Absolute risk prediction model for endometrial cancer in white women aged 50 years or older: Derivation and validation from population-based cohort studies

Ruth Pfeiffer

Advances in Endometrial Cancer Epidemiology and Biology Symposium
Boston, March 17, 2014

Biostatistics Branch
Division of Cancer Epidemiology and Genetics
National Cancer Institute, National Institutes of Health
Motivation: Absolute Risk for Cancer Incidence

\[ r(x, a, \tau) = P(T \leq a + \tau, \text{cause} = C \mid T > a; x) \]

\[ = \int_a^{a+\tau} h_C(t, x) \exp \left[ -\int_a^t \{ h_C(u, x) + h_D(u, x) \} \, du \right] \, dt \]

\( T \) - event time
\( X \) - individual risk or protective factors
\( a \) - age
\( \tau \) - length of projection
\( h_C(t, x) \) – cancer hazard at age \( t \)
\( h_D(t, x) \) - mortality hazard from competing risks
Combine Data from Different Sources to Estimate Absolute Risk

model \( h_c(t,x) = h_{c0}(t) \text{rr}(\beta_c x) \)
Combine Data from Different Sources to Estimate Absolute Risk

\[ h_c(t,x) = h_{c0}(t) \cdot rr(\beta_c x) \]

- **Cohort, nested case-control, case cohort or case-control data**
- Estimate **relative risk**, \( rr(\beta_c x) \) and **attributable risk**, \( AR(x) \)
Combine Data from Different Sources to Estimate Absolute Risk

model $h_c(t,x) = h_{c0}(t) \cdot rr(\beta_c x)$

Cohort, nested case-control, case cohort or case-control data

Estimate relative risk, $rr(\beta_c x)$ and attributable risk, $AR(x)$

SEER Cancer Registries

$h_{c0}^*(t)$, composite cancer hazard (age spec.)
Combine Data from Different Sources to Estimate Absolute Risk

\[ h_c(t, x) = h_{c0}(t)rr(\beta_c x) \]

- Cohort, nested case-control, case cohort or case-control data
- Estimate **relative risk**, \( rr(\beta_c x) \) and **attributable risk**, \( AR(x) \)

**SEER Cancer Registries**
- \( h_{c0}^*(t) \), composite cancer hazard (age spec.)

\[ h_c(t, x) = h_{c0}^*(t)(1-AR_c)rr(\beta_c x) \]
Combine Data from Different Sources to Estimate Absolute Risk

Model \( h_c(t,x) = h_{c0}(t) \cdot \text{rr}(\beta_c x) \)

Estimate relative risk, \( \text{rr}(\beta_c x) \)
and attributable risk, \( \text{AR}(x) \)

SEER Cancer Registries

\( h_{c0}^*(t) \), composite cancer hazard (age spec.)

Cohort, nested case-control, case cohort, case-control data

\[ \hat{h}_c(t,x) = h_{c0}^*(t)(1 - \text{AR}_c) \cdot \text{rr}(\beta_c x) \]

\[
\hat{r}(x,a,\tau) = \int_a^{a+\tau} \hat{h}_C(t,x) \exp \left[ -\int_a^t \left\{ \hat{h}_C(u,x) + \hat{h}_D(u,x) \right\} du \right] dt
\]
We combined data on white non-Hispanic women ages 50+ from two large cohorts (PLCO, NIH-AARP) to estimate relative risks for endometrial cancer.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Size of final analytic cohort</th>
<th>Total # cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIH-AARP</td>
<td>113,746</td>
<td>1185</td>
</tr>
<tr>
<td>PLCO</td>
<td>42,350</td>
<td>471</td>
</tr>
<tr>
<td>Total numbers</td>
<td>156,096</td>
<td>1656</td>
</tr>
</tbody>
</table>
## Relative Risk Estimates

