Estimating Absolute and Relative Case Fatality Ratios from Infectious Disease Surveillance Data

Nicholas G. Reich,1,2,∗ Justin Lessler,3 Derek A. T. Cummings,3 and Ron Brookmeyer4

1Division of Biostatistics and Epidemiology, University of Massachusetts, Amherst, Massachusetts 01002, U.S.A.
2Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.
3Department of Epidemiology, Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.
4Department of Biostatistics, University of California, Los Angeles, California 90095, U.S.A.

∗email: nick@schoolph.umass.edu

SUMMARY. Knowing which populations are most at risk for severe outcomes from an emerging infectious disease is crucial in deciding the optimal allocation of resources during an outbreak response. The case fatality ratio (CFR) is the fraction of cases that die after contracting a disease. The relative CFR is the factor by which the case fatality in one group is greater or less than that in a second group. Incomplete reporting of the number of infected individuals, both recovered and dead, can lead to biased estimates of the CFR. We define conditions under which the CFR and the relative CFR are identifiable. Furthermore, we propose an estimator for the relative CFR that controls for time-varying reporting rates. We generalize our methods to account for elapsed time between infection and death. To demonstrate the new methodology, we use data from the 1918 influenza pandemic to estimate relative CFRs between counties in Maryland. A simulation study evaluates the performance of the methods in outbreak scenarios. An R software package makes the methods and data presented here freely available. Our work highlights the limitations and challenges associated with estimating absolute and relative CFRs in practice. However, in certain situations, the methods presented here can help identify vulnerable subpopulations early in an outbreak of an emerging pathogen such as pandemic influenza.

KEY WORDS: Case fatality ratio; EM algorithm; Generalized linear models; Infectious disease; Influenza; Surveillance.

1. Introduction and Background

The case fatality ratio (CFR), a measure of the virulence of an infectious disease, is the fraction of cases that die after contracting the disease. A relative CFR is defined as the CFR of one group divided by that of a reference group. However, incomplete reporting of the number of infected individuals, both recovered and dead, makes it difficult to accurately estimate the CFR.

In public health response, both the absolute and the relative CFRs have important roles to play. The absolute CFR has an obvious role—it provides a measure of the severity of the disease. For setting public health priorities, an order-of-magnitude estimate of the absolute CFR may be adequate, and sophisticated methods may not be needed. Once the severity of the disease has been established, the relative CFR takes on primary importance as it becomes the guiding principle for targeting interventions to those populations that are most at risk of severe outcomes. In the 2009 pandemic of influenza A (H1N1) we have a prime example. Early on, limited supplies of antivirals were the only option for prophylaxis and treatment of the disease, and public health agencies needed to decide how these should best be deployed. Later, as limited vaccine became available, officials had to prioritize subpopulations for vaccination. In many cases, targeting decisions came down to the relative severity of disease in different populations, i.e., the relative CFR. Hence, accurately estimating this quantity early in an epidemic in a changing surveillance environment is of great importance.

Because an accurate and precise estimate of the relative CFR depends on complete observation of the number of cases and deaths in an outbreak, reliable data can often be difficult to find or collect. In many settings, nonfatal cases will go undetected either because of mild symptoms or insufficient public health surveillance infrastructure. Fatalities may be misreported as well, perhaps due to poor surveillance or incorrect attribution to another cause. Furthermore, depending on the survival times of cases, many deaths will be unreported simply because they have not yet occurred.

Due to variation in the host and the disease, the CFR may vary across a population. For example, one group of individuals may be more or less likely to succumb to a disease than others because of previous exposure or an underlying condition.

Although estimation of the CFR has been the focus of several papers in recent years, none of them address directly the challenges posed by an incompletely observed epidemic. Using data collected during the 2002–2003 severe acute respiratory syndrome (SARS) outbreak, Ghani et al. (2005) developed methods to address the challenge of estimating the CFR in real time when the survival distribution of the
disease is not known and when not all infected cases have died. Also motivated by the SARS epidemic, Jewell et al. (2007) developed nonparametric methods to estimate the CFR using a competing-risks framework from survival analysis. However, because the SARS outbreak was assumed to be fully observed—an assumption verified by subsequent serological analysis (Leung et al., 2004)—neither of these papers addresses issues of incomplete case ascertainment or changes in case reporting rates.

In the 2009 H1N1 influenza pandemic, a few attempts were made to estimate the CFR. Garske et al. (2009) gave a concise summary of the issues surrounding accurate estimation of the CFR. They proposed an estimator that adjusts for the survival distribution of influenza but not for changes in reporting. Nishiura et al. (2009) developed methods to estimate the CFR in the middle of an epidemic. This method adjusts a naive estimator (which divides the total number of observed deaths by the number of observed cases) based on information about the (assumed known) survival distribution of the particular disease. These two papers proposed analogous estimators, similar to what we propose as the E-step of the expectation maximization (EM) algorithm for the lag-adjusted estimator (see Section 3.3 below). However, neither of these recent works investigated the effects of changes in reporting rates on their estimate of the CFR. Other papers have taken more direct approaches to estimate the CFR, using surrogate measures for the number of individuals who become infected and who die (Presanis et al., 2009) or simply by making ad hoc adjustments to the observed case counts to account for underreporting (de Silva et al., 2009; Donaldson et al., 2009; Wilson and Baker, 2009).

