BIOMETRICS 58, 684–688
September 2002

READER REACTION

Estimating the Variance of Disease-Prevalence Estimates from Population-Based Registries

Limin X. Clegg,* Mitchell H. Gail, and Eric J. Feuer

National Cancer Institute, National Institutes of Health,
6116 Executive Boulevard, MSC 8316, Bethesda, Maryland 20892-8316, U.S.A.
*email: clegg@mail.nih.gov

SUMMARY. We propose a new Poisson method to estimate the variance for prevalence estimates obtained by the counting method described by Gail et al. (1999, *Biometrics* **55**, 1137–1144) and to construct a confidence interval for the prevalence. We evaluate both the Poisson procedure and the procedure based on the bootstrap proposed by Gail et al. in simulated samples generated by resampling real data. These studies show that both variance estimators usually perform well and yield coverages of confidence intervals at nominal levels. When the number of disease survivors is very small, however, confidence intervals based on the Poisson method have supranominal coverage, whereas those based on the procedure of Gail et al. tend to have below-nominal coverage. For these reasons, we recommend the Poisson method, which also reduces the computational burden considerably.

KEY WORDS: Approximation; Cancer registry; Censoring; Kaplan–Meier estimator; Poisson distribution; Prevalence estimation; Variance and confidence interval for prevalence.

1. Introduction

Disease prevalence, the proportion of individuals in a population with a history of a particular disease at a point in time, is important for assessing the public health impact of the disease and to project medical care needs. Gail et al. (1999) described two general approaches for estimating disease prevalence using registry data: the counting method and the transition rate method. They also developed variance estimators for these prevalence estimators. Their proposed variance estimator for the counting method obtains one component of the variance by a bootstrap procedure. We propose a computationally simpler Poisson method and compare its performance to that of the variance estimator given by Gail et al. in simulations based on data from the Surveillance, Epidemiology, and End Results (SEER) program at the National Cancer Institute (NCI).

2. Methods

2.1 *Counting Method Prevalence Estimator*

Following Gail et al. (1999), we denote the age- and time-specific disease prevalence at calendar time s as $\pi(c_1, c_2, a_1, a_2, s)$, the probability that an individual who is alive at calendar time s and in the age range $[a_1, a_2)$ had disease incident in the age interval $[c_1, c_2)$, where $c_1 \leq a_1$ and $c_2 \leq a_2$. Note that Gail et al. (1999) used the notation $\pi(c_1, c_2, a, s)$ because they estimated prevalence in the age range $[a, a+1)$. Let (t, x, d) represent (calendar time at onset of disease, age at onset of disease, duration time from disease onset to calendar time s). Denote the disease incidence intensity as $\alpha(t, x)$ for an individual who has not previously been diagnosed with the disease at age x at calendar time t. Assume that an individual who was diagnosed with the disease at calendar time t at age x is at risk of death with intensity $\lambda(t, x+d, d)$ at calendar time $(t+d)$. Then

$$S(d; x; t) = \exp\left\{-\int_0^d \lambda(t, x+u, u)du\right\}$$

is the probability that an individual who develops disease at age x and date t will survive beyond duration d after disease incidence. Hence, the prevalence

$$\begin{aligned}\pi(c_1, c_2, a_1, a_2, s) &= N(a_1, a_2, s)^{-1} \int_{a_1}^{a_2} \int_{c_1}^{c_2} n^*(x, s-v+x) \\ &\quad \times \alpha(s-v+x, x) \\ &\quad \times S(v-x; x; s-v+x) \\ &\quad \times dxdv, \qquad (1)\end{aligned}$$

where $n^*(x,t)dtdx$ is the number of individuals who are at risk of first developing disease at calendar time $[t, t+dt)$ and age $[x, x+dx)$ and $N(a_1, a_2, s)$ is the number alive in the study population at calendar time s in the age range $[a_1, a_2)$.

