides: www.unc.edu/~colesr/
Acknowledgements

Collaboration with David B. Richardson, Haitao Chu, and Ashley I. Naimi

Supported by NIH grant R01CA117841

Expert advice form James M. Robins and Daniel Westreich
Background

- Exposure to asbestos fibers has been associated with an increased risk of lung cancer mortality in a number of prospective occupational studies, e.g. (Hein, Stayner et al. 2007)

- However, this evidence-base (which is used to set national asbestos standards) is subject to the *healthy worker survivor bias* (Eisen and Robins 2002)
Healthy Worker Survivor Bias

- Employment status is a \textit{time varying confounder} (Robins, Hernán et al. 2000)
  - Associated with subsequent workplace asbestos exposure and mortality
- Employment status is \textit{affected by prior exposure}
- Exposure does not occur after leaving the workplace, inducing \textit{non-positivity}
Generalized (G) Methods

- Methods to estimate exposure effects defined by potential outcomes (Robins and Hernán 2008)

1. Parametric G formula, “usually” fit by Monte Carlo simulation (Robins 1986)


3. Marginal structural model, usually fit by inverse probability weighting (Robins 1999)
Objective

Employ the parametric G formula to explore lung cancer mortality in a cohort of individuals occupationally-exposed to asbestos fibers in the textile manufacturing industry.
South Carolina Asbestos Cohort

- 3072 men and women employed at a textile production plant for at least one month between 1 January 1940 and 31 December 1965 (Dement, Brown et al. 1994)
- Follow up from 18 years of age; therefore we excluded 70 of 3072 individuals who left work before age 18
- Date of birth, year of study entry, sex, race (i.e., Caucasian vs. non-Caucasian), prevalent asbestos exposure at entry, and a time varying indicator of employment status were ascertained from employment records
- Followed until: loss to follow up (297, 10%), death (195+1703, 63%), or administrative censoring at 90 years of age or 31 December 2001 (807, 27%)
Lung Cancer Mortality

- 3002 individuals at risk for lung cancer mortality over 117,471 person-years between 1940-01, and cases of lung cancer death (as red lines) by year of age

- Vital status from NDI or state vital records, lung cancer deaths coded according to ICD in effect at the time of death
Exposure Assessment

- Asbestos exposure, in fiber-years per milliliter of air (f-y/ml), by age; 3002 individuals over 117,471 person-years between 1940-01
- Average asbestos concentrations were estimated using a department, operation, and calendar time specific job exposure matrix (Dement, Harris et al. 1983)
Notation

- 3002 participants are indexed by $i = 1$ to $N$ (index sometimes omitted)
- Age is indexed by $j = 18$ to 90, with $S_i$ denoting the age at entry
- Let $V_{i(S_i)}$ be the vector of time fixed covariates (entry calendar year, sex, race)
- Let $W_{ij} = 1$ indicate participant $i$ was at work during age $j$, 0 otherwise
- Let $X_{ij}$ represent asbestos exposure during age $j$, measured in f-y/ml
- Let $C_{ij} = 1$ indicate censoring due to drop-out at age $j$
- Let $D_{ij} = 1$ indicate death at age $j$ due to causes other than lung cancer
- Finally, let $Y_{ij} = 1$ indicate death at age $j$ due to lung cancer

- We assume the temporal order: $V_{i(S_i)}, W_{ij}, X_{ij}, C_{ij}, D_{ij}, Y_{jj}$
Cumulative Incidence

\[ I(j) \]

\[
= \sum_{k=18}^{j} \sum_{v} \sum_{w_j} \sum_{\bar{x}_j} P(Y_k = 1 | V = v, \bar{W}_k = \bar{w}_k, \bar{X}_k = \bar{x}_k, \bar{Y}_{k-1} = \bar{D}_k = \bar{C}_k = 0, S \leq k)
\]

