A misclassification model is presented for the assessment of bias in rate ratios estimated by person-time analyses of automated medical care databases. The model allows for misclassification of events and person-time and applies to both differential and nondifferential errors. The focus is on medical care exposures that occur at discrete points in time (e.g., vaccinations) and on adverse events that are closely associated in time. Bias corrections for rate ratios and binomial tests of equality of event rates during exposed and unexposed person-time are developed and illustrated. For nondifferential under- or over-ascertainment of events, the observed rate ratio (r) is unbiased at the null hypothesis (true rate ratio \( R = 1 \)), negatively biased when \( R > 1 \), and positively biased when \( R < 1 \) (i.e., biased toward the null). Differential under-ascertainment of unexposed events and differential over-ascertainment of exposed events positively bias \( r \) when \( R = 1 \). Differential event sensitivities cause larger biases in rate ratios than differential false event rates. False positive exposures bias observed event rate ratios more than false negative exposures. Biases are small when event sensitivities are nondifferential and when less than 10% of database exposures and events are false. The usefulness of the model for critical sensitivity analysis is illustrated by an example from a linked database study of childhood vaccine safety. Greater dissemination of data quality assessments, sensitivity analyses, and methods used to supplement automated databases are needed to further our understanding of the appropriate role of medical care databases in epidemiologic research. Am J Epidemiol 1996;144:782–92.

A critical question in such safety evaluations is whether medical care databases developed for administrative and clinical versus research purposes can meet the quality standards required for valid assessments of the causality of adverse events. For example, vaccine safety evaluations require very high data quality standards because of extremely limited margins for error (4, 5). Erroneous associations can undermine confidence in vaccines, reduce vaccine acceptance, and increase vaccine-preventable diseases; and failure to detect true associations can result in false confidence. There has been extensive discussion in the pharmacoepidemiologic literature of the use of automated databases in research and the need to identify and overcome their weaknesses (6–9). Graham and Smith (10) discussed issues raised by misclassification in a large managerial database, and they simulated the biases in person-based associations between drug exposures and adverse medical events.

In this paper, I develop a misclassification model for cohort event rates that allows for measurement error in the number of events and person-time at risk. Previous studies of misclassification bias have focused on person-based analyses of 2 × 2 tables of counts of exposure and disease status from cohort or case-control
designs (11–15). Modeling misclassification bias in person-time analysis of event rates requires new methodology. I develop event misclassification and person-time misclassification model components and link them through person-time exposure specificity and sensitivity parameters.

This model can be used to define database quality standards required for valid assessment of exposure-outcome associations, to perform critical sensitivity analyses, and to correct rate ratios and hypothesis tests for misclassification bias. The model focuses on discrete exposures (e.g., vaccinations) and on medical events that are closely associated in time (so-called adverse events).

**MODEL**

The rate model describes the experience of an open cohort of individuals who are observed for variable periods of time. I assume that exposures occur at discrete times and that exposed individuals are at elevated risk of an adverse event for a short period of time, \( W \) (e.g., 14 days).

Denote the proportion of person-days exposed by \( H \), the event rates during unexposed and exposed periods by \( \lambda_u \) and \( \lambda_e \), and the numbers of events occurring during unexposed and exposed periods by \( N_u \) and \( N_e \). The rate ratio \( R = \lambda_e/\lambda_u \) can be expressed as

\[
R = (((1-H)/H)(N_e/N_u)).
\]

If \( N_e \) and \( N_u \) are independent Poisson variates, the null hypothesis of no association between exposure and event \( (R = 1) \) is tested against the simple alternative \( (R > 1) \) by referring \( N_e \) to a binomial distribution with \( N = N_u + N_e \) trials and "success" probability \( H \). The standard normal test provides an adequate approximation when \( N \) is moderately large.

The observational model is derived by incorporating misclassification into the underlying rate model. The observed exposures and events are assumed to be obtained from automated medical care databases that are subject to measurement error. Sources of measurement error include: medical care not sought for events; incorrect diagnoses of events or non-events; and event and exposure dates incorrectly entered into automated databases.

