SAND2011-3904C

Dynamic Casualty Estimation from Biosurveillance Data

Karen Cheng, David Crary

Applied Research Associates, Inc.

Jaideep Ray, Cosmin Safta, Sophia Lefantzi

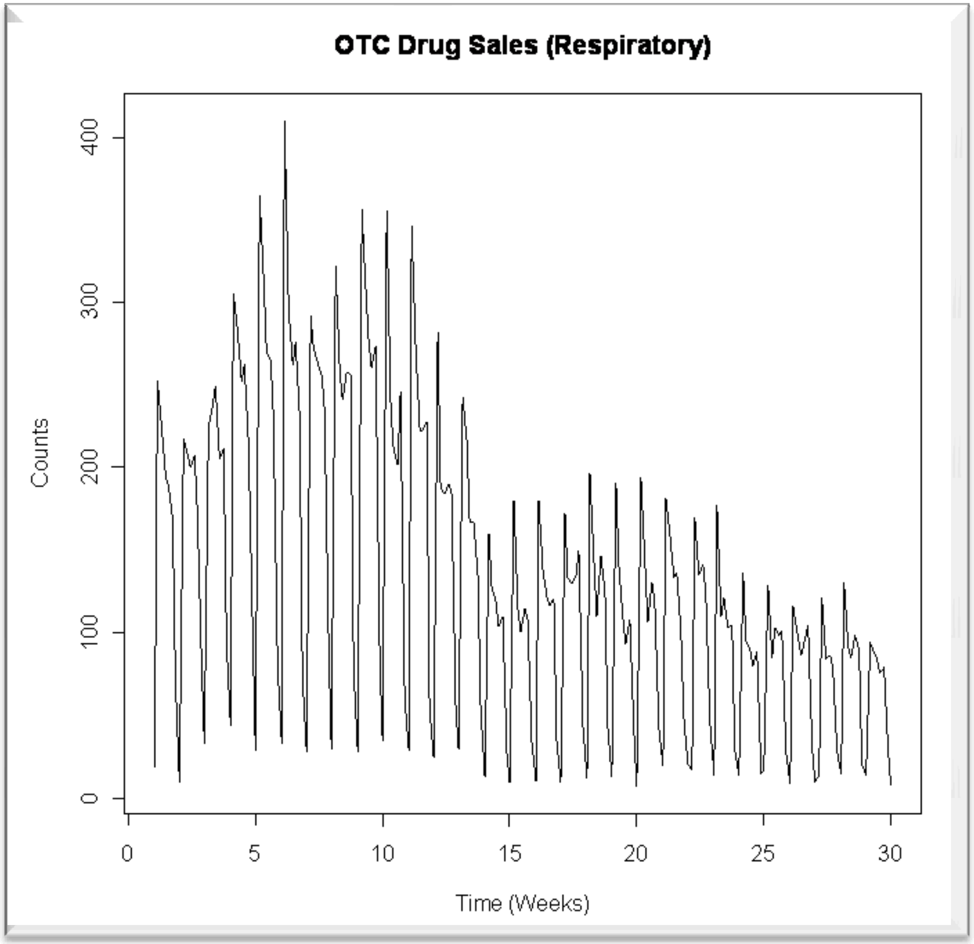
Sandia National Laboratories

Contact Info: Ms. Karen Cheng, kcheng@ara.com, 703-816-8886 x 138

Estimating Disease Parameters Using Biosurveillance Data

- **Biosurveillance Data:** Time series (counts/day) related to syndromic data
  - Our present work uses ICD-9 codes from hospitals as well as disease models to simulate a epidemic or bioterrorist event within a population
- **Goal of this Work:** To develop statistical techniques to characterize ongoing epidemics from initial/partial biosurveillance data
  - Estimate disease parameters : index cases, time of infection, infection rate
  - Do so early in the outbreak, with minimal data
  - Quantify the confidence in the estimates of disease parameters
    - Useful for bracketing outcomes in forward prediction
- **Motivation:**
  - To provide initial conditions for disease models, to be used for planning medical interventions, resource allocation etc.

Biosurveillance Data is Complex



- Biosurveillance data shows a broad range of structures (spikes, weekly cycles, seasonal trends, random walk properties, missing data)
- “Normal” cycles and trends must be ***discovered*** dynamically
- Any outbreak will be superimposed on this background, and must be detected and subtracted from the background for analysis
- Background must be accurately modeled to differentiate outbreak counts from background counts in the data

Bloom, Buckeridge and Cheng, Jour. Am. Med. Informatics. Assoc. (2007) conclude: “7 day moving average filter suppress exactly the short scale features that were the intended object of study” More sophisticated methods are required.

Steps Used in Our Analysis

- The components of the procedure are:
  - **Background Modeling/Outbreak Detection** from time-series data
    - Data contains the outbreak and background/endemic morbidity
  - **Extraction** of the outbreak from the background
    - Endemic component needs to be separated from the epidemic component
  - **Characterization** of the outbreak
    - Estimation of index cases, time/rate of infection
  - **Identification** of the outbreak
    - What was the disease that caused it, given a few competing guesses

Steps for Detection and Characterization

- **Background Modeling/Outbreak Detection** from time-series data
  - Data contains the outbreak and background/endemic morbidity
- **Extraction** of the outbreak from the background
  - Endemic component needs to be separated from the epidemic component
- **Characterization** of the outbreak
  - estimation of index cases, time/rate of infection
- **Identification** of the outbreak
  - What was the disease that caused it, given a few competing guesses

Modeling the Background

“Observation w/ noise” $x_t = \mu_t + \gamma_t + \varepsilon_t \quad \varepsilon_t \sim N(0, \sigma_\mu^2)$

“Random Walk” $\mu_{t+1} = \mu_t + \nu_t + \xi_t \quad \xi_t \sim N(0, \sigma_\xi^2)$

“Cyclic Term” $\gamma_{t+1} = -(\gamma_t + \gamma_{t-1} + \ldots + \gamma_{t-5}) + \omega_t \quad \omega_t \sim N(0, \sigma_\omega^2)$