Our goal with this article is to define situations where we can identify the relative CFR and to develop methods to estimate it. We define a typology of reporting rates for fatal and nonfatal cases that can be used to classify surveillance systems. We use this typology to determine conditions under which the relative CFR is identifiable. When it is, we utilize log-linear regression to estimate the relative CFR—a method analogous to standard relative risk estimation (Frome and Checkoway, 1985).

Section 2 presents the notation and surveillance system typology. Section 3 introduces methods to estimate the relative CFR that can control for time-varying reporting and disease-delayed mortality. Section 4 presents results from a simulation study. Section 5 demonstrates the methods with an analysis of data from the 1918 influenza pandemic.

2. The Structure of Case Fatality Data
2.1 Observed Data and Notation

During an outbreak, health organizations periodically report the number of incident (or cumulative) cases and deaths. These counts may only represent a fraction of the actual cases and actual deaths. Suppose we have \( T \) time periods, and a covariate (e.g., age) with \( J \) levels. We define \( N_{ij} \) be the total number of reported cases with symptom onset at time \( t \) for covariate level \( j \). We define \( D_{ij} \) as the number of reported deaths with symptom onset at time \( t \) for covariate level \( j \). (For now, we assume that the onset time of all deaths is known and that survival time is always shorter than the reporting interval, i.e., deaths are reported in the same interval in which they fall sick. We relax this assumption in our discussion of varying survival times in Section 3.3). Also, \( S_{ij} \) is the reported number of recovered cases with onset at time \( t \) for covariate level \( j \). Therefore \( N_{ij} = D_{ij} + S_{ij} \). However, underlying the reported data are recovered cases and deaths that go unreported. Let \( N_{ij}^* \) be the total number of cases, both reported and unreported, at time \( t \) for covariate level \( j \).

The CFR, \( p_{ij} \), is the probability of death, conditional on being a case. We assume the following reporting rates for time \( t \) and group \( j \): \( \phi_{ij} \) is the probability a recovered case is reported, and \( \psi_{ij} \) is the probability a dead case is reported. Table 1 illustrates the probabilistic structure of the observed and unobserved data.

We wish to clarify our use of the term reported. If a recovered case is reported then by that we mean it is counted as a case and is included in the denominator but not the numerator of the CFR. On the other hand, if a dead case is reported, by that we mean it is counted as both a case and a death and accordingly is included in both the numerator and the denominator of the CFR. By allowing different reporting probabilities for dead cases and recovered cases, we are making the event of being reported depend on the vital status of the case. We suggest in Section 6 alternative set-ups that may also be useful to consider.

2.2 A Typology of Reporting Rate Scenarios

We assume that \( \phi_{ij} \) and \( \psi_{ij} \) vary independently across groups and times and that \( p_{ij} = p_{i} \), i.e., the group-specific CFR stays constant over time. We return to our central question: what conditions or assumptions allow us to estimate an absolute or relative CFR?

Observed relative CFRs and reporting rates for cases and deaths are mathematically intertwined. For example, say we calculate an observed relative CFR for group A over group B at two different times. An increase in this observed relative CFR over time could be due to any number of factors: an increase in death reporting in group A, an increase in case reporting in group B, a decrease in case reporting in group A, a decrease in death reporting in group B, an actual change in the relative CFR, or some combination of all these factors. Without some assumptions on how the case reporting rates

| Table 1 |
| --- |
| Cell probabilities for calculating the CFR, for covariate level \( j \) and time \( t \) |
| Recovered | Died | Total |
| --- | --- | --- |
| Reported | \( \pi_{j1} = (1 - p_{ij})\phi_{ij} \) | \( \pi_{j2} = p_{ij}\psi_{ij} \) | \( (1 - p_{ij})\phi_{ij} + p_{ij}\psi_{ij} \) |
| Not reported | \( \pi_{j3} = (1 - p_{ij})(1 - \phi_{ij}) \) | \( \pi_{j4} = p_{ij}(1 - \psi_{ij}) \) | \( (1 - p_{ij})(1 - \phi_{ij}) + p_{ij}(1 - \psi_{ij}) \) |
| Total | \( (1 - p_{ij}) \) | \( p_{ij} \) | 1 |
vary over group and time, the relative CFR itself is unidentifiable.

We developed a typology of reporting rate scenarios that classifies surveillance systems by how reporting rates for recovered and dead cases vary across time and covariate group.

In a general model, we make no assumptions about how the reporting rates vary by $t$ and $j$. In this scenario, both the absolute and relative CFRs are unidentifiable. In a covariate-independent model, the reporting rates do vary across time but, at a given $t$, are constant across covariate $j$. For example, reporting rates may improve over time as an outbreak develops and public health surveillance teams are mobilized. However, at any given time, the detection rates may be identical for, say, men and women. In a constant proportion model there is a constant factor which relates the reporting rates of cases to that of deaths. This type of model might be appropriate in nonoutbreak or endemic disease contexts. Finally, in a complete fatality reporting model, we assume that the reporting rate for recovered cases does not vary by covariate group and that the reporting rate for fatal cases is nearly 100%. This may be a feasible model to assume in an enclosed population or when there is good reason to believe that a surveillance system is identifying all deaths.

