A Statistical Framework for Individualizing Health
Scott L. Zeger, Zhenke Wu, Yates Coley, PERCH Study Team
Johns Hopkins Department of Biostatistics
Seattle Symposium on Health Care Data Analytics
Play Doctor (even if not on TV)

• 40 year old man, no family history of disease X, tests “positive” in a screening test
• What is his disease state?
• What action do you, his doctor recommend to him?

<table>
<thead>
<tr>
<th>Test result</th>
<th>True cancer status</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Yes</td>
<td>15</td>
<td>985</td>
</tr>
<tr>
<td>Negative</td>
<td>No</td>
<td>5</td>
<td>8,995</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>20</td>
<td>9,980</td>
</tr>
</tbody>
</table>
Variability is the law of life, and as no two faces are the same, so... no two individuals react alike and behave alike under the abnormal conditions which we know as disease. – William Osler
“…the framework sets a moral presumption in favor of learning, in which health professionals and institutions have an affirmative obligation to conduct learning activities and patients have an affirmative obligation to contribute to these activities.”
Common Questions about Patient and Population Health

1. What is the person’s health state given health measurements?

2. What is the person’s health “trajectory”?

3. Does a particular intervention improve health – on average; for a particular person?

4. Is the intervention being used optimally? How much does it the population’s health at what cost?
Boole -a- Bayes

George Boole 1815-1864

Thomas Bayes 1701-1761
Radar to Land Aircraft Safely at BWI

Measurements

Initial Radar Measurements (radar echoes)

Filtered (smoothed) “track” updates

Smoothed track updates

Symbol of type of aircraft

Indicates expected future direction and speed

Display of altitude

Data:
- Track No. 1234
- Position (x,y,z)
- Speed and direction
- Acceleration
- Identification

September 30, 2014

Seattle Meeting
Complementary Approaches

• Expand biomedical knowledge
  – Discovery of mechanisms
  – Novel measurements of underlying processes

• Use existing science and measurement more intelligently
# Statistical Model Components

## State Equations
1. Health state model
   1a. Health state definition
   1b. Health state trajectory
   1c. Covariate and intervention effects on health state

2. Intervention model

## Observation Equations
3. Measurement model

## Embedding within Relevant Population
Health State/Trajectory ($\eta_{it}$) with Person-specific Indicator ($\delta_i$)
Effects of Exogenous (X) and Endogenous (Rx) Covariates on Health State/Trajectory with Person-specific Regression Coefficients ($\beta_i$)
Observations \((Y)\) that Inform about Health State through Coefficients \((\varphi_i)\)
1. What is the person’s health state given health measurements to date? $E(\eta_{it} | Y, X, \ldots)$
2. What is the person’s health “trajectory”? $E(\delta_i|Y_i, X, \ldots)$
3. Does a particular intervention improve health – on average; for a person with specific set of characteristics?

Benefit = E(η_{it} | Y, Rx=1) - E(η_{it} | Y, Rx=0)
4. Is the intervention being used optimally? If so, how much difference does it make to the population’s health?

$$\Pr(Rx_{it}=1|X_{it}, Y_i)$$ causing increase in benefit?
Hierarchy of Hopkins *in*Health Aspirations

- Measure or predict state of health
- Predict trajectory of health
- Estimate average intervention effect
- Estimate subgroup-specific intervention effects
- Predict optimal use of intervention in population?
Observations ($Y$) that Inform about Health State through Coefficients ($\varphi_i$)
Observations \((M)\) that Inform about Health State \((I)\) through Coefficients \((\varphi_i)\)
Pneumonia Etiology Research for Child Health (PERCH) study

* PERCH coordinating center

* Core team members

---

2. Levine O. et al. (2012)
Childhood pneumonia
Measurements and their features

Lung State $I^L$

- NP PCR Measure $M^{NP}$
- Blood Culture Measure $M^B$
- LA Measure $M^L$

controls cases

Bronze Silver Gold

*NP-nasopharyngeal; PCR-polymerase chain reaction; LA-lung aspirate
Proposed method: partially-latent class models (pLCM)
Model structure (bronze-standard only)
Local Identifiability

• Absent prior knowledge, bronze standard data alone are not sufficient to identify both the true positive rates ($\theta_j$) and the etiologic fractions ($\pi_j$)
  – observed frequency for pathogen j = $\theta_j \pi_j + \psi_j(1-\pi_j)$

• Prior knowledge from lab experiments and vaccine trials about $\theta_j$

• Etiology is estimable given priors; uncertainty about $\theta_j$’s only reduced by SS, GS data; is reflected in posterior distributions for etiologic fractions

• Prior elicitation among international investigators as a check for what our JHU colleagues have proposed for priors
Respiratory syncytial virus (RSV)

bronze-standard  silver-standard  posterior

BrS  SS  $\frac{\Lambda}{\pi}$

conditional OR

ctrl case

0.8% 25.2%

21.8 175.9

$\Lambda_{11}^{\pi} = 27.3\%$

RSV_A_B
Viral, or Bacterial pneumonia?