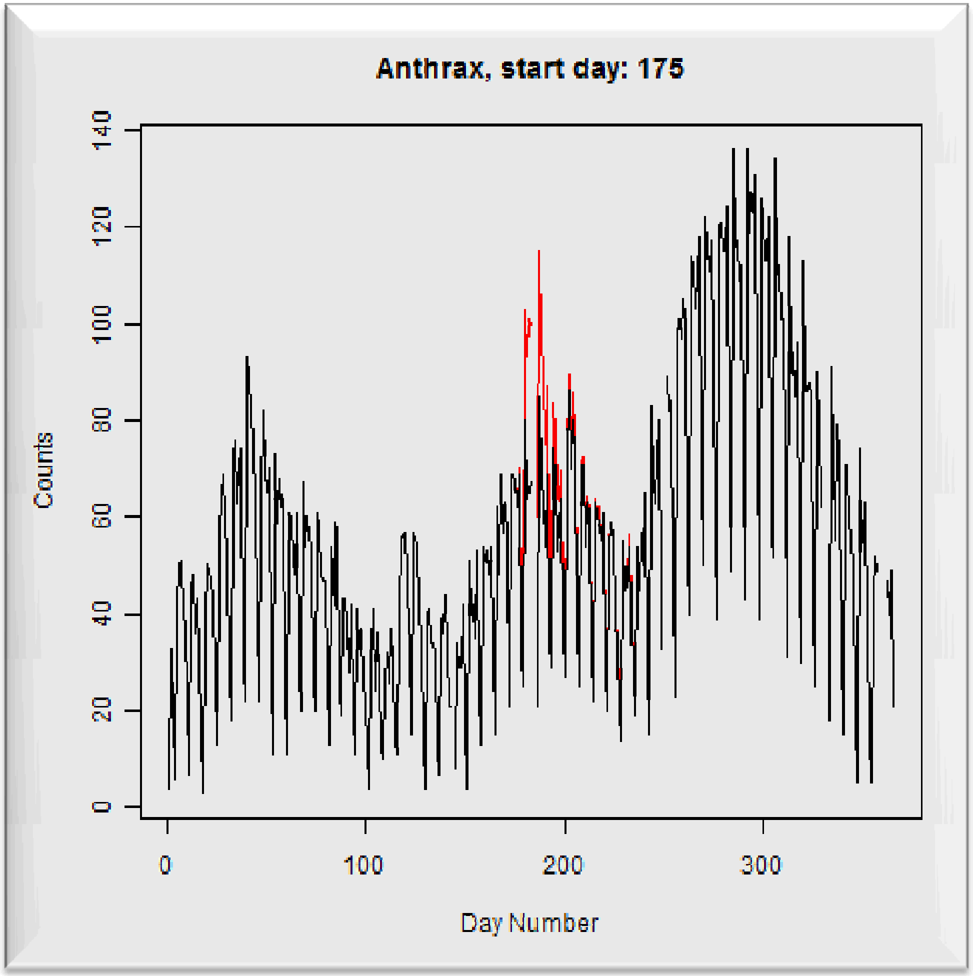
- Background Model included a random walk term for long term trends, a zero mean weekly cycle, and additive noise
- Model is fit to data by MLE techniques using Kalman filter to calculate the likelihood
- Kalman filter provides both 1-day ahead prediction and the prediction uncertainty

This model provides the basis for both statistical anomaly detection and background subtraction capabilities

Test of Anomaly Detection Using Anthrax Outbreak Data

- Background data is from Miami of daily counts of ILI-related codes:
  - 487.0 Influenza with Pneumonia
  - 487.1 Influenza with other respiratory manifestations
  - 487.2 Influenza with other manifestations
- Total outbreak size is 500
  - Anthrax outbreak is calculated using a realistic model with dose dependent incubation time ("Wilkening A2" model)
  - Time to seek care model is also included in the model
- Detection threshold set to 3σ
  - Kalman filter determines one-step ahead prediction $\hat{x}_{t+1}$, as well as the error in this prediction $\hat{\sigma}_{t+1}$
  - Detection occurs if standardized residual $(x_{t+1} - \hat{x}_{t+1}) / \hat{\sigma}_{t+1} > 3$

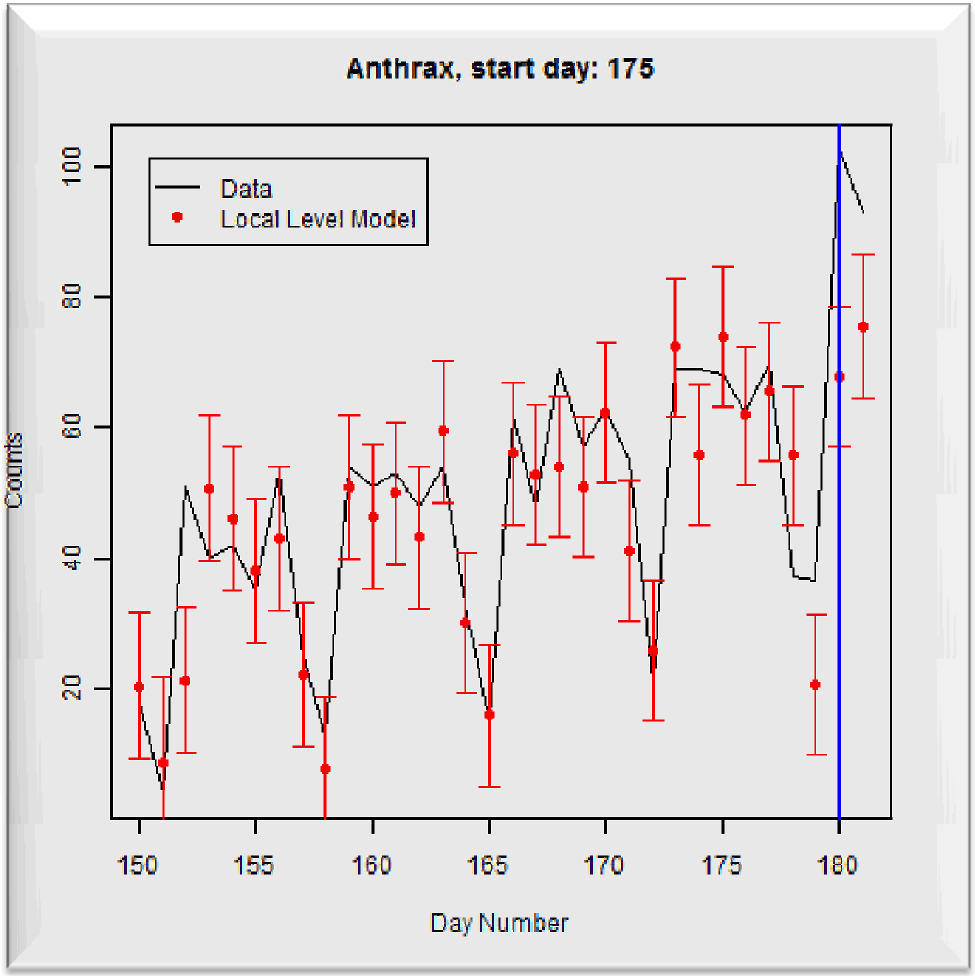
Anthrax Data: Start Day = 175



- Background: ILI ICD-9 codes from Miami data
- Red Line: Calculated anthrax outbreak from Wilkening A2 model, plus visit delay; 500 index cases

Can we detect an anomaly in this noisy data, and how early?