Table 2 displays this classification that can be applied to disease surveillance systems and Figure 1 depicts the relationships between the different models. In the three models with assumptions about reporting rates, the relative CFR is identifiable. The absolute CFR for group $j$ will only be identifiable if the reporting rates for deaths and recovered cases are identical for all $t$.

3. Estimation of the Absolute and Relative Case Fatality Ratios

3.1. The Naïve Estimator

The naïve estimator for the absolute CFR is based on the observed case counts. We define the naïve estimator for group $j$ as $e_j := \frac{S_j}{D_j}$.

In Web Appendix A, we show that in covariate-independent models $e_j$ is an inconsistent estimator for $p_j$ unless $\phi_j = \psi_j$, $\forall t$. To do this, we assume a structure based on Table 1:

\[
(S_{ij}, D_{ij}, N_{ij}^* - D_{ij} - S_{ij})
\]

\[
\sim \text{Multinomial}(N_{ij}^*, \{\pi_{ij1}, \pi_{ij2}, \pi_{ij3} + \pi_{ij4}\}).
\]

Although this model gives insight into the behavior of the naïve estimator, it does not enable us to estimate $p_j$. The

### Table 2

| Model type            | Reporting rate assumptions | Cases | Fatalities $p_{ij}$ identifiable |
|-----------------------|---------------------------|-------|-------------------------------|
| General               | $\phi_{ij} = \phi_t$ $\psi_{ij} = \psi_t$ | No    |                               |
| Covariate independent | $\phi_{ij} = \phi_t$ $\psi_{ij} = \psi_t$ | Yes   |                               |
| Constant proportion   | $k \cdot \phi_{ij} = \psi_{ij}$ | Yes   |                               |
| Complete fatality     | $\phi_{ij} = \phi_t$ $\psi_{ij} = 1$ | Yes   |                               |

Figure 1. Venn diagram showing which models are subsets of the others.

identifiability problems are twofold. First, the model specified by equation 1 has more parameters than observations. Second, the multinomial probabilities are nonlinear functions of the parameters of interest. Under certain conditions, a slightly modified version of this model can lead to estimates of the relative CFR. We explore these scenarios in the upcoming sections.

3.2. Adjusting for Time-Varying Reporting

In the context of an outbreak, where surveillance and public awareness of a disease change over time, a major weakness of the naïve estimator is that it does not adjust for time-varying rates of case reporting. In this section, we develop a model which, under certain conditions, provides accurate estimates of the relative CFR.

We begin by introducing a conditional binomial approach that is derived from the multinomial model in equation 1. By conditioning on $N_{ij}$, we can model $D_{ij}$ directly as realizations from a binomial distribution. At each $(t,j)$ pair, we say that $D_{ij}$ follows a Binomial distribution of size $N_{ij}$ and success probability $p_j \psi_{ij}/(1-p_j)\psi_{ij}$. However, the sum in the denominator of the probability of death will make this conditional binomial model hard to fit in practice because the likelihood contains nonlinear functions of the parameters.

A second approach provides a more practical alternative while closely approximating the conditional binomial model in many situations. Assume that the $p_j$ are small. Then $N_{ij}^*$ is well approximated by $N_{ij}$ (see derivation in Web Appendix B) and the multinomial framework from equation 1 gives us that $E[D_{ij} \mid N_{ij}] = N_{ij}^* \cdot p_j \psi_{ij} \approx \frac{N_{ij}}{\phi_{ij}} \cdot p_j \psi_{ij}$. Because death is assumed to be a rare event, we can use the Poisson approximation to formulate the model as

\[
D_{ij} \mid N_{ij} \sim \text{Poisson} \left( \frac{N_{ij}}{\phi_{ij}} p_j \psi_{ij} \right).
\]
Although the conditional binomial model represents the data-generating model that we assume to be true, the Poisson formulation in equation 2 will only be accurate when the \( p_j \) are small (see Section 4.3). The Poisson formulation can be useful in log-linear form:

\[
\log E(D_{ij} \mid N_{ij}) = \log N_{ij} - \log \phi_i + \log p_j + \log \psi_j. \tag{3}
\]

We can now merge this model with the typology of reporting rates outlined in Section 2.2.

**Constant proportion of reporting rates**

In this class of models, we assume that \( \psi_j = k \cdot \phi_i \). Under these assumptions about reporting rates, the model in equation 3 reduces to

\[
\log E(D_{ij} \mid N_{ij}) = \log N_{ij} + \log p_j + \log k.
\]

We can reparameterize this model as

\[
\log E(D_{ij} \mid N_{ij}) = \log N_{ij} + \beta_0 + \gamma_j \tag{4}
\]

where \( \gamma_j = \log \frac{p_j}{k} \) is the log relative CFR comparing group \( j \) and group 1. Also, \( \beta_0 = \log(p_0k) \).

