Methodology

Reporting incidence from a surveillance system with an operational case definition of unknown predictive value positive

Scott R Kegler*

Address: Office of Statistics and Programming, National Center for Injury Prevention and Control, Centers for Disease Control and Prevention, 4770 Buford Highway NE, Mailstop K59, Atlanta GA 30341-3724, USA

Email: Scott R Kegler* - skegler@cdc.gov

* Corresponding author

Abstract

When reporting incidence rate estimates for relatively rare health conditions, associated case counts are often assumed to follow a Poisson distribution. Case counts obtained from large-scale electronic surveillance systems are often inflated by the presence of false positives, however, and adjusted case counts based on the results of a validation sample will have variances which are hyper-Poisson. This paper presents a simple method for constructing interval estimates for incidence rates based on case counts that are adjusted downward using an estimate of the predictive value positive of the surveillance case definition.

Introduction

Large-scale surveillance for selected medical or health conditions often relies on electronic data sources which provide comprehensive coverage of a given population. For example, the Centers for Disease Control and Prevention conduct surveillance of brain injuries involving hospitalization or death, based on electronic hospital discharge and vital statistics data received from twelve to fifteen states each year [1]. To identify cases, electronic records are scanned for specified diagnosis codes which collectively form the operational case definition. The resulting case counts are subsequently combined with population data to estimate incidence rates.

As with most surveillance methods, an operational case definition as described above may admit some records that do not represent true cases under a strict clinical definition ("false positives") and may also fail to capture some records representing true cases ("false negatives"). The customary terms reflecting these aspects of an operational case definition are predictive value positive (PVP)

and sensitivity, defined in the present context as the conditional probabilities [2]:

PVP = Pr{case meets clinical definition | case detected under operational definition};

sensitivity = Pr{case detected under operational definition | case meets clinical definition}.

Depending on the extent to which false positives and/or false negatives are believed to influence the surveillance process, it may be appropriate to use estimates of PVP and/or sensitivity to adjust incidence rate estimates accordingly. It is not generally possible to assess PVP or sensitivity using electronic surveillance data alone. The most direct approach to obtaining the additional data required for estimation of PVP involves manual review of medical records for a random sample of provisional cases identified via the operational case definition. Obtaining the additional data necessary for estimation of sensitivity may be more labor-intensive, particularly when consider-

ing an uncommon condition. Without additional "markers" (apart from the operational case definition) to narrow the scope of review, it may be necessary to select a very large sample of general medical records in order to identify enough true cases to support a stable estimate of sensitivity.

The methodology described in this paper is oriented to surveillance of relatively rare health conditions. Because validation data quantifying the influence of false positives will typically be easier to obtain than data quantifying the influence of false negatives in this setting, the development concentrates on incidence rate estimates reflecting adjustments for PVP. This emphasis is not intended to diminish the potential influence of false negatives; rather, it reflects the logistical difficulties associated with obtaining data on false negatives as part of ongoing surveillance. If there is sufficient doubt surrounding the sensitivity of case ascertainment for any particular surveillance process, the proposed methodology should be applied with due caution.

Analysis

For a given surveillance period, it is assumed that case confirmation data are available for a random sample (selected without replacement) of provisional cases. Data obtained through such validation efforts allow estimation of PVP as well as adjustments to case counts to eliminate the bias due to false positives. To illustrate, suppose that for a set period (e.g., one year) of observation:

N = size of the at-risk population covered by the surveillance system;

M = count of provisional cases detected under the operational case definition;

M_T = count of true cases (unknown) among the provisional cases;

M_F = count of false positive cases (unknown) among the provisional cases = M - M_T;

S = number of provisional cases sampled for case confirmation;

C_T = count of confirmed true cases among those sampled;

C_F = count of cases determined to be false positives among those sampled = S - C_T.