The counting method obtains estimates of prevalence by dividing the estimated number of diseased persons by the study population size, taking loss to follow-up into account. Let X_i be the age at disease incidence for the ith case of a registry and T_i be the calendar time of the disease diagnosis for that case. Let Y_i be the potential calendar time of death and U_i the potential calendar time at loss to follow-up for the ith case. Unless both U_i and Y_i exceed s, we get to observe the minimum of Y_i and U_i. Let $I(\cdot)$ present an indicator function equaling one when the argument is true and zero otherwise. The counting method prevalence estimator at calendar time s for (1) is (Gail et al., 1999)

$$\begin{aligned}\hat{\pi}(c_1, c_2, a_1, a_2, s) &= N(a_1, a_2, s)^{-1} \\ &\times \Bigg[\sum I(c_1 \le X_i < c_2, s \le Y_i, s \le U_i, \\ &\qquad a_1 \le X_i + s - T_i < a_2) \\ &\quad + \sum \{ I(c_1 \le X_i < c_2, U_i < Y_i, s > U_i, \\ &\qquad a_1 \le X_i + s - T_i < a_2) \} \\ &\qquad \times \frac{\hat{S}(s - T_i; X_i; T_i)}{\hat{S}(U_i - T_i; X_i; T_i)} \Bigg], \end{aligned} \tag{2}$$

where summations are over all disease cases in the registry and $\hat{S}(d; X; T)$ is an estimator of $S(d; X; T)$.

2.2 *Gail's Bootstrap Variance Estimator*

Let M be the number of incident cases diagnosed in the age range $[c_1, c_2)$ who would be in the age range $[a_1, a_2)$ at calendar time s if they survived and let $\hat{\xi}_M$ be the estimated proportion of such cases who were alive at s. In particular, $M\hat{\xi}_M$ is the term in square brackets in equation (2). Assume that M follows a Poisson distribution. As M increases, $\hat{\xi}_M$ converges to ξ and $M\text{var}(\hat{\xi}_M)$ tends to the limiting variance σ^2, say. Gail et al. (1999) derived

$$\text{var}\{\hat{\pi}(c_1, c_2, a_1, a_2, s)\} \doteq N(a_1, a_2, s)^{-2} M \left[\sigma^2 + \xi^2\right]. \tag{3}$$

They estimated $\text{var}\{\hat{\pi}(c_1, c_2, a_1, a_2, s)\}$ by substituting $\hat{\xi}_M$ for ξ and obtaining a bootstrap estimate of $\text{var}(\hat{\xi}_M) = \sigma^2/M$. To reduce the computational burden, Gail et al. (1999) described how σ^2/M could be estimated from bootstrap resampling of samples of size $m < M$ cases.

2.3 *Proposed Method Based on a Poisson Approximation*

Suppose that we use the method of Kaplan and Meier (1958) to estimate the survival functions $S(d; X; T)$ separately within strata defined by demographic and disease-related variables (e.g., age at diagnosis, year of diagnosis, gender, race, and stage of disease). Suppose that all cases in a stratum have a common unknown survival distribution and that there are K such strata. Let $d_i = s - T_i$ denote the duration time from diagnosis to prevalence time and $m_i = U_i - T_i$ denote the duration of survival from diagnosis to loss to follow-up. We reexpress equation (2) as

$$\begin{aligned}\hat{\pi}(c_1, c_2, a_1, a_2, s) &= N(a_1, a_2, s)^{-1} \left\{ A + \sum_{k=1}^{K} \sum_{i=1}^{N_k} \hat{S}_k(d_i \mid m_i) \right\}, \end{aligned} \tag{4}$$

where A is the first sum in equation (2), i.e., the number of cases diagnosed in the age interval $[c_1, c_2)$ who were known to be alive and in the age interval $[a_1, a_2)$ at calendar time s, where the second sum in equation (4) is over all $L = \Sigma_{k=1}^{K} N_k$ patients lost to follow-up before s, and where N_k is the number of cases in stratum k who are lost to follow-up. To be precise, $L = \Sigma\{I(c_1 \le X_i < c_2, U_i < Y_i, s > U_i, a_1 \le X_i + s - T_i < a_2)\}$. We use $\hat{S}_k(d_i \mid m_i) = \hat{S}_k(d_i)/\hat{S}_k(m_i)$ to estimate the probability $S_k(d_i \mid m_i)$ that patient i will survive at least a duration d_i beyond cancer incidence (and therefore be alive at calendar time s) given that patient was diagnosed at calendar time T_i and lost to follow-up at calendar time $(T_i + m_i)$. Hence, $\Sigma_{k=1}^{K} \Sigma_{i=1}^{N_k} \hat{S}_k(d_i \mid m_i)$, hereafter denoted by B, is the estimated number of patients alive at calendar time s among those lost to follow-up. Note that the variables X_i and T_i are not included in $\hat{S}_k(d_i \mid m_i)$ because we assume, without loss of generality, that X_i and T_i are stratification variables that are used in determining the stratum index k. Our task is to estimate the variance of $\hat{\pi}(c_1, c_2, a_1, a_2, s)$ given by (4) and to construct a 95% confidence interval for the prevalence.