**Covariate-independent reporting**

Assume \( \phi_i = \phi_0 \) for all \( j \), that is, we assume that the reporting probability depends only on time and not the covariate; also assume that \( \psi_j = \psi_0 \) for all \( j \). Then the model in equation 3 reduces to

\[
\log E(D_{ij} \mid N_{ij}) = \log N_{ij} + \log \psi_0 + \log p_j,
\]

which can be reparameterized as

\[
\log E(D_{ij} \mid N_{ij}) = \log N_{ij} + \beta_0 + \alpha_j + \gamma_j, \tag{5}
\]

where \( \alpha_j \) is defined for \( t = 2, \ldots, T \) and \( \gamma_j \) is defined for \( j = 2, \ldots, J \). Then, \( \gamma_j \) is again the log relative CFR, while \( \beta_0 = \log \left( \frac{p_0}{k} \right) \) and \( \alpha_j = \log \frac{\psi_j}{\psi_0} \).

**Complete fatality reporting**

In this model, we assume that \( \phi_i = \phi_0 \) and \( \psi_j = \psi_0 = 1 \). This model is a subset of the covariate-independent model, but it represents an interesting special case. The model in equation 3 reduces to

\[
\log E(D_{ij} \mid N_{ij}) = \log N_{ij} + \log \phi_0 + \log p_j,
\]

which can be reparameterized as

\[
\log E(D_{ij} \mid N_{ij}) = \log N_{ij} + \beta_0 + \alpha_j + \gamma_j. \tag{6}
\]

Again, \( \gamma_j \) is the log relative CFR, while \( \beta_0 = \log \frac{\phi_0}{\phi_1} \) and \( \alpha_j = \log \frac{\psi_j}{\psi_0} \).

Models 4, 5, and 6 can be fit with standard software by declaring \( \log N_{ij} \) as an offset. The relative CFR will be estimated by \( \exp(\hat{\gamma}_j) \). We will call this quantity the reporting-rate-adjusted estimator of the relative CFR.

### 3.3 Adjusting for Survival Time

If reporting rates change over time and deaths are observed only at the time of death and not at the time of infection, the methods outlined in the previous sections may yield biased estimates of the relative CFR. These methods can be extended to account for incomplete observation of deaths due to the variable survival times from the time of infection. Real-time application of this method is dependent on death occurring in response to an acute infection, as with a disease such as influenza, where most deaths would be expected to occur within a fixed number, \( L \), weeks of infection. With diseases such as HIV/AIDS, where the survival time can be years or decades, this model may be hard to apply in real-time settings.

During an outbreak, at time \( t + L \), we can estimate the relative CFR using cases with symptom onset through time \( t \). We add to our existing notation as follows, ignoring any other delays in reporting:

- \( \eta_i \) = given death, known probability of dying \( i \) time units after symptom onset.
- \( \hat{D}_{ij} \) = number of deaths whose symptom onset occurred at time \( t \) in group \( j \).

We propose a model for \( E(\hat{D}_{ij} \mid N_{ij}) \) that is similar to the earlier models for \( E(D_{ij}) \):

\[
\log E(\hat{D}_{ij} \mid N_{ij}) = \log N_{ij} + \beta_0 + \alpha_j + \gamma_j. \tag{7}
\]

However, the \( \hat{D}_{ij} \) are not explicitly observed during an outbreak. We have developed a method to reconstruct an estimate of \( \hat{D}_{ij} \) from the observed data, which allows us to then fit the model in equation 7. To reconstruct \( \hat{D}_{ij} \) we need to estimate how many of the dead cases observed at times \( t + 1, \ldots, t + L \) would be expected to have symptom onset date of time \( t \). Assuming a known survival distribution for the disease, our method employs the EM algorithm (Dempster, Laird, and Rubin, 1977) to generate estimates of the relative CFR.

We propose the following EM algorithm, assuming a covariate independent model:

(1) Fix \( \alpha(0) \), a vector of starting values for the \( \alpha_t \), for \( t = 2, \ldots, T \).
(2) Fix \( \text{iter} = 0 \) and choose a tolerance \( \delta \) to determine convergence.
(3) Fix \( \text{iter} = \text{iter} + 1 \).
(4) (E-step) Find the expected reported deaths with symptom onset at time \( t \), using equation 8, for each covariate group \( j \), conditioning on \( \eta \) (the vector of survival probabilities, assumed known) and \( \alpha(\text{iter} - 1) \).
(5) (M-step) Fit the model from equation 7 using the E-step results as the outcome.
(6) Store \( (\hat{\beta}_0(\text{iter})^T; \hat{\alpha}_0(\text{iter})^T; \hat{\eta}_0(\text{iter})^T) = \hat{\theta}(\text{iter})^T \) as the fitted coefficients from the model.
(7) Repeat steps 3.3.3-3.3 until the parameters from the covariate-independent model converge, i.e., until each component of \( (\hat{\theta}_0(\text{iter}) - \hat{\theta}_0(\text{iter} - 1))^T \) is less than the tolerance \( \delta \).
(8) Use the supplemented EM algorithm (Meng and Rubin, 1991) to calculate the standard errors for parameter estimates.

