Statistical Approach to Biosurveillance in Crisis: What is Next?
Ernest S. Shtatland, eStatConsulting, Stoneham, MA
Timur Shtatland, eStatConsulting, Stoneham, MA

ABSTRACT
Motivated by the threat of bioterrorism, biosurveillance / syndromic surveillance systems are now in crisis: with the original purpose of early detection, and more than 10 years in existence, no health department has reported using them for this purpose. This has led to a shift away from only early detection of bioterrorist attacks. The goal has been expanded in two directions: firstly, to include both early event detection and situational awareness, so that the focus is not simply on detection, but also on timely response and consequence management; and secondly, to switch the emphasis from bioterrorism only to detecting and responding to natural disease outbreaks such as seasonal and especially pandemic flu. Even with this expansion, early detection capacity is problematic. The reason is uncontrolled alert rates: there is an alarm nearly every day and most health monitors learned to ignore alarms. It results in distrust in statistical methods and in biosurveillance itself.

In this presentation, we propose a new approach that has a unique potential to successfully combine capabilities for early event detection (with more effective control of alert rates) and situational awareness including monitoring and predicting outbreaks magnitude, rate of change and duration. No existing biosurveillance systems provide this capability. Our approach is based on epidemiological models, both deterministic and stochastic, and their linear approximations – first-order autoregression models, in combination with a standard statistics toolkit: parameters estimating, confidence intervals constructing and hypotheses testing. The approach is originated from our previous research presented at the recent NESUG and SGF presentations. The proposed simple models provide us with the ability to detect an outbreak and simultaneously predict the timing, size of the epidemic outbreak peak, and also the final proportion of the affected population, which is critical for choosing optimal epidemic control strategies and estimating health resources needed.

The intended audience: SAS users of all levels who work with SAS/STAT® and SAS/ETS® (e.g. ARIMA and MODEL procedures).

INTRODUCTION: WHAT IS BIOSURVEILLANCE?
Definition of biosurveillance, epidemiological surveillance as a particular type of biosurveillance related to human health, and syndromic surveillance as a specific type of epidemiological surveillance based on pre-diagnostic medical data collected in real time, are given in Homeland Security Presidential Directive 21 (HSPD-21) (see U.S. Government (2007), Fricker (2011a) and Fricker (2011c)). In the context of our presentation, biosurveillance, epidemiological surveillance and syndromic surveillance will be used as synonyms. Driven by the threat of bioterrorism (especially after 9/11/2001), biosurveillance systems have been developed and implemented around the world. In the USA alone, we can mention the following nationwide systems (see in detail Chen (2010)): BioSense, ESSENCE, EARS, BioALIRT, RODS, RSVP, BioDefend, BloStorm, National Bioterrorism Syndromic Surveillance Demonstration Program, BioPortal, B-SAFTER, and INFERNO. This list is far from being complete. Also, there exist numerous regional and communitywide programs. In a word, first decade of XXI century was a time of mushrooming biosurveillance systems. And now, we see biosurveillance/syndromic surveillance in crisis. First has come understanding that biosurveillance systems based on statistical algorithms are of little value in early detection of bioterrorist attacks. At least, no health department has reported using them for this purpose (see Fricker (2011a) and references therein). As a result, the objective of biosurveillance has been expanded in two directions: (1) to include both early event detection (EED) and situational awareness (SA), so that our attention is concentrated not only on detection, but also on effective response and consequence management; (2) to switch the emphasis from bioterrorism solely to detecting and responding to natural disease outbreaks such as seasonal and especially pandemic flu. In spite of this expansion, early detection capacity looks problematic.
The reason is excessive and uncontrolled alert rates - there is nearly an alarm every day and most health monitors learned to ignore alarms, which results in distrust in statistical methods and in biosurveillance itself. That is why Fricker (2011a) concludes his recent review article in such a pessimistic way: “Returning to the original question of whether statistical methods are useful for early event detection, I suggest that we really don’t know yet. That is, whether the systems and their associated detection algorithms can be modified so that they appropriately minimize false positive signals while maintaining sufficient sensitivity to actual outbreaks is still an open question”. In the rejoinder to the discussion of his review article (see Fricker (2011b)), the author says: “And, in spite of my research interest in EED methods, I would suggest that situational awareness is probably the more important function for biosurveillance, since it enhances public health surveillance and management before, during, and after an outbreak.” Thus, it becomes more and more evident that emphasis is shifting from the early detection objective to supporting public health situational awareness. Since the definition of EED is self-evident, we need to specify public health SA. Usually, SA is defined as the ability to utilize detailed, real-time health data to confirm or refute the existence of an outbreak, and in case of the outbreak to provide a timely and effective response and consequence management. To be effective and comprehensive, response and consequence management should include implementing interventions (for example, vaccinations and other epidemic control strategies). To this definition, we can add monitoring activity: SA also is used to monitor magnitude and spread of an outbreak, rate of change, duration etc. This monitoring should not be limited to retrospective observations like comparison of today’s data with the corresponding data a week or a month ago and so on. To be effective, monitoring should have a strong predictive, preventive component which would allow us make projections about hospital admissions, the number of infected at the epidemic peak, critical response resources needed etc. It is obvious that providing adequate projections and efficient response is impossible without theoretical, epidemiological models.

Note that the original objective of syndromic surveillance systems - early event detection - could be partly met without epidemiological modeling, and no existing biosurveillance system uses epidemiological models for this purpose at all. Instead, these systems utilize regression-based methods applied to the raw data, e.g., daily counts of emergency department visits and Statistical Process Control (SPC) charts applied to the residuals in mentioned above regression methods by using, for example, SHEWHART, CUSUM and MACONTROL procedures (SAS/QC®). As to situational awareness, with elements of prediction of future developments of outbreaks, using epidemiological models looks not only helpful, but necessary.