Diagram showing distributions of bacteria and virus with posterior and prior distributions indicated.
What about conditional dependence of multiple binary measurements for a child?

Can we borrow the dependence structure from the controls to the cases (e.g., similar handling of specimens)?
Extension: Nested partially-latent class models (npLCM)

Example: 5 pathogens, 2 subclasses
Encourage a small number of subclasses: stick-breaking prior

\( v_j \sim \text{Beta}(1, \alpha); \) Example: \( K = 10, \alpha = 1 \)

\[ v_1 \downarrow 1^{\text{st}} \text{break} \]
\[ v_2 \downarrow 2^{\text{nd}} \text{break} \]

After 10\(^{\text{th}}\) break

First several segments receive most weights
npLCM vs pLCM

PERCH data $\Rightarrow$ Population etiology ($\pi$) estimate
Individual’s etiology prediction
Extensions

PERCH Lab Tests:
NP PCR, BCx

Conditional Independence of pathogen measurements (pLCM)

Conditional Dependence of pathogen measurements (npLCM)

Population etiology
Individual Diagnosis

Other covariates: seasonality, age, HIV status.
Observations ($Y$) that Inform about Health State through Coefficients ($\varphi_i$)
Observations ($Y$) that Inform about Health State through Coefficients ($\varphi_i$)
Prostate Screening and Treatment Effects
### Biohealth Pilot Projects

- **Cancer screening**
- **Interventional cardiology**
- **Genomics of cystic fibrosis**
- **Telomere biology chronic diseases**

### Population Health Demonstration

- **Cancer screening**
- **Cardiovascular disease**
- **Age-related cognitive loss**

### Methodology Cores

- **Health measures**
- **Bioethics**
- **Data and software solutions**
- **Statistical design and analysis**
- **Behavior change**
- **Finance/ org. models**
Open Source Learning Environment for Research \textit{inHealth}

\textbf{OSLER \textit{inHealth}}

- Concept for R-package
- Primitives
  - Input
    - access data from standard (e.g. EPIC/Cogito; TransMart) data warehouses
  - Data structures
    - Encounter $<$ subject $\times$ clinician(s) $<$ practice group $<$ population
  - Functionality
    - Embed (individual, within "otherwise similar(x)" population, distance metric and limit (d)
    - Specifying local sub-models; integrating results
- Your ideas and developers welcome
Main ideas one again

• Individualized and population health are two sides of same data coin
• Low hanging fruit is measuring health state and trajectory
• Presented simple model for estimating health state and trajectory, and intervention effects on patients and populations within a health system
• Integration of biomedical knowledge is essential to make progress in many/most problems; used with or without statisticians.
• Can improve patient and population health with novel measurements and/or analytic methods
Hopkins inHealth

Thank you
Improved Health at More Affordable Costs

Clinical and Public Health Discovery

Scale and replicate

Biohealth Pilot Projects
- Cancer screening
- Cardiovascular disease diagnosis and treatment
- Genomics of cystic fibrosis
- Telomere biology and chronic diseases
- OncoSpace in Radiation Oncology

Population Health Demonstration
- Cancer screening and early diagnosis
- Cardiovascular disease
- Age-related cognitive loss
- Obesity and Diabetes
- Children’s asthma prevention and control

Methodology Cores
- Health measurement
- Bioethics
- Big Data and software solutions
- Statistical design and analysis
- Behavior change and dissemination
- Finance – organization models
Is BrS data useful?

Simulation

<table>
<thead>
<tr>
<th>Population etiology (π)</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>67%</td>
<td>26%</td>
<td>7%</td>
</tr>
<tr>
<td>TPR</td>
<td>.90</td>
<td>.90</td>
<td>.90</td>
</tr>
<tr>
<td>FPR</td>
<td>.60</td>
<td>.02</td>
<td>.05</td>
</tr>
</tbody>
</table>

