Methodological Issues for Biosurveillance

Ronald D. Fricker, Jr.
12th Biennial CDC & ATSDR Symposium on Statistical Methods
April 6, 2009
• Associate professor, Naval Postgraduate School, Monterey, CA

• Research interests
  – Industrial quality control and statistical process control (SPC) methods
  – Developing and evaluating SPC methods for biosurveillance

• Contact information
  – Phone: 831-656-3048
  – E-mail: rdfricker@nps.edu
  – My NPS website: http://faculty.nps.edu/rdfricke/
  – Course site: http://faculty.nps.edu/rdfricke/Biosurveillance.htm
“That’s how it’s gonna be, a little test tube with a-a rubber cap that’s deteriorating... A guy steps out of Times Square Station. Pshht... Smashes it on the sidewalk... There is a world war right there.”

“Josh”
West Wing, 1999
The New Status Quo?

THEN

Atchoo!

Bless you

NOW

Atchoo!

Hello CDC?
What is Biosurveillance?

  - “The term ‘biosurveillance’ means the process of active data-gathering … of biosphere data … in order to achieve early warning of health threats, early detection of health events, and overall situational awareness of disease activity.” [1]
  - “The Secretary of Health and Human Services shall establish an operational national epidemiologic surveillance system for human health…” [1]

- Syndromic surveillance:
  - “…surveillance using health-related data that precede diagnosis and signal a sufficient probability of a case or an outbreak to warrant further public health response.” [2]

Two Purposes of Biosurveillance

• **Early event detection (EED):** Gathering and analyzing data in advance of diagnostic case confirmation to give early warning of a possible outbreak.

• **Situational awareness (SA):** The real-time analysis and display of health data to monitor the location, magnitude, and spread of an outbreak.

Idea of Biosurveillance: Leverage Secondary Health Data

- Ideal is automatic or near real-time data analysis
- Use data, methods to allow for identification of subtle trends not visible to individual MD’s
- Provide indicators to trigger detection, investigation, quantification, localization, and outbreak management

Derived from “Emerging Health Threats and Health Information Systems: Getting Public Health and Clinical Medicine to Real Time Response,” John W. Loonsk, M.D., Associate Director for Informatics, CDC
One System: BioSense

BioSense Hospitals (356)
DoD Facilities (366)
VA Facilities (886)
• In a review of the literature, Bravata et al. (2004) identified 115 health surveillance systems, including 9 syndromic surveillance systems

• Examples:
  – *Early Aberration Reporting System (EARS)* developed by the CDC
  – *Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE)* developed by the Department of Defense
  – *Real-time Outbreak Detection System (RODS)* developed by the University of Pittsburgh

• Monterey county public health department uses EARS to monitor trends from local hospitals and clinics

In 2007-2008, Buehler et al. surveyed public health officials in 59 state, territorial, and large local jurisdictions

- 52 responded (88% response rate), representing areas comprising 94% of US population
- 83% reported conducting syndromic surveillance for a median of 3 years
- ER data most commonly used (84%), followed by:
  - Outpatient clinic visits (49%)
  - OTC medication sales (44%)
  - Calls to poison control centers (37%)
  - School absenteeism (37%)
- Two-thirds said they are “highly” or “somewhat” likely to expand use of biosurveillance in next 2 years

Explore flu trends across the U.S.

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity in your state up to two weeks faster than traditional systems. Read more »

United States flu activity: Moderate
How Good is Google Flu Trends?

- Google search results correspond to CDC sentinel physician data
- Google says it is able to accurately estimate flu levels 1-2 weeks faster than published CDC reports

Illustrative Biosurveillance Data

Respiratory Data From “Hospital C”
Hospital C: Different Syndromes over Time

(Smoothed) Numbers of Cases

Red = Resp, Orange = Gastro, Green = Unspecified Infection, Blue = Neuro, Purple = Rash
“To date no bio-terrorist attack has been detected in the United Kingdom, or elsewhere in the world using syndromic surveillance systems.”[1]
• Are statistical methods useful for / effective at early event detection?
• Can excessive false alarm rates be controlled (without compromising detection capabilities)?
• Which algorithms perform best and under what conditions?
• What are the appropriate metrics and standards for judging algorithm performance?
• Can the barriers keeping SPC researchers from fully engaging be surmounted?
Issue: Are Statistical Methods Useful for Early Event Detection?

• Aerosol release with windborne spread occurred afternoon of April 2, 1979
  – 6 admitted to hospital on April 4
  – By end of first week after April 4, 28 onsets
  – Out of 96 cases, 64 people eventually died[1]

• Possible conclusions:
  – First signs of an outbreak will be either by a large increase in patients presenting or seriously ill patients who will be diagnosed rapidly[2]
  – Thus:
    • Syndromic surveillance systems, based on statistical algorithms, will be of little value in early detection of bioterrorist outbreaks
    • Early on in the outbreak, there will be cases serious enough to alert physicians and be given definitive diagnoses[2]

Clinicians vs. Biosurveillance: A Simple Simulation

- On average, 100 per day go to area emergency rooms with flu-like symptoms
  - Standard deviation is 20 people
- CUSUM monitors average number presenting daily
  - False signal rate fixed at once per 30 days
- Bio-agent exhibits flu-like symptoms early-on
  - Results in increase in number of people presenting at ERs with flu-like symptoms
  - For those exposed to bio-agent, with probability $p$ some people develop extreme symptoms that a clinicians can easily diagnose

- **Question:** *What is the probability clinician diagnoses a case of the bio-agent before CUSUM signals?*
~ 90-95 percent chance clinician detects first if $p=0.05$, $n$ between 10 and 50/day

~ 50 percent chance clinician detects first if probability of an extreme case $p=0.01$ and number presenting from bio-agent $n=50$/day

~ 75 percent chance clinician detects first if $p=0.025$, $n$ between 8 and 50/day

~ 50 percent chance clinician detects first if $p=0.01$, $n$ between 8 and 50/day
~ 88-95 percent chance clinician detects first if \( p = 0.05 \), \( n \) between 10 and 50/day

~ 46 percent chance clinician detects first if probability of an extreme case \( p = 0.01 \) and number presenting from bio-agent \( n = 50 \)/day

