Ochsner Journal 19:220–226, 2019

DOI: 10.31486/toj.18.0166

Investigating the Source of a Disease Outbreak Based on Risk Estimation: A Simulation Study Comparing Risk Estimates Obtained From Logistic and Poisson Regression Applied to a Dichotomous Outcome

Chanapong Rojanaworarit, PhD,[1] Jason J. Wong, BA[2]

[1]Department of Health Professions, School of Health Professions and Human Services, Hofstra University, Hempstead, NY [2]Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Hofstra University, Hempstead, NY

Background: In epidemiologic investigations of disease outbreaks, multivariable regression techniques with adjustment for confounding can be applied to assess the association between exposure and outcome. Traditionally, logistic regression has been used in analyses of case-control studies to determine the odds ratio (OR) as the effect measure. For rare outcomes (incidence of 5% to 10%), an adjusted OR can be used to approximate the risk ratio (RR). However, concern has been raised about using logistic regression to estimate RR because how closely the calculated OR approximates the RR depends largely on the outcome rate. The literature shows that when the incidence of outcomes exceeds 10%, ORs greatly overestimate RRs. Consequently, in addition to logistic regression, other regression methods to accurately estimate adjusted RRs have been explored. One method of interest is Poisson regression with robust standard errors. This generalized linear model estimates RR directly vs logistic regression that determines OR. The purpose of this study was to empirically compare risk estimates obtained from logistic regression and Poisson regression with robust standard errors in terms of effect size and determination of the most likely source in the analysis of a series of simulated single-source disease outbreak scenarios.

Methods: We created a prototype dataset to simulate a foodborne outbreak following a public event with 14 food exposures and a 52.0% overall attack rate. Regression methods, including binary logistic regression and Poisson regression with robust standard errors, were applied to analyze the dataset. To further examine how these two models led to different conclusions of the potential outbreak source, a series of 5 additional scenarios with decreasing attack rates were simulated and analyzed using both regression models.

Results: For each of the explanatory variables—sex, age, and food types—in both univariable and multivariable models, the ORs obtained from logistic regression were estimated further from 1.0 than their corresponding RRs estimated by Poisson regression with robust standard errors. In the simulated scenarios, the Poisson regression models demonstrated greater consistency in the identification of one food type as the most likely outbreak source.

Conclusion: Poisson regression with robust standard errors proved to be a decisive and consistent method to estimate risk associated with a single source in an outbreak when the cohort data collection design was used.

Keywords: *Case-control studies*, *cohort studies*, *logistic models*, *odds ratio*, *Poisson distribution*, *risk ratio*

Address correspondence to Chanapong Rojanaworarit, PhD, Department of Health Professions, School of Health Professions and Human Services, Hofstra University, 101 Hofstra Dome, 220 Hofstra University, Hempstead, NY 11549-2200. Tel: (516) 463-6673. Email: chanapong.rojanaworarit@hofstra.edu

INTRODUCTION

The primary objective of an outbreak investigation is to identify the source to (1) control the epidemic and (2) prevent future occurrences. To control the epidemic, efficiency is typically prioritized in an outbreak investigation to identify the potential origins in a timely fashion. However, to identify the source for the purpose of preventing future occurrences, a more precise analytical approach is required. Risk estimation is one method used to identify the source.[1] By evaluating different exposures and comparing the risk of developing disease attributed to each exposure, the most likely causative agent can be identified.

After an outbreak has been detected, a typical way to proceed with the investigation is to identify the cases (those

with the disease outcome) and the non-cases (those who have not yet developed disease). Past exposures are then ascertained in the same way for both groups. In this practical approach, the case-control design is the applicable design for epidemiologic data collection.[1] Alternatively, the data collection design can be conceptualized as a retrospective cohort study. In this design, all individuals at risk of developing the disease could be conceptualized as an inception cohort—a group of individuals who gathered at an event where they were potentially exposed to putative risk factors and could be followed to identify whether they developed the disease at the time of the outbreak investigation.[2] After the data have been compiled through either of these approaches, the association between exposures and outcome can be determined.

