Bayesian nonparametric and semiparametric inferences for disease risk, ROC curves, and prevalence

Adam Branscum*, Ian Gardner, Chun-lung Su, Mark Thurmond
University of California, Davis
Tim Hanson
University of New Mexico
and
Wesley Johnson
University of California, Irvine
Two common epidemiology problems

1. Evaluate the performance of a diagnostic test
   - Quantify the discriminatory ability of the test when applied to diseased and non-diseased individuals, respectively

2. Disease risk assessment and prevalence estimation
   - Inferences for the predictive probability of disease for given test results
   - Estimate the prevalence of disease in a population
Setting

- Obtain a simple random sample of $n$ individuals from a population with disease prevalence $\pi$.
- Apply a continuous diagnostic test to each sampled individual.
- The ‘serology score’ measures concentration in the serum of antibodies specific to the antigen or agent.
- Larger for diseased individuals and smaller for non-diseased individuals.
- Let $Y_i$ denote the (possibly transformed) serology score for subject $i$ so that $Y_i = h(S_i)$.
Setting

- Let $D^+$ and $D^-$ denote disease positive and negative
- Denote the distribution of serology scores for $D^-$ individuals by $G_0$ and the distribution for $D^+$ individuals by $G_1$
- True disease status is unknown eg. no gold standard
Dichotomize the serology scores using a cutoff value $k$
Standard Approach

- $Y > k$ corresponds to $T^+$ and $Y < k$ corresponds to $T^-$
- $Se = \Pr(T^+|D^+) = \Pr(Y > k|D^+)$
- $Sp = \Pr(T^-|D^-) = \Pr(Y < k|D^-)$

Dichotomizing the data is an inefficient use of the available information

Results in an inherent information loss
The predictive probability of disease is the same for all $T^+$ (or $T^-$) individuals regardless of their actual serology scores.

Subjects with serology scores far above (or below) $k$ are viewed as identical to subjects whose serologic values are barely above (or below) $k$.

We generally expect the predictive probability values to increase with increasing serology scores.
Inferences Using Standard Approach

- Test accuracy estimation
  - The performance of the diagnostic test based on dichotomized data is completely characterized by $Se$ and $Sp$.

- Disease risk assessment
  - The predictive probability of disease for all $T^+$ individuals is the so-called predictive value positive:

$$PVP = \Pr(D^+|T^+) = \frac{\pi Se}{\pi Se + (1 - \pi)(1 - Sp)}$$
Inferences Using Standard Approach

- The predictive probability of disease for all $T^-$ individuals is the predictive value negative:

$$\text{PVN} = \Pr(D^- | T^-) = \frac{\pi(1 - Se)}{\pi(1 - Se) + (1 - \pi)Sp}$$

- Branscum et al. (2004, 2005) present Bayesian approaches to inference
Modeling the continuous serology scores allows for predictive inference of the probability of disease given any test result.

Provides inference for the test sensitivity and specificity across all possible cutoff values, thus yielding inferences for the receiver operating characteristic (ROC) curve.

The ROC curve provides a graphical measure of test performance. The curve plots \((1 - Sp(k), Se(k))\) for all cutoff values \(k\) used to dichotomize the data into test positive (\(\geq k\)) or negative (\(< k\)).
Letting $H_j = 1 - G_j$ for $j = 0, 1$, the ROC curve is given by

$$ROC(t) = H_1\{H_0^{-1}(t)\}$$

for $t \in (0, 1)$

The area under the curve (AUC) equals the probability that a diseased subject test score is $>$ than that for a healthy subject:

$$AUC = P(X > Y)$$

where $X|G_1 \sim G_1$ and $Y|G_0 \sim G_0$. 
Distributions of Serology Scores
ROC Curve

Se(k) vs. 1 - Sp(k)

- p. 13/56
Modeling Serology Scores

- Let $Z_i$ denote the latent binary indicator of disease status where $Z_i = 1$ if subject $i$ is diseased.
- The general model is

$$Y_i \mid G_0 \sim G_0, \ Z_i = 0$$
$$Y_i \mid G_1 \sim G_1, \ Z_i = 1$$
$$Z_i \mid \pi \sim \text{Bern}(\pi)$$

- How to specify flexible models for $G_0$ and $G_1$ when we don’t know who is diseased and who is not?
Parametric Approach

- \( G_0 \) and \( G_1 \) are parametric probability models:
  \[
  G_0 \in \{ G_{\theta_0} : \theta_0 \in \Theta_0 \} \\
  G_1 \in \{ G_{\theta_1} : \theta_1 \in \Theta_1 \}
  \]

- Transformations to normality (e.g., Box-Cox) could be investigated, but difficult to implement with no gold-standard data.