\[
\times \prod_{m=18}^{k} \left[
\begin{array}{l}
P(D_m = 0 | V = v, \bar{W}_m = \bar{w}_m, \bar{X}_m = \bar{x}_m, \bar{Y}_{m-1} = \bar{D}_{m-1} = \bar{C}_{m-1} = 0, S \leq m) \times \\
P(C_m = 0 | V = v, \bar{W}_m = \bar{w}_m, \bar{X}_m = \bar{x}_m, \bar{Y}_{m-1} = \bar{D}_{m-1} = \bar{C}_{m-1} = 0, S \leq m) \times \\
f(X_m = x_m | V = v, \bar{W}_m = \bar{w}_m, \bar{X}_{m-1} = \bar{x}_{m-1}, \bar{Y}_{m-1} = \bar{D}_{m-1} = \bar{C}_{m-1} = 0, S \leq m) \times \\
P(W_m = 1 | V = v, \bar{W}_{m-1} = \bar{w}_{m-1}, \bar{X}_{m-1} = \bar{x}_{m-1}, \bar{Y}_{m-1} = \bar{D}_{m-1} = \bar{C}_{m-1} = 0, S \leq m) \times \\
f(V = v) \times \\
P(Y_{m-1} = 0 | V = v, \bar{W}_{m-1} = \bar{w}_{m-1}, \bar{X}_{m-1} = \bar{x}_{m-1}, \bar{Y}_{m-2} = \bar{D}_{m-1} = \bar{C}_{m-1} = 0, S \leq m - 1)
\end{array}
\right]
\]

for \( j = 18 \) to \( 90 \), where history of a time varying variable is denoted using an overbar, e.g. \( \bar{X}_j = \{X_{S_i}, X_{(S_i+1)}, \ldots, X_{(j-1)}, X_j\} \)
Cumulative Incidence as Plinko

\[ V_s, W_s = 1, X_s, C_s = D_s = Y_s = 0, m = 0 \]

\[ W_{S+m} \]

- Yes: \[ X_{S+m} > 0 \]
- No: \[ X_{S+m} = 0 \]

\[ X_{S+m} > 0 \]

- Yes: \[ C_{S+m} \]
- No: \[ D_{S+m} \]

\[ C_{S+m} \]

- Yes: \[ Y_{S+m} \]
- No: \[ Y_{S+m} \]

\[ Y_{S+m} \]

- Yes: \[ Y_{S+m} \]
- No: \[ Y_{S+m} \]

- If \( S+m = 90 \) then end, else
G formula as modified Plinko

\[ V_s, W_s = 1, X_s, C_s = D_s = Y_s = 0, m = 0 \]

\[ m = m + 1 \]

\[ W_{S+m} \]

\[ X_{S+m} > 0 \quad X_{S+m} = 0 \]

if \( X_{S+m} > q \) then \( X_{S+m} = q \)

\[ C_{S+m} \]

\[ D_{S+m} \]

\[ Y_{S+m} \]

end

if \( S + m = 90 \) then end, else
G Formula

\[ I(j) \tilde{x}_j, \tilde{c}_j = 0 \]

\[
= \sum_{k=18}^{j} \sum_v \sum_{\bar{w}_j} \left\{ P(Y_k = 1|V = v, \bar{W}_k = \bar{w}_k, \bar{x}_k, \bar{Y}_{k-1} = \bar{D}_k = \bar{c}_k = 0, S \leq k) \right\} \\
\times \prod_{m=18}^{k} \left[ \begin{array}{l}
P(D_m = 0|V = v, \bar{W}_m = \bar{w}_m, \bar{x}_m, \bar{Y}_{m-1} = \bar{D}_{m-1} = \bar{c}_{m-1} = 0, S \leq m) \times 1 \\
P(W_m = 1|V = v, \bar{W}_{m-1} = \bar{w}_{m-1}, \bar{x}_{m-1}, \bar{Y}_{m-1} = \bar{D}_{m-1} = \bar{c}_{m-1} = 0, S \leq m) \times f(V = v) \times P(Y_{m-1} = 0|V = v, \bar{W}_{m-1} = \bar{w}_{m-1}, \bar{x}_{m-1}, \bar{Y}_{m-2} = \bar{D}_{m-1} = \bar{c}_{m-1} = 0, S \leq m - 1) \end{array} \right] \]
Exposure Scenarios

- The annual asbestos exposure is capped at $q$:
  - 5 f-y/ml, 1971 OSHA standard
  - 2 f-y/ml, 1976 OSHA standard
  - 0.1 f-y/ml, 2011 (current) OSHA standard, or
  - 0.05 f-y/ml asbestos exposure

- Compared to the natural course, we estimate the total effect of the dynamic regime: allow no censoring by drop out $c_j = 0$; and if at work, receive asbestos exposure no greater than $\bar{x}_j = q$ fibers/ml in any year; and if not at work, receive no exposure
Steps in the Parametric G Formula (for each exposure scenario)