~ 74-77 percent chance clinician detects first if \( p = 0.025 \), \( n \) between 8 and 50/day

~ 46-55 percent chance clinician detects first if \( p = 0.01 \), \( n \) between 8 and 50/day
Simulations suggest there is a role for statistical algorithms in biosurveillance when pathogen is hard to diagnose and/or when small numbers are presenting.
Simulations Indicate Strengths and Limitations of Biosurveillance

• These are just simple, illustrative simulations
  – However, they suggest biosurveillance for early event detection has a role in some situations:
    • As a primary detection tool for rare, hard to diagnose diseases/agents
    • As a back-up to clinicians for moderately sized outbreaks that are moderately hard to diagnose
• Seems to me that rigorous, scientific studies could help clearly define/refine that role, as well as the limitations of biosurveillance
  – Added benefit: Surveillance can focus on particular outcomes/events…more on that topic to follow
Example: Is Biosurveillance Useful for Detecting Anthrax Attack?

- Nordin et al. (2005) used Sverdlovsk to model anthrax attack on Mall of America
  - Modeled rate of physician visits for respiratory symptoms
  - SatScan used
- Would other methods have been faster?
- Would astute clinicians be faster?

• A brief introduction to **statistical process control (SPC)** – from an industrial quality control perspective

• A **control chart** is a statistical tool to detect “assignable causes of variation”

• **Advantages of control charts**
  – Graphically displays performance
  – Accounts for natural randomness
  – Removes subjective decision making

How Industry Uses Control Charts

Goal: detect a shift before not capable

A Measure of Quality

capable

not capable

observations over time

USL

LSL

WWW.NPS.EDU
• Choose control limits to guide actions
  – If points fall within control limits, assume process in control → No action required
  – If point or points fall outside control limits, evidence process out of control → Look for “assignable causes”

• Competing requirements for control limits
  – When in-control, want small chance of point falling outside control limits (i.e., low false alarm rate)
  – When out-of-control, want high chance of falling out of control limits (i.e., high power)
Univariate Statistical Process Control (SPC) Methods

• Shewhart (1931)
  – Stop when observation (or statistic) exceeds pre-defined threshold
  – Better for detecting large shifts/changes

• CUSUM (Page, 1954)
  – Stop when cumulative sum of observations exceeds threshold
  – Better for detecting small shifts/changes

• EWMA (Roberts, 1959)
  – Stop when weighted average of observations exceeds threshold
  – Very similar in performance to CUSUM
Shewhart ("X-bar") Charts

• Observations follow an in-control distribution $f_0(x)$, for which we often want to monitor the mean of the distribution

• If interested in detecting both increases and decreases in the mean, choose thresholds $h_1$ and $h_2$ such that

$$\int \left\{ x: x \geq h_1 \text{ or } x \leq h_2 \right\} f_0(x) dx = p$$

• Sequentially observe values of $x_i$; stop and conclude the mean may have shifted at time $i$ if $x_i \geq h_1$ or $x_i \leq h_2$
Example of a Shewhart (\( \bar{X} \)) Chart

<table>
<thead>
<tr>
<th>Sample, ( i )</th>
<th>(a) ( x_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.45</td>
</tr>
<tr>
<td>2</td>
<td>7.99</td>
</tr>
<tr>
<td>3</td>
<td>9.29</td>
</tr>
<tr>
<td>4</td>
<td>11.66</td>
</tr>
<tr>
<td>5</td>
<td>12.16</td>
</tr>
<tr>
<td>6</td>
<td>10.18</td>
</tr>
<tr>
<td>7</td>
<td>8.04</td>
</tr>
<tr>
<td>8</td>
<td>11.46</td>
</tr>
<tr>
<td>9</td>
<td>9.20</td>
</tr>
<tr>
<td>10</td>
<td>10.34</td>
</tr>
<tr>
<td>11</td>
<td>9.03</td>
</tr>
<tr>
<td>12</td>
<td>11.47</td>
</tr>
<tr>
<td>13</td>
<td>10.51</td>
</tr>
<tr>
<td>14</td>
<td>9.40</td>
</tr>
<tr>
<td>15</td>
<td>10.08</td>
</tr>
<tr>
<td>16</td>
<td>9.37</td>
</tr>
<tr>
<td>17</td>
<td>10.62</td>
</tr>
<tr>
<td>18</td>
<td>10.31</td>
</tr>
<tr>
<td>19</td>
<td>8.52</td>
</tr>
<tr>
<td>20</td>
<td>10.84</td>
</tr>
<tr>
<td>21</td>
<td>10.90</td>
</tr>
<tr>
<td>22</td>
<td>9.33</td>
</tr>
<tr>
<td>23</td>
<td>12.29</td>
</tr>
<tr>
<td>24</td>
<td>11.50</td>
</tr>
<tr>
<td>25</td>
<td>10.60</td>
</tr>
<tr>
<td>26</td>
<td>11.08</td>
</tr>
<tr>
<td>27</td>
<td>10.38</td>
</tr>
<tr>
<td>28</td>
<td>11.62</td>
</tr>
<tr>
<td>29</td>
<td>11.31</td>
</tr>
<tr>
<td>30</td>
<td>10.52</td>
</tr>
</tbody>
</table>

• If only interested in detecting increases in the mean, can use a **one-sided test**
  – Sequentially observe values of $x_i$; stop and conclude the mean may have shifted at time $i$ if $x_i > h$

• Industrial applications often set thresholds as multiples of process standard deviation

• Can also use Shewhart charts to monitor process variation along with mean
  – In industrial SPC, called “s-charts” or “R-charts”
• ARL is a measure of chart performance
  – In-control ARL or ARL₀ is expected number of observations between false signals
    • Assuming \( f_0(x) \) known, time between false signals is geometrically distributed, so \( ARL_0 = 1/p \)
    • Larger ARL₀ are preferred
  – Out-of-control ARL or ARL₁ is expected number of observations until a true signal for a given out-of-control condition
    • For a one-sided test and a particular \( f_1(x) \),
      \[
      ARL_1 = \left[ \int_{\{x:x\geq h\}} f_1(x)\,dx \right]^{-1}
      \]
Example: Monitoring a Process with $X_i \sim N(\mu, \sigma^2)$

