The impact of outcome misclassification on estimation of prediction accuracy

Benjamin French, PhD
Department of Biostatistics and Epidemiology
University of Pennsylvania
bcfrench@upenn.edu

WNAR 2014
17 June 2014
Chronic heart failure

- Progressive disorder in which structural damage to the heart impairs its ability to provide adequate blood flow to the body
- One of the most common causes for admission and readmission
- One million hospitalizations and \(~60,000\) deaths per year in US
- Large heterogeneity in risk of adverse outcomes among patients, even after accounting for key factors known to impact risk

**Prognostic models:** Inform predictions of adverse outcomes
  - Combine novel markers with other relevant information
  - Quantify risk heterogeneity among all patients
  - Inform treatment strategies for individual patients

- Under the Affordable Care Act, hospitals may accrue financial penalties if patients with heart failure are readmitted before 30 days
- Models for 30-day readmission have limitations and perform poorly
• Data from **electronic health records** (EHRs) facilitate the development and evaluation of prognostic models
  ▶ Our focus: Readmission to the hospital among heart failure patients
  ▶ Potentially large and diverse patient populations
  ▶ Longitudinal data on potential risk factors
  ▶ Information regarding possibly recurrent events (hospitalization)
  ▶ Information regarding competing terminal events (death)

• **Outcome misclassification** may arise if events are not captured
  ▶ Patients may be readmitted to a different health system
  ▶ Deaths may not occur in the hospital

★ How does misclassification affect prediction accuracy summaries?
★ How to adjust for potential outcome misclassification in EHR data?
ROC analysis

- Modern methods for discrimination based on fundamental concepts of sensitivity and specificity for a binary disease outcome $D$
  - Sensitivity: Proportion of diseased individuals correctly classified
  - Specificity: Proportion of non-diseased individuals correctly classified

<table>
<thead>
<tr>
<th></th>
<th>$D = 1$</th>
<th>$D = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{D} = 1$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{D} = 0$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- ROC curve: Plot of the sensitivity versus $1 -$ specificity across all possible dichotomizations of a continuous prognostic marker

- AUC measures prediction accuracy of continuous prognostic marker
  - AUC = 0.5 $\Leftrightarrow$ uninformative prognostic marker
  - AUC = 1.0 $\Leftrightarrow$ perfect prognostic marker

- Extensions to time-dependent outcomes [Heagerty et al., 2000]
Risk reclassification methods

- ROC criticized for relative insensitivity to detect clinically important risk differences and for lack of direct clinical relevance

- **Risk reclassification**: Focus on differences in predicted absolute risk between ‘nested’ models [Cook, 2007; Cook and Ridker, 2009]

- Degree to which model of interest (alternative, $A$) more accurately classifies ‘cases’ as higher risk and ‘controls’ as lower risk relative to a comparison model (null, $N$)

- **Integrated discrimination improvement (IDI)** [Pencina et al., 2008]

$$\hat{\text{IDI}}(N, A) = \left( \hat{p}_{A, D=1} - \hat{p}_{N, D=1} \right) - \left( \hat{p}_{A, D=0} - \hat{p}_{N, D=0} \right)$$

- Improvement among cases

- Improvement among controls

- Extensions to time-dependent outcomes [Pencina et al., 2011]

B French (bcfrench@upenn.edu)
Properties of methods

• [Pepe, 2011]
  ▶ Reclassification summaries may mask contributions of components
  ▶ Reclassification does not necessarily imply improved performance

• [Kerr et al., 2011]
  ▶ IDI has a Normal distribution only for large values of the IDI
  ▶ IDI test statistic may not have a standard Normal distribution under the null hypothesis that IDI = 0

• [French et al., 2012]
  ▶ ROC curves and risk reclassification have similar power to detect improvements in prediction accuracy

★ How does misclassification affect prediction accuracy summaries?
Simulation study

Outcome misclassification: Events are potentially not captured in EHR

- Markers $X$ (old) and $Z$ (new), bivariate Normal, correlation = 0.3
  \[
  P[D_i = 1] = \expit(-1.4 + 1.0X_i + 1.4Z_i)
  \] (1)
  with an event rate of 30%
- Misclassification independent of marker values
  \[
  P[M_i = 1] = \begin{cases} 
  0.05, 0.1, 0.2, 0.4 & \text{if } D_i = 1 \\
  0 & \text{if } D_i = 0 
  \end{cases}
  \] (2)
- Misclassification dependent on marker values
  \[
  P[M_i = 1] = \begin{cases} 
  \expit(\gamma_0 + \gamma_1X_i + \gamma_2Z_i) & \text{if } D_i = 1 \\
  0 & \text{if } D_i = 0 
  \end{cases}
  \] (3)
  with $\gamma_0$ selected for misclassification rates of 5–40%
- Estimate improvement in discrimination from adding $Z$ to $X
Simulation results

Misclassification independent of marker values

\[ \Delta \text{AUC} \]

\[ n = 200 \]
\[ n = 500 \]
\[ n = 1000 \]

\[ \text{IDI} \]

\[ n = 200 \]
\[ n = 500 \]
\[ n = 1000 \]

