American Journal of Epidemiology

Vol. 162, No. 4
Printed in U.S.A.

LETTERS TO THE EDITOR

RE: "HOW MANY FOODBORNE OUTBREAKS OF SALMONELLA INFECTION OCCURRED IN FRANCE IN 1995? APPLICATION OF THE CAPTURE-RECAPTURE METHOD TO THREE SURVEILLANCE SYSTEMS"

In a recent paper, Gallay et al. (1) used three-sources capture-recapture modeling to estimate the number of foodborne outbreaks of *Salmonella* infection that had occurred in France during the year 1995. The data provided in the article were used in a course on capture-recapture methods given in March 2004 for the Faculty of Public Health at Mahidol University in Bangkok, Thailand. The purpose of this letter is twofold: 1) to discuss some inconsistencies in the way the capture-recapture data were presented by Gallay et al. (1), leading to potentially very different analyses and conclusions, and 2) to argue for presenting capture-recapture data as completely as possible (for k sources, it should be a 2^k table—with one missing cell) to avoid the occurrence of misunderstandings such as the one outlined below.

The analysis in the article by Gallay et al. (1) was based on three French surveillance systems: the National Public Health Network (NPHN), the Ministry of Agriculture (MA), and the National *Salmonella* and *Shigella* Reference Center (NRC). A complete table describing the available information would be as provided in table 1.

Table 1 shows the multiple identifications of *Salmonella* outbreaks in the most complete form. n_1 are the outbreaks identified by all three sources, n_2 are the outbreaks identified by the NPHN and MA only, n_3 are the outbreaks identified by the NPHN and NRC only, etc. n_8 represents the outbreaks identified by none of the three sources and is the variable for missing information in the table. For the Mahidol University course, the table had to be constructed from the information provided by Gallay et al. in the text (1). According to Gallay et al.'s table 2 (1, p. 173), we have $n_1 + n_2 = 30$, $n_1 + n_3 = 59$, and $n_1 + n_5 = 39$. This is the number of outbreaks that could be matched by two sources. Consequently, $3n_1 + n_2 + n_3 + n_5 = 128$. On page 173 of Gallay et al.'s paper (1), as well as in the abstract, it is reported that 108 was the number of matches of any kind. This leads to $n_1 + n_2 + n_3 + n_5 = 108$, since n_1, n_2, n_3, and n_5 are the frequencies of all kinds of matches. Subtracting this equation from the previous one leads to $3n_1 + n_2 + n_3 + n_5 - (n_1 + n_2 + n_3 + n_5) = 128 - 108 = 20$, or $2n_1 = 20$, or $n_1 = 10$. It is now easy to construct $n_2 = 20$, $n_3 = 49$, and $n_5 = 29$. Finally, $n_4 = 35$ is found from the marginal NPHN count provided in Gallay et al.'s (1) table 2—namely, $n_1 + n_2 + n_3 + n_4 = 114$ (and similarly for n_6 ($n_1 + n_2 + n_5 + n_6 = 73$) and n_7 ($n_1 + n_3 + n_5 + n_7 = 529$)); the resulting frequencies are given as the first entries in the brackets in column 4 of table 1. One group of students in the Mahidol University course followed this route and derived the results given in table 2.

The models in table 2 are selected as follows: The first one corresponds to the best choice according to the Akaike Information Criterion; the second is the best choice

TABLE 1. Identification of *Salmonella* outbreaks in France in 1995, according to the three sources of data used by Gallay et al. (1)

Data source			Frequency
NPHN*	MA*	NRC*	
1†	1	1	n_1 [10, 20]‡
1	1	0	n_2 [20, 10]
1	0	1	n_3 [49, 39]
1	0	0	n_4 [35, 45]
0	1	1	n_5 [29, 19]
0	1	0	n_6 [14, 24]
0	0	1	n_7 [441, 451]
0	0	0	n_8 = missing

* NPHN, National Public Health Network; MA, Ministry of Agriculture; NRC, National *Salmonella* and *Shigella* Reference Center.

† A 1 indicates that the outbreak has been identified by the respective source, whereas a 0 indicates that the outbreak has not been identified by the respective source.

‡ The two numbers in brackets refer to the two data sets that could be constructed from the information provided in the paper by Gallay et al. (1).

TABLE 2. The three "best" log-linear models fitted to three sources of data on foodborne *Salmonella* outbreaks and their estimates of the total number of outbreaks occurring in France during 1995, using the first entries in the brackets of frequency data presented in column 4 of table 1

Model	Log-likelihood	AIC*,†	BIC*,‡	Estimate of n_8
Full	−18.8786	−51.7572	−82.5123	76
NPHN* × MA*	−23.2693	−56.5386	−78.5066	346
NPHN × MA, NPHN × NRC*	−21.6795	−55.3590	−81.7206	213

* AIC, Akaike Information Criterion; BIC, Bayesian Information Criterion; NPHN, National Public Health Network; MA, Ministry of Agriculture; NRC, National *Salmonella* and *Shigella* Reference Center.

† AIC = 2 log-likelihood − 2 (number of parameters in the model).

‡ BIC = 2 log-likelihood − log(n) (number of parameters in the model), where $n = n_1 + \cdots + n_7$.

according to the Bayesian Information Criterion; and the third is the second-best with respect to both criteria. Apparently, the associated estimates for the missing cell are substantially different from the ones given in the article by Gallay et al. (1) using identical models.

However, there is another way to construct the frequency information for column 4 in table 1. It is also reported on page 173 of Gallay et al.'s paper (1) that 20 was the number of matches obtained from all three sources; in other words, $n_1 = 20$, and since $n_1 + n_2 = 30$, $n_2 = 10$, and similarly, $n_3 = 39$ and $n_5 = 19$. Using the remaining information, the second entry in brackets in column 4 of table 1 can be constructed. Using this frequency column, results identical to those provided by Gallay et al. (1) could be achieved. This indicates that the second analysis is likely to correspond to the true data constellation and that the total number of matches of any kind given by Gallay et al. (1), 108, is incorrect and needs to be replaced by 88.

This analysis shows that simple errors in reported data can result in very different analyses and substantially different conclusions (here, underestimation of underreporting). This is particularly true for capture-recapture frequency data, since the log-linear modeling used is very sensitive to the observed frequencies. It also shows that it is preferable to provide the complete capture-recapture table (such as table 1) so that simple errors like the one reported above can be avoided by allowing for cross-checking.

ACKNOWLEDGMENTS

Conflict of interest: none declared.

REFERENCE

1. Gallay A, Vaillant V, Bouvet P, et al. How many foodborne outbreaks of *Salmonella* infection occurred in France in 1995? Application of the capture-recapture method to three surveillance systems. Am J Epidemiol 2000;152:171–7.

Prof. Dr. Dankmar A. Böhning
Institute for Social Medicine, Epidemiology and Health Economics, Charité Medical School, Fabeckstrasse 60–62, Berlin 14195, Germany

DOI: 10.1093/aje/kwi209