Step 1: Fit parametric models for each component, using 3002 participants (6 models for 5 components)
Step 2: Draw a Monte Carlo sample of $M = 50,000$ pseudo-participants sampled randomly with replacement from the original $N = 3002$ observed participants (inferences similar when $30,000 < M < 100,000$)
Step 3: Build out follow up on $M$ pseudo-participants using models fit in Step 1
Step 4: Compute the cumulative incidence of lung cancer mortality at 90 years of age (other deaths as competing risks) on the built-out pseudo-participants
Step 5: Estimate the risk differences and ratios at 90 years of age as measures of association, with 95% confidence intervals (CI) based on a nonparametric bootstrap of 200 resamples
Characteristics of textile workers at study entry and over follow up 1940-01

<table>
<thead>
<tr>
<th>Characteristic:</th>
<th>Study entry</th>
<th>Person-Years</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n = 3002$</td>
<td>$n = 117,471$</td>
</tr>
<tr>
<td>Age, years</td>
<td>24 (20, 31)</td>
<td>48 (36, 60)</td>
</tr>
<tr>
<td>Calendar year</td>
<td>1943 (41, 47)</td>
<td>1967 (55, 81)</td>
</tr>
<tr>
<td>Male, % ($n$)</td>
<td>58% (1749)</td>
<td>57% (67,178)</td>
</tr>
<tr>
<td>Caucasian, % ($n$)</td>
<td>81% (2436)</td>
<td>83% (97,277)</td>
</tr>
<tr>
<td>Employed, % ($n$)</td>
<td>100% (3002)</td>
<td>19% (21,867)</td>
</tr>
<tr>
<td>Asbestos exposure:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exposed, % ($n$)</td>
<td>99% (2984)</td>
<td>16% (19,312)</td>
</tr>
<tr>
<td>f-y/ml, among exposed</td>
<td>1.83 (0.89, 3.51)</td>
<td>3.30 (1.49, 5.00)</td>
</tr>
</tbody>
</table>
Characteristics of person years for observed data and natural course

<table>
<thead>
<tr>
<th>Exposure:</th>
<th>Male, %</th>
<th>Employed, %</th>
<th>Asbestos:</th>
<th>f-y/ml, among exposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed data</td>
<td>57%</td>
<td>19%</td>
<td>16%</td>
<td>3.30 (1.49, 5.00)</td>
</tr>
<tr>
<td>Natural course</td>
<td>55%</td>
<td>15%</td>
<td>13%</td>
<td>4.33 (2.17, 6.70)</td>
</tr>
</tbody>
</table>
Cumulative lung cancer mortality in observed data (gray line) and simulated natural course (black line); 3002 individuals over 117,471 person-years 1940-01
Cumulative lung cancer mortality by age 90 years

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Mortality, %</th>
<th>Risk difference</th>
<th>Risk ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed data</td>
<td>9.4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Natural course</td>
<td>9.2</td>
<td>-0.2</td>
<td>0.98</td>
</tr>
</tbody>
</table>
## Characteristics of person years for observed data and simulated scenarios

<table>
<thead>
<tr>
<th>Exposure:</th>
<th>Male, %</th>
<th>Employed, %</th>
<th>Asbestos:</th>
<th>Exposed, %</th>
<th>f-y/ml, among exposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed data</td>
<td>57%</td>
<td>19%</td>
<td>16%</td>
<td>3.30 (1.49, 5.00)</td>
<td></td>
</tr>
<tr>
<td>Natural course</td>
<td>55%</td>
<td>15%</td>
<td>13%</td>
<td>4.33 (2.17, 6.70)</td>
<td></td>
</tr>
<tr>
<td>&lt; 5 f-y/ml</td>
<td>55%</td>
<td>13%</td>
<td>11%</td>
<td>3.46 (1.64, 5.00)</td>
<td></td>
</tr>
<tr>
<td>&lt; 2 f-y/ml</td>
<td>55%</td>
<td>9%</td>
<td>8%</td>
<td>2.00 (0.59, 2.00)</td>
<td></td>
</tr>
<tr>
<td>&lt; 0.1 f-y/ml</td>
<td>55%</td>
<td>6%</td>
<td>4%</td>
<td>0.10 (0.09, 0.10)</td>
<td></td>
</tr>
<tr>
<td>&lt; 0.05 f-y/ml</td>
<td>55%</td>
<td>6%</td>
<td>4%</td>
<td>0.05 (0.05, 0.05)</td>
<td></td>
</tr>
</tbody>
</table>
Cumulative lung cancer mortality under the simulated natural course (solid black line), < 5 fiber-years/ml (dashed black line), < 2 fiber-years/ml (dotted black line), < 0.1 fiber-years/ml (dash-dot black line), and < 0.05 (solid gray line)
Cumulative lung cancer mortality by age 90 years