The usual estimate of PVP is given by [3]:

$$\hat{PVP} = C_T/S = C_T/(C_T + C_F).$$

Noting that $\hat{PVP}$ is definable only when M > 0 (assuming also that S > 0) a reasonable estimate of the population of true cases which eliminates the false positive bias is:

$$\hat{M}_T = \begin{cases} 0 & , \quad M = 0 \\ \hat{PVP} \cdot M, & M > 0 \end{cases} \quad (1)$$

Case counts obtained through comprehensive surveillance may be considered inherently variable even though they are essentially census-level quantities, in the sense that a case count can be viewed as representing one observation from a hypothetically repeatable process [4-7]. For relatively rare conditions such case counts are often assumed to follow a Poisson distribution [6,7]. For example, suppose that all M provisional cases were to be reviewed so that the count of true cases M_T could be determined. When reporting the corresponding incidence rate R = M_T/N one might also make use of the variance estimate $\hat{Var}(R) = M_T/N^2$, based on the assumption that M_T represents one observation from a Poisson process [6,7]. Due to the estimation of PVP, however, the adjusted case count $\hat{M}_T$ cannot be treated in a similar fashion. Depending on the validation sample and the underlying PVP, for example, Var($\hat{M}_T$) can be well in excess of the variance that would be estimated under the assumption that $\hat{M}_T$ simply follows a Poisson distribution.

The remainder of this paper addresses three aspects of the problem outlined above: (i) a simple model for the true and false positive case counts within the defined framework, (ii) selected properties of $\hat{M}_T$ under a broadly applicable validation sample plan, and (iii) the relative frequency of coverage for interval estimates formulated using these properties.

A Case Count Model

To evaluate the proposed estimator $\hat{M}_T$, a working model characterizing the process underlying the case counts M, M_T, and M_F is needed. For a given at-risk population and surveillance period it will be assumed that the provisional case count M is generated according to a Poisson process with parameter λ. Each provisional case, independent of other provisional cases, will be assumed to be a true case with probability equal to the underlying PVP. These assumptions are reflected in the following mixture model [8]:

M ~ POI(λ);

M_T|M ~ BIN(M, PVP)

where POI denotes the Poisson distribution and BIN denotes the binomial distribution. The count of false positive cases is implicitly given by $M_F = M - M_T$. It is well-established that under this type of decomposition M_T and M_F are independent Poisson random variables such that $M_T \sim POI(\tau)$ and $M_F \sim POI(\phi)$, where $\tau = \lambda \cdot PVP$ and $\phi = \lambda \cdot (1\text{-}PVP)$ [9,10]. In this model, the parameter λ represents the average size of the recurring count of provisional cases and τ represents the average size of the recurring count of true cases among the provisional cases. The quantity $1/PVP$ can be viewed as the factor by which the count of true cases is inflated (on average) under the operational case definition. Finally, the parameters λ, τ and ϕ are implicitly dependent on the size of the at-risk population N; however, the functional form of this dependency is not important in the present development.

A Validation Sample Plan
This section examines several important properties of the estimator $\hat{M}_T$ when a fixed fraction of provisional cases are sampled for confirmation. The properties presented are derived in Appendix A. Letting $0 < f < 1$ denote the fixed sampling fraction, assume that the sample size $S = \lceil f \cdot M \rceil$ where the quantity $f \cdot M$ is rounded up. Under this procedure:

$$E[\hat{M}_T] = \tau \qquad (2)$$

and when $f \cdot \lambda$ is sufficiently large:

$$Var(\hat{M}_T) \cong \tau + \tau \cdot (1\text{-}PVP) \cdot \left(\frac{1\text{-}f}{f}\right). \qquad (3)$$

Equality (2) indicates that $\hat{M}_T$ is an unbiased estimator for the mean recurring count of true cases. The first component τ on the right-hand side of (3) represents the variance of the true case count M_T. The second component approximates the addition to variance that results from the case count adjustment based on $\hat{PVP}$. Note that for any given PVP the variance inflation factor is essentially constant as a result of holding the sampling fraction fixed.

It is noted in passing that when case populations are typically small, it may be feasible to adopt the practice of confirming all provisional cases. Under this approach $\hat{M}_T$ will be equivalent to the true case count M_T and it follows that $\hat{M}_T \sim POI(\tau)$. Based on familiar properties of the Poisson distribution [8] it follows that $E[\hat{M}_T] = Var(\hat{M}_T) = \tau$ and customary analysis methods are applicable.