WHY EPIDEMIOLOGICAL MODELING IS NECESSARY IN BIOSURVEILLANCE

There seems to be a consensus in the biosurveillance community that whether biosurveillance systems are useful for detecting bioterrorism or not, their most important contribution to public health practice is detecting and responding to natural disease outbreaks, such as seasonal and especially pandemic flu (see Fricker (2011a and 2011c)). That is why our focusing at epidemics of infectious diseases and pandemic flu in particular, is justified and using epidemiological models in this context is more than appropriate. Our model of choice is the so-called epidemic SIR model, which is defined as a nonlinear compartmental model with three compartments. Standard convention labels these three compartments S (for susceptible), I (for infectious) and R (for recovered). Therefore, this model is called the SIR model. This is a good, simple, model for many infectious diseases including measles, mumps, rubella, smallpox among others. Also, it is widely and successfully used in modeling seasonal and pandemic flu, which is of particular interest to us. See, for example, Chowell, Nishiura and Bettencourt (2007), Bettencourt et al. (2007), Wallinga and Lipsitch (2007), Bettencourt and Ribeiro (2008), and Nishiura et al. (2010). It will be seen below that, in spite of its simplicity, the SIR model and its linearization can be used for EED and SA purposes simultaneously (a unique ability for biosurveillance and epidemiology!).

Note that the SIR model, created in 1927, is one of the early triumphs of mathematical epidemiology. A natural question is whether we can limit ourselves with SIR models. Clearly, SIR models rely on some general assumptions, which are simplistic in specific situations. First, these models do not account for contact heterogeneities, resulting from spatial effects, age, and/or the
structure of social networks. Second, the models do not include latent or incubation periods. According to Heathcote (2000), latent periods are often omitted, especially when they are small, because they are not so crucial for the susceptible-infective interaction. Notwithstanding these limiting features, the SIR structure allows reliable real time parameter estimation with quantified uncertainty at very low computational overhead. It is important to us that influenza and ILI have approximately 2-day latent period, so SIR models can be safely used. At the same time, such widespread highly contagious diseases as SARS, measles, mumps, smallpox, and rubella have much longer incubation period. To work with these diseases, we have to use some extension of SIR models, for example so-called Susceptible–Exposed–Infectious–Recovered (SEIR) models.

**DETERMINISTIC SIR MODELS**

Mathematically, a SIR model can be described by the following first-order nonlinear system of difference equations:

\[
S(n+1) = S(n) - \left(\frac{\beta}{N}\right)S(n)I(n),
\]
\[
I(n+1) = I(n) + \left(\frac{\beta}{N}\right)S(n)I(n) - \delta I(n),
\]
\[
R(n+1) = R(n) + \delta I(n),
\]

where \(S(n)\), \(I(n)\) and \(R(n)\) represent the numbers of susceptible, infected and recovered individuals correspondingly on day \(n\); \(N\) is the total closed population, assumed to be constant: \(N = S(n) + I(n) + R(n)\); \(\beta\) is the infection transmission rate and \(\delta\) is the average rate of recovery from infection; \(d = 1/\delta\) can be considered as the mean duration of infectivity (in days). Both rates \(\beta\) and \(\delta\) are supposed to be constant. There is no demography (i.e., births and deaths). Since total population \(N\) is typically unknown and cannot be effectively estimated it is useful sometimes to work with population fractions: \(s(n)=S(n)/N\), \(i(n)=I(n)/N\) and \(r(n)=R(n)/N\). The corresponding system for fractions does not contain parameter \(N\):

\[
s(n+1) = s(n) - \beta s(n)i(n),
\]
\[
i(n+1) = i(n) + \beta s(n)i(n) - \delta i(n),
\]
\[
r(n+1) = r(n) + \delta i(n),
\]

The continuous-time counterparts of (1) and (1') are written in the form of the system of non-linear differential equations for counts

\[
dS/dt = -\left(\frac{\beta}{N}\right)SI,
\]
\[
dI/dt = \left(\frac{\beta}{N}\right)SI - \delta I,
\]
\[
dR/dt = \delta I,
\]

and for fractions

\[
ds/dt = -\beta si,
\]
\[
di/dt = \beta si - \delta i,
\]
\[
dr/dt = \delta i,
\]

Note that we will use discrete-time models (1) and (1') for early detection EED purposes, and their continuous-time counterparts (2) and (2') for SA goals: getting global information such as the number of infected at the epidemic peak, the total number of infected in the epidemic course, etc.

Systems (1), (1'), (2) and (2') are nonlinear and they cannot be solved analytically, in closed form, but they can be solved numerically, by using, for example, the MODEL procedure (SAS/ETS®). The simplest syntax of PROC MODEL for system (2') is as follows:

```plaintext
proc model data = one;
   dependent s s0 i i0 r r0;
```
parms beta beta0 delta delta0;
dert.s = -beta*s*i;
dert.i = beta*s*i - delta*i;
dert.r = delta*i;
solve s i r | dynamic out = two;
run;