<table>
<thead>
<tr>
<th>Factor</th>
<th>HR (95%CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Body mass index (BMI)</strong></td>
<td></td>
</tr>
<tr>
<td>(&lt;25, 25-30, 30-35, 35-40, 40+)</td>
<td>1.72 (1.65-1.80)</td>
</tr>
<tr>
<td><strong>Oral contraceptive use</strong></td>
<td></td>
</tr>
<tr>
<td>(1+years, &lt;1year)</td>
<td>1.44 (1.29-1.62)</td>
</tr>
<tr>
<td><strong>Menopausal hormone therapy (MHT) use</strong></td>
<td></td>
</tr>
<tr>
<td>(0, 1-9, 10+ years)</td>
<td>1.15 (1.05-1.26)</td>
</tr>
<tr>
<td><strong>Interaction MHT use with (BMI &lt; 25)</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.61 (1.43-1.81)</td>
</tr>
<tr>
<td><strong>Parity</strong></td>
<td></td>
</tr>
<tr>
<td>(0, 1-2, and 3+ children)</td>
<td>1.21 (1.13-1.29)</td>
</tr>
<tr>
<td><strong>Age at menopause</strong> (&lt;50 (ref), 50-54, 55+)</td>
<td></td>
</tr>
<tr>
<td>Premenopausal vs &lt;50</td>
<td>1.26 (1.17-1.35)</td>
</tr>
<tr>
<td><strong>Never vs current smokers</strong></td>
<td></td>
</tr>
<tr>
<td>Former vs current smokers</td>
<td>1.47 (1.22-1.78)</td>
</tr>
<tr>
<td></td>
<td>1.21 (1.00-1.47)</td>
</tr>
</tbody>
</table>
Recall: Combine Data from Different Sources to Estimate Absolute Risk

model $h_c(t,x) = h_{c0}(t)rr(\beta_c x)$

Cohort, nested case-control, case cohort, case-control data

SEER Cancer Registries
$h_{c0}^*(t)$, composite cancer hazard (age spec.)

Estimate relative risk, $rr(\beta_c x)$ and attributable risk, $AR(x)$

$\hat{h}_c(t,x) = h_{c0}^*(t)(1-AR_c)rr(\beta_c x)$

$\hat{r}(x,a,\tau) = \int_a^{a+\tau} \hat{h}_C(t,x) \exp \left[ -\int_a^t \left\{ \hat{h}_C(u,x) + \hat{h}_D(u,x) \right\} du \right] dt$
Computation of baseline rates

- SEER rates include women with no uterus in denominator, thus are too low
- Adjusted age-specific SEER rates by dividing them by % of women who had not had hysterectomy estimated from Behavioral Risk Factor Surveillance System (BRFSS) survey for same areas included in SEER
Recall: Combine Data from Different Sources to Estimate Absolute Risk

model \( h_c(t,x) = h_{c0}(t) \text{rr}(\beta_c x) \)

- **Cohort data**
  - Estimate **relative risk**, \( \text{rr}(\beta_c x) \)
  - Estimate **attributable risk**, \( \text{AR}(x) \)

- **SEER Cancer Registries**
  - Estimate \( h_{c0}^*(t) \), composite cancer hazard (age spec.)

- **BRFSS Survey**
  - Hysterectomy rates

\[
\hat{h}_c(t,x) = h_{c0}^*(t)(1-\text{AR}_c) \text{rr}(\beta_c x)
\]

\[
\hat{r}(x,a,\tau) = \int_a^{a+\tau} \hat{h}_C(t,x) \exp \left[ -\int_a^t \left\{ \hat{h}_C(u,x) + \hat{h}_D(u) \right\} du \right] dt
\]
Absolute endometrial cancer risk estimates: two 50 year old women

<table>
<thead>
<tr>
<th></th>
<th>Woman 1</th>
</tr>
</thead>
<tbody>
<tr>
<td># of life births</td>
<td>3</td>
</tr>
<tr>
<td>BMI, kg/m²</td>
<td>24</td>
</tr>
<tr>
<td>Menopausal</td>
<td>no</td>
</tr>
<tr>
<td>Oral contraceptive use</td>
<td>0</td>
</tr>
<tr>
<td>Smoking</td>
<td>Current smoker</td>
</tr>
<tr>
<td>HRT use+ duration</td>
<td>No (0)</td>
</tr>
</tbody>
</table>
### Absolute endometrial cancer risk estimates: two 50 year old women

<table>
<thead>
<tr>
<th></th>
<th>Woman 1</th>
</tr>
</thead>
<tbody>
<tr>
<td># of life births</td>
<td>3</td>
</tr>
<tr>
<td>BMI, kg/m²</td>
<td>24</td>
</tr>
<tr>
<td>Menopausal</td>
<td>no</td>
</tr>
<tr>
<td>Oral contraceptive use</td>
<td>0</td>
</tr>
<tr>
<td>Smoking</td>
<td>Current smoker</td>
</tr>
<tr>
<td>HRT use+ duration</td>
<td>No (0)</td>
</tr>
<tr>
<td>10 year absolute risk estimate</td>
<td>0.4%</td>
</tr>
<tr>
<td>20 year absolute risk estimate</td>
<td>1.1%</td>
</tr>
</tbody>
</table>
Absolute endometrial cancer risk estimates: two 50 year old women