Anthrax Data: Start Day= 175 (Detail)



- Details show prediction (red dots) along with estimates in prediction
- Blue line shows 3σ detection on day 5

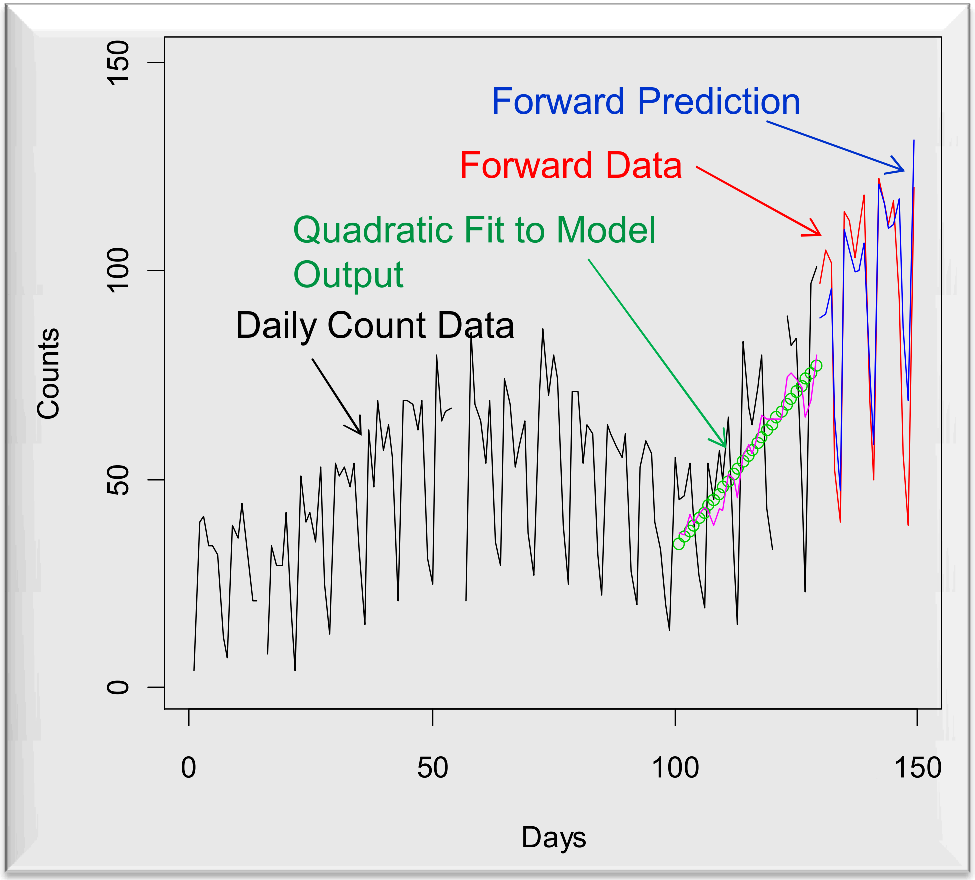
Model provides a robust method for detecting counting anomalies in a statistical framework

Steps for Detection and Classification

- The components of the procedure are:
  - **Background Modeling/Outbreak Detection** from time-series data
    - Data contains the outbreak and background/endemic morbidity
  - **Extraction** of the outbreak from the background
    - Endemic component needs to be separated from the epidemic component
  - **Characterization** of the outbreak
    - estimation of index cases, time/rate of infection
  - **Identification** of the outbreak
    - What was the disease that caused it, given a few competing guesses

Forward Prediction of Background

- Goal: subtraction of background model from data, after detection, to isolate epidemic
- Classification module
  - Only fits epidemic curve
  - Requires an accurate subtraction of background from data
- A the time of a detection, background counts must be accurately predicted into the future



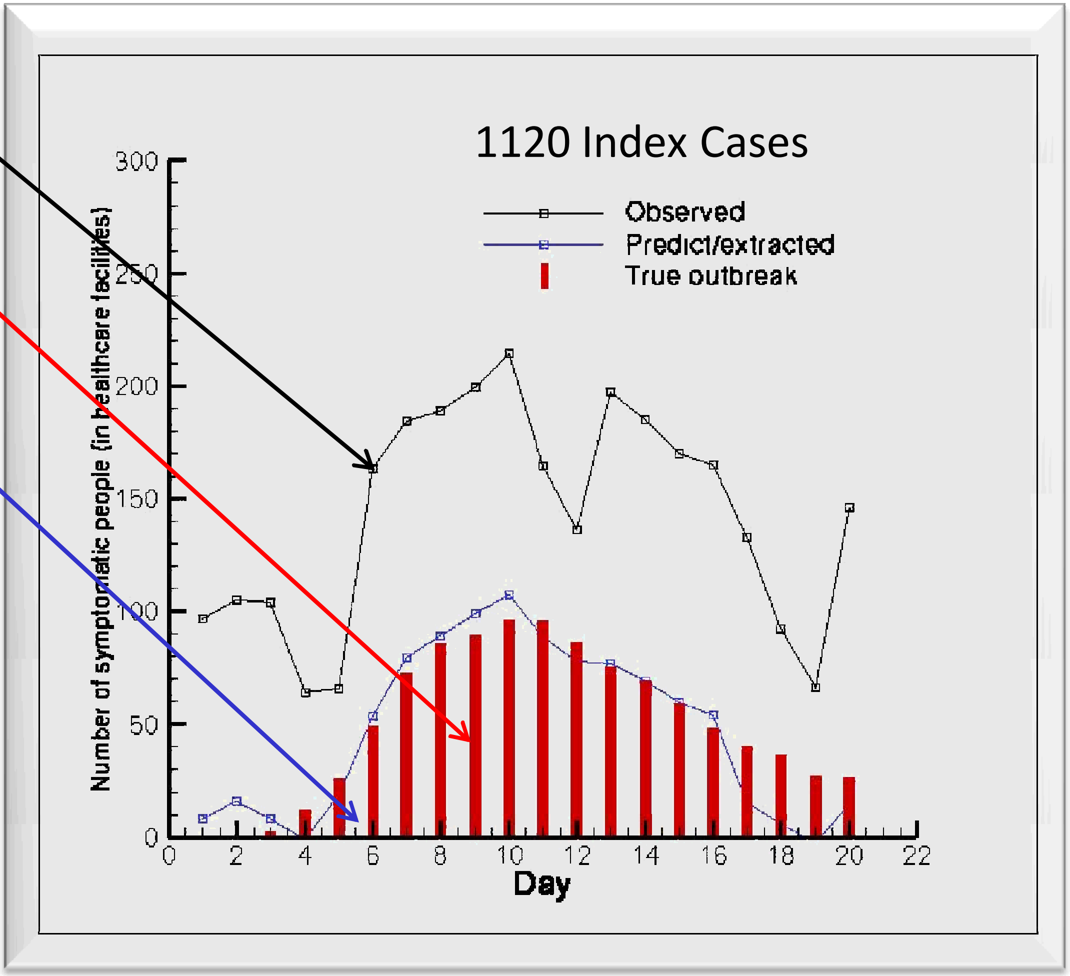
Longer-term predictions are typically valid for 2 weeks or greater. Subtracting the background model from the data yields the epidemic curve for the classification module.