For rare diseases, A and $\{N_k\}$ are approximately independent Poisson variables (Haberman, 1978; Brillinger, 1986; Keiding, 1991). Consider the contribution $B_k = \Sigma_{i=1}^{N_k} \hat{S}_k(d_i \mid m_i) \equiv \Sigma_{i=1}^{N_k} \hat{p}_{ki}$ from stratum k. Assume the quantities $\hat{p}_{ki}$ are *i.i.d.* with mean p_k and are independent of N_k, which has mean λ_k. Then $\text{E}(B_k) = \text{E}(N_k)p_k = \lambda_k p_k$ and

$$\begin{aligned}\text{var}(B_k) &= \text{E}_{N_k}\text{var}\{(B_k \mid N_k)\} + \text{var}_{N_k}\{\text{E}(B_k \mid N_k)\} \\ &= \lambda_k \text{var}(\hat{p}_{ki}) + \lambda_k p_k^2 < \lambda_k p_k.\end{aligned}$$

Thus, B_k is underdispersed compared with a Poisson variable with mean $\lambda_k p_k$. This suggests that regarding $A + B \equiv A + \Sigma_{k=1}^{k=K} B_k$ as Poisson with mean estimated as $A + B$ will yield confidence intervals on $\pi(c_1, c_2, a_1, a_2, s)$ with coverage at or above nominal levels. A positive correlation between A and B would increase the variance, but simulations indicate the correlations are small (Section 3). For large numbers of events N_k in each stratum, however, this Poisson distribution for B_k becomes more accurate because B_k is stochastically equivalent to $B_k^* = \Sigma_{i=1}^{N_k} Z_{ki}$, where Z_{ki} are independent Bernoulli variates with means $\hat{p}_{ki}$. Standard moment-generating function arguments show that B_k^* is a Poisson variable with mean $\lambda_k p_k$. Hence, we anticipate that regarding $A + B$ as Poisson with estimated mean $A + B$ will yield confidence intervals on $\pi(c_1, c_2, a_1, a_2, s)$ with nominal coverage with large N_k.

To construct a confidence interval on $\pi(c_1, c_2, a_1, a_2, s)$, we rely on the well-known relationships between the Poisson distribution and the chi-square distribution (Johnson and Kotz, 1969) to obtain lower (π_L) and upper (π_U) $1 - \alpha$ level limits as

$$\pi_L = \chi^{-2}_{2(a+b),\alpha/2}/\{2N(a_1, a_2, s)\}$$

and

$$\pi_U = \chi^{-2}_{2(a+b+1),1-\alpha/2}/\{2N(a_1, a_2, s)\}. \tag{5}$$

In this equation, $a + b$ is the observed value of $A + B$ and $\chi^{-2}_{f,p}$ is the pth quantile of the chi-square distribution with f d.f., where f may be noninteger.

2.4 *Adaptation for a Definition of Prevalence Based on Calendar Year of Diagnosis*

Cases contributing to the prevalence $\pi(c_1, c_2, a_1, a_2, s)$ are those who developed the disease between age c_1 and c_2 and were aged $[a_1, a_2)$ at calendar time s (Section 2.1). It is also possible to define the eligible cases as those whose disease developed within $\tau \in [\tau_1, \tau_2)$ years before s and who were aged $[a_1, a_2)$ at s. Using this definition of eligible cases, one could define a prevalence $\pi(\tau_1, \tau_2, a_1, a_2, s)$ analogous to $\pi(c_1, c_2, a_1, a_2, s)$. To estimate $\pi(\tau_1, \tau_2, a_1, a_2, s)$ by the counting method, one modifies equation (2) by replacing the condition $c_1 \leq X_i < c_2$ by $\tau_1 \leq s - T_i < \tau_2$. The methods in Sections 2.2 and 2.3 can be used with this modification, as in the examples and simulations. These two definitions of prevalence are not the same because the eligible cases differ somewhat, as can be appreciated by drawing the corresponding parallelograms in Figure 1 of Gail et al. (1999).