<table>
<thead>
<tr>
<th></th>
<th>Woman 1</th>
<th>Woman 2</th>
</tr>
</thead>
<tbody>
<tr>
<td># of life births</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>BMI, kg/m²</td>
<td>24</td>
<td>40</td>
</tr>
<tr>
<td>Menopausal</td>
<td>no</td>
<td>yes (age 50)</td>
</tr>
<tr>
<td>Oral contraceptive use</td>
<td>0</td>
<td>&gt;1</td>
</tr>
<tr>
<td>Smoking</td>
<td>Current smoker</td>
<td>Never smoker</td>
</tr>
<tr>
<td>HRT use+ duration</td>
<td>No (0)</td>
<td>yes (1yr)</td>
</tr>
<tr>
<td>10 year absolute risk estimate</td>
<td>0.4%</td>
<td></td>
</tr>
<tr>
<td>20 year absolute risk estimate</td>
<td>1.1%</td>
<td></td>
</tr>
</tbody>
</table>
Absolute endometrial cancer risk estimates: two 50 year old women

<table>
<thead>
<tr>
<th></th>
<th>Woman 1</th>
<th>Woman 2</th>
</tr>
</thead>
<tbody>
<tr>
<td># of life births</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>BMI, kg/m²</td>
<td>24</td>
<td>40</td>
</tr>
<tr>
<td>Menopausal</td>
<td>no</td>
<td>yes (age 50)</td>
</tr>
<tr>
<td>Oral contraceptive use</td>
<td>0</td>
<td>&gt;1</td>
</tr>
<tr>
<td>Smoking</td>
<td>Current smoker</td>
<td>Never smoker</td>
</tr>
<tr>
<td>HRT use+ duration</td>
<td>No (0)</td>
<td>yes (1yr)</td>
</tr>
<tr>
<td>10 year absolute risk estimate</td>
<td>0.4%</td>
<td>5.9%</td>
</tr>
<tr>
<td>20 year absolute risk estimate</td>
<td>1.1%</td>
<td>15.1%</td>
</tr>
</tbody>
</table>
Absolute endometrial cancer risk estimates: two 50 year old women

<table>
<thead>
<tr>
<th></th>
<th>Woman 1</th>
<th>Woman 2</th>
</tr>
</thead>
<tbody>
<tr>
<td># of life births</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>BMI, kg/m²</td>
<td>24</td>
<td>40</td>
</tr>
<tr>
<td>Menopausal</td>
<td>no</td>
<td>yes (age 50)</td>
</tr>
<tr>
<td>Oral contraceptive use</td>
<td>0</td>
<td>&gt;1</td>
</tr>
<tr>
<td>Smoking</td>
<td>Current smoker</td>
<td>Never smoker</td>
</tr>
<tr>
<td>HRT use+ duration</td>
<td>No (0)</td>
<td>yes (1yr)</td>
</tr>
<tr>
<td>10 year absolute risk estimate</td>
<td>0.4%</td>
<td>5.9%</td>
</tr>
<tr>
<td>20 year absolute risk estimate</td>
<td>1.1%</td>
<td>15.1%</td>
</tr>
<tr>
<td>10 year absolute breast cancer risk*</td>
<td>1.8%</td>
<td>6.8%</td>
</tr>
<tr>
<td>20 year absolute breast cancer risk*</td>
<td>4.2%</td>
<td>15.3%</td>
</tr>
</tbody>
</table>

* For specific choices of BC risk factors
External Validation: Independent Population for Assessment of Model Performance

Assume population of $N$ individuals followed over time period $t$ (cohort data)
Observe disease outcome at end of follow-up
$Y_i = \begin{cases} 
1, & \text{if ith woman develops event during } t \\
0, & \text{otherwise} 
\end{cases}$
External Validation: Independent Population for Assessment of Model Performance

Assume population of N individuals followed over time period $t$ (cohort data)
Observe disease outcome at end of follow-up

$$Y_i = \begin{cases} 
1, & \text{if $i$th woman develops event during $t$} \\
0, & \text{otherwise} 
\end{cases}$$

$$r(x_i) = \hat{P}(Y = 1| x_i)$$ absolute risk estimate for $i$th person with baseline covariates $x_i$ including age $a$ over time $t$

Risk model $r$ known

Risk estimates $r$ have distribution $F$ on $[0, 1]$
Distribution $F$ of Risk in Nurses’ Health Validation Cohort (N=37,241)

Endometrial cancer absolute risk estimates
Assesses model bias

Model $r(x)$ is **well calibrated** if for each $x$

$$P(Y = 1 | r(x) = r) \approx r$$

Model $r$ is unbiased (well calibrated) in population if

$$\frac{1}{N} \sum_{i=1}^{N} Y_i \approx \frac{1}{N} \sum_{i=1}^{N} r_i$$
Assess Performance of Models in Nurses’ Health Validation Cohort