Application
The remaining objective is the formulation of a simple method for constructing interval estimates for τ and the corresponding incidence rate. From (2) it is already known that $\hat{M}_T$ is an unbiased estimator of τ. In Appendix B it is shown that the following estimator is nearly unbiased for the right-hand side of (3):

$$\hat{Var}(\hat{M}_T) = \begin{cases} 0 & , \quad M = 0 \\ \hat{M}_T + \hat{M}_T \cdot (1\text{-}\hat{PVP}) \cdot \left(\frac{1\text{-}f}{f}\right), & M > 0 \end{cases} \qquad (4)$$

Based on (4) an approximate $(1\text{-}\alpha) \cdot 100\%$ confidence interval (adjusted for the false positive bias) for the recurring case count τ is given by:

$$\hat{M}_T \pm z_{\alpha/2} \cdot \sqrt{\hat{Var}(\hat{M}_T)}$$

where $z_{\alpha/2}$ represents the appropriate quantile of the standard normal distribution. The corresponding interval estimate for the population-based incidence rate is:

$$\hat{M}_T/N \pm z_{\alpha/2} \cdot \sqrt{\hat{Var}(\hat{M}_T)/N^2} \qquad (5)$$

where it will be recalled that N is the size of the at-risk population under surveillance. As an example, suppose that an interval estimate providing 95% relative frequency of coverage is desired for the population-based incidence rate. Table 1 shows the relative frequency with which interval (5) covers the underlying incidence rate in repeated Monte Carlo simulations involving various underlying values of PVP, λ, and f. For several cells $f \cdot \lambda$ is small and the coverage is below the nominal (95%) level, providing an illustration of where the interval estimation procedure begins to break down. In the remaining table cells coverage is close to the nominal level.

Table 1: Estimated Relative Coverage Frequencies of a Nominal 95% Interval with Variance Correction.

f	PVP = 0.70			PVP = 0.80			PVP = 0.90		
	λ = 100	λ = 500	λ = 1000	λ = 100	λ = 500	λ = 1000	λ = 100	λ = 500	λ = 1000
0.10	0.92	0.94	0.95	0.92	0.95	0.95	0.94	0.95	0.95
0.25	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
0.50	0.95	0.95	0.95	0.95	0.95	0.95	0.94	0.95	0.95

Table 2: Estimated Relative Coverage Frequencies of a Nominal 95% Interval w/o Variance Correction.

f	PVP = 0.70			PVP = 0.80			PVP = 0.90		
	$\lambda = 100$	$\lambda = 500$	$\lambda = 1000$	$\lambda = 100$	$\lambda = 500$	$\lambda = 1000$	$\lambda = 100$	$\lambda = 500$	$\lambda = 1000$
0.10	0.73	0.70	0.68	0.78	0.77	0.76	0.86	0.84	0.84
0.25	0.84	0.85	0.85	0.87	0.88	0.88	0.92	0.91	0.91
0.50	0.91	0.92	0.91	0.92	0.93	0.93	0.93	0.94	0.94

To illustrate the importance of the correction to the variance, Table 2 shows the relative coverage frequencies (again based on repeated simulations) if the adjusted case counts are simply assumed to follow a Poisson distribution. It is apparent that for smaller sampling fractions, coverage is well below the nominal level even with the larger case populations.

Extensions to independent subgroups (e.g., age groups) and aggregates (e.g., age-adjusted rates) are straightforward. Provided that subgroup boundaries do not divide the surveillance population too finely, the error associated with the interval estimation method described above should remain minimal.

Conclusion

This paper was motivated by considerations related to analysis of data from the brain injury surveillance system mentioned in the introduction. Beginning with surveillance year 2000, a number of participating states identified provisional cases which were subsequently determined to be false positives upon in-depth review. Preliminary estimates of PVP were observed to fall close to 0.9 for some states, suggesting the need for adjusted incidence rate estimates. This issue is also relevant in a broader context, as a wide range of PVP estimates have been reported for other surveillance systems [11].

Adjustments to incidence rate estimates to eliminate the false positive bias are straightforward. However, since the PVP estimates used to make such downward adjustments are subject to random variation, the adjusted rates have an additional source of variation beyond what is usually assumed. Interval estimates failing to account for this fact may have coverage frequencies well below the nominal level. This paper presents a simple method of interval estimation for rates that have been adjusted to remove the bias due to false positives, applicable in large-scale surveillance settings.