- With $3\sigma$ control limits, when in-control, probability an observation is outside the control limits is $p = 0.0027$, so $ARL_0 = 1/0.0027 = 370$
  - If sampling at fixed times, says will get a false signal on average once every 370 time periods

- For out-of-control condition where mean shifts up or down $1\sigma$, probability an observation is outside the control limits is $p = 0.0227$, so $ARL_1 = 1/0.0227 \approx 44$
  - For a $2\sigma$ shift, $ARL_1 = 1/0.1814 \approx 5.5$
  - Etc.
• The two-sided CUSUM plots two statistics:

\[ C_i^+ = \max \left\{ 0, C_{i-1}^+ + (x_i - \mu_0) - k \right\} \]

\[ C_i^- = \min \left\{ 0, C_{i-1}^- + (x_i - \mu_0) + k \right\} \]

typically starting with \( C_0^+ = C_0^- = 0 \)

– Stop when either \( C_0^+ > h \) or \( C_0^- > -h \)

– A one-sided test only uses one of the statistics

• Must choose both \( k \) and \( h \)

– E.g., Setting \( h = 5\sigma \) and \( k = \sigma / 2 \) works well for a 1\( \sigma \) shift in the mean:
  • \( \text{ARL}_0 \) approximately 465 and \( \text{ARL}_1 = 8.4 \) (Shewhart: 44)
(Two-Sided) CUSUM Chart Example

<table>
<thead>
<tr>
<th>Sample, $i$</th>
<th>(a) $x_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.45</td>
</tr>
<tr>
<td>2</td>
<td>7.99</td>
</tr>
<tr>
<td>3</td>
<td>9.29</td>
</tr>
<tr>
<td>4</td>
<td>11.66</td>
</tr>
<tr>
<td>5</td>
<td>12.16</td>
</tr>
<tr>
<td>6</td>
<td>10.18</td>
</tr>
<tr>
<td>7</td>
<td>8.04</td>
</tr>
<tr>
<td>8</td>
<td>11.46</td>
</tr>
<tr>
<td>9</td>
<td>9.20</td>
</tr>
<tr>
<td>10</td>
<td>10.34</td>
</tr>
<tr>
<td>11</td>
<td>9.03</td>
</tr>
<tr>
<td>12</td>
<td>11.47</td>
</tr>
<tr>
<td>13</td>
<td>10.51</td>
</tr>
<tr>
<td>14</td>
<td>9.40</td>
</tr>
<tr>
<td>15</td>
<td>10.08</td>
</tr>
<tr>
<td>16</td>
<td>9.37</td>
</tr>
<tr>
<td>17</td>
<td>10.62</td>
</tr>
<tr>
<td>18</td>
<td>10.31</td>
</tr>
<tr>
<td>19</td>
<td>8.52</td>
</tr>
<tr>
<td>20</td>
<td>10.84</td>
</tr>
<tr>
<td>21</td>
<td>10.90</td>
</tr>
<tr>
<td>22</td>
<td>9.33</td>
</tr>
<tr>
<td>23</td>
<td>12.29</td>
</tr>
<tr>
<td>24</td>
<td>11.50</td>
</tr>
<tr>
<td>25</td>
<td>10.60</td>
</tr>
<tr>
<td>26</td>
<td>11.08</td>
</tr>
<tr>
<td>27</td>
<td>10.38</td>
</tr>
<tr>
<td>28</td>
<td>11.62</td>
</tr>
<tr>
<td>29</td>
<td>11.31</td>
</tr>
<tr>
<td>30</td>
<td>10.52</td>
</tr>
</tbody>
</table>

The EWMA (exponentially weighted moving average) plots or tracks

\[ z_i = \lambda x_i + (1 - \lambda) z_{i-1} \]

- \( x_i \) is the observation at time \( i \)
- \( 0 < \lambda \leq 1 \) is a constant that governs how much weight is put on historical observations
  - \( \lambda = 1 \): EWMA reduces to the Shewhart
  - Typical values: \( 0.1 \leq \lambda \leq 0.3 \)

With appropriate choice of \( \lambda \), can be made to perform similar to Shewhart or CUSUM
EWMA Chart Example

<table>
<thead>
<tr>
<th>Sample, ( i )</th>
<th>( x_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.45</td>
</tr>
<tr>
<td>2</td>
<td>7.99</td>
</tr>
<tr>
<td>3</td>
<td>9.29</td>
</tr>
<tr>
<td>4</td>
<td>11.66</td>
</tr>
<tr>
<td>5</td>
<td>12.16</td>
</tr>
<tr>
<td>6</td>
<td>10.18</td>
</tr>
<tr>
<td>7</td>
<td>8.04</td>
</tr>
<tr>
<td>8</td>
<td>11.46</td>
</tr>
<tr>
<td>9</td>
<td>9.20</td>
</tr>
<tr>
<td>10</td>
<td>10.34</td>
</tr>
<tr>
<td>11</td>
<td>9.03</td>
</tr>
<tr>
<td>12</td>
<td>11.47</td>
</tr>
<tr>
<td>13</td>
<td>10.51</td>
</tr>
<tr>
<td>14</td>
<td>9.40</td>
</tr>
<tr>
<td>15</td>
<td>10.08</td>
</tr>
<tr>
<td>16</td>
<td>9.37</td>
</tr>
<tr>
<td>17</td>
<td>10.62</td>
</tr>
<tr>
<td>18</td>
<td>10.31</td>
</tr>
<tr>
<td>19</td>
<td>8.52</td>
</tr>
<tr>
<td>20</td>
<td>10.84</td>
</tr>
<tr>
<td>21</td>
<td>10.90</td>
</tr>
<tr>
<td>22</td>
<td>9.33</td>
</tr>
<tr>
<td>23</td>
<td>12.29</td>
</tr>
<tr>
<td>24</td>
<td>11.50</td>
</tr>
<tr>
<td>25</td>
<td>10.60</td>
</tr>
<tr>
<td>26</td>
<td>11.08</td>
</tr>
<tr>
<td>27</td>
<td>10.38</td>
</tr>
<tr>
<td>28</td>
<td>11.62</td>
</tr>
<tr>
<td>29</td>
<td>11.31</td>
</tr>
<tr>
<td>30</td>
<td>10.52</td>
</tr>
</tbody>
</table>