Ages 51-70 at baseline  (N=37,241)

Calibration: \[ O = \sum_{i=1}^{N} Y_i, \quad E = \sum_{i=1}^{N} r(X_i) \]

<table>
<thead>
<tr>
<th>Observed # cases</th>
<th>Expected from model</th>
<th>E/O (95%CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>O 532</td>
<td>E 637</td>
<td>1.20 (1.11–1.29)</td>
</tr>
</tbody>
</table>
## Validation of RRs for Endometrial Model in NHS cohort

<table>
<thead>
<tr>
<th></th>
<th>AARP/PLCO HR (95%CI)</th>
<th>NHS HR (95%CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BMI</strong></td>
<td>1.72 (1.65-1.80)</td>
<td>1.78 (1.64-1.93)</td>
</tr>
<tr>
<td>Oral contraceptive use</td>
<td>1.44 (1.29-1.62)</td>
<td>1.30 (1.07-1.57)</td>
</tr>
<tr>
<td><strong>MHT use</strong></td>
<td>1.15 (1.05-1.26)</td>
<td>2.43 (1.97-2.99)</td>
</tr>
<tr>
<td>Interaction MHT use with (BMI &lt; 25)</td>
<td>1.61 (1.43-1.81)</td>
<td>1.41 (1.10-1.80)</td>
</tr>
<tr>
<td><strong>Parity</strong></td>
<td>1.21 (1.13-1.29)</td>
<td>1.26 (1.11-1.44)</td>
</tr>
<tr>
<td><strong>Age at menopause</strong></td>
<td>1.26 (1.17-1.35)</td>
<td>1.30 (1.12-1.51)</td>
</tr>
<tr>
<td>Premenopausal vs &lt;50</td>
<td>1.29 (1.01-1.63)</td>
<td>2.16 (1.57-2.97)</td>
</tr>
<tr>
<td><strong>Never vs current smokers</strong></td>
<td>1.47 (1.22-1.78)</td>
<td>1.82 (1.36-2.44)</td>
</tr>
<tr>
<td>Former current smokers</td>
<td>1.21 (1.00-1.47)</td>
<td>1.30 (0.96-1.76)</td>
</tr>
</tbody>
</table>
Age specific incidence of Corpus Uteri and Uterus, NOS, per 100,000 person years in white women from SEER and the NHS cohort

<table>
<thead>
<tr>
<th>Age</th>
<th>SEER</th>
<th>SEER corrected for hysterectomy</th>
<th>NHS women who had uterus during follow-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>50-54</td>
<td>50.5</td>
<td>76.24</td>
<td>40.7</td>
</tr>
<tr>
<td>55-59</td>
<td>77.5</td>
<td>127.55</td>
<td>42.8</td>
</tr>
<tr>
<td>60-64</td>
<td>97.44</td>
<td>174.72</td>
<td>40.2</td>
</tr>
<tr>
<td>65-69</td>
<td>106.5</td>
<td>193.67</td>
<td>45.0</td>
</tr>
<tr>
<td>70-74</td>
<td>109.02</td>
<td>199.96</td>
<td>49.7</td>
</tr>
</tbody>
</table>
Evaluating Model Performance: Discrimination

Area under the receiver operating characteristic (ROC) curve (AUC)

\[ AUC = P(r_{Y=1} > r_{Y=0}) \]
Distribution of absolute risk estimates in NHS by endometrial cancer status
Distribution of absolute risk estimates in NHS by endometrial cancer status

AUC=0.68
Criteria that Assess Model Performance for Screening/Follow-up Applications

Wish to make screening/follow-up recommendations for population over next five years based on baseline risk assessment from absolute risk model

Compute 5-year cancer risk from model for every woman given baseline covariates $X$, $r_i = r(X_i)$, $i=1,…,N$

Rank risks from lowest to highest risk: $r_{(1)} \leq r_{(2)} \leq \ldots \leq r_{(N)}$
Criteria that Assess Model Performance for Screening/Follow-up Applications

Wish to make screening/follow-up recommendations for population over next five years based on baseline risk assessment from absolute risk model

Compute 5-year cancer risk from model for every woman given baseline covariates $X$, $r_i = r(X_i)$, $i=1,…N$

Rank risks from lowest to highest risk: $r_{(1)} \leq r_{(2)} \leq \ldots \leq r_{(N)}$

1. Proportion of cases followed, PCF($q$): proportion of cases followed-up in program that screens proportion $q$ of population at highest risk
Criteria that Assess Model Performance for Screening/Follow-up Applications

Wish to make screening/follow-up recommendations for population over next five years based on baseline risk assessment from absolute risk model