To solve system (2’) we have to know parameters $\beta$ and $\delta$, and initial values of susceptible $s(0)$, infected $i(0)$ and recovered $r(0)$. In case of system (2), we need to know total population $N$. For simulation purposes, $S(0), I(0)$ and $R(0)$ are considered known and we just have to solve the equations. In real life, the situation is different: usually we do not know parameters $\beta$ and $\delta$ (they should be estimated from the observed data). In the most interesting case of emerging diseases and potential new pandemics, it is epidemiologically reasonable to assume about initial values $S(0), I(0)$ and $R(0)$ that: (1) Almost all population is susceptible at the beginning of an epidemic, (i.e. $S(0) \approx N$ or $s(0) \approx 1$); (2) There are no immune to the new disease ($R(0) = 0$ or $r(0) = 0$); (3) There is a small number (at least one) of infectious ($I(0)$ is a small number, i.e. $I(0) = 1$, or $i(0)$ is very small, $i(0) = 0$). Obviously, to have epidemic we need some number (usually small) of infectious invaders. Is it enough to develop the epidemic? To answer this question we have to introduce a central concept in epidemiology – the basic reproduction number or basic reproduction ratio, $R_0$, which is usually defined as the expected number of secondary cases that would arise from a single primary infectious case in a fully susceptible population. For SIR models, there is a simple formula for $R_0$

$$R_0 = \frac{\beta}{\delta} \quad (3)$$

It is easy to see that if $R_0 < 1$ then each successive generation of infectious is smaller than its predecessor, and as a result, the infection cannot persist. When $R_0 > 1$ then next generations of infectious become larger than their predecessors, and the number of infectious cases will grow. This increase does not continue indefinitely. The infection process reduces the pool of susceptible, and hence reduces the probability that an infectious individual contacts a susceptible one within its period of infectiousness. We can ignore this non-linear effect only at the start of an epidemic. This behavior can be easily seen from analyzing the second equation in (2) and (2’). Indeed, from equations $di/dt = \beta si - \delta i$ and $dl/dt = (\beta/N)SI - \delta l$ we see that the number of infected $I$ (the fraction of infected $i$) increases if and only if $S > N\delta/\beta$ (s > $\delta/i(\beta)$). More details can be found in the next two sections. Thus, $R_0$ is a threshold parameter, determining whether or not there will be an epidemic. As such, $R_0$ is of the prime importance in epidemiology. Estimating this parameter as soon as possible is the main goal for epidemiologists and a public health authority, which is related to EED. In addition to indicating whether we will have an epidemic ($R_0 > 1$) or not ($R_0 < 1$), $R_0$ provides information about such global characteristics of the epidemic as the number of infected at the epidemic peak and the total number of infected in the course of an epidemic, which is related to early SA. This is why we care so much about $R_0$. It is more convenient for us to formulate basic results of the next two sections in terms of fractions for the continuous-time SIR model (2’).

THE EPIDEMIC PEAK IN SIR MODELS

Suppose that $R_0 > 1$, so we have an epidemic. Let $i_{max}$ be the maximum value of $i(t)$ and $s_{max}$ and $r_{max}$ are the values of $s(t)$ and $r(t)$ correspondingly at the moment when $i_{max}$ is attained. A necessary (and sufficient in our case) condition for $i(t)$ to attain its maximum is $di/dt = 0$, which is equivalent to $\beta si - \delta i = 0$ and hence to $\beta s - \delta = 0$ since $i(t)$ has to be positive at the moment when it reaches maximum. Thus, we can conclude that

$$s_{max} = \frac{\delta}{\beta} = 1/ R_0 \quad (4)$$

We see also from above that fraction of infected $i(t)$ is increasing as long as the fraction of susceptible $s(t)$ is large enough ($s(t) > 1/ R_0$), at the moment when $s(t) = 1/ R_0$, $i(t)$ attains its
maximum and then decreases to 0. As to the formula for $i_{max}$, we refer to Theorem 2.1 in Hethcote (2000), which states that under our assumptions ($s(0) \approx 1$ and $i(0) = 0$)

$$i_{max} = 1 - 1/R_0 - ln(R_0)/R_0 \quad (5)$$

Obviously,

$$r_{max} = 1 - s_{max} - i_{max} = ln(R_0)/R_0 \quad (6)$$

Note that if $R_0 = e \approx 2.718...$, which is close to 3.0 and belongs to the interval of typical for influenza values 1.5 – 4.0, then $ln(R_0) = 1$ and as a result $i_{max} = s_{max} = r_{max} = 1/3$, i.e. all fractions are equal at the epidemic peak.

**WHAT HAPPENS IN THE LONG RUN (ASYMPTOTICALLY)?**

Let $s(\infty)$, $i(\infty)$ and $r(\infty)$ are limiting values of $s(t)$, $i(t)$ and $r(t)$ correspondingly as $t$ increases to infinity. Dividing first equation in (2') by third one, we have

$$ds/dr = -\beta s / \delta = -R_0 s.$$ 

Integrating this equation with respect to $r$ and assuming $r(0) = 0$, we arrive at

$$s(t) = s(0)exp(- R_0 r(t)).$$

Since $r(t)$ is always less than 1 as a fraction, we have a fundamental inequality $s(t) \geq \exp(- R_0)$ and consequently $s^{(\infty)} \geq \exp(- R_0)$. Thus, each epidemic eventually breaks down due to lack of infected, not lack of susceptible (which sounds counter-intuitive!). According to Theorem 2.1 in Hethcote (2000) limiting value $s^{(\infty)}$ is the unique root in interval $(0, 1/R_0)$ of the equation

$$i(0) + s(0) - s^{(\infty)} + ln (s^{(\infty)}/s(0))/R_0 = 0,$$

which simplifies under our assumptions ($i(0)=0$ and $s(0)=1$) to the equation

$$1 - s^{(\infty)} + ln (s^{(\infty)}/s(0))/R_0 = 0 \quad (7)$$

Taking into account that $i^{(\infty)} = 0$ and consequently $r^{(\infty)} = 1 - s^{(\infty)}$, we have the following equation for $r^{(\infty)}$:

$$1 - r^{(\infty)} = exp(- R_0 r^{(\infty)}) \quad (8)$$

Equations (7) and (8) cannot be solved analytically, in a closed form, but they can be solved numerically. Below, see numerical results for some values of $R_0$:

- If $R_0 = 2$, then $r^{(\infty)} = 0.8$ and $s^{(\infty)} = 0.2$;
- If $R_0 = 3$, then $r^{(\infty)} = 0.95$ and $s^{(\infty)} = 0.05$;
- If $R_0 = 5$, then $r^{(\infty)} = 0.993$ and $s^{(\infty)} = 0.007$.