Some Multivariate SPC Methods

- **Hotelling’s $T^2$ (1947)**
  - Stop when statistical distance to observation $T^2 = (X - \mu)'\Sigma^{-1}(X - \mu)$ exceeds threshold $h$
  - Like Shewhart, good at detecting large shifts

- **Lowry et al.’s MEWMA (1992)**
  - Multivariate generalization of univariate EWMA
    - At each time, calculate $z_i = \lambda(x_i - \mu) + (1 - \lambda)z_{i-1}$
    - Stop when $E_i = z_i'\Sigma^{-1}z_i \geq h$

- **Crosier’s MCUSUM (1988)**
  - Cumulates vectors componentwise
  - As with CUSUM, good at detecting small shifts
Crosier’s Multivariate CUSUM

• Crosier (1988) proposed various MCUSUMs; His preferred defines

\[ k \text{ so that } k \Sigma^{-1} k = k^2, \text{ for some distance } k, \text{ and } \]

\[ C_i = \sqrt{(S_{i-1} + X_i - \mu)' \Sigma^{-1} (S_{i-1} + X_i - \mu)}. \]

Calculate

\[ \begin{cases} 
S_i = 0 \text{ if } C_i \leq k \\
S_i = (S_{i-1} + X_i - \mu)(1 - k / C_i) \text{ if } C_i > k 
\end{cases} \]

where \( S_0 = 0 \) and \( k > 0. \)

Let \( Y_i = \sqrt{S_i' \Sigma^{-1} S_i} \) and stop when \( Y_i > h. \)
• Motivation: In both industrial SPC and in biosurveillance, goal is to detect anomalies

• In industrial setting, control charts used to monitor production and test for a change level of quality
  – Have parameter(s) of quality characteristic shifted?

• In biosurveillance, goal is to monitor for indications of changes in population health
  – Has distribution of leading indicators shifted in some meaningful (i.e., worrisome) way?
Issue: SPC Methods Don’t Translate Directly to Biosurveillance Problem

- Dependent data
  - Industrial methods assume independence
- Nonstationary data
  - No control over “in-control” distribution
- Systematic effects
  - Seasonal, day-of-the-week and other effects in data
- Transient “out-of-control” conditions
  - Outbreaks/attacks begin, peak, and subside
- Vague alternative hypotheses
  - Detect only bioterrorism or natural diseases too?
  - Which diseases and/or outbreak manifestations?
Classical epidemiology is largely retrospective while biosurveillance is a prospective problem. Retrospective is hard enough while prospective detection provides new challenges.

Original map by Dr. John Snow showing clusters of cholera cases in London epidemic of 1854.\[1\]

Identify as early as possible when an outbreak occurs...

Lots of New Methods Have Been Proposed, Most Illustrated with Data

• For example:
  – RTR-based methods: see Fricker and Chang, 2008
  – CUSUM-based methods with adaptive regression: see Fricker et al., 2008
  – Directional MEWMA and MCUSUM: see Joner et al., 2008, and Fricker, 2007
  – Bayesian network-based methods: see for example Rolka et al., 2007
  – Distance-based methods: see Forsberg et al., 2006
  – Bayesian dynamic models: see, for example, Sebastiani et al., 2006
  – Wavelet-based methods: see Shmueli, 2005, Zhang et al., 2003, etc.
  – Point process model-based methods: see Brookmeyer and Stroup, 2004
  – Rule-based methods: see, for example, Wong, 2003
  – Hidden Markov models: see Le Strat and Carrat, 1999

• And some methods are in use, such as:
  – SatScan: see, for example, Kulldorff 2001 and subsequent literature
  – GLM-based methods: see, for example, Kleinman et al., 2004
  – EARS’ C1, C2, and C3 methods: see Hutwagner, et al., 2005
• Though many methods have been proposed, the sheer plethora raises questions:
  – Under what conditions do the various methods work best?
  – Is a method more sensitive than others to detecting a particular type of outbreak?
  – Conversely, is a method overly sensitive to particular assumptions about the data?
  – How to compare the methods to determine?
Side Comment: More Sophisticated Methods Aren’t Always Better

• Common criticism of traditional SPC methods is jump change in mean is artificial
• Chang and Fricker (1999) assessed what happens when mean is monotonically increasing
  – Compared performance of standard SPC methods (CUSUM and EWMA) to likelihood ratio test (LRT)
• Result: *LRT explicitly designed for the problem often outperformed by SPC methods designed for jump change in mean*

Issue: Looking for Everything Means It’s Harder to Find Any One Thing
It’s a Hard Problem Even When You Know What You’re Looking For...
• To greatest extent possible, specify characteristics of events to be detected
  – Where’s Waldo: Only look for red and white stripes
  – In biosurveillance, only signal when rates of disease increase
    • E.g., MCPHD tell me EARS signals on decreases

• Think about it as follows:
  – Restricted focus should decrease false positives
  – Thus, can lower thresholds for greater detection
    • In SPC terms, restricting focus to increases results in smaller $\text{ARL}_1$ for fixed $\text{ARL}_0$ or larger $\text{ARL}_0$ for same $\text{ARL}_1$
• Dembeck, Kortepeter, and Pavlin (2007) identified eleven “clues to a deliberate epidemic”:
  1. A highly unusual event with large numbers of casualties
  2. Higher morbidity or mortality than expected
  3. Uncommon disease
  4. Point-source outbreak
  5. Multiple epidemics
  6. Lower attack rates in protected individuals
  7. Dead animals
  8. Reverse or unnatural spread
  9. Unusual disease manifestation
  10. Downwind plume pattern
  11. Direct evidence