More specifically, the interval τ between date of diagnosis and prevalence time s, age a at s, and the age c at disease diagnosis are linearly dependent: $a = c + \tau$. Because a_1 and a_2 are determined by the age grouping of the (available) population data at s, we specify either τ_1 and τ_2 or c_1 and c_2, depending on the aims of the study. For example, for the prevalence of childhood cancer survivors, we specify c_1 and c_2 as ages that define childhood. Thus, only the patients who were diagnosed at age $c \in [c_1, c_2)$ will be eligible to be included in the calculation. From the linear dependency, we know that these eligible childhood cancer cases were diagnosed within τ years before s, where $\tau \in [a_1 - c_2, a_2 - c_1)$ and $\tau \geq 0$. In contrast, in the $\pi(\tau_1, \tau_2, a_1, a_2, s)$ specification, only patients who were diagnosed within τ years before s will be the eligible cases, where $\tau \in [\tau_1, \tau_2)$; and the linear dependency dictates that these eligible cases were diagnosed at age $c \in [a_1 - \tau_2, a_2 - \tau_1)$ for all $c \geq 0$. Therefore, the only difference between the $\pi(c_1, c_2, a_1, a_2, s)$ and the $\pi(\tau_1, \tau_2, a_1, a_2, s)$ specifications is the determination of inclusion of eligible cases, which does affect the definition of prevalence.

3. Examples and Simulations

As an illustration, age-specific cancer prevalences and their variance estimates were calculated using population-based cancer registry data collected by the SEER Program at the NCI. Based on the 1990 census data, the SEER program covers about 14% of the U.S. population (see www.seer.cancer.gov for detailed information). We used 24 age-specific datasets (Table 1) of breast and brain cancers diagnosed in the original nine SEER geographic areas between 1987 and 1996. As pointed out by Gail et al. (1999), the survival distribution following breast cancer incidence is relatively favorable and the incidence rate is relatively high, whereas the survival distribution following brain cancer incidence is relatively unfavorable and the incidence rate is low. We used a single stratum to estimate $S(d; x; t)$ (i.e., $K = 1$). For the procedure of Gail et al. (1999), we randomly sampled a subset of size $m = 500$ cases with replacement from the original M cases for each bootstrap. We used $B = 100$ repetitions to estimate σ^2 in (3).

We also evaluate the performance of the Poisson procedure and the Gail estimator in realistic simulated samples obtained by resampling the data in Table 1. For each of the 24 data sets in Table 1, we resampled 500 data sets with replacement, assuming that the number of incident cancers follows a Poisson distribution. More specifically, for each of the 24 settings, the following steps were used to generate 500 independent simulated data sets:

(1) Generate M, the number of incident cases from a Poisson random variable with mean equal to the observed number of cases in the original data set.
(2) Resample M incident cases with replacement from the original cases.
(3) Repeat steps (1) and (2) 500 times to generate 500 data sets.

For each of the 500 simulations, we calculated $\hat{\pi}$, $\widehat{\text{var}}(\hat{\pi})$ from the Gail estimator (3) and from the Poisson approximation $(A + B)/\{N(a_1, a_2, s)\}^2$, and 95% confidence intervals based on equation (5) and on the normal approximation in the Gail method. From the 500 simulations, we calculated the mean of $\hat{\pi}$, the standard error (SE) of $\hat{\pi}$, the mean of the estimates of standard error corresponding to the variance estimates above, the empirical coverages of the confidence intervals, namely, the percentage of times the confidence intervals included the $\hat{\pi}_0$ obtained from the original data, and the correlation coefficient between A and B. All simulations and computations were performed in SAS (SAS Institute, Cary, North Carolina).