Therefore, increasing $R_0$ has a dramatic effect on the proportion that escape the epidemic outbreak and consequently the proportion of eventually infected. Instead of exact nonlinear equation (7), for large $R_0$ we can use a very simple approximation

$$s^{(\infty)} \approx exp(- R_0) \quad (9)$$

Even for not very large $R_0$ such as $R_0 = 3$ or $R_0 = 5$, formula (9) presents a surprisingly good approximation. Note that the SIR model is the simplest model that more or less realistically describes real-life epidemics. In spite of the fact that the SIR model ignores incubation or latent periods (among other things) it can produce quite a good fit in practice.
Summarizing, we can see that $R_0$ provides fundamental insights into the dynamics of infectious diseases:

1. $R_0$ is the threshold parameter, determining whether or not there will be an epidemic (EED goal).
2. $R_0$ determines the initial rate of increase of an epidemic, i.e., during its exponential growth phase (both EED and SA goals).
3. $R_0$ determines the fractions (and the numbers) of susceptible, infected and recovered at the epidemic peak (SA goal).
4. $R_0$ determines the final size of the epidemic, i.e., what fraction of susceptible will ultimately be infected over the course of the outbreak (SA goal).
5. $R_0$ determines the critical vaccination threshold ($\approx 1/R_0$), which shows the optimal choice of epidemic control strategies (SA goal).

Also, knowing the $R_0$ is a prerequisite for designing public health measures against a potential pandemic using simulation techniques. Though we are interested in epidemics only, we should add that $R_0$ determines the endemic equilibrium fraction of susceptible in the population ($\approx 1/R_0$), which could serve as a critical fraction of susceptible for a future epidemic startup. Let us remind that for a novel pandemic we assume that almost all population is susceptible at the beginning of an epidemic, i.e. $S(0) \approx N$ or $s(0) = 1$.

Thus, we see that $R_0$ is a “magic number”, all-in-all parameter for SIR models and that $R_0$ has to play a key role in decision making in epidemiology. Consequently, timely and adequate estimating $R_0$ is of greatest importance. There exist several methods of estimating $R_0$. We will mention four of them:

Method 1. Estimating $R_0$ by fitting epidemic SIR models to epidemic curve data. This is probably the most widespread approach in classic epidemiology, but it is useless for the early detection purposes.

Method 2. Estimating $R_0$ based on the final state of epidemic (evaluating the total number of infected and recovered $r(\infty)$, and solving Equation 8 for $R_0$: $R_0 = -ln(1 - r(\infty))/r(\infty)$). Again, it is useless in the early detection context.

Method 3. Estimating $R_0$ based on the epidemic peak data (evaluating $l_{max}$ and solving Equation 5 for $R_0$). This method does not meet early detection requirements also.

Method 4. The method is based on the initial phase of the epidemic, i.e., it works with a rather short portion of data related to the very beginning of a potential epidemic.

Obviously, Method 4 provides the solely appropriate approach in the early detection framework. In the next section, we will show that difference $\beta - \delta$ can also be utilized as a threshold parameter and it will prove to be even more useful than $R_0$ in the context of early detection. Note that after the epidemic detection, Methods 1 - 3 can be used for getting more specific estimates of $R_0$, and consequently of duration and the total number of infected and recovered.

FROM CLASSICAL EPIDEMIOLOGY TO MODERN BIOSURVEILLANCE THROUGH REAL TIME EPIDEMIOLOGY

Usually, classical epidemiology is considered to be largely retrospective while biosurveillance should be prospective by definition. Since SIR models are very simple and effective instruments in epidemiology, the natural question arises – why they are practically not used for early detecting in biosurveillance/syndromic surveillance. The sole explanation of this is the fact that syndromic surveillance systems operate only with counts of Emergency Departments, clinics and physician offices visits and not with numbers of susceptible, infectious and recovered, which are basic variables of SIR models. The problem is that it is not easy to transform the original counts of visits to these variables. Fortunately, in the case of influenza (and influenza-like illness (ILI) in general) it can be effectively done. This is an additional reason for using influenza as a background disease. But the main reason is that influenza is the most widespread naturally occurring infection disease in recent human history that has caused countless deaths worldwide. There were three influenza pandemics in the 20th century – the “Spanish” flu of 1918-19 (“the mother of all pandemics”), the “Asian” flu of 1957-58, and the “Hong Kong” flu of 1968-69. The
1918 flu, caused by a strain of H1N1, was by far the most deadly. Between 50 and 100 million people died totally as a result of the Spanish flu, possibly more than during the entire course of The Black Death. It makes the Spanish flu the deadliest natural disaster in human history. The 1957 pandemic was due to a new H2N2 strain of influenza virus and killed two million people, while the 1968 pandemic resulted from an H3N2 strain and killed one million. In addition, there were 3 so-called flu pandemic scares (unrealized pandemics). The first pandemic of 21st century is the Swine flu (April 2009 – July 2010). Fortunately, this H1N1 Swine flu pandemic was mild. Actually, pandemics happen every few decades. They occur when a new subtype of influenza A arises that has either never circulated in the human population or has not circulated for a very long time (so that most people do not have immunity against the virus), and it can spread easily through the human population. Taken together with the fact that influenza virus is readily accessible and may be causing more deaths than previously suspected, the possibility for genetic engineering and aerosol transmission suggests an enormous potential for bioterrorism. There is an opinion in the biosurveillance community that CDC should advance influenza as a critical agent in priority as a bioterrorism threat. Thus, the H1N1 Swine flu is over, but what is next? The uncertainty and concern about it makes us consider influenza and influenza-like-illnesses (ILI) with a particular attention, and justifies using influenza as a background infectious disease.