In epidemiologic investigations, binomial or dichotomous outcome variables, such as the occurrence and nonoccurrence of disease, are common. For example, in an investigation of a *Staphylococcus aureus* food poisoning outbreak that occurred in Oswego County, NY, the two states for the dichotomous variable (disease) were ill and not-ill.[3] For the Oswego outbreak, the classic analytic approach of calculating attack rate was used. The food-specific attack rate was calculated by dividing the number of people who ate a specific food and became ill by the total number of people who ate that food. However, attack rate only provides the risk of getting the disease solely among those exposed to a specific factor. The major limitation of attack rate is that it does not allow hypothesis testing of the association between each food with the disease.

In the contemporary approach using cohort or case-control data collection designs, multivariable regression techniques with adjustment for confounding can be applied to assess the association between exposure and outcome.[4-7] Multivariable logistic regression techniques are the most commonly used, especially for binomial outcome variables.[8-10] Traditionally, logistic regression is used in analyses of case-control studies to determine the odds ratio (OR) as the effect measure.[4,11] Yet logistic regression has also been applied to the dichotomous outcomes in cohort studies and randomized controlled trials (RCTs).[12] In cohort studies and RCTs, logistic regression can serve as a valuable tool to estimate risk and assess the association between exposure and outcome in certain situations. For rare outcomes (incidence of 5% to 10%), an adjusted OR can be rationally used to approximate the risk ratio (RR).[10-12] Therefore, some epidemiologists have advocated for the use of ORs in cohort studies.[13,14] Nonetheless, concern has been raised about using logistic regression to estimate RR because how closely the calculated OR approximates the RR depends largely on the outcome rate.[15,16] The literature shows that when the incidence of outcomes exceeds 10%, ORs greatly overestimate RRs.[8-9] Consequently, in addition to logistic regression, other regression methods to accurately estimate adjusted RRs in cohort studies and RCTs have been explored.[8,17] Two methods of interest are Poisson regression with robust standard errors and log-binomial regression. These generalized linear models estimate RR directly vs logistic regression that determines OR.[8]

Even though the overestimation of RR by OR in a cohort study has been illustrated,[8] the extent of the difference between the OR and RR in an outbreak investigation with high-incidence dichotomous outcomes has not been examined. In this current study, estimates of gastroenteritis risk attributed to consumption of certain foods were obtained from analyzing simulated outbreak data using logistic and Poisson regression. The purpose of these analyses was to compare the ORs and RRs and to examine how these two models led to different conclusions regarding the most likely source of the outbreak. In addition, we examined these differences in a set of simulated scenarios with varying attack rates to assess whether there was a situation in which a certain statistical technique was more applicable.

METHODS

Dataset

We created a dataset to simulate a hypothetical foodborne disease outbreak following a public event with 75 people in attendance. This scenario included 14 types of food (foods 1 to 14). The dichotomous outcome of interest was gastroenteritis that occurred after ingestion of contaminated food. To model an outbreak (common outcome), the simulated data had an overall attack rate (incidence of disease) of 52.0%. Food 12 was designated as the most likely source of this single-source outbreak.

Theoretically, this situation could be conceptualized either as a retrospective case-control study or as a retrospective cohort study. If the data could be practically collected by a case-control approach, the cases would consist of people who developed gastroenteritis and sought medical treatment, and the controls would be those who did not develop the outcome. These two groups of people would be traced backwards to identify the types of food they ingested. In this case, the OR could be estimated to approximate the RR of the outcome attributed to food ingestion. In contrast, if the situation were considered a retrospective cohort study, the cohort would be people who were at risk of the outcome, and food ingestion at the event would be the exposure that preceded the gastroenteritis occurrence. In this case, RR could be directly estimated.

In our investigation, the aim was to identify the food that most likely contributed to the gastroenteritis outbreak to prevent future occurrences. We used an analytic approach that measured the independent effect of each explanatory variable, controlling for the confounding effect of other factors.