The methodology presented does not address the potential bias associated with false negatives. In situations where validation data also support estimation of sensitivity, surveillance case counts could be further adjusted to

reduce or eliminate such bias. This in turn would introduce another source of variation in the adjusted case counts and associated rates. Other types of sampling plans might also be considered. For example, a fixed sample size s^* might be preferred, in which case $S = \min(s^*, M)$ and an alternate expression for $\text{Var}(\hat{M}_T)$ would result. Technical details aside, the essential point is that data available from validation samples can have a nontrivial influence on point and interval estimates, and should be factored into surveillance statistics whenever feasible.

Appendix A. Moments of the Estimator $\hat{M}_T$

In the sampling procedure considered, the size of the validation sample depends on the provisional case count M. To make the analysis generic, the sample size will be denoted by s(M) where s($\cdot$) depends on the particular sampling procedure but is assumed positive whenever M > 0. The PVP-adjusted case count (1) can then be defined more precisely as:

$$\hat{M}_T \mid M = \begin{cases} 0 & , \quad M = 0 \\ (C_T/s(M)) \cdot M, & M > 0 \end{cases} \qquad (A.1)$$

where implicitly $\hat{\text{PVP}} = C_T/s(M)$. When M > 0 the distribution of C_T conditional on M and M_T is hypergeometric [12], that is, $C_T \mid M, M_T \sim \text{HYP}(s(M), M_T, M)$. It is not difficult to show that when M > 0 the distribution of C_T conditional on M only is binomial, that is, $C_T \mid M \sim \text{BIN}(s(M), \text{PVP})$. It follows that $E[\hat{M}_T \mid M] = M \cdot \text{PVP}$ for $M \geq 0$. Applying principles of conditional expectation [8] it is readily established that $\hat{M}_T$ is an unbiased estimator of $\tau = \lambda \cdot \text{PVP}$:

$$E[\hat{M}_T] = E[E[\hat{M}_T \mid M]] = E[M \cdot \text{PVP}] = \lambda \cdot \text{PVP}.$$

To determine $\text{Var}(\hat{M}_T)$ it is convenient to employ the following variance decomposition [8]:

$$\text{Var}(\hat{M}_T) = E[\text{Var}(\hat{M}_T \mid M)] + \text{Var}(E[\hat{M}_T \mid M]).$$

Since $E[\hat{M}_T|M] = M \cdot PVP$ it follows that $Var(E[\hat{M}_T|M]) = \lambda \cdot PVP^2$. Evaluation of the first component of variance is more complicated. Defining:

$$g(M) = \begin{cases} 0 & , \quad M = 0 \\ M^2/s(M), & M > 0 \end{cases}$$

it follows from (A.1) and the fact that $C_T|M \sim BIN(s(M), PVP)$ when $M > 0$ that:

$$Var(\hat{M}_T|M) = PVP \cdot (1\text{-}PVP) \cdot g(M).$$

The task is thus reduced to determining $E[g(M)]$. When $s(M) = \lceil f \cdot M \rceil$ it holds that $g(M) \leq M/f$ and hence that $E[g(M)] \leq E[M/f] = \lambda/f$. Given fixed f the upper bound is a good approximation provided that λ is sufficiently large, so that $E[g(M)] \cong \lambda/f$ and $E[Var(\hat{M}_T|M)] \cong PVP \cdot (1\text{-}PVP) \cdot \lambda/f$. Combining variance components and simplifying results in:

$$Var(\hat{M}_T) \cong \tau + \tau \cdot (1\text{-}PVP) \cdot \left(\frac{1\text{-}f}{f}\right). \qquad (A.2)$$

Numerical calculation of $Var(\hat{M}_T)$ across a range of values for PVP, λ, and f shows that for $f \geq 0.01$ and $f \cdot \lambda \geq 50$, the relative error of (A.2) is less than 0.01.

