DOI: 10.33470/2379-9536.1243

PERSPECTIVE Volume 5 Issue 4

Statistical Methods Used in Clinical Simulation and Medical Education Scholarship

Zuber D. Mulla, PhD, CPH[1], J. Hector Aranda, BS, CHSOS[1], Donovan Rojas, BS[1], Sanja Kupesic Plavsic, MD, PhD[1]

Author affiliations are listed at the end of this article.

Correspondence to:
Zuber Mulla, PhD, CPH
Texas Tech University
Health Sciences Center
zuber.mulla@ttuhsc.edu

ABSTRACT

The objective of this paper is to introduce selected statistical and epidemiologic topics that are of interest to interdisciplinary teams of healthcare quality professionals, educators, technical staff, and researchers who participate in clinical simulation scholarship. Four research vignettes in the setting of a hypothetical clinical simulation training workshop are presented. The first vignette illustrates the utility of exact logistic regression when analyzing a small dataset. The second underscores the importance of using an appropriate method to account for the repeated measurement of an outcome. The third illustrates the use of the intraclass correlation coefficient to measure inter-rater reliability. The final vignette demonstrates the benefits of creating a causal diagram known as a directed acyclic graph.

KEYWORDS
Education, Simulation, Quality improvement, Data analysis, Epidemiology

INTRODUCTION

Simulation offers opportunities to improve the quality of healthcare.[1] Simulation in health care education has a long history with roots dating back to the early 20th century.[2] In the past several decades, clinical simulation has been rapidly adopted in multiple areas including undergraduate, graduate, and continuing medical education.[3] Simulation also plays an integral role in effective faculty development initiatives.[4] As clinical simulation has increased in popularity, so has the interest in health care simulation research.[5] Reporting guidelines for health care simulation research were recently promulgated.[5]

The objective of this paper is to introduce selected statistical topics that are of interest to healthcare quality professionals, educators, technical staff, and researchers who participate in clinical simulation scholarship. The goal is not to provide a detailed review of statistical concepts but rather to give an overview of four important methods in the setting of a hypothetical clinical simulation training workshop. These methods will prepare the clinical simulation research team for their consultations with a statistician or epidemiologist during the design and analysis phases of a study.

The four methods are exact logistic regression (for sparse data), generalized estimating equations (for the analysis of longitudinal data), the intraclass correlation coefficient (for measuring reliability), and causal diagrams (for the building of statistical models). Other methods such as Poisson regression (for count data) and quantile regression (a method that may be preferred to linear regression in certain situations) are also of interest to clinical simulation scholars and educators in the health sciences; however, we chose to focus on the four techniques listed above given that the evaluation of simulation-based training sessions may involve small sample sizes that are prone to sparse data bias and may utilize non-randomized study designs that are especially vulnerable to confounding. Additionally, we have noted that causal diagrams are underutilized in clinical simulation and medical education research.

MATERIALS AND METHODS

Our project did not involve data from human subjects and hence it did not require approval by our institutional review board. The data presented in this article are fictitious (they were generated by the authors). Our hypothetical data were analyzed using SAS 9.4 software (SAS Institute, Inc., Cary, Nor Carolina).

Framework of the hypothetical workshop

A team of educators is designing a one-day clinical simulation training workshop that will address the diagnosis and management of gestational diabetes. The workshop will have multiple training stations. The attendees of the workshop will be students, post-doctoral trainees, and professionals from several disciplines in the health sciences including medicine, nursing, and pharmacy.

Each of the learners will be administered a pre-workshop clinical knowledge examination focusing on the topics that will be addressed in the workshop. During the workshop two faculty members will evaluate the learners on the performance of a task, and, to ensure quality, a measure of inter-rater reliability will be calculated. Immediately after the completion of the one-day workshop, the participants will be administered the same clinical knowledge examination, and this knowledge examination will also be completed by the participants three months and six months after the workshop.