Table 1 presents results for the estimated prevalences on January 1, 1997 ($\tau_1 = 0, \tau_2 = 11$), from the original and simulated data sets. The examples cover a wide range of estimated prevalences, numbers of incident cases, numbers of cases lost before the date of prevalence estimation, and percentages lost. For example, among black women aged <55 years, $236/(236 + 2855) = 7.6\%$ were lost compared with $2/(2+4) = 33\%$ of black males aged 65–74 with brain cancer (Table 1). Estimates of prevalence per 100,000 persons ranged from 3175 for breast cancer in white women aged ≥75 years to 6 for black men with brain cancer aged ≥75 years. Estimates of SE($\hat{\pi}$) from the Poisson method and from the Gail method (3) agreed well in most cases in the original data.

In simulations, the average estimates of SE($\hat{\pi}$) from the two methods also agreed well with each other and with the empirical standard error estimate of $\hat{\pi}$ (Table 1). It is noteworthy, however, that the average estimate from the Poisson method exceeded that from the Gail method in 22 of the 24 settings examined, as suggested by the discussion of underdispersion in Section 2. The average estimated standard errors were within 0.01 units of each other in the two remaining cases.

In most cases, the estimated coverages of both methods fell within the interval (93.1%, 96.9%) that would be expected to include 95% of coverage estimates if the procedures had nominal 95% levels. In fact, the bootstrap method of Gail et al. (1999) yielded coverages in that range except for three age groups of black men with brain cancer, for whom the numbers of survivors were very small. The coverages were subnominal in these cases, and there is a suggestion of slight subnominal coverage for black women with brain cancer, for whom the numbers of survivors were again very small. In these cases, the coverage of the Poisson confidence interval exceeded the nominal level (i.e., was conservative). Note that the coverage

Table 1

Estimated prevalence with standard errors on January 1, 1997, of breast and brain cancers diagnosed in the previous 10 years and simulated results on the average estimated standard errors and on coverage of the 95% confidence intervals

Race gender, cancer, age	Original observed data				Simulations[a]				
			Estimated SE		SE of $\hat{\pi}$			Coverage (%)	
	A/L[c]	Estimated prevalence[b] $\hat{\pi}_0$	Gail method	Poisson method	Empirical	Mean Gail	Mean Poisson	Gail method	Poisson method
Black female breast cancer									
<55	2855/236	237	4.33	4.28	4.21	4.28	4.28	95.2	95.4
55–64	1482/94	1624	41.54	41.16	39.26	40.90	41.20	95.2	95.6
65–74	1519/48	2097	51.63	53.10	53.38	52.52	53.06	94.4	94.8
≥75	1253/43	2271	63.11	63.30	64.28	62.61	63.37	94.8	94.6
White female breast cancer									
<55	23633/2124	329	1.98	2.06	2.09	2.06	2.06	95.4	95.8
55–64	17333/1161	2226	16.50	16.40	16.09	16.23	16.40	95.4	95.8
65–74	22303/476	2999	19.60	19.89	20.49	19.60	19.89	93.2	94.0
≥75	23874/613	3175	19.57	20.34	20.68	19.98	20.33	95.0	95.0
Black female brain cancer									
<55	109/17	10	0.88	0.86	0.88	0.87	0.87	94.6	95.8
55–64	14/1	15	3.70	3.90	3.64	3.87	3.88	93.4	98.0
65–74	7/1	10	3.45	3.71	3.70	3.62	3.66	93.6	97.0
≥75	9/0	16	4.78	5.29	5.16	5.14	5.19	93.8	95.2
White female brain cancer									
<55	1220/191	18	0.50	0.48	0.48	0.48	0.48	95.8	95.4
55–64	169/19	22	1.62	1.62	1.68	1.62	1.62	94.4	94.4
65–74	124/8	17	1.63	1.49	1.48	1.48	1.49	93.6	94.8
≥75	77/7	10	1.15	1.16	1.11	1.16	1.16	95.6	97.6
Black male brain cancer									
<55	134/31	13	1.06	1.05	1.09	1.06	1.05	94.0	95.0
55–64	13/0	17	5.12	4.77	4.89	4.75	4.76	92.4	96.8
65–74	4/2	7	3.45	3.58	3.43	3.38	3.40	90.4	97.8
≥75	2/0	6	4.30	4.34	4.19	3.89	3.90	87.8	98.8
White male brain cancer									
<55	1676/214	23	0.55	0.54	0.56	0.54	0.54	94.2	94.4
55–64	199/16	27	1.93	1.84	1.89	1.84	1.84	93.8	95.0
65–74	147/7	24	1.81	1.95	1.96	1.95	1.94	94.2	96.0
≥75	51/4	12	1.59	1.62	1.60	1.61	1.62	96.2	97.4