Now, let us return to the SIR model and its variables S, I, and R. At this point, it is convenient for us to work with counts in discrete-time models. As a rule, variables \( S(n) \) and \( R(n) \) cannot be observed or measured systematically. Only \( I(n) \) can be estimated, though indirectly, through some approximations, by using the observable daily number of patient visits \( v(m) \) to a hospital emergency department or a clinic or a physician office, on day \( m \). Mohtashemi et al. (2006) propose the following approximate formula for \( I(n) \):

\[
I(n) = \sum_{m=n-d+1}^{n} v(m),
\]

(10)

where \( d \) is an average number of days of infectivity per patient, i.e. \( d = 1/\delta \). According to formula (10), it is assumed that the overall number of infected on day \( n \) can be approximated by the sum of the number of visits to the emergency department (clinic or doctor office) during the past \( d \) days. This is a reasonable approximation if parameter \( d \) is adequately estimated or chosen. Mohtashemi et al. (2006) suppose that for influenza, the mean duration of infectivity \( d \) is equal to 7 days. Indeed, most epidemiologists agree that influenza infectivity begins the day before illness onset and can persist for up to 7 days, although some persons may shed virus for longer periods, particularly young children and severely immuno-compromised patients. See, e.g. CDC Report (2010), Mohtashemi et al (2006), and Spicer and Lawrence (1984). Thus, assumption \( d = 1/\delta = 7 \) is clinically realistic. Formula (10) plays a fundamental role in our approach. It transforms visit counts, which are basic observable variables in syndromic surveillance, into the main epidemiological SIR variable – the number of infected. It is precisely the point where integration of syndromic surveillance with epidemiological predictive modeling is taking place. In this fusion, syndromic surveillance is an information provider (in terms of daily visit counters) and real-time epidemiology contributes to analysis and decision-making. As an additional bonus, 7-day summation in formula (10) allows to compensate for the day-of-week (DOW) effect in \( v(m) \) variability. Note that the DOW effect is a primary systematic feature of the data in all recent biosurveillance systems, which drastically influences their performance. The simple 7-day procedure is very effective in removing weekly patterns. There was some concern that averaging across a week could dampen the signal and cause delays in detection i.e. loss of timeliness (see Shmueli and Burkom (2010) and references therein). Though, taking into account the mentioned above excessive false alarm rate that is typical of most systems (an alarm nearly every day!) we can conclude that this concern is grossly exaggerated. We can mention also the results of Reis, Pagano and Mandl (2003) and Mohtashemi et al (2006) which suggest that methods based on 7-day averaging are superior to those based only on the current-day observation. Also it can be shown (see, e.g. Lotze, Murphy and Shmueli (2008), and Shmueli and Burkom (2010)) that 7-day averaging is quite competitive with more sophisticated methods such as a linear regression with 6 daily dummy variables (ESSENCE) or a Poisson regression (BioSense). To obtain accurate estimates, both regression methods require large amounts of data, which are unavailable in the early detection situation. With
our simple 7-day procedure, we adhere to an important principle: keeping balance between simplicity and performance.

With approximation (10) to the number of infected \( I(n) \) and our assumption that at the very beginning of the new pandemic or emerging epidemic \( S(0) \approx N \) we immediately arrive at the following linearization of the 2\(^{nd} \) equation of the system (1)

\[
I(n+1) = I(n) + (\beta - \delta)I(n), \quad (11)
\]

or for the continuous-time version,

\[
\frac{dI}{dt} \approx (\beta - \delta)I(t). \quad (11')
\]

Equations (11) and (11') describe exponential growth if \( \beta - \delta > 0 \) and exponential decay if \( \beta - \delta < 0 \). Thus, the difference \( \beta - \delta \) can be considered as a threshold parameter alternative to \( R_0 \) (and equivalent to it): if \( \beta - \delta > 0 \) then \( R_0 > 1 \) and vice versa if \( \beta - \delta < 0 \) then \( R_0 < 1 \). In the early detection context, the advantage of the threshold parameter \( \beta - \delta \) over \( R_0 \) is obvious: we have a linear parameter in linear equations (11) and (11') as opposed to nonlinear parameter \( R_0 = \beta / \delta \) in a nonlinear setting.

Note, that in our previous SAS presentations we derive the equation equivalent to (9) in the following way: (a) We eliminate unobservable variable \( S(n) \) from the first two equations in (1) and arrive at the closed equation of the 2\(^{nd} \) order for \( I(n) \) alone (see equation (3) in Shtatland & Shtatland (2008a)); (b) Through some regrouping, we transform this equation into a nonlinear equation with time/state-dependent, slowly varying coefficients (see equations (4) – (6) in Shtatland & Shtatland (2008a)); (c) As a result of some approximations (omitting some negligible terms etc), we arrive at the 1\(^{st} \) order linear equation similar to (11); (d) These approximations generate some small errors, and it is not easy to take those errors into account individually. Combining all these sources of uncertainty in one we arrive at the stochastic equation

\[
I(n+1) = I(n)(1 + (\beta - \delta)) + w(n), \quad (12)
\]

where \( w(n) \) is a Gaussian white noise. The intensity of white noise \( w(n) \), i.e. \( \text{Var} w(n) \) cannot be specified at this point, so the simplest way is to assume that \( \text{Var} w(n) = \text{constant} \). It is easy to see that (12) is the equation of a first-order autoregression process (AR(1)). Thus, in the approach developed in our previous presentations, the stochastic element in the form of white noise \( w(n) \) appears only as an error in approximations. We will see below that the stochastic roots in equation (12) are much deeper than mere approximation errors.