Misclassification rate, %
Percent bias
5 10 20 40
−20
−15
−10
−5
0

B French (bcfrench@upenn.edu)
Simulation results

Misclassification dependent on marker values

\( D_i = 1 \): \( P[M_i = 1] = \text{expit}(\gamma_0 + \gamma_1 X_i + \gamma_2 Z_i) \)

<table>
<thead>
<tr>
<th>Rate</th>
<th>( \Delta \text{AUC} )</th>
<th>IDI</th>
<th>( \Delta \text{AUC} )</th>
<th>IDI</th>
<th>( \Delta \text{AUC} )</th>
<th>IDI</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>4.2</td>
<td>-0.2</td>
<td>-7.2</td>
<td>-6.3</td>
<td>-0.7</td>
<td>-5.1</td>
</tr>
<tr>
<td>10%</td>
<td>5.1</td>
<td>-2.4</td>
<td>-18.8</td>
<td>-16.5</td>
<td>-7.0</td>
<td>-14.1</td>
</tr>
<tr>
<td>20%</td>
<td>15.4</td>
<td>-1.3</td>
<td>-29.5</td>
<td>-27.3</td>
<td>-14.3</td>
<td>-27.4</td>
</tr>
<tr>
<td>40%</td>
<td>30.3</td>
<td>-3.4</td>
<td>-60.0</td>
<td>-49.9</td>
<td>-28.5</td>
<td>-49.3</td>
</tr>
</tbody>
</table>

\( n = 500 \)
Simulation results: $\{\eta_1, \eta_2\} = \{1, 1\}$

$$\Delta\text{AUC} \mid \text{True} = 0.112 = 0.869 - 0.757$$

$$\Delta\text{AUC} \mid \text{Miss} = 0.096 = 0.790 - 0.694$$

$$\text{IDI} \mid \text{True} = 0.200 = (0.566 - 0.427) - (0.188 - 0.249)$$

$$\text{IDI} \mid \text{Miss} = 0.145 = (0.508 - 0.397) - (0.238 - 0.272)$$
Simulation results

Misclassification dependent on marker values

\[ D_i = 1: \ P[M_i = 1] = \expit(\gamma_0 + \gamma_1 X_i + \gamma_2 Z_i) \]

Percent bias

<table>
<thead>
<tr>
<th>Rate</th>
<th>(\Delta AUC)</th>
<th>IDI</th>
<th>(\Delta AUC)</th>
<th>IDI</th>
<th>(\Delta AUC)</th>
<th>IDI</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>-7.4</td>
<td>-2.3</td>
<td>6.8</td>
<td>3.8</td>
<td>4.6</td>
<td>4.7</td>
</tr>
<tr>
<td>10%</td>
<td>-19.0</td>
<td>-6.6</td>
<td>9.8</td>
<td>6.2</td>
<td>4.6</td>
<td>7.5</td>
</tr>
<tr>
<td>20%</td>
<td>-32.1</td>
<td>-12.2</td>
<td>18.0</td>
<td>10.4</td>
<td>8.0</td>
<td>14.7</td>
</tr>
<tr>
<td>40%</td>
<td>-59.0</td>
<td>-22.1</td>
<td>32.4</td>
<td>20.5</td>
<td>8.9</td>
<td>24.9</td>
</tr>
</tbody>
</table>

\(n = 500\)
Simulation results: \( \{\eta_1, \eta_2\} = \{-1, -1\} \)

\[
\begin{align*}
\Delta \text{AUC} \mid \text{True} &= 0.112 = 0.869 - 0.757 \\
\Delta \text{AUC} \mid \text{Miss} &= 0.121 = 0.902 - 0.781
\end{align*}
\]

\[
\begin{align*}
\text{IDI} \mid \text{True} &= 0.200 = (0.566 - 0.427) - (0.188 - 0.249) \\
\text{IDI} \mid \text{Miss} &= 0.223 = (0.626 - 0.452) - (0.199 - 0.254)
\end{align*}
\]
Simulation summary

Outcome misclassification arising from ‘missed’ events

- Independent of marker values: Bias toward the null
- Dependent on marker values: Bias toward null or alternative
  - Direction of bias depended on direction of marker associations
  - Amount of bias depended on whether ‘new’ and/or ‘old’ marker was associated with misclassification
Readmission in heart failure

- 4548 patients admitted with a primary diagnosis of heart failure to the University of Pennsylvania Health System (UPHS), 2005–2012, alive at discharge, excluding those discharged to hospice

- Collected data on sociodemographic factors, secondary diagnoses, and number of admissions in the previous year [Baillie et al., 2013]
  - Null: Sociodemographic factors + secondary diagnoses
  - Alternative: Null + number of admissions in the previous year

- Built and validated models for 30-day readmission
  - Readmission to UPHS: Collected from UPHS EHR
  - Readmissions elsewhere in Pennsylvania: Collected from Pennsylvania Health Care Cost Containment Council (PHC4)

- UPHS outcomes are a misclassified version of PHC4 outcomes; readmissions elsewhere in Pennsylvania are not captured by UPHS
### Results