→ Perhaps a starting point?
Performance Comparison #1

- $F_0 \sim N(0,1)$ and $F_1 \sim N(\delta,1)$
Performance Comparison #2

- $F_0 \sim N(0, 1)$ and $F_1 \sim N(0, \sigma^2)$
Performance Comparison #3

- $F_0 \sim \mathcal{N}(0, 1)$
- $F_1 \sim \text{[Graph]}$

<table>
<thead>
<tr>
<th>Distribution #</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>0.0</td>
<td>0.435</td>
<td>0.6</td>
<td>0.715</td>
<td>0.8</td>
<td>0.866</td>
<td>0.917</td>
<td>0.954</td>
<td>0.980</td>
<td>0.995</td>
<td>1.0</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>1.0</td>
<td>0.9</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
<td>0.4</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
<td>$\sim$ 0</td>
</tr>
</tbody>
</table>
• $F_0 \sim N_2((0,0)^T, \mathbf{I})$

• $F_1$ mean shift in $F_0$ of distance $\delta$
Performance Comparison #5

- $F_0 \sim N_2((0,0)^T, I)$
- $F_1 \sim N_2((0,0)^T, \sigma^2 I)$

Graph showing ARL vs $\sigma^2$ with different markers representing different methods:
- Hotelling
- MEWMA
- MCUSUM
- RTR
Examples: “One-sided” MEWMA and MCUSUM

• Joner et al. (2008) modified the MEWMA to only signal increases in the mean vector:
  \[ z_i = \max \left( 0, \lambda (x_i - \mu) + (1 - \lambda) z_{i-1} \right) \]

• Similarly, Fricker (2007) modified Crosier’s MCUSUM by using
  \[ S_{i,j} = \max \left[ 0, (S_{i-1,j} + X_{i,j} - \mu_j)(1 - k / C_i) \right] \] for
  \[ S_i = (S_{i,1}, \ldots, S_{i,d}) \]

• However, it’s not as simple as turning omni-directional methods into “one-sided” tests
  – The tests above are better for the biosurveillance problem
  – But more precise alternatives would allow even more focused (i.e., more sensitive/powerful) tests to be developed

Issue: What Are the Appropriate Metrics for Biosurveillance?

• SPC methods are **sequential hypothesis tests**
  – At each time period, do a simple hypothesis test on a set of data – but then repeat test over and over
  – “Many papers have addressed the problem of on-line surveillance, but the mistake of not noting the sequential type of decision situation is quite common.”[1]

• Issue: Concepts from standard hypothesis testing – such as sensitivity and specificity – do not translate well to this type of problem
  – Yet most common biosurveillance metrics are “sensitivity, specificity, and timeliness”

Sensitivity and Specificity Classically Defined

• Sensitivity and specificity are statistical metrics for binary classification tests
• Consider a test for a disease applied to both sick and health people where test outcome can be positive (sick) or negative (healthy)
  • **Sensitivity**: a measure of how well a test correctly classifies sick people as sick
  • **Specificity**: a measure of how well a test correctly classifies healthy people as healthy
Calculating the Sensitivity and Specificity of a Binary Test

- E.g., administer $N$ independent tests and classify each outcome:
  - True positives (TP) are sick people correctly diagnosed
  - False positives (FP) are healthy people wrongly diagnosed
  - True negatives (TN) are healthy people correctly diagnosed
  - False negatives (FN) are sick people wrongly diagnosed

<table>
<thead>
<tr>
<th>Test Outcome</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sick</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td></td>
<td>(Type II error)</td>
<td>(Type I error)</td>
</tr>
<tr>
<td>Healthy</td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

Sensitivity = \[
\frac{\#TP}{\#TP + \#FN}
\]

Specificity = \[
\frac{\#TN}{\#TN + \#FP}
\]
ROC Curves Depict How Sensitivity and Specificity Trade-Off for a Test

• With a classical hypothesis test, with one observation or set of observations we must decide whether Ho or Ha is true
• ROC curve shows relationship between sensitivity and specificity for all choices of a “threshold”

\[
\text{Sensitivity} = \Pr(\text{reject Ho} \mid \text{Ha})
\]

\[
1-\text{Specificity} = \Pr(\text{accept Ha} \mid \text{Ho})
\]
But What Happens When Hypothesis Test Repeatedly Applied?

• Rather than administer the test to $N$ independent people, what if we kept administering the test to the same person over and over?
  – What does sensitivity and specificity mean now?
  – Can we use an ROC curve to describe test performance?

• Defining sensitivity of a surveillance system test:

  “The sensitivity of a surveillance system can be considered on two levels. First, at the level of case reporting, sensitivity refers to the proportion of cases of a disease (or other health-related event) detected by the surveillance system. Second, sensitivity can refer to the ability to detect outbreaks, including the ability to monitor changes in the number of cases over time.”

• “Sensitivity is defined as the number of days with true alarms divided by the number of days with outbreaks.”

• “Sensitivity can be assessed by estimating the proportion of cases of a disease or health condition detected by the surveillance system. Sensitivity can also be considered as the ability of the system to detect unusual events.”

• “Sensitivity is the probability that a public health event of interest will be detected in the data given the event really occurred.”

• “Sensitivity is the probability of an alarm given an outbreak.”


Two Methods with Same Specificity But Very Different Performance

• Consider the following performance of two methods:

![Graphs showing outbreak signals from Method 1 and Method 2](image)

- Based on the table → both have sensitivity equal to 4/15
- But Method 2 is clearly better

<table>
<thead>
<tr>
<th>TABLE 2</th>
<th>Classification Used to Calculate Sensitivity and PVP in Example</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Condition present</td>
</tr>
<tr>
<td>Detected</td>
<td>4</td>
</tr>
<tr>
<td>No signal</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>15</td>
</tr>
</tbody>
</table>

“Evaluation by the significance level, power, specificity, and sensitivity which is useful for a fixed sample is not appropriate in a surveillance situation without modification since they have no unique value unless the time period is fixed. Also, a formulation of an optimality criterion for surveillance must naturally take into account the delay time in detection, since the aim of a surveillance method is quick detection.”