Background Subtraction Uses Model Fit For Anomaly Detection

Simulated Anthrax Attack + Background

Simulated Anthrax Attack

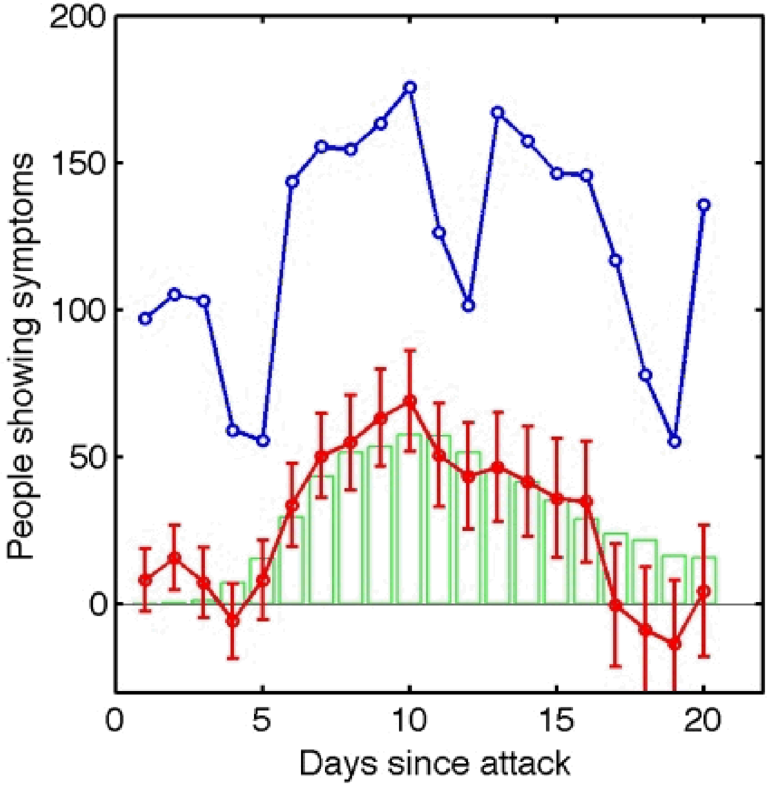
Estimated Anthrax Attack = Simulated Data – Background Model



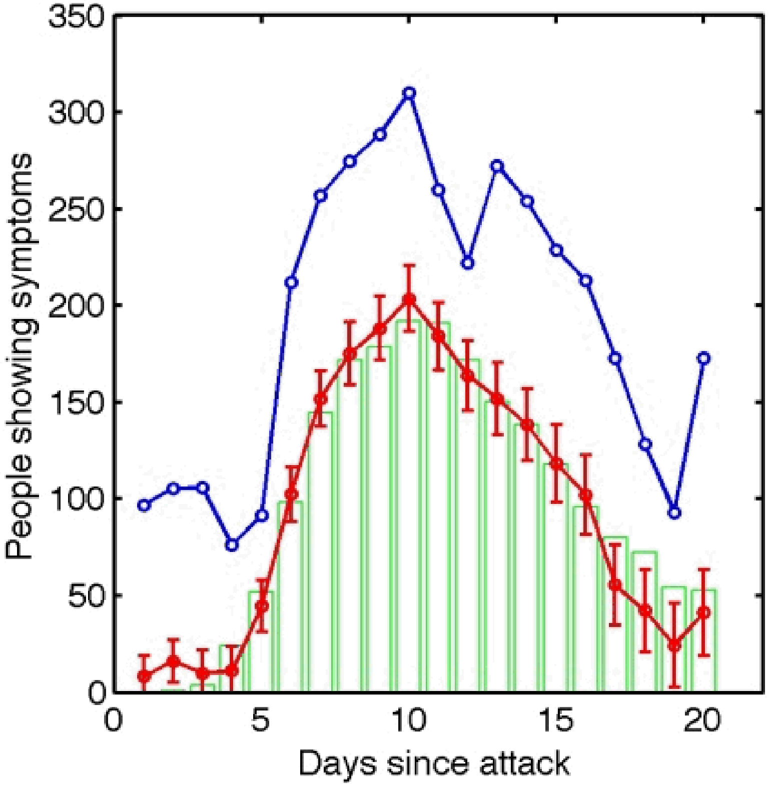
For this case:
Day 0 = Start of attack
Day 5 = Detection
Anthrax incubation period = 3-4 days

Background subtraction accurate for approximately 16 days, as required for Classification Module

Background Subtraction For Different Sized Attacks



680 Index Cases



2250 Index Cases

Steps for Detection and Classification

- The components of the procedure are:
  - **Background Modeling/Outbreak Detection** from time-series data
    - Data contains the outbreak and background/endemic morbidity
  - **Extraction** of the outbreak from the background
    - Endemic component needs to be separated from the epidemic component
  - **Characterization** of the outbreak
    - estimation of index cases, time/rate of infection
  - **Identification** of the outbreak
    - What was the disease that caused it, given a few competing guesses

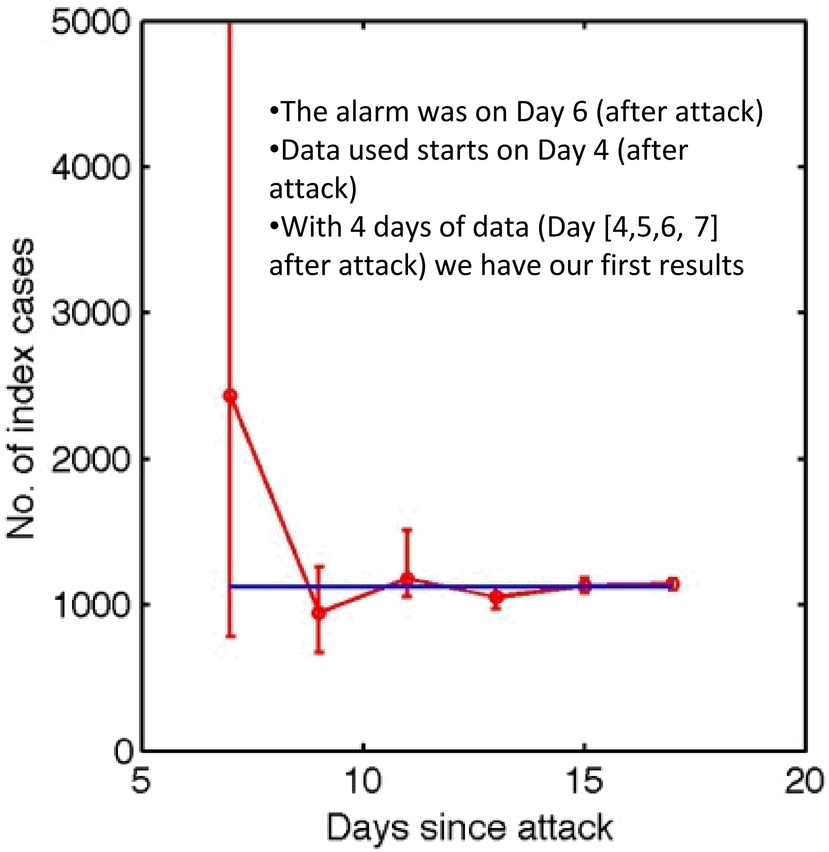
Characterization of the Anthrax Epidemic

- **Characterization:**
  - Estimation of the number of index cases, time of release, an average dose, and some parameters of the visit-delay model
- **Hypothesis:**
  - An anthrax incubation period model + a visit delay model can reproduce the epidemic curve
    - The quantities of interest are all parameters/inputs into this epidemic model
  - So given a partial epidemic curve, fitting an anthrax model should reveal the necessary model parameters
- **Questions:**
  - How much data is needed to estimate these parameters?
    - i.e., is less than 15 days of (good, normal background extracted) data sufficient?
  - What is the level of uncertainty in parameter estimates, as a function of (quantity of) data?