**WHY WE NEED STOCHASTIC MODELS**

SIR models (1), (1'), (2) and (2') incorporate the underlying mechanism of transmission and recovery dynamics and have been able to account for experimental data in many cases. However, transmission and recovery are intrinsically stochastic processes, and deterministic models (1), (1'), (2) and (2') do not account for fluctuations. These fluctuations are especially important at the early stages of epidemics when numbers of infected are very small. That is why we need stochastic epidemiological models, and model (12) in particular, for early detection purposes. Let us emphasize again that all diseases are subject to stochasticity in terms of the chance nature of transmission and recovery, and so, in principle, a stochastic model is always more realistic than a deterministic one. On the other hand, the relative magnitude of stochastic fluctuations reduces as the number of cases increases; therefore, in large populations, with a high level of disease incidence, a deterministic model may be a good approximation. Also, an important advantage of stochastic modeling approach concerns estimation. Knowledge about uncertainty requires a stochastic model, and an estimate is not of much use without some
knowledge of its uncertainty. To conclude, stochastic models are to be preferred when their analysis is possible, otherwise deterministic ones should be used.

Here we will mention the simplest example: a stochastic version of the second equation of SIR model (2)

\[
\frac{dI(t)}{dt} = \beta(S(t)/N)I(t) - \delta I(t) + (\beta S(t)/NI(t) + \delta I(t))^{1/2} dW(t)/dt
\]  

(13)

where \( dW(t)/dt \) is a generalized derivative of a Wiener process \( W(t) \), i.e. a Gaussian white noise, which manifests combined transmission and recovery stochasticity. Sometimes this equation is written by using two independent Wiener processes for transmission and recovery separately. Since we assume \( S(0) \approx N \) at the beginning of a new pandemic the equation above is simplified to

\[
\frac{dI(t)}{dt} = (\beta - \delta)I(t) + ((\beta + \delta)I(t))^{1/2} dW(t)/dt
\]  

(13')

Equation (13') is a continuous-time version of (12) with the variance of noise that depends on the number of infected. This form of noise mimics event-driven (demographic) stochasticity, and is derived from the fact that events (i.e. occurrences of infection and recovery) form a Poisson process. Demographic stochasticity is variation arising because individual outcomes are not certain. Equation (13') is a particular case of so-called Ornstein-Uhlenbeck processes. The discrete-time version of (13') can be written as follows

\[
l(n+1) = l(n)(1 + (\beta - \delta)) + ((\beta + \delta)l(n))^{1/2} w(n),
\]  

(14)

where unlike (12), the variance of noise depends on the number of infected. Obviously, the version (14) is more fundamental than (12). Nevertheless, we will show that in practice we can use the simpler and more manageable model (12) instead of the more adequate, but more sophisticated model (14).

**WHAT CAN WE DO WITH AR(1) MODEL**

**POINT ESTIMATES**

Equation (12) can be rewritten as AR(1)

\[
l(n+1) = a l(n) + w(n),
\]  

(15)

where \( a = 1 + (\beta - \delta) \). Our main task is making statistical inference about parameter \( a = 1 + (\beta - \delta) \). This inference includes estimating parameter \( a \), constructing confidence intervals (CI), and testing the null hypothesis \( H_0: a < 1 \), which is equivalent to \( \beta - \delta < 0 \), i.e. \( R_0 < 1 \) (no outbreak) against the alternative hypothesis \( H_A: a > 1 \), which is equivalent to \( \beta - \delta > 0 \), i.e. \( R_0 > 1 \) (outbreak). There exists a well-developed theory of estimating the AR(1) parameter. It includes ordinary least-squares, Yule-Walker, Burg, and various modified least-squares estimators (see, for example, Provost and Sanjel (2005) and references therein). The most widely used estimator is the ordinary least-squares (OLS) method that provides the following estimate of parameter \( a \) based on the time series \( l(1), l(2), ..., l(T) \):

\[
\hat{a}(T) = \frac{\sum_{n=2}^{T} l(n) l(n-1)}{\sum_{n=1}^{T-1} l^2(n)}
\]  

(16)

\( T \) can be considered as a baseline of historical data used for estimating parameters of the model and making decisions. Here and further in the paper, a typical value of \( T \) is 7, though larger values (e.g., \( T = 14 \)) can be useful. Note that for general ARMA processes, a *nonlinear iterative* least-squares procedure must be used for estimating ARMA parameters (SAS/ETS® User’s Guide (2002)). In case of AR(1), we have a very simple, explicit formula (16), which can be easily
used and interpreted. The properties of estimate (16) are well known. It is consistent, i.e. \( \hat{a}(T) \to a \) (the real value of the autoregressive parameter) as \( T \to \infty \). At the same time \( \hat{a}(T) \) is a biased estimator, it always underestimates parameter \( a \). See corresponding formulas (11), (11') and (12) for bias correction in Shtatland and Shtatland (2008a). Instead of using these formulas, we can also use a modified least-squares estimator defined by the following formula

\[
\hat{a}(T) = \frac{\sum_{n=2}^{T} l(n) l(n-1)}{\sum_{n=2}^{T-1} I^2(n)}
\]