[a] Based on 500 independent simulations for each row of the table. The empirical SE is the standard error of $\hat{\pi}$ in the 500 simulations, whereas Mean Gail and Mean Poisson denote the average estimated SE. The coverage is the percentage of times in 500 simulations that the confidence interval included the true $\hat{\pi}_0$ obtained from the original data. The nominal level of these confidence intervals is 95%.

[b] Prevalence per 100,000 persons.

[c] Number of cases alive at the end of follow-up, A, and number of cases lost before the end of follow-up, L.

of the Poisson confidence intervals exceeded those of the Gail method in 20 of the 22 untied cases (Table 1) despite the fact that standard errors from the two methods were very similar. Thus, the higher coverage of the Poisson method may reflect skewness in the Poisson model for the prevalence distribution with low prevalence in addition to slightly larger variance.

The correlation coefficients between A and B ranged from −0.09 to 0.10, with most being near zero (data not shown). To determine if similar results obtained with more severe loss to follow-up, we also used data sets that have nearly double the percentage of loss to follow-up seen in Table 1. Results very similar to those in Table 1 were obtained (data not shown).

4. Discussion

Both the Gail method and the Poisson method performed well in realistic simulations whenever the numbers of survivors exceeded 10. The Poisson method had the advantage of having coverage equaling or exceeding nominal levels when the numbers of survivors were small, whereas the procedure of Gail et al. had subnominal coverage in this circumstance. In addition, the Poisson method reduced the computational burden considerably by avoiding the bootstrap procedure used by Gail et al. This is an advantage for applications requiring annual evaluation of prevalence in multiple subgroups in a database with millions of cases. Thus, we recommend the

Poisson method and are planning to incorporate it into public software (SEER*STAT) for analysis of registry data (see www.seer.cancer.gov).

Acknowledgements

The authors wish to thank the editor and two referees for their critical reading of the manuscript and helpful suggestions, which lead to considerable improvement of earlier versions. The authors also thank Steve Scoppa and Michael Depry of IMS for their assistance in computation.

Résumé

Nous proposons une nouvelle méthode poissonienne pour estimer la variance d'estimations de prévalence par la méthode de comptage décrite par Gail et al. (1999) et pour construire un intervalle de confiance de la prévalence. Nous évaluons simultanément la procédure poissonienne et la procédure basée sur la méthode bootstrap proposée par Gail et al. sur des échantillons simulés engendrés en rééchantillonnant des données réelles. Ces études montrent qu'à la fois la variance des estimateurs est généralement bonne et fournit un recouvrement des intervalles de confiance aux niveaux nominaux. Quand le nombre de survivants d'une maladie est très faible, cependant, les intervalles de confiance basés sur la méthode poissonienne a un recouvrement supérieur au niveau nominal, alors que pour ceux basés sur la procédure de Gail et al. il est inférieur. Pour ces raisons, nous recommandons la méthode poissonienne, qui réduit aussi la charge de calcul.

References

Brillinger, D. R. (1986). The natural variability of vital rates and associated statistics (with discussion). *Biometrics* **42,** 693–734.

Gail, M. H., Kessler, L., Midthune, D., and Scoppa, S. (1999). Two approaches for estimating disease prevalence from population-based registries of incidence and total mortality. *Biometrics* **55,** 1137–1144.

Haberman, S. (1978). Probabilistic treatment of the incidence and prevalence of disease. *Social Science and Medicine, Series A* **12,** 159–161.

Johnson, N. L. and Kotz, S. (1969). *Distributions in Statistics: Discrete Distributions.* New York: Wiley.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53,** 457–481.

Keiding, N. (1991). Age specific incidence and prevalence: A statistical perspective. *Journal of the Royal Statistical Society, Series A* **154,** 371–412.

Received February 2001. *Revised May* 2002.
Accepted May 2002.