Figures 1 and 2 illustrate the misclassification model. Upper-case letters are used for true quantities and lower-case letters for observed quantities. \( D_e \) and \( D_u \) denote the true numbers of exposed and unexposed person-days. The observed numbers of exposed and unexposed person-days are denoted by \( d_e \) and \( d_u \). The total number of observed person-days is assumed to be free of measurement error, i.e.,

\[
D = D_e + D_u = d_e + d_u.
\]

\( N_e \) and \( N_u \) denote the true numbers of exposed and unexposed events; \( n_e \) and \( n_u \) denote the observed numbers of exposed and unexposed events; and \( m_e \) and \( m_u \) denote the numbers of exposed and unexposed false events.

The relations between true and observed quantities are expressed in terms of the following model param-

\[
\begin{align*}
D_e & \quad \lambda_e \quad \text{Ne} \quad a_2b_2 \quad \text{Ne} \quad b_2 \quad \text{me} \quad a_2(1-b_2) \\
D_u & \quad \lambda_u \quad \text{Nu} \quad a_1b_1 \quad \text{nue} \quad b_1 \quad \text{mu} \quad a_1(1-b_1)
\end{align*}
\]

**FIGURE 1.** Event misclassification model.
MISCLASSIFICATION PARAMETERS

INTERPRETATION OF PERSON-TIME

As can be seen from figures 1 and 2, the person-time exposure sensitivity \( b_2 \) and specificity \( b_1 \) parameters are included in the event misclassification model and in the person-time misclassification model. In figure 1, \( b_2 \) is defined as the probability that an exposed event is observed as exposed, and \( b_1 \) is defined as the probability that an unexposed event is observed as unexposed. In figure 2, \( b_2 \) is defined as the probability that an exposed time interval is observed as exposed, and \( b_1 \) is defined as the probability that an unexposed time interval is observed as unexposed. These definitions are equivalent when person-time is expressed in units of risk windows \( W \) and when the probability of observing an exposed time interval as exposed is independent of the occurrence of an event in the time interval.

An exposure database with a person-time exposure sensitivity parameter of \( b_2 = 0.90 \) captures 90 percent of true exposures. Interpretation of \( b_1 \) is facilitated by relating it to the proportion of false exposures in a database. The number of true and false observed exposures are given by \( b_2 D_e \) and \( (1 - b_1) D_u \), respectively, and the proportion of false observed exposures is given by

\[
G = \frac{(1 - b_1) D_u}{b_2 D_e + (1 - b_1) D_u}.
\]

Substitution of \( H D \) and \((1 - H) D\) for \( D_e \) and \( D_u \) gives:

\[
b_1 = \frac{((1 - H) - (1 - (1 - b_2) H)) G}{((1 - H)(1 - G))}.
\]

\( b_1 \) equals 1 when \( G = 0 \) and decreases with slope \(-b_2 H(1 - G)^2\) as \( G \) increases. For small \( G \), \( b_1 \) is only slightly less than 1 when \( H \) is small. For example, if 5 percent of exposures are false, and \( H = 0.10 \), and \( b_2 = 0.90 \), then

\[
b_1 = 0.9 - (1 - 0.1(0.1))(0.05))/((0.9)(0.95)) = 0.995.
\]

When \( G \) is increased to 0.10 and 0.20, \( b_1 \) decreases to 0.989 and 0.975, respectively. These would be extremely high values for person-based definitions of exposure specificity, but are modest levels for person-time exposure specificity parameters. Because \( G \) relates more directly to the occurrence of false exposures in the database than does \( b_1 \), we present numerical illustrations in terms of \( b_2 \) and \( G \). We restrict attention to moderate and high quality exposure databases with \( b_2 \geq 0.80 \) and \( G \leq 0.10 \).