<table>
<thead>
<tr>
<th>Exposure:</th>
<th>Mortality, %</th>
<th>Risk difference, %</th>
<th>Risk ratio</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed data</td>
<td>9.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Natural course</td>
<td>9.2</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>&lt; 5 f-y/ml</td>
<td>8.6</td>
<td>-0.6</td>
<td>0.93</td>
<td>0.88, 0.99</td>
</tr>
<tr>
<td>&lt; 2 f-y/ml</td>
<td>7.5</td>
<td>-1.7</td>
<td>0.81</td>
<td>0.71, 0.93</td>
</tr>
<tr>
<td>&lt; 0.1 f-y/ml</td>
<td>7.0</td>
<td>-2.2</td>
<td>0.76</td>
<td>0.62, 0.94</td>
</tr>
<tr>
<td>&lt; 0.05 f-y/ml</td>
<td>6.9</td>
<td>-2.3</td>
<td>0.75</td>
<td>0.61, 0.92</td>
</tr>
</tbody>
</table>
Assumptions/Limitations

- *Exchangeability*, complete set of confounders (e.g. smoking)
- *Correct models specification*, employment, competing and lung cancer deaths (e.g. empirical induction period)

- *Positivity*, exposed/unexposed at every level of confounders; non-positivity alters the questions we can ask
- *Well-defined interventions*, that versions of the treatments yield the same potential outcome (e.g. set exposures at cap)
- *No interference*, that a participant’s potential outcomes do not depend on other participant’s exposure
- *Computationally intense*, 5 min to obtain the set of point estimates (Hernán macro faster)
An Application of the G-formula to Asbestos and Lung Cancer

Stephen R. Cole

Epidemiology, UNC Chapel Hill

Slides: www.unc.edu/~colesr/
References


* Details: Step 1: Fit parametric models, using 3002 participants, for:

a) Probability of remaining at work conditional on baseline covariates, age, and past time varying asbestos exposure among those who remained at work in the prior year, as well as alive, under follow up and less than 90 years of age in the current year. Specifically, we fit the model

$$\text{logit} \ P \left( W_j = 1 \mid V_S = v_S, \bar{W}_{j-1} = 1, \bar{X}_{j-1} = \bar{x}_{j-1}, \bar{Y}_{j-1} = \bar{D}_{j-1} = \bar{C}_{j-1} = 0 \right) = \alpha_0 + \sum_{p=1}^{7} \alpha_p g(v_S) + \sum_{q=8}^{15} \alpha_q g(\bar{x}_{j-1})$$

where logit = log(.)/[1 - log(.)], $g(v_S)$ included a cubic polynomial for age, a linear term for prevalent asbestos exposure, indicators for female sex and nonwhite race as well as a linear term for year of study entry, and $g(\bar{x}_{j-1})$ included 2 indicators for any asbestos exposure in the prior 2 years as well as cubic polynomials for asbestos dose in each of those years
Step 1, continued

b) Probability of any asbestos exposure conditional on baseline covariates, age, and past time varying asbestos exposure among those who remained at work, alive, and under follow up in the current year. Specifically, we fit the model

\[
\text{logit } P \left( I[X_j > 0] = 1 \mid V_S = v_S, \bar{W}_j = 1, \bar{X}_{j-1} = \bar{x}_{j-1}, S \geq j, \bar{Y}_{j-1} = \bar{D}_{j-1} = \bar{C}_{j-1} = 0 \right) = \\
\beta_0 + \sum_{p=1}^{7} \beta_p g(v_s) + \sum_{q=8}^{15} \beta_q g(\bar{x}_{j-1}),
\]

where the indicator function \( I(.) = 1 \) if the condition is true, and \( g(.) \) are as described in Step 1a, above.
Step 1, continued

c) Density of asbestos exposure in f-y/ml conditional on baseline covariates, age, and past time varying asbestos exposure among those who were at work and exposed, alive, and under follow up in the current year. Specifically, we fit the model