- An ad-hoc frequentist approach was presented in Thurmond et al. (2002).

- A Bayesian parametric model was considered by Choi et al. (2005).

- Computationally much easier than the frequentist approach.
Modeling

- Distribution of serology scores for infected subjects are likely multimodal and/or skewed.
- The distribution may depend on covariates not accounted for in the model (such as age).
- The subpopulation of diseased individuals may be comprised of individuals with different stages of disease.
- For diseased individuals, serology scores may be expected to be smaller for recently infected subjects because they may have a lower concentration of the agent the test is designed to measure.
Modeling Issues

- Serology scores for long term diseased individuals in an advanced disease stage are expected to be higher.
Modeling Issues

- A finite mixture model could be used as an initial attempt to broaden the parametric model.
- Need to know the number of mixture components or place a prior on the number of mixture components.
- Varying dimensional parameter space requires specialized Markov chain Monte Carlo (MCMC) methods such as reversible jump.
- We consider Bayesian nonparametric models for $G_0$ and $G_1$.
- Provides data-driven estimates of $G_0$ and $G_1$, and hence also for the ROC curve and predictive probabilities of disease.
A Nonparametric Alternative

- We consider Bayesian nonparametric models for $G_0$ and $G_1$
- Provides data-driven estimates of $G_0$ and $G_1$, and hence also for the ROC curve and predictive probabilities of disease
- Let $G_0$ and $G_1$ be arbitrary subject to the constraint that $G_1$ is stochastically greater than $G_0$
- Place a probability measure $\mathcal{P}$ on the space of probability measures

$$G_0 \mid \theta_0 \sim \mathcal{P}_{\theta_0}(dG_0)$$
$$G_1 \mid \theta_1 \sim \mathcal{P}_{\theta_1}(dG_1)$$
Bayesian Inference

Given the data \( Y = (Y_1, Y_2, \ldots, Y_n) \) and latent data \( Z = (Z_1, Z_2, \ldots, Z_n) \), the likelihood function is

\[
L(G_0, G_1, \theta_0, \theta_1, \pi \mid Y, Z) \propto \left( \prod_{i=1}^{n} [g_0(y_i)]^{1-Z_i} [g_1(y_i)]^{Z_i} \right) \times \pi \sum_{i=1}^{n} Z_i \left(1 - \pi\right)^{n-\sum_{i=1}^{n} Z_i}
\]

and probability distributions on all uncertain quantities eg.
Bayesian Inference

- Obtain the joint posterior distribution

\[
p(dG_0, dG_1, d\theta_0, d\theta_1, d\pi | Y) \propto L(G_0, G_1, \theta_0, \theta_1, \pi | Y, Z) \times p(dG_0, dG_1, d\theta_0, d\theta_1, d\pi)
\]

or an approximation to it using an MCMC sampling algorithm e.g. Gibbs sampling

- The parameters \( G_0 \) and \( G_1 \) are infinite dimensional so in fitting the model via MCMC, \( G_0 \) and \( G_1 \) can only be partially sampled
Bayesian Inference

- Posterior inferences for

\[ \text{ROC}(t) = H_1\{H_0^{-1}(t)\} \equiv \gamma(t) \]

- Predictive inferences for the probability of disease given serology score \( y \)

\[ \Pr(Z = 1|y, Y) \]

- Require finding the induced posterior/predictive distributions for \( \gamma(t) \) and \( Z|y, Y \)
Median of the posterior distribution provides Bayesian point estimate for $\gamma(t)$

Interval with 95% content provides Bayesian interval estimate
Aspects of Bayesian Nonparametrics

- We model the data from an unspecified probability distribution, $G$, where $G$ belongs to the space of all probability distributions.

- We let $G_0$ and $G_1$ be “random” probability measures. This amounts to placing a distribution, $\mathcal{P}$, on the space of all probability measures.