To further investigate the differences in risk estimates when attack rates were altered from the initial dataset, we generated 5 additional datasets with a maintained total of 75 participants. We gradually decreased the number of ill individuals by one for each subsequent scenario by altering their status from ill to not-ill. In an effort to reduce the attack rate for the most likely food source (food 12) and maintain the attack rate for the second most likely source (food 4), we selected ill individuals who initially ate food 12 without eating food 4. Thus, the effect of food 12 on the outcome of gastroenteritis was decreased.

Because these hypothetical datasets did not involve human subjects, no human subject research ethical considerations were applicable, and institutional review board approval was not required.

Statistical Analysis

Statistical analysis was performed using Stata software, v.15.0 (StataCorp, LLC). Descriptive statistics are used to

describe general characteristics of the hypothetical subjects. The food-specific attack rates were calculated and reported as percentages. Univariable logistic regression analysis was used to estimate crude ORs and 95% confidence intervals (CIs). Crude RRs and their CIs were estimated using univariable Poisson regression with robust standard errors. Adjusted ORs and RRs were obtained through multivariable logistic regression and multivariable Poisson regression with robust standard errors.[18] The Poisson regression is a generalized linear model with a log link function and a Poisson distribution.[19] The robust standard error is estimated using the sandwich estimation method to take the incorrect assumption of Poisson distributed outcome in the Poisson regression into consideration.[8] Using this approach, Poisson regression can be applied to estimate the risk in prospective studies with binary outcomes.[18] To examine whether different analytical methods led to different conclusions regarding the food type that most likely contributed to the disease outbreak, attack rate was compared to its corresponding OR and RR that were estimated in the multivariable models. OR, RR, and their corresponding CIs for each explanatory variable were also compared to evaluate the difference in risk estimation. In addition, the feasibility of applying log-binomial regression—a generalized linear model with a log link function and a binomial distribution that also allows direct estimation of RR—to analyze this data was explored.[8]

RESULTS

In this hypothetical data, a higher sex-specific attack rate was found among females (56.8%). The age-specific attack rate was highest in the elderly age group (56.2%). Four types of food had food-specific attack rates >60%. Food 14 had the highest food-specific attack rate (66.7%) (Table 1).

Crude ORs were estimated further from 1.0 than the crude RR estimates on both sides of the scale—above and below 1.0. All the CIs estimated by the univariable Poisson regression with robust standard errors were narrower than those estimated by the univariable logistic regression (Table 2).

Multivariable logistic regression revealed two food types with adjusted ORs ≥2 and statistically significant *P* values (foods 4 and 12). Multivariable Poisson regression with robust standard errors, in contrast, specifically identified a single food type with an adjusted RR ≥2 and statistically significant *P* value (food 12) (Table 3).

When the overall attack rate of 52.0% in the prototype scenario was reduced to 50.7% and 49.3%, the multivariable logistic regression model still identified two food types—foods 4 and 12—with meaningful ORs and statistical significance (Table 4). However, in these same two scenarios, the multivariable Poisson regression consistently identified a single food type (food 12) with an RR ≥2 and statistical significance. In scenarios 4 and 5, when the overall attack rate was further reduced to 48.0% and 46.7%, respectively, the logistic regression models provided meaningful ORs for both foods 4 and 12; however, only food 4 maintained statistical significance. The Poisson regression model in scenario 4 determined one food type (food 12). Yet in scenario 5, food 12 was no longer statistically significant. Scenario 6 (reduction of the attack rate to 45.3%) diverged from scenarios 1 to 5 because the Poisson regression model suggested multiple sources of the outbreak. Thus, scenario 6 was outside the scope of our investigation into single-source outbreaks.