To compute the E-step, we use the following formula, derived in Web Appendix C:

\[
E(\hat{D}_{ij} \mid D_j, N_j, \eta, \alpha) = \sum_{i=1}^{L} D_{i+t} \cdot \sum_{A_t} \frac{N_{ij} \cdot \eta_i \cdot \exp(\alpha_i - \alpha_j)}{\sum_{A_t} \sum_{N_{ij}} N_{ij} \cdot \eta_i \cdot \exp(\alpha_i - \alpha_j)}, \tag{8}
\]

where \( A_t = \{(t', t); t' + t = t + i \} \) is a set of all \( (t, t) \) pairs that contribute to the observed death count on day \( t + i \) and \( D_j \) and \( N_j \) are the vectors of observed values of \( D_{ij} \) and \( N_{ij} \) for \( t = 1, \ldots, T \). The M-step may be run using standard GLM software. The full EM routine may be implemented using the EMforCFR() function in the coarseDataTools package for R (see Web Appendix F), freely available from CRAN.
Table 3
Results from 500 simulated outbreaks. For pairs of reporting rate step functions (defined by the first four columns), we calculate the average estimate of the absolute CFR per 1000 cases and relative CFR using the naïve estimator (\(\epsilon_i\)) and the reporting-rate-adjusted estimator (\(\epsilon^{RR}\)). The standard errors of the estimates are given in the parentheses. If the outbreak were fully observed, i.e., when both reporting rates are 100%, then the absolute CFR would be estimated as 2.00 and relative CFR would be estimated as 0.33. The first five rows of the table represent data coming from complete fatality reporting systems; the next three rows are from covariate independent models and the final two rows are from constant proportion models.

| \(\phi_a\) | \(\psi_a\) | \(\phi_b\) | \(\psi_b\) | Avg. observed counts | CFR | Relative CFR |
|-----------|-----------|-----------|-----------|----------------------|-----|-------------|
|           |           |           |           | Deaths | Cases | Naïve | Naïve | RR-adj |
| 90        | 10        | 100       | 100       | 2749   | 739,226 | 3.7 (0.07) | 1.02 (0.044) | 0.34 (0.02) |
| 10        | 90        | 100       | 100       | 2740   | 635,985 | 4.3 (0.08) | 0.09 (0.004) | 0.34 (0.02) |
| 50        | 50        | 100       | 100       | 2742   | 687,602 | 4.0 (0.08) | 0.33 (0.015) | 0.33 (0.02) |
| 70        | 1         | 100       | 100       | 2739   | 533,522 | 5.1 (0.09) | 1.37 (0.058) | 0.36 (0.02) |
| 1         | 1         | 100       | 100       | 2745   | 16,439  | 167.0 (3.14) | 0.39 (0.016) | 0.39 (0.02) |
| 90        | 10        | 30        | 100       | 1365   | 737,833 | 1.8 (0.05) | 2.30 (0.117) | 0.34 (0.02) |
| 10        | 90        | 30        | 100       | 1366   | 634,612 | 2.2 (0.06) | 0.19 (0.010) | 0.33 (0.03) |
| 70        | 1         | 30        | 100       | 1365   | 532,141 | 2.6 (0.07) | 3.10 (0.171) | 0.38 (0.03) |
| 10        | 25        | 40        | 100       | 1561   | 231,592 | 6.7 (0.17) | 0.34 (0.018) | 0.34 (0.03) |
| 5         | 30        | 10        | 60        | 660    | 224,203 | 2.9 (0.12) | 0.33 (0.027) | 0.34 (0.04) |

The results, given in Table 3, show that while the reporting-rate-adjusted estimator consistently performed well, with minimal bias across all models, the naïve estimator’s performance was more erratic. In some cases, notably the constant proportion models, the naïve estimator was close to unbiased. (These empirical results are supported by the theoretical asymptotic results in Web Appendix A and more detailed simulations in Web Appendix D.) However, in most scenarios, the naïve estimator missed the target by a wide margin. In a few cases, it reversed the direction and on average estimated the relative CFR to be over 1.

4.2 Comparison of all Estimators, Accounting for Survival Time

We ran a simulation to examine the performance of our estimators when the symptom onset for deaths is unknown but the survival distribution, dependent on a parameter \(\eta\), is assumed to be known. Details of the data-generation algorithm are provided in Web Appendix D.

Under each of 10 different discrete survival distributions (scenarios A through J in Table 4) 1000 datasets were generated. The naïve, reporting-rate-adjusted and lag-adjusted estimators were calculated for each dataset. The lag-adjusted estimator was computed three times assuming different discrete survival distributions. First, we computed the estimate using the survival distribution used to generate the data (denoted by truth in Table 4). Second, we computed the estimate assuming a symmetric survival distribution with a mean of three time units and maximum possible survival of 4 (denoted by short in Table 4). The time-unit-specific probabilities of death for times 2 through 4 were (0.3, 0.4, 0.3). Lastly, we computed the estimate assuming a symmetric survival distribution, with mean survival of 8 days and maximum survival of 11 days (denoted by long in Table 4). The time-unit-specific probabilities of death for times 5 through 11 were (0.1, 0.15, 0.15, 0.2, 0.15, 0.15, 0.15, 0.1).