Consider the Following Relevant Metrics for Sequential Testing

• If we keep applying the test to a healthy person over and over we will *eventually* get a false positive
  – One useful measure of performance is the expected time between false positives
    • The larger the better

• Assume the repeated testing used to quickly identify when a healthy person gets sick
  – Another useful measure is the expected time from when the person gets sick until the first positive test
    • The smaller the better
• Two aspects of biosurveillance differ from industrial SPC practice
  – Because outbreaks are transient, it is possible for the algorithm to miss them
    • So, it’s not clear how to calculate $\text{ARL}_1$
    • Not the case in industrial SPC, assuming persistent out-of-control conditions
  – Often algorithms not re-set after a signal
    • Sequences or clusters of signals taken as stronger evidence of outbreak
    • As a result, even less clear how to calculate $\text{ARL}_1$
Many Metrics Have Been Proposed in the Biosurveillance Literature

• “Substantially more metrics have been proposed in the public health surveillance literature than in the industrial monitoring literature.”[1]

• Examples:
  – Sensitivity, specificity, and timeliness
  – Sensitivity and predictive value positive
  – Recurrence interval
  – Area under the ROC curve, activity monitoring operating characteristic (AMOC) curve, and free response operating characteristic (FROC) curve
  – Average run length (ARL), average overlapping run length (AORL), average time to signal given an outbreak
  – Expected delay and conditional expected delay (CED)
  – Probability of successful detection (PSD)
  – Average time between signal events (ATBSE) and average signal event length (ASEL)

• The field needs a set of standard metrics
  – Without them, it’s virtually impossible to synthesize and compare results across the literature

• Recommend run length-based metrics augmented with a metric for missed outbreaks
  – Retrospective and prospective methods require different metrics

• Perhaps different metrics appropriate for systems that reset vs. not reset after a signal
  – Which then leads to questions about what the system is monitoring for: natural outbreaks versus bioterrorism…
• It’s a question about the primary purpose of a biosurveillance system
• Basic issue:
  – A system designed to detect bioterrorism will be useful for detecting natural diseases
  – But a system focused on natural disease outbreaks could miss bioterrorism
• The problem: If a system that is signaling during a natural disease outbreak is not re-set, then it cannot detect bioterrorism
  – The smoke alarm that goes off every time you use the oven is of little use detecting real fires when you’re cooking
A Smart Bioterrorist Would Attack During the Flu Season

• I’m unclear on the primary purpose
  – But the answer has implications for both choice of appropriate metrics and how the biosurveillance system is operated

• E.g., if the goal is bioterrorism detection:
  – During natural disease outbreak, should revise background incidence rate so system can look for further outbreaks
  – If so, it also implies re-setting the detection algorithm(s) after each signal
“…a general challenge for all biosurveillance research is to develop improved methods for evaluating detection algorithms in light of the fact that we have little data about outbreaks of many potential diseases that are of concern.”

“Evaluations and comparisons of statistical performance in public health surveillance often involve the use of real surveillance over a past time period of interest. The outbreak locations in time are either assumed to be known or outbreaks are artificially superimposed on the data. As pointed out by Woodall (2006), this is rarely, if ever, the case in the industrial literature where case study-type data are used only to illustrate the application of methods, not to evaluate statistical performance.”

"Reliance on the use of Monte Carlo simulation in the field of Statistics is well known. It has been this author’s experience that the technique is undervalued in the field of Public Health because it has previously not been required."[1]

At issue is breaking out of the “my data is unique” and “only real data is valid” paradigms

Monte Carlo can:
- Facilitate evaluating algorithms across many scenarios
- Eliminate unneeded/distracting real world complexities
- Allow clean and clear comparisons of algorithms
- Make it easier to get at generalizable conclusions/results

Sub-Issue: Must Be Able to Well Characterize Biosurveillance Data

- Valid Monte Carlo simulation depends on being able to appropriately characterize and simulate biosurveillance data
  - “ Appropriately” does not mean “perfectly”
  - But must understand important features of (types of) biosurveillance data
    - Both systematic and probabilistic
- Utility of Monte Carlo methods often in understanding broad conditions under which methods work better or worse
- Solution: Basic research with real data
• Little is known about which methods work best and under what conditions
  – Emphasis in biosurveillance literature is on presenting new methods illustrated on a specific set of data
  – Use of unique data does not permit comparisons across papers
  – Few papers make comparisons between methods

• In contrast, QC/SPC literature has long history of comparing methods under conditions that can be replicated
“The body of literature on health-related surveillance is smaller than that on industrial surveillance, and is somewhat less mathematical in nature.”


Solution: Foster a Culture of Studying Algorithmic Performance

- Recommend encouraging *on-going* research that conducts comparisons between methods under various conditions
- Also, promote research into characterizing data (normal background and outbreak) so that comparisons can be made on *simulated* data
- In my opinion, competitions (e.g., DARPA-sponsored Bio-ALIRT competition, 2001-2004) of limited utility
  - Problem does not lend itself to a single “solution” arising from a competition
  - Use of actual data interesting, but best performer on that data does not mean results are generalizable
Example: Comparing EARS to Alternative Based on CUSUM\cite{1}

- Early Aberration Reporting System (EARS)
  - Designed to be a drop-in surveillance system
  - Available on the web, so increasingly being used as standard health surveillance system

- EARS’ algorithms:

  \[ C_1(t) = \frac{Y(t) - \bar{Y}_1(t)}{s_1(t)} \]

  \[ C_2(t) = \frac{Y(t) - \bar{Y}_3(t)}{s_3(t)} \]

  \[ C_3(t) = \sum_{i=t}^{t-2} \max [0, C_2(i) - 1] \]

  - Sample statistics calculated from previous 7 days’ data
  - Stop when statistic > 3

  - Sample statistics calculated from 7 days’ of data prior to 2 day lag
  - Stop when statistic > 3

  - Stop when statistic > 2

Alternative: CUSUM on Residuals from “Adaptive Regression”

• Adaptive regression: regress a sliding baseline of observations on time relative to current observation
  – I.e. regress $Y(t-1),...,Y(t-n)$ on $n,...,1$

• Calculate standardized residuals from one day ahead forecast, $X(t) = R(t)/\sigma_Y$, where

$$R(t) = Y(t) - \left[ \hat{\beta}_0 + \hat{\beta}_1 \times (n+1) + \hat{\beta}_j \right]$$