Table 1. Characteristics of Hypothetical Subjects in a Foodborne Outbreak (n = 75)

Variable	Ill, n (%)[a]	Not Ill, n (%)[a]	Total, n (%)[b]
Overall	39 (52.0)	36 (48.0)	–
Sex			
Male	14 (45.2)	17 (54.8)	31 (41.3)
Female	25 (56.8)	19 (43.2)	44 (58.7)
Age, years (mean ± SD = 37.24 ± 20.9, min-max = 8-77)			
≤19	11 (45.8)	13 (54.2)	24 (32.0)
20-59	19 (54.3)	16 (45.7)	35 (46.7)
≥60	9 (56.2)	7 (43.8)	16 (21.3)
Food consumed			
Food 1	25 (54.3)	21 (45.7)	46 (61.3)
Food 2	22 (51.2)	21 (48.8)	43 (57.3)
Food 3	18 (48.7)	19 (51.3)	37 (49.3)
Food 4	18 (64.3)	10 (35.7)	28 (37.3)
Food 5	14 (60.9)	9 (39.1)	23 (30.7)
Food 6	18 (48.7)	19 (51.3)	37 (49.3)
Food 7	15 (55.6)	12 (44.4)	27 (36.0)
Food 8	2 (50.0)	2 (50.0)	4 (5.33)
Food 9	15 (48.4)	16 (51.6)	31 (41.3)
Food 10	12 (50.0)	12 (50.0)	24 (32.0)
Food 11	22 (55.0)	18 (45.0)	40 (53.3)
Food 12	34 (63.0)	20 (37.0)	54 (72.0)
Food 13	21 (44.7)	26 (55.3)	47 (62.7)
Food 14	4 (66.7)	2 (33.3)	6 (8.0)

[a]Row percentage.
[b]Column percentage.

DISCUSSION

In classical analyses, food-specific attack rates have been used as epidemiologic evidence to show the probability of foodborne infection following consumption of a certain food.[1] Nonetheless, analysis of food-specific attack rates in this scenario did not lead to a definitive conclusion regarding the source of the outbreak because four possible food types (foods 4, 5, 12, and 14) had remarkably high attack rates, and the confounding problem still existed (Table 1).

Univariable logistic regression and univariable Poisson regression with robust standard errors produced crude estimates of ORs and RRs (Table 2). The univariable logistic regression model revealed 2 food types (foods 4 and 12) with meaningful ORs (OR ≥2), and food 12 had statistically significant results ($P<0.05$). The univariable Poisson regression model revealed only 1 food type (food 12) with a meaningful RR (RR ≥2) and statistical significance. These crude associations between individual food types and the outcome of gastroenteritis seemed suggestive of causal relationships, but they were not conclusive. The univariable models did not account for the interplay among multiple exposures or individuals having eaten more than one type of food. If a

Table 2. Crude Odds Ratios (OR), Crude Risk Ratios (RR), and Corresponding Confidence Intervals (CI) of Association Between Foodborne Gastroenteritis and the Independent Variables

	Univariable Logistic Regression			Univariable Poisson Regression with Robust Standard Errors		
Variable	**OR**	**95% CI**	***P* Value**	**RR**	**95% CI**	***P* Value**
Female	1.60	0.63, 4.03	0.321	1.26	0.79, 2.01	0.337
Age, years						
20-59	1.40	0.49, 3.98	0.524	1.18	0.69, 2.02	0.535
≥60	1.52	0.43, 5.43	0.519	1.23	0.66, 2.28	0.515
Food consumed						
Food 1	1.28	0.50, 3.24	0.609	1.13	0.71, 1.79	0.616
Food 2	0.92	0.37, 2.31	0.866	0.96	0.62, 1.50	0.867
Food 3	0.81	0.32, 2.01	0.642	0.90	0.58, 1.41	0.645
Food 4	2.23	0.85, 5.84	0.103	1.44	0.94, 2.20	0.093
Food 5	1.68	0.62, 4.56	0.309	1.27	0.82, 1.96	0.288
Food 6	0.77	0.31, 1.90	0.567	0.88	0.57, 1.37	0.571
Food 7	1.25	0.48, 3.22	0.644	1.11	0.71, 1.73	0.641
Food 8	0.92	0.12, 6.89	0.934	0.96	0.35, 2.64	0.936
Food 9	0.78	0.31, 1.96	0.599	0.89	0.56, 1.40	0.606
Food 10	0.89	0.34, 2.35	0.812	0.94	0.58, 1.53	0.815
Food 11	1.29	0.52, 3.21	0.579	1.13	0.73, 1.77	0.583
Food 12	5.44	1.73, 17.11	0.004	2.64	1.19, 5.87	0.017
Food 13	0.48	0.18, 1.25	0.133	0.71	0.46, 1.09	0.120
Food 14	1.94	0.33, 11.31	0.460	1.31	0.71, 2.43	0.384