The 10 discrete survival distributions were chosen to represent a range of maximum survival times, different degrees
of skewness, and varying levels of heterogeneity. A discretized version of a Weibull distribution (shape = 1.5, scale = 8) was truncated at 15 units and used for scenarios H and I. The survival probabilities for the 15 units are (0.05, 0.08, 0.09, 0.1, 0.1, 0.09, 0.08, 0.07, 0.06, 0.05, 0.04, 0.04, 0.03, 0.02). This distribution has a mean of 6.75. Scenario H used these probabilities as the probabilities of death at time 1 through 15. Scenario I used the reversed vector of probabilities. Other distributions are described in Table 4.

Table 4 shows the simulation results and compares the estimators’ MSE and bias. Comparing MSEs of the five estimators across the different scenarios, we see that the adjustment for survival time provides a large gain in efficiency when the assumed survival distribution has roughly the same mean as the truth. This holds for distributions with short survival times (scenarios A–E) as well as those with long possible survival times (scenarios F–J). In the 10 scenarios we studied, the MSEs for the lag-adjusted estimates that assumed a known survival distribution with the same mean as the truth were reached the rough threshold of $\frac{4}{1000}$.

4.3 Sensitivity of Estimation to Large Case Fatality Ratio
As discussed earlier, approximating $N^*$ by $\frac{N}{p_i^2}$ provides an important link between a model that is unidentified and one where the relative CFR is estimable. This approximation, as discussed in Web Appendix B, relies on the $p_i$ being small. We examined the degree to which this approximation can impact the reliability of results.

We used the same data-generating process as in the previous simulations to generate simulated datasets with case and death counts for two subgroups of the population. The relative CFR between the groups was held fixed at $\frac{1}{3}$, and the larger of the two CFRs was allowed to vary from $\frac{1}{1000}$ to 1. Deaths were assumed to be completely reported and case reporting followed the step-function pattern described earlier. For a given pair of CFRs and case reporting step function, 500 datasets were simulated and the naive and reporting-rate adjusted estimators calculated.

Figure 2 shows the sensitivity of the estimates of the relative CFR to the true CFR for a particular case reporting step function. We found that until the larger of the two CFRs reached the rough threshold of $\frac{1}{20}$, the reporting-rate-adjusted estimator remained within 10% of the true value of the relative CFR. The naive estimator showed large bias for all values of the group 1 CFR. Further simulations (not shown) uphold these conclusions for other case reporting rate patterns and confirm that the bias depends largely on the magnitude of the maximum CFR.
5. Data Analysis: 1918 Influenza Case Fatality in Maryland

We analyzed data from the 1918 influenza pandemic from counties in the state of Maryland, USA. The *Annual Report of the State Board of Health of Maryland for the year ending December 31st, 1918* provides counts of influenza cases and deaths for the final 4 months of 1918 (Maryland State Board of Health, 1922). Influenza was made a reportable disease in September of 1918 and there are virtually no records of cases or deaths prior to this time.

Case and death counts were cross-tabulated for the months of September through December, 1918 and for a subset of Maryland counties. These data are presented in Web Table 1. Inclusion criteria for counties were established to create a subset of counties for which the assumption of covariate independence might be reasonably assumed to hold. Covariate independence implies that at every month \( t \), the rates of case reporting are the same across all counties and the rates of reporting of deaths are the same across all the counties. The following inclusion criteria were chosen to control for county-level factors that we believed could impact reporting rates. The county must (1) have a hospital, (2) have population density less than 100 individuals per square mile, and (3) not contain military bases or installations. Eight counties met these criteria: Carroll, Cecil, Dorchester, Frederick, Somerset, Talbot, Washington, and Wicomico counties. Data on these county-level characteristics were obtained from the state Board of Health annual report, U.S. Census data, and an article on hospitals in the United States (Anonymous., 1921; Maryland State Board of Health, 1922; Department of Commerce, USA, 1924).

Because the time period of reporting (1 month) is generally thought to be greater than most survival times for influenza, we did not additionally adjust for prolonged survival after symptom onset. Reporting-rate-adjusted relative CFRs were estimated by fitting the covariate-independent model from equation 5.

It has been shown that socioeconomic status of geographical regions was associated with mortality from influenza in 1918 (Murray et al., 2007). We sorted the eight counties by percentage of the population that was not native-born white (range: 4.8–36.6%), a proxy for socioeconomic status. We chose Somerset, the county with the lowest percentage of native-born white population, as the reference county for the data analysis.

The estimated relative CFRs and accompanying 95% confidence intervals are shown in Figure 3A. In Figure 3B, the percent nonwhite population for each county is plotted against the adjusted estimate of the relative CFR with a linear regression line drawn through the data. We find that the counties with a higher proportion of white population have on average lower CFRs than counties with higher proportions of minorities. Three of eight counties had percent native-born white population above 90% (Carroll, Frederick, and Washington counties) and these were the only three counties where we observed a significant difference in the reporting-rate-adjusted CFR when compared with Somerset county, with the lowest percent white population.