which differs from (16) only in the denominator (now summation is performed from \( n = 2 \) to \( n = T - 1 \), rather than from \( n = 1 \) to \( n = T - 1 \)). The idea behind this correction is very simple: dropping a positive term in the denominator results in overall estimate increase. According to Provost and Sanjel (2005), this simple correction can be surprisingly effective, including cases with small values of \( T \). Another interesting and more general approach is median-unbiased estimating of autoregressive parameter \( a = 1 + \beta^* - \delta \) in (14). In Zielinski (1999) and Luger (2005), it has been shown that estimator

\[
\hat{a}_{\text{med}}(T) = \text{median}(\frac{I_2}{I_1}, \frac{I_3}{I_2}, \ldots, \frac{I_T}{I_{T-1}})
\]

is median-unbiased, robust against any deviation from Gaussian distribution of the noise term \( w(n) \), including heavy tails as well as contamination with outliers. Moreover, it has been proven that \( w(n) \) in (14) are not necessarily identically distributed. Luger (2005) has showed that the results mentioned above remain true under more general distributional assumptions, without assuming statistical independence. It is important to emphasize that given the structure of estimators (16), (16') and (17), they do not depend on the unknown \( \text{Var}(w(n)) \) so that without loss in generality we may assume \( \text{Var}(w(n)) = 1 \) and work with the simpler model (12).

**CONFIDENCE INTERVALS**

Confidence intervals for autoregressive parameter \( a \) are discussed in details in Shtatland & Shtatland (2008a). See, for example, formulas (13) and (14) in Shtatland & Shtatland (2008a). Note that confidence intervals for parameter \( a \) in a non-stationary, explosive, area \((a > 1)\) are based on the standard Cauchy distribution (instead of the standard Gaussian distribution). We mention here only formula (14) for \( a > 1 \) (an epidemic case). The 100(1 - \( \alpha \))% confidence interval for \( a \) is given by the formula

\[
(\hat{a}(T) - ((\hat{a}^2(T) - 1) / \hat{a}(T))_{C_{\alpha}}, \hat{a}(T) + ((\hat{a}^2(T) - 1) / \hat{a}(T))_{C_{\alpha}})
\]

where \( C_{\alpha} \) is the two-tailed \( \alpha \) percentile critical value of the standard Cauchy distribution. For 90 and 95 percent confidence intervals, these critical values are \( C_{0.10} = 6.315 \) and \( C_{0.05} = 12.7 \), respectively. They are much larger than critical values based on the Gaussian distribution. As a result, confidence intervals are too wide to be practical for the sliding baselines commonly used in early detection. Note that \( T = 7 \) is adopted not only in this paper, but also in Reis, Pagano and Mandl (2003), Mohtashemi et al. (2006), and in most biosurveillance methods based on Statistical Process Control (Fricker (2011c)). With our assumption \( d = 1/\delta = 7 \), we have the following relationship between parameters \( R_0 \) and \( a \):

\[
R_0 = 1 + 7(a - 1).
\]

It is well-known (see, for example, Shtatland & Shtatland (2009) and references therein) that usually for seasonal flu \( 1.5 \leq R_0 \leq 3.0 \). For Spanish Influenza, the second wave (1918-1919), even higher values were observed: \( 3.0 \leq R_0 \leq 4.0 \). According to the formula above, it translates into intervals for parameter \( a \) as follows: \( 1.07 \leq a \leq 1.29 \) for seasonal flu and \( 1.29 \leq a \leq 1.43 \) for Spanish Influenza. By using formula (18) for 90% confidence intervals with \( T = 7 \), we can calculate hypothetical confidence intervals assuming \( \hat{a}(T) = 1.29 \) and \( \hat{a}(T) = 1.43 \). They are: \( 1.29 \pm 0.71 \) and \( 1.43 \pm 0.54 \), respectively. We see that in both cases confidence intervals contain
the borderline value $a = 1$ separating epidemic and non-epidemic intervals. It means that even if we have got such extreme values of estimates as $\hat{a}(T) = 1.29$ and $\hat{a}(T) = 1.43$ we still cannot take a decision that we have an epidemic at a given level of confidence. Thus, we would not be able to detect early not only a very strong seasonal flu ($R_0 = 3.0$), but even Spanish Influenza!

**HYPOTHESIS TESTING**

In Shtatland & Shtatland (2008b), we propose using so-called mid-sample and end-of-sample tests which are designed to locate a breakpoint in the middle or in the end of the sample (time window) respectively. They are based on OLS estimate $\hat{a}(T)$ though this estimate does not appear in the formulas explicitly. Implementation of mid-sample and end-of-sample tests requires massive computations: multiple executions per day of PROC ARIMA (SAS/ETS®) with ESTIMATE statement, which provides the estimates of the first-order autoregressive parameter $a$ and residuals needed for calculating test statistics. We can add to this that $p$-values for these tests can be calculated only trough bootstrapping. Since confidence intervals and hypothesis testing approaches are essentially equivalent, it is unlikely to get satisfactory significance level with the baseline of 7 days. It makes using these tests for EED purposes impractical. Our pessimism regarding confidence intervals and hypothesis testing for EED and SA is in keeping with critical views on this topic expressed recently in the biosurveillance community.