large proportion of ill individuals who ate one innocuous food type also ate the contaminated food type, the crude ORs and RRs would theoretically suggest causal relationships between both the innocuous and contaminated food exposures and the disease outcome. The innocuous food would seem to have had an effect on the outbreak. While we actually measured the effect of the contaminated food, we wrongly concluded that the innocuous food also contributed to the outbreak. This problem is known as confounding.

To account for the confounding problem, multivariable logistic regression was initially used to statistically adjust the confounding effect, estimating the risk of disease attributed to a certain food type independent of the effect from other factors. In this scenario, the multivariable logistic regression model still revealed four food types (foods 4, 11, 12, and 14) with meaningful ORs (OR ≥ 2), two of which had statistically significant results ($P<0.05$) (Table 3). Therefore, the analysis of food-specific attack rates (Table 1) and multivariable logistic regression analysis (Table 3) similarly pointed to four potential food types that likely contributed to the outbreak. Although three of the four food types (foods 4, 12, and 14) had attack rates ≥60% and ORs ≥2, conclusions regarding which of the four food types most likely contributed to the disease outbreak would differ according to analytic method. Based on the analysis of attack rates, food 5 (with an attack rate of 60.9%) would be considered in addition to foods 4, 12, and 14 (Table 1). However, food 5 failed to produce a meaningful OR in the multivariable logistic regression model. Conversely, food 11 which had an attack rate of 55% but an OR of 2 would be considered a likely source of the outbreak based on the multivariable logistic regression model.

The food-specific attack rates—an epidemiologic measure of disease frequency—indicated that food 14 was the most likely source of the outbreak (Table 1). In contrast, based on the adjusted ORs estimated by multivariable logistic regression—an epidemiologic measure of association—food 12 was identified as the most likely source of the outbreak (Table 3). Based on the use of the epidemiologic measure of association and the deconfounding principle, the adjusted OR provided more reliable epidemiologic evidence than the attack rate for identifying the likely source of outbreak in this situation.

In contrast to the multivariable logistic regression model that revealed two possible food types that potentially contributed to the outbreak, the multivariable Poisson regression with robust standard errors specifically identified food 12 as the single and most likely food type responsible for the outbreak (adjusted RR=3.09, 95% CI=1.23, 7.80). The other food types failed to obtain meaningful RRs and statistical significance.

For each of the explanatory variables in both the univariable and multivariable models, the OR was estimated further away from 1.0 than its corresponding RR. This finding from empirical analysis indicates the overestimation of RR by OR

Table 3. Adjusted Odds Ratios (OR), Adjusted Risk Ratios (RR), and Corresponding Confidence Intervals (CI) of Association Between Foodborne Gastroenteritis and the Independent Variables

	Multivariable Logistic Regression			Multivariable Poisson Regression With Robust Standard Errors		
Variable	**OR**	**95% CI**	***P* Value**	**RR**	**95% CI**	***P* Value**
Female	3.07	0.80, 11.71	0.100	1.55	0.88, 2.72	0.127
Age, years						
20-59	2.08	0.35, 12.49	0.422	1.14	0.54, 2.38	0.737
≥60	2.21	0.27, 18.04	0.458	1.37	0.54, 3.45	0.506
Food consumed						
Food 1	1.81	0.10, 31.15	0.684	1.10	0.57, 2.13	0.772
Food 2	0.18	0.02, 1.81	0.147	0.56	0.28, 1.15	0.113
Food 3	0.55	0.10, 3.04	0.497	0.77	0.44, 1.35	0.360
Food 4	5.70	1.01, 32.07	0.048	1.74	0.95, 3.22	0.075
Food 5	1.09	0.18, 6.81	0.923	0.92	0.47, 1.79	0.805
Food 6	1.83	0.20, 17.08	0.595	1.38	0.69, 2.76	0.366
Food 7	0.74	0.12, 4.48	0.742	1.07	0.60, 1.88	0.824
Food 8	0.30	0.01, 6.94	0.452	0.68	0.17, 2.77	0.586
Food 9	0.42	0.06, 3.05	0.389	0.80	0.39, 1.64	0.541
Food 10	1.75	0.31, 9.89	0.525	1.23	0.70, 2.17	0.463
Food 11	2.00	0.47, 8.52	0.349	1.23	0.63, 2.43	0.546
Food 12	7.14	1.54, 33.13	0.012	3.09	1.23, 7.80	0.017
Food 13	0.64	0.16, 2.54	0.529	0.96	0.57, 1.60	0.871
Food 14	3.54	0.30, 41.81	0.316	1.49	0.77, 2.89	0.239