- Common choices of $\mathcal{P}$ are
  1. Mixtures of Dirichlet processes (MDP)
  2. Dirichlet process mixtures (DPM)
  3. Mixtures of (finite) Polya trees
The nonparametric model we consider is specified as:

\[ Y_i \mid Z_i, G_{Z_i} \sim G_{Z_i} \]

\[ Z_i \mid \pi \sim \text{Bern}(\pi) \]

\[ G_0 \mid (c_0, \mu_0, \sigma_0) \sim FPT(N(\mu_0, \sigma_0), c_0) \]

\[ G_1 \mid (c_1, \mu_1, \sigma_1) \sim FPT(N(\mu_1, \sigma_1), c_1) \]

\[ \pi \sim \text{Beta}(a_\pi, b_\pi) \]

\[ \mu_0 \sim N(a_{\mu_0}, b_{\mu_0}) \]

\[ \sigma_0 \sim \Gamma(a_{\sigma_0}, b_{\sigma_0}) \]

\[ \mu_1 \sim N(a_{\mu_1}, b_{\mu_1}) \]

\[ \sigma_1 \sim \Gamma(a_{\sigma_1}, b_{\sigma_1}) \]
For $i = 0, 1$, let $S_i = \sum_{k=1}^{N_i-1} p_{i,k}$ and let $S_{0t} = 1 - t - S_0$. The ROC curve is given by:

$$ROC(t) = 1 - S_1 - p_{1,N_1} \left[ 2^{J_1} \Phi \left\{ \sigma_1^{-1} \right\} - \mu_1 \sigma_1^{-1} \right] - N_1 + 1$$

where $J_0$ and $J_1$ are the levels of the finite trees, $p_{i,1}, p_{i,2}, \ldots, p_{i,2^{J_i}}$ denote the $G_i$-probabilities of the sets at level $J_i$, $i = 0, 1$, and $N_i \in \{1, 2, \ldots, 2^{J_i}\}$ indexes the set at level $J_i$ containing $H_0^{-1}(t)$ ($i = 1$) or containing $t$ ($i = 0$).
Simulated Data Examples

- Several simulated data sets were fit to the model. We focus on 3 scenarios here.
- Normal / Normal with large separation
- Normal / Normal with small separation
- Normal / Mixture of Two Normals
Normal/Normal: Large Separation

- Non-diseased: \( N(50, 10) \)  Diseased: \( N(90, 15) \)
Normal/Normal: Large Separation

![Graph showing two bell curves representing different distributions of serology scores. The x-axis represents serology score ranging from 30 to 150, and the y-axis represents density ranging from 0.0 to 0.04. The two curves are well-separated, indicating a large separation between the two populations.](attachment:image.png)
Normal/Normal: Large Separation
Normal/Normal: Large Separation

AUC = 0.981 (0.942, 0.994)

True AUC = 0.988

\[ \text{Se}(k) \]

\[ 1 - \text{Sp}(k) \]
Normal/Normal: Small Separation

- Non-diseased: $N(50, 10)$  
  Diseased: $N(80, 15)$
Normal/Normal: Small Separation
Normal/Normal: Small Separation
Normal/Normal: Small Separation

AUC = 0.905 (0.806, 0.954)
True AUC = 0.925
Normal/ Mixture of 2 Normals

- Non-diseased: $N(50, 10)$
- Diseased: $0.65 \times N(90, 15) + 0.35 \times N(145, 15)$

![Graph showing two normal distributions: one for healthy (N(50, 10)) and one for diseased (0.65*N(90,15) + 0.35*N(145,15)). The graph illustrates the density of serology scores.](image-url)
Normal / Mixture of 2 Normals
Normal / Mixture of 2 Normals

Predictive Probability

Serology Score

0.0 0.2 0.4 0.6 0.8 1.0
30 40 50 60 70 80 90 100
Normal / Mixture of 2 Normals

AUC=0.974 (0.936, 0.990)
True AUC = 0.991
Normal / Mixture of 2 Normals

Serology Score

Density

30 60 90 120 150 180

0.0 0.01 0.02 0.03 0.04
Normal / Mixture of 2 Normals
Normal / Mixture of 2 Normals

AUC = 0.986 (0.950, 0.993)
True AUC = 0.991
Example: Johne’s Disease

- Paratuberculosis (Johne’s disease) is an incurable wasting disease.
- A study was conducted to estimate the performance of an ELISA for Johne’s disease in cows in Australia.
- True infection status was not known.
- 440 cows were randomly sampled from a herd and tested for Johne’s disease.
- Reasonably informative prior information available.
Johne’s Dz: Density

- Healthy
- Disease

log(Serology Score)

Density
Johne’s Dz: Predictive Probability
Johne’s Dz: ROC Curve

AUC = 0.914 (0.860, 0.952)
Semiparametric Extensions

- Incorporate covariates related to disease status
- The semiparametric model is

\[ Y_i \mid G_0 \sim G_0, \quad Z_i = 0 \]

\[ Y_i \mid G_1 \sim G_1, \quad Z_i = 1 \]