• CUSUM:

$$S(t) = \max\left[0, S(t-1) + X(t) - k\right]$$

with

$$k = \frac{1}{2} \sqrt{\frac{(n+2)(n+1)}{n(n-1)}}$$
Comparison Methodology

• Generate synthetic data:

\[ Y(t) = \max \left( 0, \left[ c + s(t) + d(t) + Z(t) + o(t) \right] \right) \]

• Scenarios:

\[
\begin{array}{c|c|c|c|c}
\text{None} & \text{Small} & \text{Large} \\
\hline
A & 0 & 20 & 80 \\
\sigma & n/a & 10 & 30 \\
\end{array}
\]

Large count: \( c = 90 \)

\[
\begin{array}{c|c|c|c|c}
\text{None} & \text{Small} & \text{Large} \\
\hline
\mu, \sigma & 0 & 2 & 6 \\
\sigma & n/a & 1.0, 0.5 & 1.0, 0.7 \\
\end{array}
\]

Small count: \( c = 0 \)

• Outbreaks
  – Linear increase & decrease
  – Characterized by duration and magnitude
Synthetic Data: Outbreaks?
Some Large Count Results
Shewhart-based Methods Not Suited for this Problem?
• CUSUMs based on adaptive regression with longer baselines performed best
• CUSUMs outperformed EARS’ methods
  – Seemingly due to Shewhart design and additional data used in adaptive regression
• Suggests “drop in” strategy of starting with CUSUM with 7-day baseline
  – As time progresses, increase baseline until long enough to allow it to slide
Issue: Developing Methods That Support Both EED and SA

- Methods that both identify and track changes in disease patterns desirable
  - Is an outbreak/attack likely occurring?
  - If so, where and how is it spreading?
- Most methods focus on either early event detection or spatial clustering using aggregated (i.e., daily count) data
- Ideal: Method that uses individual-level data in (near) real time
• ER patients come from surrounding area
  – On average, 30 per day
    • More likely from closer distances
  – Outbreak occurs at (20,20)
    • Number of patients increase linearly by day after outbreak
A Couple of Major Assumptions

• Can geographically locate individuals in a medically meaningful way
  – Data not currently available
  – Non-trivial problem

• Data is reported in a timely and consistent manner
  – Public health community working this problem, but not solved yet

• Assuming the above problems away…
• Construct kernel density estimate (KDE) of “normal” disease incidence using $N$ historical observations
• Compare to KDE of most recent $w+1$ obs

But how to know when to signal?
Solution: Repeated Two-Sample Rank (RTR) Procedure

- Sequential hypothesis test of estimated density heights
- Compare estimated density heights of recent data against heights of set of historical data
  - Single density estimated via KDE on combined data
- If no change, heights uniformly distributed
  - Use nonparametric test to assess

• Let \( X_i = \{ X_{1i}, X_{2i} \} \) be a sequence of bivariate observations
  – E.g., latitude and longitude of a case
• Assume a historical sequence \( X_{1-N}, \ldots, X_0 \) is available
  – Distributed iid according to \( f_0 \)
• Followed by \( X_1, X_2, \ldots \) which may change from \( f_0 \) to \( f_1 \) at any time
• Densities \( f_0 \) and \( f_1 \) unknown
• Consider the $w+1$ most recent data points
• At each time period estimate the density

$$f_n(x) = \begin{cases} 
\frac{1}{N+n} \sum_{i=1-N}^{n} k_h(x, X_i), & n < w+1 \\
\frac{1}{N+w+1} \sum_{i=n-w-N-1}^{n} k_h(x, X_i), & n \geq w+1 
\end{cases}$$

where $k$ is a kernel function on $\mathbb{R}^2$ with bandwidth set to $h_i = \sigma_i \left(1/(N+w+1)\right)^{1/6}$
Illustrating Kernel Density Estimation (in one dimension)
The density estimate is evaluated at each historical and new point

- For \( n < w+1 \)
  \[
  \hat{f}_n(X_{1-N}), \ldots, \hat{f}_n(X_0), \hat{f}_n(X_1), \ldots, \hat{f}_n(X_n)
  \]
  historical observations \hspace{2cm} new observations

- For \( n \geq w+1 \)
  \[
  \hat{f}_n(X_{n-w-N-1}), \ldots, \hat{f}_n(X_{n-w-1}), \hat{f}_n(X_{n-w}), \ldots, \hat{f}_n(X_n)
  \]
  historical observations \hspace{2cm} new observations
• **Theorem:** If $X_i \sim F_0$, $i \leq n$, the RTR is asymptotically distribution free
  – I.e., the estimated density heights are exchangeable, so all rankings equally likely
  – Proof: See Fricker and Chang (2009)

• Means can do a hypothesis test on the ranks each time an observation arrives
  – Signal change in distribution first time test rejects
Comparing Distributions of Heights

• Compute empirical distributions of the two sets of estimated heights:

\[
\hat{J}_n(z) = \frac{1}{w+1} \sum_{i=n-w}^{n} I \left\{ f_n(X_i) \leq z \right\},
\]

\[
\hat{H}_n(z) = \frac{1}{N} \sum_{i=n-w-N-1}^{n-w-1} I \left\{ f_n(X_i) \leq z \right\}
\]

• Use Kolmogorov-Smirnov test to assess:

\[
S_n = \max_z \left| \hat{J}_n(z) - \hat{H}_n(z) \right|
\]

– Signal at time \( t = \min \{ n : S_n > c \} \)
Illustrating Changes in Distributions (again, in one dimension)
• At signal, calculate optimal kernel density estimates and plot pointwise differences

\[ \Delta_n(x) = \hat{h}_n(x) - \hat{g}_n(x) \]

where

\[ \hat{h}_n(x) = \frac{1}{w + 1} \sum_{i=n-w}^{n} k_h(x, X_i) \]

\[ \hat{g}_n(x) = \frac{1}{N} \sum_{i=n-w-N-1}^{n-w-1} k_h(x, X_i) \]

and

\[ h_i = \sigma_i \left( \frac{1}{w + 1} \right)^{1/6} \text{ or } h_i = \sigma_i \left( \frac{1}{N} \right)^{1/6} \]
• Assess performance by simulating outbreak multiple times, record when RTR signals
  – Signaled middle of day 5 on average
  – By end of 5\textsuperscript{th} day, 15 outbreak and 150 non-outbreak observations
  – From previous example:
Same Scenario, Another Sample