RESULTS

Sparse data: Grappling with a small sample size

Categorical outcomes with two levels are common in clinical and educational research. For example, a learner either passed or failed an examination. When both the independent variable (also known as the exposure or risk factor) and the outcome are dichotomous, then the initial results are frequently displayed in a 2 x 2 table, a contingency table with two rows and two columns (Table 1).
The intersection of a row and column is a cell. The letter A represents the number of subjects who were both exposed to a certain factor (or intervention) and had the outcome of interest (Table 1). The letters B, C, and D represent the remaining three cell values. Multiplying A and D and dividing this quantity by the product of B and C results in a quantity known as the odds ratio (OR): OR=(A x D)/(B x C). The OR quantifies the strength of the association between the independent and dependent (outcome) variables. Logistic regression analysis also produces ORs. The reader is referred elsewhere for an introduction to logistic regression.[6,7]

	Outcome present	Outcome absent
Exposed	A	B
Unexposed	C	D

TABLE 1. Orientation of data in a 2 x 2 contingency table.

Categorical outcomes are typically analyzed using the chi-square test. However, the chi-square test should not be used in the presence of sparse data. The phrase "sparse data" refers to data with no subjects or few subjects at important combinations of variables, e.g., a limited number of exposed cases in a 2 x 2 table (cell A).[8] The chi-square test assumes that each expected (not observed) cell value is at least five.[9] If this assumption is violated then the chi-square test is contraindicated and a Fisher's exact test is usually performed.

Table 2 depicts a sparse data scenario in which one of the cell values is 0. This type of situation may arise in simulation-based education and research when conducting sub-group analyses.

	Passed	Failed	
Specialty A	8	2	10
Specialty B	0	10	10

TABLE 2. Association between the specialty of the resident physician and successful completion (passed vs. failed) of a simulated clinical procedure in 20 learners (hypothetical data). Both the empirical odds ratio and the relative risk for passing are undefined due to division by zero. However, the exact method yields an odds ratio (median unbiased estimator) of 33.6 (one-sided P=0.001).

The educators in the hypothetical interdisciplinary workshop described above would like to analyze data on the participants who are resident physicians. Specifically, the educators would like to determine if the performance on a task at one of the workshop stations varies by the specialty of the learner.

Hand calculations reveal that the OR is undefined due to a division by 0: (8)(10)/(2)(0). Similarly, attempting to fit a traditional logistic regression model to the data found in Table 2 results in a warning from the statistical package stating that the maximum likelihood estimate may not exist. Additionally, the relative risk is also undefined: 0.8/0. It appears that a measure of association cannot be calculated in this situation; however, clinical simulation researchers should be aware of a technique known as exact logistic regression which may be indicated when analyzing small samples. Fitting a logistic regression model when the sample size is small is a complicated version of Fisher's exact test for a 2 x 2 table.[6]

Using the values found in Table 2, exact methods for logistic regression resulted in an estimate of the OR and a test of statistical significance: learners from specialty A had 33.6 times the odds of successfully performing the simulated procedure than individuals from specialty B.[6] This value of 33.6 is a median unbiased estimate of the exact odds ratio (one-sided P=0.0004). It is not the conditional maximum likelihood estimate.

The workshop educators can also control for one or more factors using exact logistic regression. For example, the educators may want to estimate the exact OR for the relationship between the resident's specialty and the dichotomous outcome after adjusting for potential confounders such as the resident's academic rank (postgraduate year). Hosmer and Lemeshow provide a detailed explanation of exact logistic regression.[6] Fernandez and Mulla give details regarding how to perform exact logistic regression using the SAS software package.[9]

Repeated measurement of an outcome

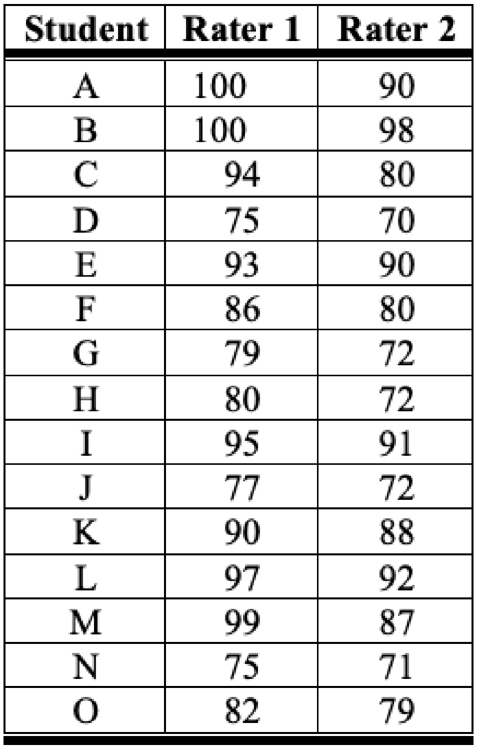
Educators in clinical simulation may measure an outcome at several points in time and therefore should be familiar with techniques to analyze correlated response data.[10] In our hypothetical scenario, the educators will measure clinical knowledge at four points in time: immediately before and after the workshop, and three months and six months after the workshop. The majority of statistical tests and methods that are familiar to non-statisticians assume independence. However, the assumption of independence will most likely be violated in our hypothetical longitudinal study since clinical knowledge examination scores within the same learner will tend to be more similar to each other than scores from different learners. If the workshop educators ignore the correlation between the repeated measurements of clinical knowledge, then the results of their statistical analysis may be biased.[11]