We postulate several possible explanations for this observed pattern. First, the existence of county-level variation in the CFR may be due in part to socioeconomic status. Second, differential case or death reporting in the counties may violate the assumption of covariate independence and lead to biased estimates of the relative CFR. For example, if the case reporting rate was higher in counties with higher white population, this may increase the denominator for estimating the CFR when compared with the other counties, leading to a reduction in the estimate of the relative CFR. Third, the variation in the estimated relative CFR appears to depend on geographical location as well. Carroll, Frederick, and Washington counties are all counties in central Maryland, near to Baltimore County and Baltimore City. Dorchester, Somerset, Talbot, and Wicomico counties are all located on the eastern shore of Maryland, a peninsula bordered by the Chesapeake Bay and the Atlantic Ocean.

6. Discussion

The CFR can play a large role in establishing the public health threat of a given disease. Accurate estimates of the relative CFR can help determine the optimal allocation of resources for surveillance, prevention, and treatment of disease. However, outbreak settings often generate data that are incomplete, where both recovered and fatal cases go unreported. In these situations, it is important to understand the assumptions necessary to identify an absolute or relative CFR, and to what extent those assumptions are realistic. We have shown that the absolute CFR is only identifiable when reporting rates for cases and deaths are equal to each other at every observation timepoint—a very unlikely scenario in practice.
Figure 3. This graphic summarizes the results of the data analysis presented in section 5. Panel A shows the estimated relative CFRs for the seven counties with respect to Somerset, the reference county. The vertical tick marks indicate the point estimates for each county and the horizontal lines indicate 95% confidence intervals for each county. The vertical tick marks have been scaled to represent the total number of cases observed in each county. Panel B plots the estimated relative CFRs against the percent of the county’s population whose race is not native-born white. The linear regression line is drawn through the points to illustrate an observed association between these two variables. The Pearson correlation coefficient for these two variables is 0.76.

Furthermore, we have shown that only with fairly stringent assumptions about the way that reporting varies over time can a relative CFR be identified. However, when it is identifiable, our reporting-rate-adjusted and lag-adjusted estimators can provide unbiased or nearly unbiased estimates of the relative CFR while the naive estimator is virtually always biased, often severely so.

Our work identifies and defines several important structural aspects of case fatality data. First, a $2 \times 2$ table (see Table 1) elucidates the structure of the observed and unobserved data. Second, our typology of reporting rates provides a simple classification scheme for disease surveillance systems. The new methods proposed in this article require case and death counts that are crosstabulated by units of time (i.e., weeks or months) and a categorical covariate such as gender. By providing such crosstabulations, surveillance reporting systems could make more data available for estimating the virulence of an infectious disease.

Although our model for a surveillance system reporting framework is general and applicable to a wide range of settings, other setups may be worthy of consideration. For example, a framework could define reporting rates without conditioning on the outcome status.

There are several limitations to this work. Some of our methods rely on the assumption that the reporting rates are independent of the covariate of interest. This is likely not true for a covariate such as age, as surveillance may more easily target school-age children than adults. But it may hold for other covariates such as gender or geographical location. Another limitation is that delays in reporting not due to survival times are not accounted for in this model. If these delays are uniform across all cases, then this may not introduce bias. However if the delays are different for different subgroups, this may impact the performance of the proposed estimators. Also, when using the lag-adjusted estimator, we assume a survival distribution for the disease in question. Our simulation results (see Section 4.2 and Web Appendix D) suggest that knowing the exact distribution is not vital as long as it the center of mass is roughly centered on the truth (Table 4), although it may increase the estimated variance of the estimates (Web Table 4). However, knowing the center of mass of a survival distribution may be difficult with an emerging pathogen. An in-depth analysis could provide more detailed information about the performance of these methods.

We also rely on the assumption that the true CFR is small in each of the population subgroups. This assumption enables us to make a key algebraic simplification. The impact of the assumption may be small for diseases such as influenza, whose CFR is thought to be on the order of 1 in 1000 (see Section 4.3 and Figure 2). However, an avenue for further research could be to find ways to adapt this method to work for diseases with larger CFRs. We also assume that the CFR does not change over time. This may not be the case with an emerging pathogen, as disease treatment and management may improve as clinical and epidemiological understanding of the disease evolves (Yip et al., 2005).

Finally, our method for calculating the lag-adjusted estimator is not computationally simple to implement. However
the code, an example dataset, and a vignette are available in the `coarseDataTools` package (see Web Appendix F).

Developing the framework to include additional covariates would be a useful extension to this work. For example, methods to estimate and compare the relative CFRs between men and women across several different countries could be useful. This could be achieved by inclusion of more covariates directly in the GLM framework or by fitting a multilevel model. Either way, the ability to test whether there is evidence that the CFR is different between two locations (while controlling for a second covariate) could be a helpful addition to this model. Further, allowing for a survival distribution that varies across levels of a covariate would be a useful addition to this framework. Finally, extending these methods to incorporate available evidence on reporting rates would be valuable.

Because disease outbreaks are often only partially observed, estimating absolute and relative CFRs remains a challenging problem for epidemiologists and public health officials. The methods developed in this article contribute a new set of tools for obtaining accurate estimates of the relative CFR in some scenarios. Such estimates could inform a timely and targeted response to an infectious disease outbreak.

7. Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 3, 4, 5 and 6 are available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org/.

Acknowledgements

NGR and RB were supported by the National Center for the Study of Preparedness and Catastrophic Event Response (PACER), which is funded by the U.S. Department of Homeland Security (N00014-06-1-0091). NGR and DATC were supported by the National Institute of General Medical Sciences (Award R01GM090204). JL and DATC were supported by grants from the NIG Fogarty Institute (1 R01 TW 0008246-01) and the Bill and Melinda Gates Foundation (705580-3). DATC holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund.

References

Anonymous (1921). Hospital service in the United States. JAMA 76, 1083–1103.

de Silva, U. C., Warachit, J., Waicharoen, S., and Chittaganpitch, M. (2009). A preliminary analysis of the epidemiology of influenza a(h1n1)v virus infection in thailand from early outbreak data, June-July 2009. Euro Surveill 14.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39, 1–38.

Department of Commerce, USA (1924). Fourteenth Census of the United States: State Compendium Maryland. Washington, D.C.: Government Printing Office.

Donaldson, L. J., Rutter, P. D., Ellis, B. M., Greaves, F. E. C., Mytton, O. T., Pebody, R. G., and Yardley, I. E. (2009). Mortality from pandemic a/h1n1 2009 influenza in England: Public Health Surveillance Study. BMJ 339, b5213, doi:10.1136/bmj.b5213.

Frome, E. and Checkoway, H. (1985). Use of Poisson regression models in estimating incidence rates and ratios. American Journal of Epidemiology 121, 309–323. ISSN 0002-9262.

Garske, T., Legrand, J., Donnelly, C., Ward, H., Cauchemez, S., Fraser, C., Ferguson, N., and Ghani, A. (2009). Assessing the severity of the novel influenza A/H1N1 pandemic. BMJ 339, 220–224.

Ghani, A., Donnelly, C., Cox, D., Griffin, J., Fraser, C., Lam, T., Ho, L., Chan, W., Anderson, R., Hedley, A., and Leung, G. M. (2005). Methods for estimating the case fatality ratio for a novel, emerging infectious disease. American Journal of Epidemiology 162, 479–486.

Jewell, N., Lei, X., Ghani, A., Donnelly, C., Leung, G., Ho, L., Cowling, B., and Hedley, A. (2007). Non-parametric estimation of the case fatality ratio with competing risks data: An application to severe acute respiratory syndrome (SARS). Statistics in Medicine 26, 1982–1998.

Leung, G., Chang, P.-H., Tsang, T., Lim, W., Chan, S., Chau, P., Donnelly, C., Ghani, A., Fraser, C., Riley, S., Ferguson, N., Anderson, R., Law, Y., Mok, T., Ng, T., Fu, A., Leung, P., Peiris, J., Lam, T., and Hedley, A. (2004). SARS-CoV antibody prevalence in all Hong Kong patient contacts. Emerging Infectious Diseases 10, 1653–1656.

Maryland State Board of Health. (1922). Annual Report of the State Board of Health of Maryland for the Year Ending December 31st, 1918. Baltimore, MD: King Brothers.

Meng, X. and Rubin, D. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. JASA 86, 899–909.

Murray, C., Lopez, A., Chin, B., Feehan, D., and Hill, K. (2007). Estimation of potential global pandemic influenza mortality on the basis of vital registry data from the 1918-20 pandemic: A quantitative analysis. The Lancet 368, 2211–2218.

Nishiura, H., Klinkenberg, D., Roberts, M., and Heesterbeek, J. (2009). Early epidemiological assessment of the virulence of emerging infectious diseases: A case study of an influenza pandemic. PLoS One 4, e6852, doi:10.1371/journal.pone.0006852.

Presanis, A., De Angelis, D., New York City Swine Flu Investigation Team, Hagy, A., Reed, C., Riley, S., Cooper, B., Finelli, L., Biedrzycki, P., and Lipsitch, M. (2009). The severity of pandemic H1N1 influenza in the United States, from April to July 2009: A Bayesian analysis. PLoS Med 6, e1000207.. 10.1371/journal.pmed.1000207.

Reed, C., Angulo, F.J., Swerdlow, D.L., Lipsitch, M., Meltzer, M.I., Jernigan, D., and Finelli, L. (2009). Estimates of the prevalence of pandemic (H1N1) 2009, United States, April–July 2009. Euro Surveill 15, 2004–2007.

Reich, N. (2010). coarseDataTools: A Collection of Functions to Help with Analysis of Coarse Infectious Disease Data. R package version 0.5.1. http://cran.r-project.org/web/packages/coarseDataTools/index.html

Wilson, N. and Baker, M. G. (2009). The emerging influenza pandemic: Estimating the case fatality ratio. Euro Surveillance 14.

Yip, P., Lam, K., Lau, E., Chau, P., Tsang, K., and Chao, A. (2005). A comparison study of realtime fatality rates: Severe acute respiratory syndrome in Hong Kong, Singapore, Taiwan, Toronto and Beijing, China. Journal of the Royal Statistical Society, Series A168, 233–243. ISSN 1467-985X.

Received November 2010. Revised September 2011.

Accepted September 2011.