<table>
<thead>
<tr>
<th></th>
<th>Not readmitted</th>
<th>Readmitted</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n = 3405$</td>
<td>$n = 810$</td>
<td>$n = 333$</td>
<td>$P$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age, years</td>
<td>69 (56, 80)</td>
<td>68 (55, 80)</td>
<td>65 (51, 76)</td>
<td>0.003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1529 (45)</td>
<td>417 (51)</td>
<td>190 (57)</td>
<td>0.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td>0.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>2299 (68)</td>
<td>530 (65)</td>
<td>226 (68)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>1037 (30)</td>
<td>260 (32)</td>
<td>99 (30)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>69 (2)</td>
<td>20 (2)</td>
<td>8 (2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insurance</td>
<td></td>
<td></td>
<td></td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicare</td>
<td>2281 (67)</td>
<td>543 (67)</td>
<td>195 (59)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicaid</td>
<td>634 (19)</td>
<td>163 (20)</td>
<td>97 (29)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Private</td>
<td>460 (14)</td>
<td>102 (13)</td>
<td>37 (11)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uninsured</td>
<td>30 (1)</td>
<td>2 (&lt;1)</td>
<td>4 (1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discharging hospital</td>
<td></td>
<td></td>
<td></td>
<td>0.003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>924 (27)</td>
<td>237 (29)</td>
<td>70 (21)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Presbyterian</td>
<td>1228 (36)</td>
<td>287 (35)</td>
<td>113 (34)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University of Pennsylvania</td>
<td>1253 (37)</td>
<td>286 (35)</td>
<td>150 (45)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Admissions in previous year, #</td>
<td>1 (0, 2)</td>
<td>2 (1, 4)</td>
<td>3 (1, 5)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results: AUC

Readmitted
- Alternative model, AUC = 0.647
- Null model, AUC = 0.559

Readmitted to UPHS
- Alternative model, AUC = 0.603
- Null model, AUC = 0.537

\[ \Delta \text{AUC} = 0.088 \]

\[ \Delta \text{AUC} = 0.066 \]
Results: IDI

- Readmitted: 0.059
- Readmitted to UPHS: 0.039

Recall: $\hat{IDI}(N, A) = (\hat{p}_A, D=1 - \hat{p}_N, D=1) - (\hat{p}_A, D=0 - \hat{p}_N, D=0)$

<table>
<thead>
<tr>
<th></th>
<th>$\hat{p}_A, D=1$</th>
<th>$\hat{p}_N, D=1$</th>
<th>$\hat{p}_A, D=0$</th>
<th>$\hat{p}_N, D=0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Readmitted</td>
<td>0.303</td>
<td>0.258</td>
<td>0.234</td>
<td>0.249</td>
</tr>
<tr>
<td>Readmitted to UPHS</td>
<td>0.288</td>
<td>0.256</td>
<td>0.243</td>
<td>0.250</td>
</tr>
</tbody>
</table>
Results: IDI

- Readmitted: **0.059**
- Readmitted to UPHS: **0.039**

Recall: \( \hat{IDI}(N, A) = (\hat{p}_A, D=1 - \hat{p}_N, D=1) - (\hat{p}_A, D=0 - \hat{p}_N, D=0) \)

<table>
<thead>
<tr>
<th></th>
<th>( \hat{p}_A, D=1 )</th>
<th>( \hat{p}_N, D=1 )</th>
<th>( \hat{p}_A, D=0 )</th>
<th>( \hat{p}_N, D=0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Readmitted</td>
<td>0.303</td>
<td>0.258</td>
<td>0.234</td>
<td>0.249</td>
</tr>
<tr>
<td>Readmitted to UPHS</td>
<td>0.288</td>
<td>0.256</td>
<td>0.243</td>
<td>0.250</td>
</tr>
</tbody>
</table>
Summary

- EHRs are a rich source of data, but (statistical) challenges remain.
- Outcome misclassification may arise if events are not captured.
  - In our example, the misclassification rate for readmission was \( \sim 30\% \).
- Misclassification may result in biased prediction accuracy summaries.
  - In our example, misclassification resulted in an understatement of overall accuracy and the accuracy improvement from a new marker: \( \Delta \text{AUC}, -25\%; \text{IDI}, -34\% \).
- To maximize the utility of EHR data in research, the potential for outcome misclassification should be considered.
Future directions

- In our example, we used PHC4 data to measure number of admissions in the previous year
  - Consider potential for exposure misclassification in risk factors

- In our example, we used a binary variable for readmission and ignored competing risks due to death
  - Consider readmission as a time-dependent outcome and accommodate competing risks due to death [Saha and Heagerty, 2010]

★ How to adjust for potential outcome misclassification in EHR data?
  - Imputation models
  - Validation subsets
Acknowledgements

Le Wang, MS
University of Pennsylvania

Hansie Mathelier, MD
University of Pennsylvania

Stephen Kimmel, MD, MSCE
University of Pennsylvania
References

dbe.med.upenn.edu/biostat-research/bcfrench