Compute 5-year cancer risk from model for every woman given baseline covariates $X, r_i = r(X_i), i=1,…,N$

Rank risks from lowest to highest risk: $r^{(1)} \leq r^{(2)} \leq \ldots \leq r^{(N)}$

1. **Proportion of cases followed, PCF(q):** proportion of cases followed-up in program that screens proportion $q$ of population at highest risk

2. **Proportion needed to follow, PNF(p):** proportion of population at highest risk that needs to be followed so that proportion $p$ of future cases will be followed
PCF and PNF Estimates for Endometrial Models in NHS Cohort

<table>
<thead>
<tr>
<th>$q$</th>
<th>Estimated proportion of cases followed (PCF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.30</td>
</tr>
</tbody>
</table>
### PCF and PNF Estimates for Endometrial Models in NHS Cohort

<table>
<thead>
<tr>
<th></th>
<th>Endometrial cancer model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q$</td>
<td>Estimated proportion of cases followed (PCF)</td>
</tr>
<tr>
<td>0.10</td>
<td>0.30</td>
</tr>
<tr>
<td>$p$</td>
<td>Estimated proportion needed to follow (PNF)</td>
</tr>
<tr>
<td>0.90</td>
<td>0.75</td>
</tr>
</tbody>
</table>
Summary

• Developed novel model that predicts absolute risk for endometrial cancer using cohort data
• Validated model in independent cohort
• Model overestimated number of cancers by 20% in NHS validation cohort
• Discriminatory accuracy adequate to use model for risk stratification
• Further validation needed to assess calibration
  – Ongoing: validation in WHI cohort
• Extended models for African American women
Reference

Pfeiffer RM, Park Y, Kreimer AR et al, **Risk predicting for breast, endometrial or ovarian cancer in white women aged 50 years or older: Derivation and validation from population-based cohort studies**, Plos Medicine, 2013
Collaborators

Yikyung Park
Aimée Kreimer
Mitchell Gail
Patricia Hartge
James Lacey
David Pee

Robert Greenlee
Saundra Buys
Sue Hankinson
Albert Hollenbeck
Bernard Rosner
### NHS Endometrial Cancer model

<table>
<thead>
<tr>
<th>BMI</th>
<th>O</th>
<th>E</th>
<th>E/O</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;25</td>
<td>190</td>
<td>235</td>
<td>1.24</td>
</tr>
<tr>
<td>25-&lt;30</td>
<td>160</td>
<td>185</td>
<td>1.16</td>
</tr>
<tr>
<td>30-&lt;35</td>
<td>96</td>
<td>117</td>
<td>1.22</td>
</tr>
<tr>
<td>35-&lt;40</td>
<td>42</td>
<td>57</td>
<td>1.37</td>
</tr>
<tr>
<td>40+</td>
<td>44</td>
<td>42</td>
<td>0.95 *</td>
</tr>
</tbody>
</table>

### Smoking

<table>
<thead>
<tr>
<th></th>
<th>O</th>
<th>E</th>
<th>E/O</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never</td>
<td>287</td>
<td>318</td>
<td>1.11*</td>
</tr>
<tr>
<td>Former</td>
<td>191</td>
<td>243</td>
<td>1.27</td>
</tr>
<tr>
<td>Current</td>
<td>54</td>
<td>76</td>
<td>1.40</td>
</tr>
</tbody>
</table>

### MHT use and duration

<table>
<thead>
<tr>
<th></th>
<th>O</th>
<th>E</th>
<th>E/O</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>304</td>
<td>388</td>
<td>1.28</td>
</tr>
<tr>
<td>1-9 yrs</td>
<td>186</td>
<td>228</td>
<td>1.23</td>
</tr>
<tr>
<td>10+ yrs</td>
<td>42</td>
<td>21</td>
<td>0.49</td>
</tr>
</tbody>
</table>

* no significant evidence of lack of fit
Distributions of Risk in Cases and Non-cases

\[ r(x) = \hat{P}(Y = 1|x) \] risk estimate

\[ F(r^*) = P(r \leq r^*) \] distribution of risk in general population
Evaluating Performance of Risk Prediction Models: Calibration

Assesses model bias

Model \( r(x) \) is **well calibrated** if for each \( x \)

\[
P(Y = 1 | r(x) = r) \approx r
\]

Then \( \mu = E(Y) = P(Y = 1) = \int_{0}^{1} r dF(r) = E(R) \)