The true proportion of person-time exposed \( (H) \) is related to the observed proportion of person-time exposed \( (h) \) and the proportion of false exposures \( (G) \) by:

\[
H = (1 - G) h / b_2.
\]

**Am J Epidemiol** Vol. 144, No. 8, 1996
In order to facilitate the interpretation of person-time rates of false events as measures of database quality, we relate them to the proportion of false events in a database. \( D_e f_e + D_u f_u \) of the \( n \) observed events are false events and the proportion of false observed events is

\[
F = (D(Hf_e + (1 - H)f_u))/n. \tag{3}
\]

When over-ascertainment of events is nondifferential (i.e., \( f_e = f_u = f \)), the person-time false event rate is the product of the proportion of false events and the observed event rate, i.e., \( f = F(n/D) \). For example, if the observed event rate is 0.0001 and the false event rate among exposed person-time is twice the false event among unexposed person-time (i.e., \( f_e = 2f_u \) and 5 percent of person-time is exposed (\( H = 0.05 \)), then a false event proportion of \( F = 0.1 \) corresponds to person-time false event rates of \( f_e = 0.00002 \) and \( f_u = 0.00001 \).

**DERIVATION OF BIAS**

Figure 1 illustrates the following system of linear equations:

\[
n_e = a_2 b_2 \lambda_e D_e + a_1 (1 - b_1) \lambda_u D_u + b_2 f_e D_e + (1 - b_1) f_e D_u \tag{4}
\]

\[
n_u = a_2 (1 - b_2) \lambda_e D_e + a_1 b_1 \lambda_u D_u + (1 - b_2) f_e D_e + b_1 f_u D_u, \tag{5}
\]

where \( m_e \) and \( m_u \) have been expressed as \( f_e D_e \) and \( f_u D_u \), and \( N_e \) and \( N_u \) have been expressed as \( \lambda_e D_e \) and \( \lambda_u D_u \). The ratio of exposed to unexposed observed events is given by

\[
n_e/n_u = [(a_2 b_2 HR + a_1 (1 - b_1)(1 - H))\lambda_u + f_e b_2 H + f_u (1 - b_1)(1 - H)]\lambda_e + f_e (1 - b_2) H + f_u b_1 (1 - H). \tag{6}
\]

Figure 2 illustrates the following system of linear equations:

\[
d_e = b_2 D_e + (1 - b_1)(D - D_e) \tag{7}
\]

\[
(D - d_e) = (1 - b_2) D_e + b_1 (D - D_e), \tag{8}
\]

where the total person-days of observation (\( D \)) is expressed in units of \( W \) and assumed free of measurement error.

The observed proportion of recently exposed person-days \( (h = d_e/D) \) is related to the true proportion of recently exposed person-days \( (H = D_e/D) \) by

\[
h = (1 - b_1) + (b_1 + b_2 - 1)H. \tag{9}
\]

Exposure databases with \( b_1 + b_2 < 1 \) are of inadequate quality even for administrative and clinical purposes, because \( h \) and \( H \) would be inversely related. No generality is lost by restricting attention to \( b_1 + b_2 > 1 \). Notice that \( h < 1 \) requires that

\[
H < b_1/(b_1 + b_2 - 1). \tag{10}
\]

The observed ratio of unexposed to exposed person-days is given by

\[
(1 - h)/h = (b_1 - (b_1 + b_2 - 1)H)/((1 - b_1) + (b_1 + b_2 - 1)H). \tag{11}
\]

The observed rate ratio

\[
r = (((1 - h)/h)(n_e/n_u) \tag{12}
\]

may be expressed as a function of the true rate ratio \( (R) \), the true event rate during unexposed time (\( \lambda_u \)), the true proportion of exposed person-days (\( H \)), and the misclassification parameters \( (a_1, a_2, f_e, f_u, b_1, b_2) \). Substitution of equations 6 and 11 into equation 12 yields

Am J Epidemiol  Vol. 144, No. 8, 1996
\[
    r = (b_1 - (b_1 + b_2 - 1)H)/(a_2(b_2 H + a_1(1 - b_1)(1 - H)) + f_e b_2 H + f_u (1 - b_1)(1 - H))
\]

Misclassification bias is the difference \((r - R)\) between the observed and true rate ratios. The following submodels are distinguished:

1. nondifferential under-ascertainment of events: \(f_e = f_u = 0\), \(a_1 = a_2 < 1\);
2. differential under-ascertainment of events: \(f_e = f_u = 0\), \(a_1 \neq a_2\);
3. nondifferential over-ascertainment of events: \(a_1 = a_2 = 1\), \(f_e = f_u > 0\); and
4. differential over-ascertainment of events: \(a_1 = a_2 = 1\), \(f_e \neq f_u\).