Bayesian Techniques to Solve the Problem

- We formulate the estimation as a Bayesian inverse problem
  - Predicated on the extracted epidemic data
- Allows one to use bounds / prior beliefs regarding the value of the parameters
  - We assumed that index cases ranged between 100-10,000
- Solved using an adaptive Markov Chain Monte Carlo sampler
  - All parameters estimated as probability density functions (PDF)
  - Used autocorrelation analysis to determine “convergence” of the Markov chain

Antrhax: Estimates of the Number of Index Cases



•The alarm was on Day 6 (after attack)
•Data used starts on Day 4 (after attack)
•With 4 days of data (Day [4,5,6, 7] after attack) we have our first results

Number of index cases bounded in 7 days after attack;

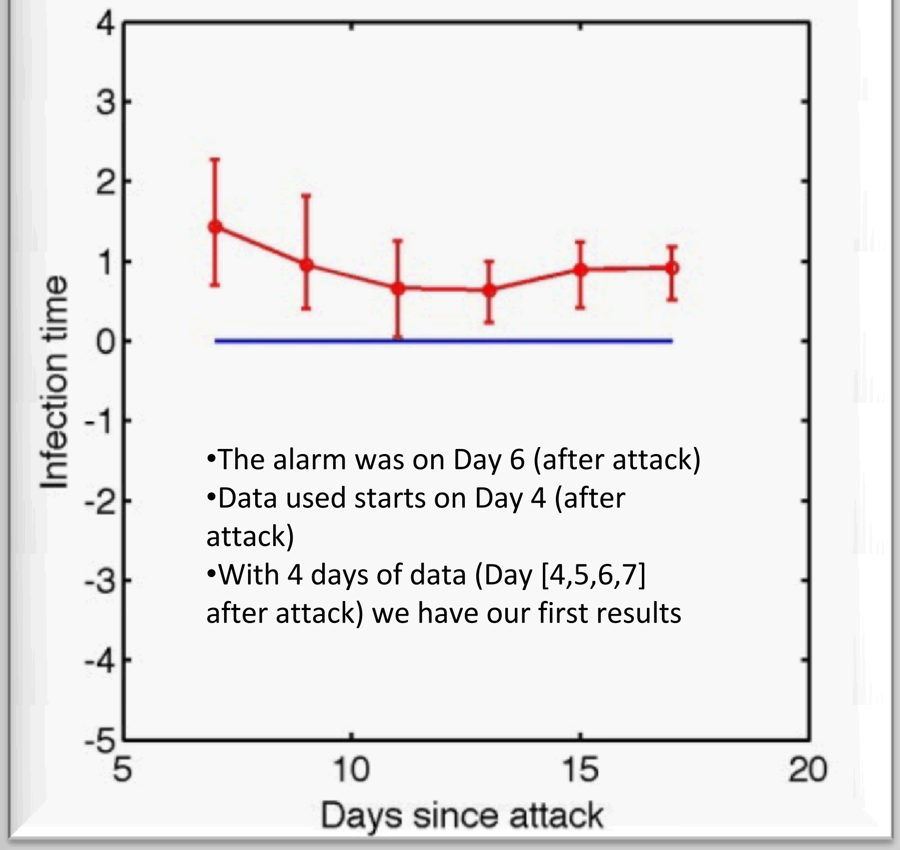
Bounded to 2250 people out of original population of 3 Million;

Accurate to 20% after 9 days, post attack.

Incubation period is 3-4 days so will not get earlier than that.

- Estimates of the number of index cases (in **red**).
- True figure in **blue. Left edge determines the day we first try to infer.**

Estimates of the Time of Infection



- The alarm was on Day 6 (after attack)
- Data used starts on Day 4 (after attack)
- With 4 days of data (Day [4,5,6,7] after attack) we have our first results

Red is the estimated release time / time of infection.

With 4 days of data, we're within a day of the actual release!

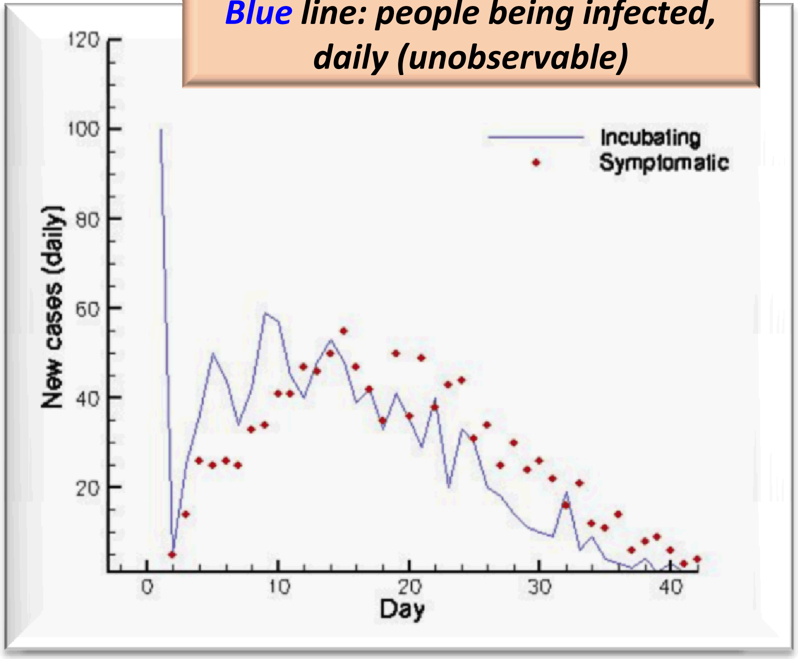
- 4 days of data, post-alarm, correctly estimate time of infection

Application to a Communicable Disease

- The technique can be applied to a communicable disease
- Apart from the “usual” quantities, have to estimate infection rate
- Assumptions for communicable diseases model
  - The infection rate increases and thereafter decreases smoothly in time
    - Model using a skewed distribution like Weibull or Gamma
  - Index cases are a small fraction of the total number of victims
- A lightweight model can be created and fit to data
  - Uses MCMC, as before
  - Estimates total size of the epidemic, visit delay parameters and infection rate parameters, all as PDFs