Statistical methods that are appropriate for analyzing longitudinal data include repeated-measures analysis of variance (ANOVA), generalized estimating equations (GEE), and mixed-effect models.[11,12] The popularity of repeated-measures ANOVA has declined over time due to its strong assumptions.[11] GEE and mixed-effect models, in contrast, are modern, flexible approaches to analyzing data that arise from repeated measures designs.[11] Mixed-effects models (also known as mixed models) contain fixed and random effects.[11]

Inter-rater reliability

Healthcare quality professionals and educators may be interested in the agreement (concordance) between two or more raters. For example, during the hypothetical one-day simulation workshop two raters who are faculty members assigned a performance score ranging from 0 to 100 to a group of 15 workshop participants (Table 3). While calculating a Pearson correlation coefficient (or the nonparametric Spearman rank correlation coefficient) is possible in this situation, both of these familiar measures of reliability do not account for the systematic difference (bias) between the two raters.[13]

In the hypothetical dataset presented in Table 3, rater 1's measurements are consistently greater than those of rater 2. A measure of agreement that combines information on both the correlation and the bias between the two readings is the intraclass



Student	Rater 1	Rater 2
A	100	90
B	100	98
C	94	80
D	75	70
E	93	90
F	86	80
G	79	72
H	80	72
I	95	91
J	77	72
K	90	88
L	97	92
M	99	87
N	75	71
O	82	79

Table 3. Hypothetical performance scores assigned by two raters who evaluated a group of 15 nursing students. The Pearson and intraclass correlation coefficients are 0.93 and 0.77, respectively.

correlation coefficient. Shrout and Fleiss presented guidelines for choosing among the six types of intraclass correlation coefficients.[14] The Pearson correlation coefficient and the intraclass correlation coefficient (treating the raters as random effects) calculated from the data reported in Table 3 are 0.93 and 0.77, respectively. The Pearson correlation coefficient indicates a strong linear association between the two sets of scores; however, it does not correct for bias (systematic differences) between the two raters.

Causal diagrams

While any clinical or educational research investigation may be affected by confounding, non-randomized studies are particularly prone to confounding bias. Confounding is a mixing of the effect of the exposure variable on the outcome with a third (extraneous) factor.[15] A confounder is a factor that is related to the exposure, and independent of that exposure variable, is a factor that also affects the outcome.[15] Finally, a potential confounder cannot be on the causal pathway between the exposure and outcome.[15] Methods for identifying confounders have been debated for some time.[16-18] Educators and researchers in the area of clinical simulation research should be aware of the utility of causal diagrams. Causal diagrams aid the researcher in selecting the proper variables for inclusion in a regression model.

Causal diagrams have a long history of use in research.[19] The popularity of causal diagrams, known as directed acyclic graphs (DAG) among epidemiologists and clinical researchers, has increased during the past 15 years.[18-21] DAGs not only allow researchers to identify predictor (independent) variables, such as confounders, that should be included in their regression models, but they also assist in avoiding overadjustment bias. Overadjustment bias occurs when the researcher controls for (adjusts for) a variable that is on the causal path between the exposure and the outcome.[22] Assume that a group of obstetricians would like to estimate the association between maternal tobacco smoking during pregnancy and the outcome of infant mortality. A DAG illustrating this relationship would most likely show that maternal smoking leads to low birth weight which in turn influences infant mortality.[23] If the obstetricians control for low birth weight (which is an intermediate variable) using regression modeling or other techniques, then the total causal effect of maternal smoking on infant mortality cannot be consistently estimated.[22]