The general model makes simultaneous allowances for false positive events and incomplete ascertainment of true events and allows for false positive and false negative errors in classifying events and person-time as exposed or unexposed.

**DIRECTION AND MAGNITUDE OF BIAS**

Expression 13 indicates that \(r\) is an increasing function of \(R\). When \(R = 0, r > 0\) and when \(R\) goes to infinity, \(r\) converges to

\[
    (b_2(b_1 - (b_1 + b_2 - 1)H))/((1 - b_2)((1 - b_1) + (b_1 + b_2 - 1)H)),
\]

which is finite if and only if \(b_2 < 1\). Hence, when \(b_2 < 1\), there is a value of \(R\), say \(R^*\), for which \(r > R\) (i.e., bias > 0) when \(R < R^*\), and \(r < R\) (i.e., bias < 0) when \(R > R^*\). This value of \(R^*\) can be determined by equating expression 13 to \(R\) and solving for the positive root of the resulting quadratic equation.

When \(R = 1, r > 1\) (i.e., bias > 0) if and only if

\[
    (a_2 - a_1) > (f_u - f_e)/\lambda_u.
\]

For nondifferential under- and over-ascertainment of events, \(r\) is unbiased when \(R = 1\), positively biased when \(R < 1\), and negatively biased when \(R > 1\). For differential under-ascertainment, \(r\) is positively biased at \(R = 1\) if and only if \(a_2 > a_1\). For differential over-ascertainment, \(r\) is positively biased at \(R = 1\) if and only if \(f_e > f_u\).

For the special case of error-free exposure data \((b_1 = b_2 = 1)\), expression 13 becomes

\[
    r = (a_2 \lambda_u R + f_e)/(a_1 \lambda_u + f_u),
\]

and bias \((r) = ((a_2 - a_1)\lambda_u R + (f_e - f_u R))/(a_1 \lambda_u + f_u),\)

from which it is evident that bias can be positive or negative.

When \((a_2 - a_1) > f_u/\lambda_u\), bias > 0 for all \(R > 0\). If \((a_2 - a_1) < f_u/\lambda_u\), bias > 0 if and only if

\[
    R < f_u/(f_u - (a_2 - a_1)\lambda_u).
\]

For example, when \(b_1 = b_2 = 1, a_2 = 0.9, a_1 = 0.75, \lambda_u = 0.001, f_u = 0.0002, f_e = 0.0003\), \(r\) is positively biased when \(R < 6\) and negatively biased when \(R > 6\). When event sensitivity is nondifferential, \((a_2 = a_1)\), \(r\) is positively biased when \(R < f_u/\lambda_u\) and negatively biased when \(R > f_u/\lambda_u\). When \(b_1 = b_2 = 1\) and \(f_u = f_e = 0\), \(r = AR\) and bias = \((A - 1) R\), where \(A = a_2/a_1\) is the ratio of exposed event sensitivity to unexposed event sensitivity. In this case, bias > 0 for all \(R\) when \(A > 1\), and bias < 0 for all \(R\) when \(A < 1\).