A Communicable Disease Example

- Example: A simulated plague epidemic
  - Performed with an agent-based model for disease spread
  - Includes visit-delay
  - Incubation is NOT dose dependent
- 100 index cases
  - Epidemic dies out in 40 days
  - 1500 victims, total
- Aim:
  - Estimate the total size of the epidemic
  - Also, the infection rate curve
  - Compare with the “true” figures from the simulation

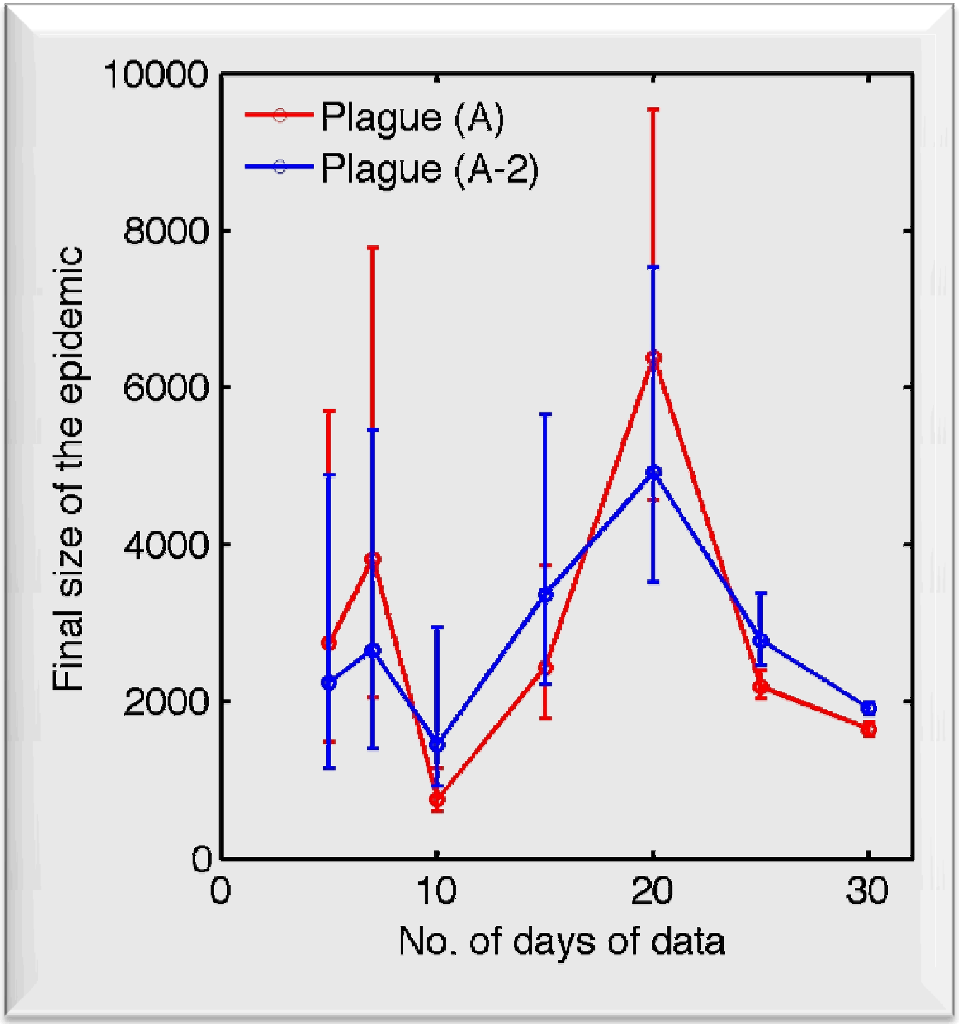


***Red** points: People turning symptomatic, daily (observed)*
***Blue** line: people being infected, daily (unobservable)*

- ***The epidemic is driven by an unknown time-variant process (infection) and we have to infer it.***
- ***Much harder!***

Estimation of the Final Epidemic Size

- The true figure is 1500
- The estimate improves (shorter error bars) with time (and data!)
- Estimates performed with data starting from
  - Day of alarm (A)
  - 2 days before alarm (A-2)
- Easier for large outbreaks

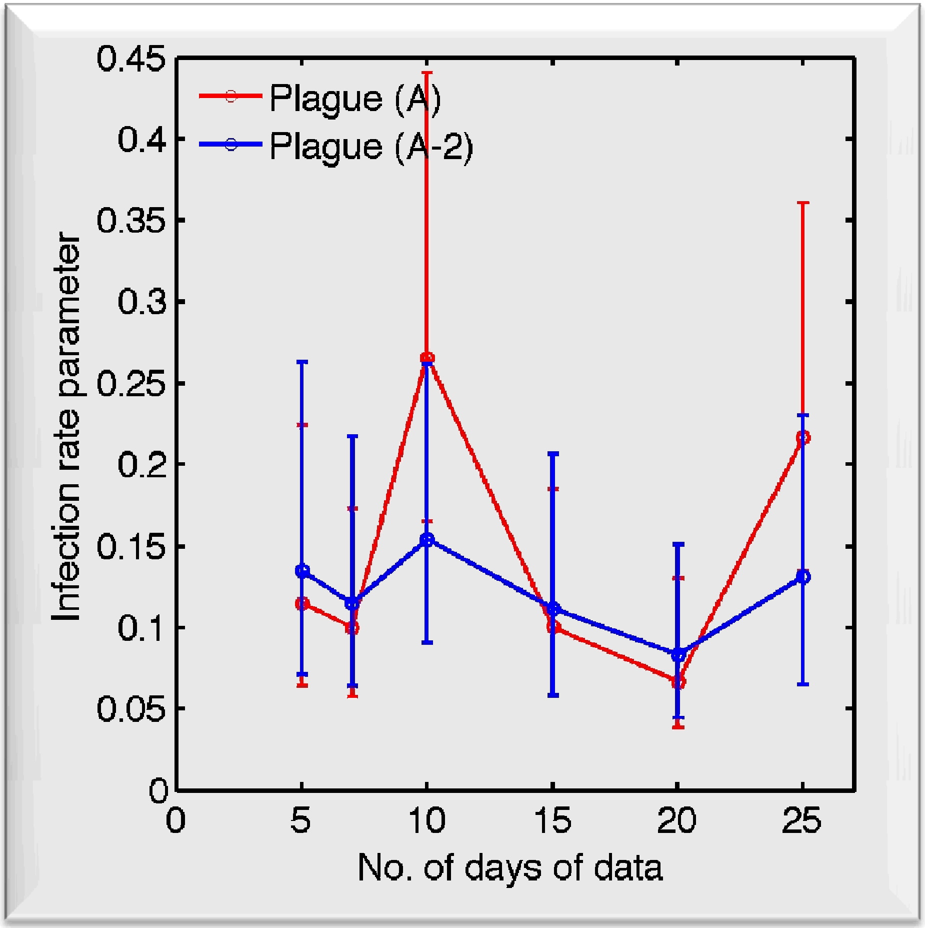


The size of the epidemic can be inferred, but the inference is noisy (no nice trend with increasing data).