Figure 1 displays a DAG which depicts the causal relationship between several variables in a group of learners who participated in the hypothetical interprofessional clinical simulation training scenario. The outcome of interest is the satisfaction of the learner with the training. The researchers believe that the age of the learner impacts their level of professional experience which in turn influences satisfaction. If the researchers assume that the associations shown in Figure 1 are true, then controlling for professional experience when attempting to estimate the association between the learner's age and learner satisfaction will cause overadjustment bias. If the researchers want to estimate the overall effect of age on satisfaction, then there is no reason to control for (condition on)

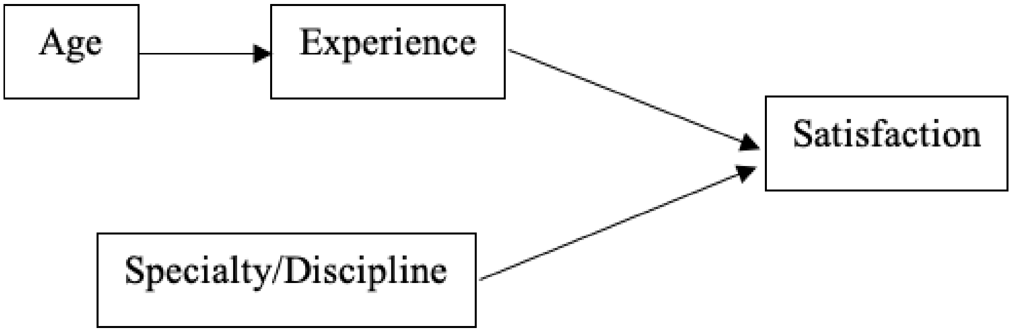


FIGURE 1. Directed acyclic graph (causal diagram) depicting the effects of various factors on the outcome of learner satisfaction during an interprofessional clinical simulation scenario.

the learner's experience, which is an intermediate.[23]

The researchers in this hypothetical interprofessional education scenario may construct several hypothesized DAGs each depicting different relationships between the variables shown in Figure 1. For example, it can be argued that the specialty of the learner has an effect on the type of professional experience possessed by the learner. If this were true, then an arrow leading from "Specialty/Discipline" to "Experience" would have to be included in Figure 1. The variable "Experience" would then be considered a collider. A collider is a variable where two arrowheads meet.[18] Controlling for a collider may result in collider-stratification bias.[18,22,23] The resulting bias may be strong enough to move the observed association between the learner's "Age" and the outcome of "Satisfaction" in the opposite direction of the true association.[18]

Using observational rather than experimental data for causal inference is necessary at times. When conducting observational studies, knowledge of the conditions/outcomes that are being studied combined with the use of DAGs are especially critical in ensuring that the proper variables are adjusted for.[18] Elegant methodological techniques are of little use if subject-matter experts are not included in the design of a study.

DISCUSSION

From a patient safety perspective, the use of simulation for training can be viewed as an ethical imperative.[1,24] To ensure that simulation-based training is effective, educators and healthcare quality professionals must be equipped with appropriate research skills. In this paper, we introduced four important techniques of interest to healthcare quality experts and educators in the health sciences: (1) Data arising from studies involving a small sample size with a binary outcome may benefit from the use of exact logistic regression. (2) Familiarity with GEE can prove beneficial when dealing with investigations in which learners are assessed at multiple points in time. (3) Assessing inter-rater reliability when measuring continuous outcomes should involve the calculation of the intraclass correlation coefficient. (4) A causal diagram known as a DAG facilitates the construction of an appropriate statistical regression model. A limitation of our report is that we discussed only four techniques, and therefore consultations with study design and statistical experts are recommended if the principal investigator does not possess a robust set of research skills. Authors of future, similar overview papers may consider using Monte Carlo simulation to illustrate the use of important statistical methods under varying scenarios.

Improper analysis of data can lead to errors in inference. Lessons learned from the disciplines of epidemiology and clinical research can inform teams of healthcare quality professionals and educators as they engage in simulation scholarship. A recent example involved a popular publicly-available health database, the National Inpatient Sample (NIS).[25] Proper statistical analysis of the NIS requires that researchers account for the complex survey design of this dataset including clustering. Khera et al. randomly selected 120 studies from a population of 1082 studies that used the NIS.[25] The majority (85%) of these published studies on the NIS did not adhere to one or more of the statistical practices that are required to properly analyze and interpret NIS data.[25]

The strength of simulation is its ability to advance the expertise of both individuals and teams.[26] Improving clinical simulation via research similarly requires a multi-disciplinary approach. Experts from the public health and social sciences are valuable additions to the modern simulation-based research team.