The percent relative bias in \(r\) when exposure data are error-free is graphed in figure 3 for ranges of values of \(a_1, a_2, f_e/\lambda_u, \text{and } f_u/\lambda_u\). It is evident from figure 3 that differential event sensitivities \((a_1 \neq a_2)\) impact rate ratio bias more than do differential false event rates \((f_e \neq f_u)\). The percent relative bias in \(r\) when event data are error-free is graphed in figure 4 for ranges of \(A = a_2/a_1, G, b_2, \text{and } H\). Notice from figure 4 that the relative bias is strongly dependent on the ratio of event sensitivities. Note also that increases in the proportion of false exposures (i.e., reductions in exposure specificity) have greater effects on relative bias than reductions in exposure sensitivity.
HYPOTHESIS TESTS

The observed number of exposed events is given by

\[ n_e = a_1(1 - b_1)N + (a_2b_2 - a_1(1 - b_1))N_e + f_e b_2 D_e + f_u(1 - b_1)D_u. \]  \hspace{1cm} (15)

The observed total number of events \( n = n_e + n_u \) is related to the true numbers of exposed \( (N_e) \) and unexposed events \( (N_u) \) by

\[ n = a_1N + (a_2 - a_1)N_e + f_e D_e + f_u D_u. \]  \hspace{1cm} (16)

Solving equations 15 and 16, we have

\[ N_e = (n_e - (1 - b_1)n - f_e D_e(b_1 + b_2 - 1))(a_2(b_1 + b_2 - 1)) \]  \hspace{1cm} (17)

and

\[ N = (n - f_e D_e - f_u D_u)a_1 - ((A - 1)a_1 A)(n_e - (1 - b_1)n - f_e D_e(b_1 + b_2 - 1))(b_1 + b_2 - 1). \]  \hspace{1cm} (18)

When estimates of the misclassification model parameters \( (a_1, a_2, b_1, b_2, f_e, f_u) \) are available, expressions 17 and 18 can be used to compute the corrected true number of exposed and total events from the observed numbers of events and observed proportion of person-days exposed. The corrected proportion of events expected to be exposed under the null hypothesis \( (R = 1) \) is given by

\[ H = (h - (1 - b_1))/(b_1 + b_2 - 1). \]  \hspace{1cm} (19)

As an example, suppose that 6.4 percent of 500,000 person-periods are observed to be exposed (i.e., \( h = 0.064 \)) and that \( n = 57 \) total events and \( n_e = 7 \) exposed events are observed. The observed rate ratio is

\[ r = ((1 - 0.064)(7))/(0.064)(50) = 2.05. \]

Under the null hypothesis, \((57)(0.064) = 3.65\) exposed events are expected and the variance of \( n_e \) is \((57)(0.064)(0.936) = 3.42\). Applying the conditional binomial test to the uncorrected data, the \( Z \) statistic,
FIGURE 4. Percent relative bias in rate ratio: no false events ($f_e = f_u = 0$). $R$ = true rate ratio; $f_e(f_u)$ = false event rate for exposed (unexposed) person-time; $a_2(a_1)$ = sensitivity of exposed (unexposed) events; $G$ = proportion of false exposures; $b_2$ = sensitivity of exposure; $H$ = exposure prevalence.
\[ Z = \frac{\text{obs} - \text{exp}}{\sqrt{\text{var}}} = \frac{7 - 3.65}{1.85} = 1.81, \]

has a significance probability of \( p = 0.035 \) and the null hypothesis is rejected in favor of an increased risk during exposure periods.

To illustrate the effect of correcting for misclassification bias, we suppose that the database captures 90 percent of true events, independent of exposure status (i.e., \( a_1 = a_2 = 0.90 \)), and that 10 percent of events are false (\( F = 0.10 \)). We further suppose that 10 percent of exposures in the database are false (\( G = 0.10 \)) and that exposures are 80 percent sensitive (\( b_2 = 0.80 \)). Assuming nondifferential false event rates (\( f_e = f_u \)), the false event rate is
\[ f = F(n/D) = (0.1)\frac{57}{500,000} = 0.00001 \text{ per person-period}. \]

Using equation 2, which is equivalent to equation 19, the proportion of exposed person-time is corrected to
\[ H = \frac{(1 - 0.1)(0.064)}{0.8} = 0.072. \]

For these values of \( b_2, G, \) and \( H, \) the exposure specificity parameter (expression 1) is
\[ b_1 = \frac{(0.928 - (1 - (0.2)(0.072))(0.1))/(0.928)(0.9)}{0.928}, \]
and \( (b_1 + b_2 - 1) = 0.792. \) The corrected numbers of exposed and unexposed person-periods are
\[ D_e = (0.072)(500,000) = 36,000, \text{ and } D_u = (0.928)(500,000) = 464,000. \]