But the uncertainty does decrease with data.

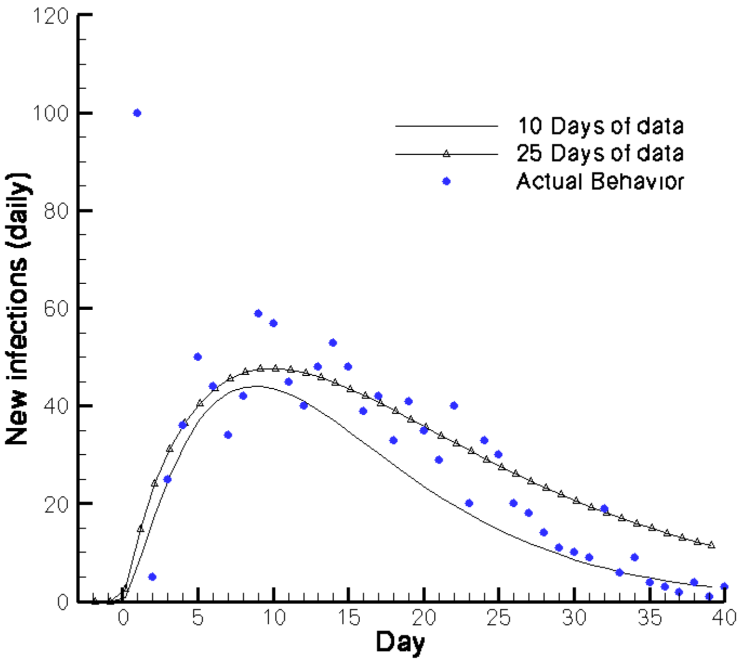
Estimation of the Parameter in Infection Rate Model

- Infection rate modeled as a $\Gamma(k, \theta^{-1})$ function
  - θ^{-1} (rate parameter) estimated from data; k set to 2
- Results: PDFs of θ^{-1}
  - About 15 days of data provide a good estimate of θ^{-1}
- But what does the infection rate look like over time?
  - Next slide ….



Estimates of θ^{-1} as a function of amount of data. Developed with data starting from day of detection as well as 2 days pre-detection.

Estimation of the Infection Rate (Over Time)



We actually manage to capture the hidden infection process, and its variation in time. The blue dots are how the infection rate actually behaved; the smooth line is our inference.

And we capture its decay too!

- Best estimate of the variation of infection rate over time
- Developed using θ^{-1}_{MAP} (after 25 days of data)
  - MAP = Maximum A Posteriori ~ best estimate

Steps for Detection and Classification

- The components of the procedure are:
  - **Background Modeling/Outbreak Detection** from time-series data
    - Data contains the outbreak and background/endemic morbidity
  - **Extraction** of the outbreak from the background
    - Endemic component needs to be separated from the epidemic component
  - **Characterization** of the outbreak
    - estimation of index cases, time/rate of infection
  - **Identification** of the outbreak
    - What was the disease that caused it, given a few competing guesses

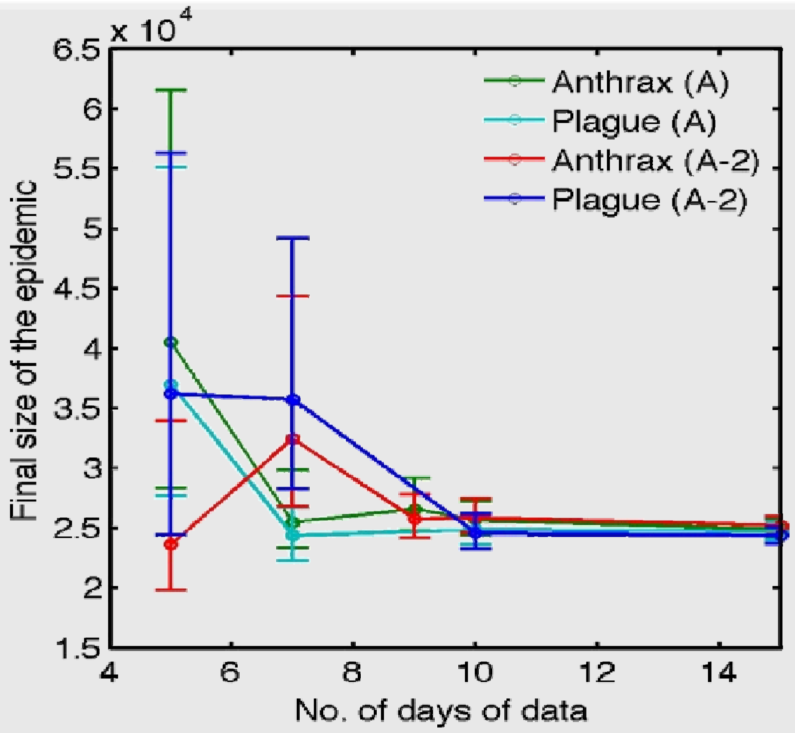
Identification of the Causative Agent

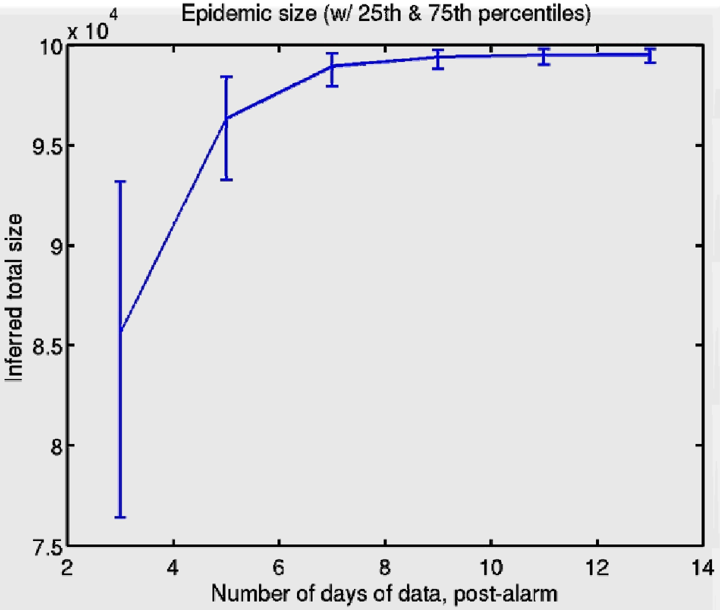
- Nobody told us the epidemic was an anthrax epidemic
  - Could be plague or flu
- In characterization step, we saw that both communicable and non-communicable diseases could be fit to data
- We will compete the anthrax, plague and flu models
  - The best fit model is probably the real cause of the disease
- Test
  - Start with an anthrax attack
  - Characterize using the 3 models
  - Show what the final size of the epidemic looks like
  - Compete the model
    - More on this later – involves AIC and BIC