Table 4. Adjusted Odds Ratios (OR), Adjusted Risk Ratios (RR), and Corresponding Confidence Intervals (CI) of Association Between Gastroenteritis and the Most Likely Food Types in Simulated Scenarios With Decreasing Attack Rates

			Multivariable Logistic Regression						Multivariable Poisson Regression					
			Food 12			Food 4			Food 12			Food 4		
	Overall Attack Rate, %	**Attack Rate for Food 12, %**	**OR**	**95% CI**	***P* Value**	**OR**	**95% CI**	***P* Value**	**RR**	**95% CI**	***P* Value**	**RR**	**95% CI**	***P* Value**
Prototype scenario														
1	52.0	63.0	7.14	1.54, 33.13	0.012	5.70	1.01, 32.07	0.048	3.09	1.23, 7.80	0.017	1.74	0.95, 3.22	0.075
Additional scenarios														
2	50.7	61.1	6.14	1.28, 29.50	0.023	7.05	1.12, 44.21	0.037	2.84	1.18, 6.87	0.020	1.74	0.94, 3.19	0.076
3	49.3	59.3	5.10	1.08, 24.13	0.040	8.91	1.37, 58.15	0.022	2.66	1.06, 6.67	0.036	1.86	0.96, 3.61	0.066
4	48.0	57.4	4.63	1.00, 21.50	0.050	8.17	1.32, 50.74	0.024	2.57	1.01, 6.52	0.048	1.89	0.97, 3.70	0.061
5	46.7	55.6	3.84	0.79, 18.66	0.095	9.86	1.49, 65.28	0.018	2.50	0.95, 6.59	0.063	1.99	0.95, 4.19	0.070
6	45.3	53.7	4.63	0.89, 24.16	0.069	22.56	2.50, 203.37	0.005	2.70	0.93, 7.79	0.066	2.90	1.29, 6.50	0.010

Notes: The attack rate for food 12, the most likely source of the outbreak, was intentionally reduced in scenarios 2 to 6 to decrease its effect on outcome. The attack rate for food 4, the second most likely source, was constant at 64.3%.
In all scenarios, foods 12 and food 4 were the two strong contenders for the most likely source, except for scenario 6. In scenario 6, food 14 was an additional strong contender for most likely food type (Logistic: OR=18.75, 95% CI=0.92, 381.18, *P*=0.056; Poisson: RR=2.52, 95% CI=1.32, 4.82, *P*=0.005).

in the analysis of a foodborne outbreak conceptualized as a retrospective cohort study with a common outcome, consistent with the theory proposed in several research methodology articles.[4,5,9,10]