$$E \left( X_j \mid V_s = v_s, \bar{W}_j = 1, X_j > 0, \bar{X}_{j-1} = \bar{x}_{j-1}, \bar{Y}_{j-1} = \bar{D}_{j-1} = \bar{C}_{j-1} = 0 \right) =$$

$$\gamma_0 + \sum_{p=1}^{7} \gamma_p g(v_s) + \sum_{q=8}^{15} \gamma_q g(\bar{x}_{j-1}),$$

where $g(.)$ are again as described in Step 1a, above; (aside: a log transformation did not fit the data better than the raw values)
Step 1, continued

d) Probability of drop out conditional on baseline covariates and age. Specifically, we fit the model
\[
\text{logit } P \left( C_j = 1 \mid V_S = v_S, \bar{X}_j = \bar{x}_j, \bar{Y}_{j-1} = \bar{D}_{j-1} = \bar{C}_{j-1} = 0 \right) = \delta_0 + \sum_{p=1}^{7} \delta_p g(v_S) + \sum_{q=8}^{10} \delta_q g(\bar{x}_{j-1}), \]
where \( g(v_S) \) was as described in Step 1a above, and \( g(\bar{x}_{j-1}) \) was a cubic polynomial of cumulative asbestos exposure.

e) Probability of a competing death conditional on baseline covariates and age. Specifically, we fit the model
\[
\text{logit } P \left( D_j = 1 \mid V_S = v_S, \bar{X}_j = \bar{x}_j, \bar{Y}_{j-1} = \bar{D}_{j-1} = \bar{C}_{j-1} = 0 \right) = \phi_0 + \sum_{p=1}^{7} \phi_p g(v_S) + \sum_{q=8}^{10} \phi_q g(\bar{x}_{j-1}), \]
where \( g(\cdot) \) were as described above.
Step 1, continued

f) Probability of lung cancer mortality conditional on baseline covariates, age, and past time varying work status and asbestos exposure among those who were alive, and under follow up in the current year. Specifically, we fit the model

$$\text{logit} \ P \left( Y_j = 1 \mid V_S = v_S, \bar{W}_j = \bar{w}_j, \bar{X}_j = \bar{x}_j, \bar{Y}_{j-1} = \bar{D}_{j-1} = \bar{C}_{j-1} = 0 \right) = $$

$$\theta_0 + \sum_{p=1}^{7} \theta_p g(v_S) + \sum_{q=8}^{10} \theta_q g(\bar{x}_j) + \sum_{r=11}^{13} \theta_r g(\bar{w}_j),$$

where $g(.)$ were as described above in Step 1, model 1), except $g(\bar{x}_j)$ was a cubic polynomial for cumulative asbestos exposure in f-y/ml, and $g(\bar{w}_j)$ was three indicators for work status over the prior three years.
Step 2

- Draw a Monte Carlo sample of 50,000 pseudo-participants sampled randomly with replacement from the original 3002 observed participants.

- Monte Carlo samples of size 10,000 through 100,000 were explored: results above 40,000 were indistinguishable.
Step 3

For each Monte Carlo sampled pseudo-participant, do the following:

a) At \( j = S_i \), set \( W_{S_i} = 1 \) and \( D_{S_i} = C_{S_i - 1} = Y_{S_i - 1} = 0 \), and assign the resampled baseline and first time varying covariates as observed.

b) At \( j > S_i \), assign the time varying covariate, drop out, competing risk and outcome values by drawing, in temporal order, from the appropriate conditional probability or density with parameters estimated in Step 1 above, but evaluated at previously simulated covariate values through age \( j \).

c) If \( \max(D_j, Y_j) = 1 \), then stop data generation for that pseudo-participant, else continue to generate data until \( j = 90 \).

- Note that for asbestos exposure scenarios we alter the drawn asbestos exposure replacing values greater than \( \{5, 2, 0.1, 0.05\} \) with that cap value.
Step 4

- Compute the cumulative incidence function (allowing for competing causes of death) on the fully generated pseudo-participant data and obtain the cumulative proportion with lung cancer mortality at 90 years of age.

- Standard errors for the cumulative proportion is calculated as the standard deviation of estimates obtained by conducting the above steps on 200 nonparametric bootstrap random samples with replacement from the original data set, each of size 3002.