Model \( r \) is unbiased (well calibrated) in population if

\[
\frac{1}{N} \sum_{i=1}^{N} Y_i \approx \frac{1}{N} \sum_{i=1}^{N} r_i
\]
Distributions of Risk in Cases and Non-cases

\[ r(x) = \hat{P}(Y = 1 | x) \] risk estimate

\[ F(r^*) = P(r \leq r^*) \] distribution of risk in general population

\[ G(r^*) = P(r \leq r^* | Y = 1) \] distribution of risk in cases
Distributions of Risk in Cases and Non-cases

\[ r(x) = \hat{P}(Y = 1|x) \] risk estimate

\[ F(r^*) = P(r \leq r^*) \] distribution of risk in general population

\[ G(r^*) = P(r \leq r^*|Y = 1) \], distribution of risk in cases

\[ K(r^*) = P(r \leq r^*|Y = 0) \], distribution of risk in non-cases
Proportion Cases Followed, 
PCF(q)

Proportion of individuals who will develop disease who are included in proportion $q$ of population at highest risk.

$$\xi_{1-q} = F^{-1}(1-q) : (1-q)\text{th quantile of population distribution } F$$

$G$- distribution of risk in cases

$$\text{PCF}(q) = 1 - G(\xi_{1-q}) = 1 - G \circ F^{-1}(1-q)$$
Comparing two risk models

Two risk models $r^1$, $r^2$ evaluated on same population: bivariate risk estimates $(r^{i1}, r^{i2})$, $i=1,\ldots,N$

Model that for a given value of $q$ has larger PCF($q$), iPCF, or, for given value of $p$, smaller PNF($p$), iPNF, better separates distributions of risk in diseased and non-diseased individuals.
Comparing two risk models

Test $H_0$: PCF$^1(q)$=PCF$^2(q)$ or $H_0$: PNF$^1(p)$=PNF$^2(p)$

$H_0$: iPCF$^1$=iPCF$^2$ or $H_0$: iPNF$^1$=iPNF$^2$

$$T_{PCF} = \frac{n(\hat{PCF}_1 - \hat{PCF}_2)^2}{\hat{V}_{PCF}} \sim \chi^2$$

$$T_{iPCF} = \frac{n(i\hat{PCF}_1 - i\hat{PCF}_2)^2}{\hat{V}_{iPCF}} \sim \chi^2$$

$$T_{PNF} = \frac{n(\hat{PNF}_1 - \hat{PNF}_2)^2}{\hat{V}_{PNF}} \sim \chi^2$$

$$T_{iPNF} = \frac{n(i\hat{PNF}_1 - i\hat{PNF}_2)^2}{\hat{V}_{iPNF}} \sim \chi^2$$

Estimate V using influence functions or bootstrap

Compare new breast cancer model to BCRAT: $T_{PCF}$ P-value=0.02
PCF, risk distribution $F$ in population is $\text{Beta}(8.5, 161.5)$ (AUC=0.59)

$\text{PCF}(0.1) = 0.17$
PCF, risk distribution $F$ in population is Beta(2.3, 43.7) (AUC=0.68)
PCF, risk distribution $F$ in population is Beta(1, 19) (AUC=0.76)

$PCF(0.1) = 0.32$
**Integrated Proportion of Cases Followed (iPCF)**

\[ \xi_{1-q} = F^{-1}(1-q) : (1-q)\text{th quantile of population distribution } F \]

G- distribution of risk in cases

\[ \text{PCF}(q)=1-G(\xi_{1-q}) = 1- G \circ F^{-1}(1-q) \]

\[ \text{iPCF}(q^*)= \int_{q^*}^{1} \text{PCF}(q) dq = 1-q^* - \int_{0}^{\xi_{1-q^*}} G(u) dF(u) \]

\[ = 1-q^* - \frac{1}{1-q^*} P(R_G \leq R_F \mid R_F \in (0, \xi_{1-q^*})) \]

\[ q^* = 0: \text{iPCF}(0)=P(R_G > R_F) \]
Comparison with ROC curve

F- distribution of risk in population

G- distribution of risk in cases

\[ \text{PCF}(q) = 1 - G \circ F^{-1}(1-q) \]

\[ \text{iPCF}(0) = \int_0^1 \text{PCF}(q) \, dq = P(R_G > R_F) \]

K- distribution of risk in non-cases

\[ \text{ROC}(q) = 1 - G \circ K^{-1}(1-q) \]

\[ \text{AUC} = \int_0^1 \text{ROC}(q) \, dq = P(R_G > R_K) \]
ROC and PCF curves, risk distribution $F$ in population is Beta(8.5, 34), AUC=0.60, iPCF=0.58 (mu=0.20)
ROC and PCF curves, risk distribution $F$ in population is Beta(2.3, 9.2), AUC=0.70, iPCF=0.66 (mu=0.20)
Proportion Needed to Follow, PNF(p)

Fraction of the general population with highest risks that needs to be screened (followed up) to assure that a given fraction $p$ of all cases in population receive screen.