The number of exposed events (expression 17) is corrected to
\[ N_e = (7 - (0.008)(57) - (0.360)(0.792))/((0.9)(0.792)) = 8.78. \]

The total number of true events (expression 18) is corrected to
\[ N = (57 - 5)/(0.9) = 57.78. \]

The corrected rate ratio is
\[ R = (0.928)(8.78)/(0.072)(49) = 2.31, \]
which is 13 percent larger than the observed rate ratio \( (r = 2.05). \) The expected value and the variance of the corrected numbers of exposed events are
\[ (57.78)(0.072) = 4.16 \text{ and } (57.78)(0.072)(0.928) = 3.86, \text{ respectively}. \]

Corrected for bias toward the null, the \( Z \) statistic for the conditional binomial test becomes
\[ (8.78 - 4.16)/1.96 = 2.36. \]

The corrected significance probability \( p = 0.009 \) provides stronger evidence of increased risk during exposed periods.

Correction for positive bias is required when exposed events were captured at a higher rate than unexposed events, say
\[ A = a_2/a_1 = 0.90/0.60 = 1.5. \]

In this case, the values of \( H \) and \( b_1 \) and the corrected number of exposed events in this example are unchanged, but the corrected number of total true events (expression 18) increases:
\[ N = (57 - 5)/(0.6) - (0.5/0.9)(7 - (0.008)(57) - (0.360)(0.792))/0.792 = 86.67 - 4.39 = 82.28. \]

The corrected rate ratio becomes
\[ R = (0.928)(8.78)/(0.072)(73.5) = 1.54, \]
which is 25 percent smaller than the observed rate ratio. The expected value and variance of \( N_e \) are now
\[ (82.28)(0.072) = 5.92 \text{ and } (82.28)(0.072)(0.928) = 5.50, \text{ respectively}. \]
The Z statistic becomes

\[ Z = \frac{8.78 - 5.92}{2.35} = 1.22 \]

and the significance probability \( p = 0.11 \) is consistent with no elevated risk during exposed periods.

**SENSITIVITY ANALYSIS**

We now present an example of sensitivity analysis based on the misclassification model that uses preliminary data from a linked database study of childhood vaccine safety being conducted in a large HMO (17). Automated vaccination and diagnostic data were linked for 58,572 child-years of observation time. The following screening data were observed for the association between recent diphtheria-tetanus-pertussis (DTP) vaccine exposure and acute asthma/bronchitis.

Thirty-one of a total of 786 acute asthma/bronchitis events occurred within 14 days of DTP vaccination. Of total observed child-days, 3.33 percent were exposed to DTP vaccination within \( W = 14 \) days. The computed outcome event rate ratio for recently vaccinated versus not recently vaccinated child-days is

\[ r = \frac{(0.967)(31)}{(0.033)(755)} = 1.20 \]

for the DTP-asthma/bronchitis association.

The results of sensitivity analyses are illustrated in figure 5, where the true rate ratio \( (R) \) is graphed over plausible ranges of the proportions of false vaccine exposures \( (G) \) and false events \( (F) \), for the best case scenario of nondifferential event sensitivity \( (a_2 = a_1) \) and nondifferential false outcome event rates \( (f_u = f_d) \). Graphs are shown for \( b_2 = 0.90 \). It is evident from figure 5 that when \( G \) and \( F \) are as large as 0.15, the true rate ratio is only 8 percent larger than the observed rate ratio. Increasing \( b_2 \) to 1.0 shifts the curves in figure 5 down very slightly.