**CRITICISM OF STATISTICAL APPROACHES IN MODERN BIOSURVEILLANCE**

Here it is appropriate to return to the question asked in Fricker (2011a) about whether statistical methods are useful for early event detection and the author’s suggestion that he really does not know yet. Why so? First of all, because of the sequential nature of EED, such fundamental concepts as significance level, power, specificity, and sensitivity cannot be used directly, without nontrivial modification. They are useful only for a fixed sample (Fricker (2011b)). Secondly, biosurveillance data are usually autocorrelated, and even if such autocorrelation can be removed via modeling, the signaling statistics for EED methods that use historical data in a moving baseline, are still strongly autocorrelated. As a result, again it is difficult to interpret specificity and sensitivity. In Fricker (2011c), the author mentions 4 different definitions of sensitivity that have appeared in biosurveillance. None of these definitions overcome the fundamental problem of the lack of independence. Our approach is fundamentally different from the conventional approaches to early detection. The mainstream approaches are based on removing autocorrelation from time series of daily counts and then applying Statistical Process Control (SPC) charts. Thus, the mainstream biosurveillance community considers autocorrelation as a nuisance. On the contrary, in our approach autocorrelation is a major player: our key parameter $a = 1 + (\beta - \delta)$ is actually the first-order autocorrelation coefficient. Note that formula above can be transformed into a relationship between our basic parameters $R_0$ and $a$: $R_0 = 1 + (a - 1)/\delta$. Equation (19) is a particular case of this formula with $\delta = 1/7$.

**STATISTICAL SIGNIFICANCE VS. PRACTICAL (EPIDEMIOLOGICAL) SIGNIFICANCE**

From our disappointing results regarding confidence intervals and hypothesis testing, and above mentioned general criticism, we conclude that we cannot use in our decision making neither the concept of statistical significance with significance level and $p$-value nor level of confidence at least for EED purposes. It is known that statistical significance depends mostly on the sample size. For example, in very large samples, even very small effects will be significant, whereas in very small samples even very large effects cannot be considered significant. Instead, we propose to use the concept of practical, epidemiological, significance. Actually, what really matters is estimating the magnitude of effects, not testing whether they are zero. In most cases, the meaningfulness of a finding is reflected in the effect size or parameter estimate, and these estimates can have large or small $p$-values, depending on the sample size. We have a unique parameter, $R_0$, in terms of which epidemiologists determine both the risk of an epidemic and the effort required to control it. Accurate estimating $R_0$ is extremely important for detecting and planning for control of an infection. We have discussed earlier 4 methods of estimating $R_0$. Due to linear relationship (19) between parameters $R_0$ and $a$, the problems of their estimation are equivalent. At the same time $R_0$ is the nonlinear parameter ($R_0 = \beta/\delta$) in the nonlinear
deterministic model (1) whereas \( a = 1 + (\beta - \delta) \) is the linear parameter in the simplest linear stochastic model (15) with the well-developed estimation theory. It is natural to build our scale of epidemiological importance in terms of \( R_0 \) or equivalently, in terms of autoregressive parameter \( a \). A high level of epidemiological importance entails that the problem warrants some action to be taken to alleviate it. Since we are particularly interested in flu epidemics and novel pandemics let us consider possible actions for a potential flu epidemic. As we know, \( 1.5 \leq R_0 \leq 3.0 \) for seasonal flu and potentially \( 3.0 \leq R_0 \leq 4.0 \) for most aggressive novel flu epidemics like Spanish influenza. Thus the smallest epidemiologically important value of \( R_0 \) is 1.5 or equivalently \( a = 1.07 \) (7% daily growth rate). The largest value of \( R_0 \) for seasonal flu is 3.0 or equivalently \( a = 1.29 \) (29% growth rate). The historically largest value of \( R_0 \) (Spanish Influenza, second wave) is 4 or correspondingly \( a = 1.43 \) (43% growth rate). Thus, if we get the estimate \( \hat{a}(T) \geq 1.07 \), then we should be on alert. If we observe similar values on the next two days, we should report our findings as a significant risk of epidemic. If we have the same with \( \hat{a}(T) \geq 1.29 \), we should report as a high risk. Finally, with \( \hat{a}(T) \geq 1.43 \) we should report the situation as a severe risk.

For simplicity, we can round off critical values of growth rates: 7%, 29% and 43% to the nearest tens percent multiples, greater than those rates. As a result, the new approximate critical rates are: 10%, 30% and 50%. Such cut-offs are merely guidelines, and should not be applied rigidly. It is interesting that exactly these benchmarks are provided in Cohen (1988) for interpreting the practical importance of the Pearson correlation coefficient sometimes called "Pearson's r". As a rule of thumb, Cohen suggests that \( r = 0.1 \), \( r = 0.3 \) and \( r = 0.5 \) are considered small, medium, and large effects respectively. Such a coincidence is not accidental: in the stationary case (\( a < 1 \)), the autoregressive parameter \( a \) is equal to the 1st autocorrelation. Thus, both \( r \) and \( a \) are correlation measures.

SUMMARY AND CONCLUSIONS

In this presentation, it has been shown that:

1) Conventional statistical approaches to biosurveillance, based on such fundamental concepts as statistical significance, \( p \)-value, sensitivity, specificity etc do not work at least for EED purposes;

2) Instead, we have to use practical, epidemiological significance, based on epidemiological modeling;

3) A most likely way out of the biosurveillance crisis is integration of current syndromic surveillance efforts with epidemiological predictive modeling. This is a possible answer to the question in the title: **What is Next**;

4) Our proposed approach, based on epidemiological SIR models and their linear approximations, provide us with the unique ability to detect an outbreak and simultaneously predict the timing, size of the epidemic outbreak peak, and also the final proportion of the affected population. It is crucial for choosing optimal epidemic control strategies (vaccination up to critical vaccination threshold, hospitalization, etc), and estimating health resources needed.

REFERENCES


SAS/STAT® and SAS/ETS® are registered trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

**CONTACT INFORMATION**

Ernest S. Shtatland
eStatConsulting
200 Park Terrace Drive, Suite 234
Stoneham, MA 02180
tel: (781) 662-9578
e-mail: eshtatland@yahoo.com