Solve

$$1 - G \circ F^{-1}(1 - PNF(p)) = p$$

Thus

$$PNF(p) = 1 - F \circ G^{-1}(1 - p)$$
Integrated Proportion Needed to Follow (iPNF)

Letting $\gamma_{1-p} = G^{-1}(1 - p)$

$$iPNF(p^*) = \int_0^{p^*} PNF(p)dp = 1 - \int_0^{\gamma_{1-p^*}} F(u)dG(u)$$

$$= 1 - p^* - \frac{1}{1-p^*} P(R_F \leq R_G \mid R_G \in (0, \gamma_{1-p^*}))$$

$p^* = 0$: $iPNF(0) = P(R_F > R_G)$
PNF, risk distribution $F$ in population is Beta(8.5, 161.5) (AUC=0.59)
PNF, risk distribution $F$ in population is Beta(2.3, 43.7) (AUC=0.68)
PNF, risk distribution $F$ in population is Beta(1, 49) (AUC=0.76)
Non-parametric estimates of PCF(q), iPBF, PNF(p), iPBF using three different types of data

- Risks $r_1, \ldots, r_N$ in a cohort under assumption of a well-calibrated model

- Risks in case-control study, $r_1, \ldots, r_{N_0}$ in controls ($Y=0$) and $r_1, \ldots, r_{N_1}$ in cases ($Y=1$) with known disease prevalence $\mu$

- Risks and outcomes in a cohort, $(r_1, Y_1), \ldots, (r_N, Y_N)$
Estimate PCF and PNF under assumption of well calibrated model

If model well calibrated: \( P(Y \mid r) \sim r \) and \( P(Y = 1) = E(R) = \mu \)

Using \( P(R \leq r^*, Y = 1) = \int_0^{r^*} rdF(r) \), distribution of risk in cases \( G \) is

\[
G(r^*) = P(R \leq r^* \mid Y = 1) = \frac{1}{\mu} \int_0^{r^*} rdF(r)
\]
Estimate PCF and PNF under assumption of well calibrated model

If model well calibrated: \( P(Y \mid r) \sim r \) and \( P(Y = 1) = E(R) = \mu \)

Using \( P(R \leq r^*, Y = 1) = \int_{0}^{r^*} rdF(r) \), distribution of risk in cases \( G \) is

\[
G(r^*) = P(R \leq r^* \mid Y = 1) = \frac{1}{\mu} \int_{0}^{r^*} rdF(r)
\]

\( \xi_{1-q} = F^{-1}(1-q) : (1-q)th \) quantile of population distribution \( F \)

\[
PCF(q) = 1 - G(\xi_{1-q}) = 1 - \frac{1}{\mu} \int_{0}^{\xi_{1-q}} rdF(r) = 1 - L(1-q)
\]

where \( L \) denotes the Lorenz curve for \( F \)

\[
PNF(p) = 1 - F \circ G^{-1}(1-p) = 1 - L^{-1}(1-p)
\]
Model well calibrated: estimate PCF(q) and PNF(p) using only observed risks

\[ r_{(1)} \leq \ldots \leq r_{(n)} \]

\[ S_i = \sum_{k=1}^{i} r_{(k)} \]

\[ L_n(p) = \frac{S_{[np]}}{S_n} \]

\[ L_n^{-1}(p) = \frac{i}{n}, \quad S_i/S_n < p \leq S_{i+1}/S_n, i = 0, \ldots, n \]

\[ \hat{PCF} = 1 - L_n(1 - q) \]

\[ \hat{PNF} = 1 - L_n^{-1}(1 - p) \]

Goldie, 1977
Estimate PCF and PNF from risks in case-control study with known disease prevalence $\mu$

Observe $r$ in sample of $N_1$ cases and $N_0$ controls

$$
\hat{K}(r^*) = \frac{1}{N_0} \sum_{k=1}^{N_0} I(r_k \leq r^*, Y_k = 0)
$$

$$
\hat{G}(r^*) = \frac{1}{N_1} \sum_{k=1}^{N_1} I(r_k \leq r^*, Y_k = 1)
$$

$$
\hat{F}^R(r^*) = \mu \hat{G}(r^*) + (1 - \mu) \hat{K}(r^*)
$$

$$
PCF = 1 - \hat{G} \circ (\hat{F}^R)^{-1}(1 - q)
$$

$$
PNF = 1 - \hat{F}^R \circ \hat{G}^{-1}(1 - p)
$$
Estimate PCF(q) and PNF(p) using observed risks and outcomes in cohort