Characterize with Plague and Flu (continued)

- Simulate an attack with anthrax
  - Atmospheric release of a population of 3,000,000
  - 22,000 infected; dosage variable, depending upon population density distribution in space and wind direction
- Fit the three models to data (anthrax, flu and plague)
  - Infer index cases, time of infection
  - For communicable disease, also estimate time-dependent infection rate and final size of epidemic
- A word about flu
  - Very interesting differences in civilian and military populations – but that is the subject of another talk!
  - So we have a “civilian” and “military” flu models

Characterize with Plague and Flu (continued)

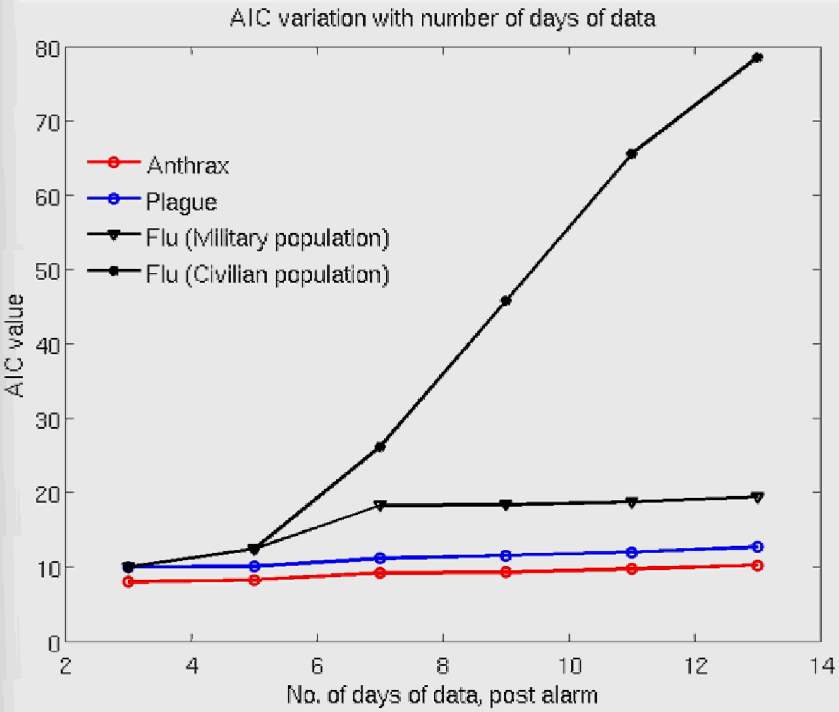




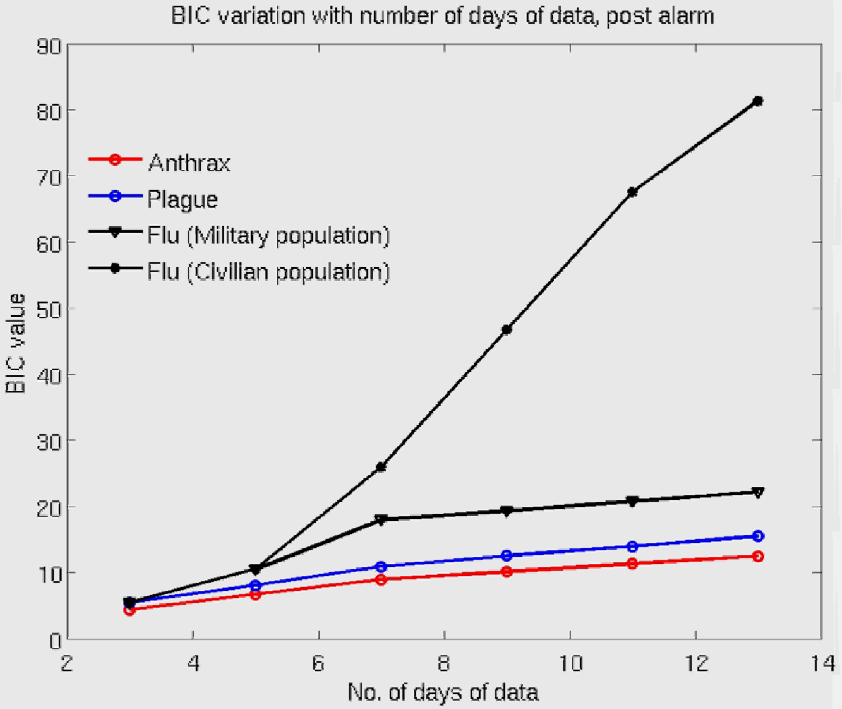
Inference performed with “civilian” flu; little difference with “military” flu model

- ***The plague and anthrax epidemic are both reasonable fits.***
- ***Flu over-estimates the final size of epidemic (it spreads).***
  - ***And the epidemic size error bars shrink with data (more later…).***

Compete the models!



AIC



BIC

- How?
  - Compute AIC & BIC for all 3 models and compare
  - Large AIC & BIC mean bad fits

- ***With 5 days of data anthrax is identified as the correct causative agent.***
- ***Basically, anthrax model fits data best.***
- ***Identification / model selection worked.***

AIC and BIC Capture Best Fit Model

- If the flu model has such a bad fit to data, how come the N_{tot} estimates have tight error bounds?
  - While being so wrong in its estimates?
- Reason: The flu model gets “fit” to a local minimum
  - Way worse than the global minimum, but flu parameters are not consistent with the global minimum
    - For example, the global minimum requires infection spread-rate to be zero
- With data, the local minimum steepens and narrows
  - Error bars shrink
  - But the maximum likelihood becomes worse and worse
  - And model fitting becomes harder and harder
- But the AIC and BIC capture the worsening likelihoods, and so no harm done

Lesson: When fitting models to data, track the error bars and the maximum likelihood. Adding more data could shrink error bars, but worsen the model fit.

Conclusions

- Techniques appear promising to construct and integrate automated detect-characterize-identify technique for epidemics
  - Working off biosurveillance data
  - Provides information on the particular/ongoing outbreak
- Parameter estimation capability ideal for providing the input parameters into an agent-based model
  - Index Cases, Time of Infection, Total Epidemic Size
- Non-communicable diseases are easier than communicable ones
  - Small anthrax can be bounded with 5 days of data, post-detection; plague and flu takes longer
  - Larger attacks can be bounded with ~3 days of data, post-detection

Conclusions (Continued)

- Identification tests (model selection) with anthrax, plague and flu were successful
- Characterization techniques are highly useful even if sentinel physicians identify the disease
  - Determines disease parameters
  - Allows medical countermeasures planning

Classification provides answers to the situational awareness puzzle created by an outbreak.

Acknowledgements

- DTRA, under contract HDTRA1 -09 C 0034

- Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.