<table>
<thead>
<tr>
<th>Nonparametric</th>
<th>Semiparametric</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \pi ) ( \rightarrow ) ( \pi_i )</td>
<td></td>
</tr>
<tr>
<td>( Z_i \mid \pi \ \text{iid} \sim \text{Bern}(\pi) ) ( \rightarrow ) ( Z_i \mid \pi_i \ \perp \sim \text{Bern}(\pi_i) )</td>
<td></td>
</tr>
<tr>
<td>( \pi_i = F(x_i\beta) )</td>
<td></td>
</tr>
</tbody>
</table>
Semiparametric Extension 1

- $F(w) = \frac{e^w}{1+e^w}$ yields logistic regression
- The ROC curve is exactly as in the no covariate model
- Get covariate-specific predictive inferences for the probability of disease for a subject with serology score $y$ and covariate $x$:

$$\Pr(Z = 1 \mid y, Y, x)$$
Covariates related to disease status and serology score

For diseased individuals,

\[
\log(Y_i) = x_i^* \alpha + \epsilon_i
\]

\[\epsilon_i | G_1 \sim G_1\]

\[G_1 | \sigma_1^2 \sim \text{FPT}(\mathcal{N}(0, \sigma_1^2), c_1)\]

The model for non-diseased individuals does not depend on covariates (although easy to incorporate covariate information here too if needed)
Inferences for covariate-specific ROC curves, $\text{ROC}(t|x^*)$

Inferences for covariate-specific predictive probabilities of disease for a subject with serology score $y$ and covariates $x$ (disease status covariate) and $x^*$ (serology score covariate)

$$\Pr(Z = 1 | y, Y, x, x^*)$$
The preceding discussion focused on non-dichotimized data from a single population.

In human and veterinary medicine, health surveys are often conducted for a region, e.g. an entire country.

Disease freedom surveys conducted in India for leprosy (Murthy et al., 2001)

Ranking regional villages in South America as to priority treatment for human onchocerciasis (Carabin et al., 2003)

The region is comprised of several subpopulations (e.g. villages)
Regional Prevalence Estimation

- Each subpopulation has a disease prevalence, $\pi_i$
- The goal is to characterize the distribution of prevalences in the region
- Vaccination efforts can be focused in subpopulations where the prevalence exceeds a maximum acceptable threshold
- Knowledge of high and low risk subpopulations is important in directing surveillance programs, including targeting sampling size and frequency to high risk subpopulations
The complete nonparametric model specification is:

\[ y_i | \pi_i, Se, Sp \sim \text{Binomial}(n_i, p_i) \]
\[ p_i = \pi_i Se + (1 - \pi_i)(1 - Sp) \]
\[ \pi_i | \tau, G \simiid \tau G + (1 - \tau)\delta\{0\} \]
\[ G | \mu, \psi \sim \text{PT}(\text{Beta}(\mu \psi, \psi(1 - \mu)), c) \]
\[ \tau | \gamma \sim \gamma H + (1 - \gamma)\delta\{0\} \]
\[ \mu \sim \text{Beta}(a_\mu, b_\mu) \quad \psi \sim \text{Gamma}(a_\psi, b_\psi) \]
\[ Se \sim \text{Beta}(a_{Se}, b_{Se}) \]
\[ Sp \sim \text{Beta}(a_{Sp}, b_{Sp}) \]
\[ \gamma = \gamma_0 \]
Regional Prevalence Estimation

Prior and True PD

\[ g(x) \]

- \( n = 500 \)
- \( n = 100 \)
- \( n = 50 \)
Semiparametric Prevalence Est

- Model the link function in GLMs nonparametrically
- Incorporate subpopulation-specific covariates

\[ y_i | \pi_i, Se, Sp \sim \text{Binomial}(n_i, p_i) \]

\[ p_i = \pi_i Se + (1 - \pi_i)(1 - Sp) \]

\[ \pi_i = F(x_i; \beta) \]

\[ F \sim FPT(F_0, c) \]

\[ Se \sim \text{Beta}(a_{Se}, b_{Se}) \]

\[ Sp \sim \text{Beta}(a_{Sp}, b_{Sp}) \]

\[ \beta \sim \mathcal{N}_{p+1}(0, \Sigma_0) \]
Semiparametric Prevalence Estimation

The graph illustrates the function $F(x)$, which represents the cumulative distribution function. The $x$-axis ranges from 0 to 1, and the $y$-axis ranges from 0 to 1, showing the probability of events occurrence as $x$ varies.