Prior
- BrS
- GS
- BrS+GS
<table>
<thead>
<tr>
<th></th>
<th>BrS</th>
<th>SS</th>
<th>$\hat{\pi}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>prior</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>posterior</td>
</tr>
<tr>
<td>C1</td>
<td>0.9%</td>
<td>25.2%</td>
<td>$\hat{\pi}_1$ = 31.7%</td>
</tr>
<tr>
<td>C2</td>
<td>1.5%</td>
<td>2.5%</td>
<td>$\hat{\pi}_2$ = 17.6%</td>
</tr>
<tr>
<td>C3</td>
<td>2.5%</td>
<td>5.5%</td>
<td>$\hat{\pi}_3$ = 13.5%</td>
</tr>
<tr>
<td>C4</td>
<td>1.8%</td>
<td>2.5%</td>
<td>$\hat{\pi}_4$ = 10.2%</td>
</tr>
<tr>
<td>C5</td>
<td>5.8%</td>
<td>2.5%</td>
<td>$\hat{\pi}_5$ = 6.6%</td>
</tr>
<tr>
<td>C6</td>
<td>1.2%</td>
<td>2.5%</td>
<td>$\hat{\pi}_6$ = 6.3%</td>
</tr>
<tr>
<td>C7</td>
<td>0.7%</td>
<td>2.5%</td>
<td>$\hat{\pi}_7$ = 3.6%</td>
</tr>
<tr>
<td>C8</td>
<td>1.2%</td>
<td>0.7%</td>
<td>$\hat{\pi}_8$ = 3.3%</td>
</tr>
<tr>
<td>C9</td>
<td>1.2%</td>
<td>0.7%</td>
<td>$\hat{\pi}_9$ = 3.2%</td>
</tr>
<tr>
<td>C10</td>
<td>1.2%</td>
<td>0.7%</td>
<td>$\hat{\pi}_{10}$ = 2.8%</td>
</tr>
<tr>
<td>C11</td>
<td>1.2%</td>
<td>0.7%</td>
<td>$\hat{\pi}_{11}$ = 1%</td>
</tr>
</tbody>
</table>
Likelihood and Bayesian computing

- **Bronze-standard**
  \[
  P_{i',i}^{1, BrS} = \Pr(M_{i',i}^{BrS} = m | \pi, \theta^{BrS}, \psi^{BrS}) = \sum_{j=1}^{J} \pi_j \cdot \left( \theta_j^{BrS} \right)^{m_j} \left( 1 - \theta_j^{BrS} \right)^{1 - m_j} \prod_{l \neq j} \left( \psi_l^{BrS} \right)^{m_l} \left( 1 - \psi_l^{BrS} \right)^{1 - m_l}, m = m_{i'}^{BrS}.
  \]

- **Silver-standard**
  \[
  P_{i',i}^{1, SS} = \Pr(M_{i',i}^{SS} = m | \pi, \theta^{SS}) = \sum_{j=1}^{J'} \pi_{j'} \cdot \left( \theta_{j'}^{SS} \right)^{m_{j'}} \left( 1 - \theta_{j'}^{SS} \right)^{1 - m_{j'}} 1_{\{\sum_{l=1}^{J'} m_l \leq 1\}}.
  \]

- **Gold-standard**
  \[
  P_{i',i}^{1, GS} = \Pr(M_{i',i}^{GS} = m | \pi) = \prod_{j=1}^{J} \pi_j \cdot 1_{\{m_j = 1\}} 1_{\{\sum_j m_j = 1\}}, m = m_{i'}^{GS}.
  \]
The data

- **Mortality:**

- **PM2.5:**

- **Census:**
  - Census 2000
  - inter-census estimates 1993-2000
  - projections 2000-2002
The data

- Initially 250 counties were chosen based on PM2.5 availability and population size.
- Mortality counts and population estimates were available by age, gender and race.
- Final analysis uses three age groups, 65-74, 75-84, 85+. Analyses suggest effects are similar across gender and race.
- County level covariates were taken from the 2000 census and included:
  - proportion of population completing high school, proportion with a degree, proportion unemployed, proportion living in poverty, and medium family income.
- COPD standardized mortality ratios were used as a surrogate measure of smoking prevalence. Ten years of county COPD deaths were indirectly standardized to the U.S. 2000 population. Adjustment was made by age, race and gender.
- Smoothed county specific PM2.5 (mg/m$^3$) averages 2000-2002.
- Smoothing done using penalized thin-plate spline (lambda=0.05).
• Directly adjusted 10-year mortality rates by county (per 1000), number of counties, and population over 65 years (in millions)
• Rates are for those 65 years or older and adjusted for age, race and gender and standardized to the 2000 U.S. population.
• Contour plot of SMR’s of all-cause mortality (Medicare 2000-2003) and residual mean PM2.5 (2000-2002) from the West Coast and the Eastern U.S.
• SMR’s and smoothed plot are for ages 65-74 and rates are per thousand population.
“...the framework sets a moral presumption in favor of learning, in which health professionals and institutions have an affirmative obligation to conduct learning activities and patients have an affirmative obligation to contribute to these activities.”