Although a meaningful measure of effect could be obtained from using logistic regression in a cohort study,[8,11] seeing an effect in OR without an effect in RR and drawing a different conclusion would also be possible.[12] This phenomenon can be illustrated by comparing the statistically significant adjusted OR of 5.70 to the statistically nonsignificant adjusted RR of 1.74 for food 4 (Table 3). The adjusted OR leads to the conclusion that food 4 is a strong candidate for contributing to the outbreak. However, the markedly smaller adjusted RR without statistical significance does not support that conclusion. Thus, a regression technique that allows direct estimation of adjusted RR should be considered first when analyzing a cohort study with a relatively common outcome rather than a regression technique that estimates adjusted OR. In addition, the considerably narrower CIs obtained from the Poisson model also improve precision in the parameter estimation of effect. One limitation of the Poisson regression model is the inability to directly estimate probabilities. In this scenario, the estimated means from the Poisson regression model were used as surrogates for probabilities. As a result, obtaining individual predicted probabilities beyond the bounds of 0 and 1.0 was possible. These unrealistic predicted probabilities >1.0 could be problematic when the research objective is to obtain individual predicted probabilities of disease in predictive research—diagnostic and prognostic research.[4] However, in an etiologic study with the focus on estimating a valid RR, the probabilities >1.0 would not pose a problem.[4]

We simulated additional scenarios to assess the change in risk estimates when attack rates were reduced (Table 4). In scenarios 1 to 4, the Poisson regression models consistently indicated a single food type (food 12), leading to the decisive conclusion that food 12 was the sole source. In contrast, when the attack rates were altered, the risk estimates produced by the logistic regression models were highly variable. In scenarios 1 to 3, a decisive conclusion about the primary source for the outbreak could not be drawn from the logistic regression models because they yielded more than one potential food type, a finding that would prompt additional investigations into the alternative sources. Furthermore, when the logistic regression model indicated one food type in scenario 4, this finding contradicted the results of the Poisson regression model that directly estimated RR. In scenarios 4 and 5, the logistic models indicated food 4 as the most likely single source, while the Poisson regression models indicated food 12 as the primary source based on the meaningful RR; however, in scenario 5, none of the RRs for the food types remained statistically significant. As a result, the Poisson regression model for this scenario could no longer lead to a decisive conclusion about the single food type.

In multiple simulated scenarios, the Poisson regression model led to a more decisive conclusion about the single source of the outbreak. When the attack rates were altered (Table 4), the Poisson regression model consistently indicated food 12 as the most likely source. In terms of generalizability, Poisson regression with robust standard errors should be the statistical method of choice when incidence of disease can be obtained from cohort data collection design[8,11]; however, this design requires relatively complete data collection that would be burdensome in large-scale outbreaks. For such outbreaks, the case-control data collection design is commonly used. For this data collection approach, the investigation usually starts by encountering a cluster of ill individuals who seek medical care. Then investigators identify a control group of non-cases to estimate the risk. OR is then calculated by logistic regression to estimate risk. In the scenarios presented in this study, the ORs led to inconsistent conclusions regarding the primary food type responsible for the single-source outbreak. Thus, in situations where the outcome is common (>10%), the OR overstates the effect size and potentially leads to misleading conclusions as shown in Table 4 and in the literature.[8,9,16,20] Sheldrick and colleagues[16] have shown how OR is mathematically related to RR in the following equation where p_A and p_B represent the probabilities of events A and B, respectively:

$$OR \cdot \frac{1 - p_A}{1 - p_B} = \frac{p_A}{p_B} = RR$$

In situations where p_A and p_B are close to zero (probability of rare event), the OR closely estimates the RR. However, in the case of a common outcome when p_A and/or p_B are considerably greater than zero, $(1 - p_A)/(1 - p_B)$ no longer approximates 1.0, and OR noticeably overestimates RR.

In this study, we also assessed the feasibility of applying log-binomial regression by applying this regression technique to estimate the adjusted RRs in the multivariable model. However, the multivariable model did not concave to yield any estimates. The literature suggests that this convergence problem could be encountered when the incidence of outcome is high.[4,8]

CONCLUSION

This study illustrates that Poisson regression with robust standard errors is a decisive and consistent method to estimate risk associated with a single source in an outbreak when the cohort data collection design is used. However, in outbreak investigations that use the case-control data collection design, ORs obtained from logistic regression could overestimate the risk and potentially influence the conclusion regarding the source. Consequently, maintaining awareness of this overestimation when interpreting ORs is important.

ACKNOWLEDGMENTS

The authors have no financial or proprietary interest in the subject matter of this article.