Observe \((r_i, Y_i), i = 1, .. N\)

\[
N_1 = \sum_{k=1}^{N} Y_k
\]

\[
\hat{F}^R(r^*) = \frac{1}{N} \sum_{k=1}^{N} I(r_k \leq r^*)
\]

\[
\hat{G}(r^*) = \frac{1}{N_1} \sum_{k=1}^{N} I(r_k \leq r^*, Y_k = 1)
\]

\[
P\hat{C}F = 1 - \hat{G} \circ (\hat{F}^R)^{-1} (1 - q)
\]

\[
P\hat{N}F = 1 - \hat{F}^R \circ \hat{G}^{-1} (1 - p)
\]
Inference

• Derived asymptotic distributions for non-parametric estimates of PCF, iPCF, PNF and iPNF for three designs
  – Observed risks alone under assumption of well calibrated model
  – Case-control data with known disease prevalence
  – Observed risks and outcomes in a population

• Variance estimation:
  – Taylor linearization using influence functions
  – Bootstrap procedure
  – Estimation under assumption of well calibrated model much more efficient
PCF, risk distribution $F$ in population is $\text{Beta}(2.3, 43.7)$, $\text{AUC}=0.68$, $\text{PCF}(0.1)=0.24$
Overview

• Motivation: absolute risk models for endometrial and breast cancer

• Evaluating performance of risk prediction models
  – Probabilistic framework & notation
  – Standard criteria
    • Calibration
    • Discrimination
  – Extension of criteria for screening applications
  – Estimating novel criteria from various designs

• Summary
Evaluating Performance of Risk Prediction Models (Validation)

Internal validation: *reusing same dataset* on which model was developed to assess overfitting

External validation: evaluation of model performance in sample *independent* of that used to develop the model
Model well calibrated: estimate iPCF and iPNF using only observed risks

\[ r_{(1)} \leq \ldots \leq r_{(N)} \]

\[ S_i = \sum_{k=1}^{i} r_{(k)} \]

\[ i\hat{PCF}(p^*) = 1 - p^* - \frac{1}{NS_N} \sum_{k=1}^{(1-p^*)N} \left( [(1 - p^*)N] - k + 1 \right) r_{(k)} \]

\[ i\hat{PNF}(q^*) = 1 - q^* - \frac{1}{NS_N} \sum_{i=1}^{k^*} ir(i+1), \quad S_{k^*}/S_N < q^* \leq S_{k^*+1}/S_N \]
Estimate iPCF and iPNF from risks in case-control study with known disease prevalence $\mu$

$$K(r^*) = \frac{1}{N_0} \sum_{k=1}^{N_0} I(r_k \leq r^*, Y_k = 0)$$

$$\hat{G}(r^*) = \frac{1}{N_1} \sum_{k=1}^{N_1} I(r_k \leq r^*, Y_k = 1)$$

$$\hat{F}(r^*) = \mu \hat{G}(r^*) + (1 - \mu) \hat{K}(r^*)$$

$$i\hat{PCF}(p^*) = (1 - p^*) - \frac{\mu}{2} \hat{G}^2 \circ F^{-1}(1 - p^*) - \frac{1 - \mu}{N_1 N_0} \sum_{i,j} I(\tilde{r}_i^G \leq \tilde{r}_j^K, \tilde{r}_j^K \in (0, F^{-1}(1 - q^*))$$

$$i\hat{PNF}(q^*) = (1 - q^*) - \frac{\mu}{2} (1 - q^*)^2 - \frac{1 - \mu}{N_1 N_0} \sum_{i,j} I(\tilde{r}_i^G > \tilde{r}_j^K, \tilde{r}_i^G \in (0, G^{-1}(1 - q^*))$$
Estimate iPCF and iPNF using observed risks and outcomes in cohort

Observe \((r_i, Y_i), i = 1, ..N\)

\[ N_1 = \sum_{k=1}^{N} Y_k \]

\[ \hat{F}^R(r^*) = \frac{1}{N} \sum_{k=1}^{N} I(r_k \leq r^*) \]

\[ \hat{G}(r^*) = \frac{1}{N_1} \sum_{k=1}^{N} I(r_k \leq r^*, Y_k = 1) \]

\[ iP\hat{CF}(p^*) = 1 - p^* - \frac{1}{N_1 N} \sum_{i,j} I(r_i^G \leq r_j^F, r_j^F \in (0, F^{-1}(1 - p^*))) \]

\[ iP\hat{NF}(q^*) = 1 - q^* - \frac{1}{N_1 N} \sum_{i,j} I(r_i^G > r_j^F, r_i^G \in (0, G^{-1}(1 - q^*))) \]
PCF, risk distribution $F$ in population is Beta(2.3, 43.7) (AUC=0.68)

PCF(0.1)=0.24