Daily Data

Outbreak Signaled on Day 5 (obs’n # 165)
Another Example

- Normal disease incidence ~ $N(\{0,0\}^t, \sigma^2 I)$ with $\sigma=15$
  - Expected count of 30 per day
- Outbreak incidence ~ $N(\{20,20\}^t, 2.2d^2 I)$
  - $d$ is the day of outbreak
  - Expected count is $30+d^2$ per day

Unobserved outbreak distribution

Daily data

Outbreak signaled on day 1 (obs’n # 2)
(On average, signaled on day 3-1/2)
• Normal disease incidence $\sim N(\{0,0\}^t,\sigma^2 I)$ with $\sigma=15$
  – Expected count of 30 per day
• Outbreak sweeps across region from left to right
  – Expected count is 30+64 per day

Outbreak signaled on day 1 (obs’n # 11)
(On average, signaled 1/3 of way into day 1)
Advantages and Disadvantages

• Advantages
  – Methodology supports both biosurveillance goals: early event detection and situational awareness
  – Incorporates observations sequentially (singly) so can be used for real-time biosurveillance
    • Most other methods use aggregated data

• Disadvantage?
  – Can’t distinguish increase distributed according to $f_0$
    • Won’t detect a general increase in background disease incidence rate
      – E.g., Perhaps caused by an increase in population
      – In this case, advantage not to detect
    • Unlikely for bioterrorism attack?
• Classical hypotheses tests set up so that Type I error rate explicitly controlled
• Thus, Type I error is usually the more serious of the two possible errors
  – Example: In criminal trials, possible errors are either convicting an innocent person or letting a guilty person go free
    • Our society feels sending an innocent person to prison is the more serious error
    • Hence, the “null hypothesis” is a person is presumed innocent and must be proven guilty
• Type II error is then a function of the observed alternative (and test design)
Errors in Biosurveillance

• In biosurveillance, the possible errors are
  – Failing to detect an outbreak/attack (false negative)
  – Incorrectly signaling when there is no outbreak/attack (false positive)

• Presumably, the first is a more significant error
  – Suggests biosurveillance systems should be structured to presume an outbreak exists unless proven otherwise

• Trial example: What if the person incorrectly let free would release smallpox in US?
  – Should the null still be innocent until proven guilty – or should it now be guilty until proven innocent?
• But always assuming an outbreak exists unless proven otherwise is impractical:
  – Would consume far too many resources
  – How to prove everything is normal?

• Alternate hypothesis testing design approach: Make the alternative hypothesis the outcome that requires empirical proof
  – But with Type II error so serious, that implies must have test with high sensitivity
  – Equivalent condition for sequential SPC methods, must have low ARL₁s
Current practice seems to try to mitigate problem by lowering detection thresholds to make detection time as low as possible

- Often without regard to the fact that making algorithms more sensitive to detecting outbreaks also results in more false positives
  
  - “…most health monitors… learned to ignore alarms triggered by their system … due to the excessive false alarm rate that is typical of most systems - there is nearly an alarm every day!”[1]

Alternatives:

- Develop more sensitive methods (i.e., that achieve same ARLO₁s for larger ARLO
- Use existing tests/systems more selectively

Possible Solution: Make Biosurveillance Systems “Tunable”

• Can’t watch for everything, everywhere, all the time and still maintain a tolerable false positive error rate
  – Instead, design systems to be “tunable”
• One approach: set detection thresholds to make most likely events most detectable
  – As threats change, can change thresholds
  – Also, set thresholds so that Type I error rate constrained at tolerable level
• A preview of my Wednesday talk…
Optimizing a County-level System
Problem Set-up

- Regions (counties) are spatially independent
- Biosurveillance system monitoring standardized residuals from an “adaptive regression” model using Shewhart charts
  - Model removes systematic effects in the data
  - Result: Reasonable to assume $F_0=\mathcal{N}(0,1)$
- An outbreak will result in a 2-sigma increase in the mean of the residuals, so $F_1=\mathcal{N}(2,1)$
- Then, maximize probability of detection subject to constraint on average number of false signals:

\[
\max_{\hat{h}} \sum_i \left[ 1 - F_1(h_i) \right] p_i \\
\text{s.t.} \sum_i \left[ 1 - F_0(h_i) \right] \leq \kappa
\]

Fricker, R.D., Jr., and D. Banschbach, Optimizing Biosurveillance Systems that Use Threshold-based Event Detection Methods, in submission.
Optimizing a County-level System
Thresholds Chosen as a Function of Probability of Attack

Counties with low probability of attack → high thresholds
- Unlikely to detect attack
- Few false signals

Counties with high probability of attack → lower thresholds
- Better chance to detect attack
- Higher number of false signals
In Summary…

- Goal was to discuss some current issues in biosurveillance detection algorithms
  - Informed by an industrial SPC viewpoint
- In my opinion, biosurveillance research has yet to fully tap industrial SPC literature and expertise
- Other disciplines have much to offer as well:
  - Operations research – optimizing biosurveillance system performance is a non-trivial problem
  - Systems engineering – these are complex systems that require careful design
  - Game theory – in a bioterrorism context, there is an autonomous, willful adversary to be accounted for
Biosurveillance is a Hard Problem

• Posed more problems than solutions
• Purpose was to highlight some of the open issues, including
  – Lack of standard evaluation methods and metrics in the literature
  – Need to move beyond inappropriate metrics
  – Benefits of better defining events to be detected
  – Utility of using more Monte Carlo methods for algorithm evaluation
But if all I’ve done is demonstrate how sequential tests differ from classical hypothesis testing, then I declare victory!
Selected References

Background Information:


Detection Algorithm Development and Assessment:


Biosurveillance System Optimization:

• Fricker, R.D., Jr., and D. Banschbach, Optimizing Biosurveillance Systems that Use Threshold-based Event Detection Methods, in submission.

→ See http://faculty.nps.edu/rdfricke/Biosurveillance.htm for links to all papers cited in this talk