Appendix B. An Estimate of Var($\hat{M}_T$)
The following is proposed as an estimator of the right-hand side of (A.2):

$$\hat{Var}(\hat{M}_T) = \begin{cases} 0 & , \quad M = 0 \\ \hat{M}_T + \hat{M}_T \cdot (1\text{-}P\hat{V}P) \cdot \left(\frac{1\text{-}f}{f}\right), & M > 0 \end{cases} \qquad (B.1)$$

Defining:

$$h(M) = \begin{cases} 0 & , \quad M = 0 \\ M/s(M), & M > 0 \end{cases}$$

it follows from the treatment in Appendix A that the expected value of the variance estimator (B.1) conditioned on M is:

$$E[\hat{Var}(\hat{M}_T)|M] = M \cdot PVP \cdot \left(\frac{1}{f}\right) \cdot (PVP \cdot (1\text{-}PVP) \cdot h(M) + M \cdot PVP^2) \cdot \left(\frac{1\text{-}f}{f}\right).$$

Then, since $E[\hat{Var}(\hat{M}_T)] = E[E[\hat{Var}(\hat{M}_T)|M]]$ it follows that:

$$E[\hat{Var}(\hat{M}_T)] = \lambda \cdot PVP \cdot \left(\frac{1}{f}\right) \cdot \left(PVP \cdot (1\text{-}PVP) \cdot E[h(M)] + \lambda \cdot PVP^2\right) \cdot \left(\frac{1\text{-}f}{f}\right).$$

When $s(M) = \lceil f \cdot M \rceil$ it holds that $h(M) \leq 1/f$ and hence that $E[h(M)] \leq 1/f$. Given fixed f the upper bound is a good approximation provided that λ is sufficiently large. Substituting 1/f in place of $E[h(M)]$ results in:

$$E[\hat{Var}(\hat{M}_T)] \cong \lambda \cdot PVP \cdot \left(\frac{1}{f}\right) \cdot \left(PVP \cdot (1\text{-}PVP) \cdot \left(\frac{1}{f}\right) + \lambda \cdot PVP^2\right) \cdot \left(\frac{1\text{-}f}{f}\right).$$

Algebraic simplification results in:

$$E[\hat{Var}(\hat{M}_T)] \cong \lambda \cdot PVP + \lambda \cdot PVP \cdot (1\text{-}PVP) \cdot \left(\frac{1\text{-}f}{f}\right) \cdot \left(1 - \frac{1}{f \cdot \lambda}\right)$$

$$= \tau + \tau \cdot (1\text{-}PVP) \cdot \left(\frac{1\text{-}f}{f}\right) \cdot \left(1 - \frac{1}{f \cdot \lambda}\right).$$

As $f \cdot \lambda$ becomes large, approximation (A.2) results.

Competing interests
The author(s) declare that they have no competing interests.

References
1. Thurman DJ, Sniezek JE, Johnson D, Greenspan A, Smith SM: *Guidelines for Surveillance of Central Nervous System Injury* Atlanta: Centers for Disease Control and Prevention; 1995.
2. Greenland S: **Basic methods for sensitivity analysis and external adjustment.** In *Modern Epidemiology* 2nd edition. Edited by: Rothman KJ, Greenland S. Philadelphia: Lippincott-Raven Publishers; 1998:343-357.
3. Romaguera RA, German RR, Klaucke DN: **Evaluating public health surveillance.** In *Principles and Practice of Public Health Surveillance* Edited by: Teutsch SM, Churchill RE. New York: Oxford University Press; 2000:176-193.
4. Keyfitz N: **Sampling variance of standardized mortality rates.** *Human Biology* 1966, **38:**309-317.
5. Brillinger DR: **The natural variability of vital rates and associated statistics.** *Biometrics* 1986, **42:**693-734.
6. Greenland S, Rothman KJ: **Introduction to categorical statistics.** In *Modern Epidemiology* 2nd edition. Edited by: Rothman KJ, Greenland S. Philadelphia: Lippincott-Raven Publishers; 1998:231-252.
7. Anderson RN, Minino AM, Fingerhut LA, Warner M, Heinen MA: *Deaths: Injuries, 2001. National Vital Statistics Reports Volume 52.* Issue 21 Hyattsville MD: National Center for Health Statistics; 2004.
8. Casella GC, Berger RL: *Statistical Inference* Belmont CA: Duxbury Press; 1990.
9. Ross SM: *Introduction to Probability Models* 4th edition. San Diego: Academic Press; 1989.
10. Taylor HM, Karlin S: *An Introduction to Stochastic Modeling* Orlando: Academic Press; 1984.
11. German RR: **Sensitivity and predictive value positive measurements for public health surveillance systems.** *Epidemiology* 2000, **11:**720-727.
12. Cochran WG: *Sampling Techniques* 3rd edition. New York: Wiley; 1977.