REFERENCES

1. Dwyer DM, Strickler H, Goodman RA, Armenian HK. Use of case-control studies in outbreak investigations. *Epidemiol Rev*. 1994;16(1):109-123.
2. Rodrigues L, Kirkwood BR. Case-control designs in the study of common diseases: updates on the demise of the rare disease assumption and the choice of sampling scheme for controls. *Int J Epidemiol*. 1990 Mar;19(1):205-213.
3. Gross M. Oswego County revisited. *Public Health Rep*. 1976 Mar-Apr;91(2):168-170.

4. Lee J, Tan CS, Chia KS. A practical guide for multivariate analysis of dichotomous outcomes. *Ann Acad Med Singapore*. 2009 Aug;38(8):714-719.
5. Hidalgo B, Goodman M. Multivariate or multivariable regression? *Am J Public Health*. 2013 Jan;103(1): 39-40. doi: 10.2105/AJPH.2012.300897.
6. McNutt LA, Wu C, Xue X, Hafner JP. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am J Epidemiol*. 2003 May 15;157(10):940-943.
7. Greenland S. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Am J Epidemiol*. 2004 Aug 15;160(4):301-305.
8. Knol MJ, Le Cessie S, Algra A, Vandenbroucke J, Groenwold RH. Overestimation of risk ratios by odds ratios in trials and cohort studies: alternatives to logistic regression. *CMAJ*. 2012 May 15;184(8):895-899. doi: 10.1503/cmaj.101715.
9. Janani L, Mansournia MA, Nourijeylani K, Mahmoodi M, Mohammad K. Statistical issues in estimation of adjusted risk ratio in prospective studies. *Arch Iran Med*. 2015 Oct;18(10):713-719.
10. Cook TD. Advanced statistics: up with odds ratios! A case for odds ratios when outcomes are common. *Acad Emerg Med*. 2002 Dec;9(12):1430-1434.
11. Wilber ST, Fu R. Risk ratios and odds ratios for common events in cross-sectional and cohort studies. *Acad Emerg Med*. 2010 Jun;17(6):649-651. doi: 10.1111/j.1553-2712.2010.00773.x.
12. Diaz-Quijano FA. A simple method for estimating relative risk using logistic regression. *BMC Med Res Methodol*. 2012 Feb 15;12:14. doi: 10.1186/1471-2288-12-14.
13. Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA*. 1998 Nov 18;280(19):1690-1691.
14. Persoskie A, Ferrer RA. A most odd ratio: interpreting and describing odds ratios. *Am J Prev Med*. 2017 Feb;52(2):224-228. doi: 10.1016/j.amepre.2016.07.030.
15. Austin PC. Absolute risk reductions, relative risks, relative risk reductions, and numbers needed to treat can be obtained from a logistic regression model. *J Clin Epidemiol*. 2010 Jan;63(1):2-6. doi: 10.1016/j.jclinepi.2008.11.004.
16. Sheldrick RC, Chung PJ, Jacobson RM. Math matters: how misinterpretation of odds ratios and risk ratios may influence conclusions. *Acad Pediatr*. 2017 Jan-Feb;17(1):1-3. doi: 10.1016/j.acap.2016.10.008.
17. Wiest MM, Lee KJ, Carlin JB. Statistics for clinicians: an introduction to logistic regression. *J Paediatr Child Health*. 2015 Jul;51(7):670-673.
18. Zou G. A modified Poisson regression approach to prospective studies with binary data. *Am J Epidemiol*. 2004 Apr 1;159(7):702-706.
19. Ravani P, Parfrey P, Murphy S, Gadag V, Barrett B. Clinical research of kidney diseases IV: standard regression models. *Nephrol Dial Transplant*. 2008 Feb;23(2):475-482. doi: 10.1093/ndt/gfm880.
20. Davies HT, Crombie IK, Tavakoli M. When can odds ratios mislead? *BMJ*. 1998 Mar 28;316(7136):989-991.

This article meets the Accreditation Council for Graduate Medical Education and the American Board of Medical Specialties Maintenance of Certification competencies for Patient Care and Medical Knowledge.