The effect of differential event sensitivity is easily assessed when false event events are nondifferential by multiplying the plotted values of \( R \) in figure 5 by the ratio of unexposed to exposed event sensitivities \( (a_1/a_2) \). For example, when exposed events are 25 percent more likely to receive medical attention and to be correctly captured by the database than unexposed events, the ratio \( a_1/a_2 = 0.8 \) reduces the largest \( R \) in

\[ R = 1.20 \]

\[ 0.9940 \quad 0.9959 \quad 0.9979 \quad 1.0 \]

\[ 0.05 \quad 0.10 \quad 0.15 \]

**FIGURE 5.** Sensitivity analysis of DTP-asthma/bronchitis association; nondifferential event errors \( (a_2 = a_1, f_d = f_u) \). \( R \) = true rate ratio; \( a_1 \) = sensitivity of events; \( f = \) false event rate; \( F = \) proportion of false events; \( G = \) proportion of false exposures; \( b_1 = \) specificity of exposure; \( b_2 = \) sensitivity of exposure = 0.9.
Misclassification Model for Person-Time Analysis of Databases

The misclassification model developed in this paper is useful for assessing information bias in person-time analyses of automated medical care databases. It applies to associations between point exposures such as vaccinations and acute adverse events that are closely related in time. Conditional binomial tests are used to compare observed numbers of adverse events in narrow risk windows following exposures with expectations under null hypotheses. Methods for correcting tests of Poisson rate ratios and for performing critical sensitivity analyses of observed associations have been developed and illustrated. Application of the misclassification model to homogeneous population strata with a common rate ratio has also been discussed.

The quality of medical care databases used in screening analyses may vary among strata (j) of the study cohort, defined by secular time, age groups, or HMO sites in multi-site studies. The model can be used to correct for measurement errors in each strata. The common rate ratio can then be estimated by the Mantel-Haenszel estimator,

\[ R_{mh} = \frac{\sum N_j (1 - H_j)}{\sum N_j H_j}, \]

which is unbiased even for sparse person-time data (16).

Misclassification bias was found to be small when event sensitivities are nondifferential and less than 10 percent of database exposures and events are false. Differential event sensitivities are the most serious threat to the validity of analyses of medical care data whether they are abstracted from medical charts or captured by automated databases. Observed exposure/event associations are biased whenever exposed individuals are more or less likely to seek medical care for events than are unexposed individuals. This type of information bias is not unique to the cohort person-time design discussed in this paper; it can affect any study design that ascertains outcomes from medical care or billing data. This potential bias can be avoided by restricting screening analyses to serious medical conditions that require hospitalization or other medical care regardless of exposure status. Population surveys with valid instruments and adequate power would be necessary to test for differential event sensitivities (a1, a2) when medical care is not an absolute requirement. Positive exposure/event cohort screening associations that persist after correcting for possible differential event ascertainment would be candidates for more definitive controlled studies.

Because exposures such as drugs, vaccines, and other biologies administered by medical care providers must be documented in patients' medical records, the specificity (b1) and sensitivity (b2) parameters for automated exposure data can be estimated by comparing medical chart abstractions with database records. Estimating false event rates (fE, fU) requires medical record reviews of database events to validate diagnoses against objective diagnostic criteria.

The simple misclassification model for person-time analyses of Poisson rates can be extended to Poisson regression with misclassification of events and exposures by allocating person-time and event experience to strata based on exposure history and covariates measured without error (e.g., age and calendar time). Whittmore and Gong (18) have described a method for incorporating supplemental quality control data into disease rate Poisson regression models.

In addition to assessing the robustness of screening results over plausible ranges of misclassification errors, critical sensitivity analyses are useful in designing exposure and event validation studies and in focusing fieldwork necessary for definitive controlled studies. Greater dissemination of data quality assessments, critical sensitivity analyses, and methods used to supplement automated databases with standardized medical record abstractions and survey data (such as reported in Hemmelgarn et al. (19)) are needed to further our understanding of the appropriate role of automated medical care databases in epidemiologic research.

ACKNOWLEDGMENTS

This work was partially supported by CDC contract no. 200–90–0867.

The Vaccine Safety Datalink Project provided the intellectual motivation for this work. The author thanks Dr. Philip H. Rhodes of CDC's National Immunization Program for helpful comments on an earlier version of the manuscript.

Am J Epidemiol Vol. 144, No. 8, 1996
REFERENCES