AUTHOR AFFILIATIONS

1. Texas Tech University Health Sciences Center El Paso, El Paso, Texas

REFERENCES

1. Brunette V, Thibodeau-Jarry N. Simulation as a tool to ensure competency and quality of care in the cardiac critical care unit. Can J Cardiol. 2017;33(1):119-27.
2. Aebersold M. The history of simulation and its impact on the future. AACN Adv Crit Care. 2016;27(1):56-61.
3. Khanduja PK, Bould MD, Naik VN, Hladkowicz E, Boet S. The role of simulation in continuing medical education for acute care physicians: a systematic review. Crit Care Med. 2015;43(1):186-93.
4. Spencer J. Faculty development research: the ‘state of the art’ and future trends (Chapter 17). In: Steinert Y, editor. Faculty development in the health professions: a focus on research and practice. Dordrecht, Netherlands: Springer; 2014. p.353-374.
5. Cheng A, Kessler D, Mackinnon R, Chang TP, Nadkarni VM, Hunt EA, et al. Reporting guidelines for health care simulation research: extensions to the CONSORT and STROBE statements. Simul Healthc. 2016;11(4):238-48.
6. Hosmer D, Lemeshow S. Applied logistic regression. 2nd ed. New York: John Wiley & Sons, Inc.; 2000.
7. Allison PD. Logistic regression using the SAS System: theory and application. Cary, North Carolina: SAS Institute, Inc.; 1999.
8. Greenland S. Bayesian perspectives for epidemiological research. II. Regression analysis. Int J Epidemiol. 2007;36(1):195-202.
9. Fernandez NP, Mulla ZD. Avoiding sparse data bias: an example from gynecologic oncology. J Registry Manag. 2012;39(4):167-71.
10. Crawford SB, Monks SM, Mendez M, Quest D, Mulla ZD, Plavsic SK. A simulation-based workshop to improve residents’ collaborative clinical practice. J Grad Med Educ. 2019;11(1):66-71.
11. Schober P, Vetter TR. Repeated measures designs and analysis of longitudinal data: if at first you do not succeed-try, try again. Anesth Analg. 2018;127(2):569-575.
12. Dunn G, Pickles A. Longitudinal data analysis, overview. In: Armitage P, Colton T, editors. Encyclopedia of biostatistics. 2nd ed. West Sussex, England: John Wiley & Sons, Ltd.; 2005, p.2906-2918.
13. Szklo M, Nieto FJ. Epidemiology beyond the basics. Gaithersburg, Maryland: Aspen Publishers, Inc.; 2000.
14. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420-8.
15. Hennekens CH, Buring JE. Epidemiology in medicine. Boston: Little, Brown and Company; 1987.
16. Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. Am J Epidemiol. 1989;129(1):125-37.
17. Greenland S. Modeling and variable selection in epidemiologic analysis. Am J Public Health. 1989;79(3):340-9.
18. Bandoli G, Palmsten K, Flores KF, Chambers CD. Constructing causal diagrams for common perinatal outcomes: benefits, limitations and motivating examples with maternal

antidepressant use in pregnancy. Paediatr Perinat Epidemiol. 2016;30(5):521-8.

19. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10(1):37-48.
20. Cole SR, Hernán MA. Fallibility in estimating direct effects. Int J Epidemiol. 2002;31(1):163-5.
21. Bodnar LM, Davidian M, Siega-Riz AM, Tsiatis AA. Marginal structural models for analyzing causal effects of time-dependent treatments: an application in perinatal epidemiology. Am J Epidemiol. 2004;159(10):926-34.
22. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology. 2009;20(4):488-95.
23. VanderWeele TJ, Mumford SL, Schisterman EF. Conditioning on intermediates in perinatal epidemiology. Epidemiology. 2012;23(1):1-9.
24. Ziv A, Wolpe PR, Small SD, Glick S. Simulation-based medical education: an ethical imperative. Acad Med. 2003;78(8):783-8.
25. Khera R, Angraal S, Couch T, Welsh JW, Nallamothu BK, Girotra S, Chan PS, Krumholz HM. Adherence to methodological standards in research using the National Inpatient Sample. JAMA. 2017;318(20):2011-2018.
26. Salas E. Reporting guidelines for health care simulation research: where is the learning? Simul Healthc. 2016;11(4):249.