Education

Improving the Quality and Design of Retrospective Clinical Outcome Studies that Utilize Electronic Health Records

Oliwier Dziadkowiec, PhD,[1-3] Jeffery Durbin, MS,[2] Vignesh Jayaraman Muralidharan, MS, ME,[1] Megan Novak, MA,[3] Brendon Cornett, MPH[3]

Author affiliations are listed at the end of this article.

Correspondence to:
Oliwier Dziadkowiec, Ph.D.
HCA Healthcare, Graduate Medical Education
4900 S Monaco St
Denver, CO 80237
(oliwier.dziadkowiec@hcahealthcare.com)

Abstract

Description

Electronic health records (EHRs) are an excellent source for secondary data analysis. Studies based on EHR-derived data, if designed properly, can answer previously unanswerable clinical research questions. In this paper we will highlight the benefits of large retrospective studies from secondary sources such as EHRs, examine retrospective cohort and case-control study design challenges, as well as methodological and statistical adjustment that can be made to overcome some of the inherent design limitations, in order to increase the generalizability, validity and reliability of the results obtained from these studies.

Keywords

electronic health records (EHRs); retrospective studies; cohort studies; retrospective cohort studies; case-control studies; comparative effectiveness research

Background

Electronic health records (EHRs) have primarily been developed to allow for more efficient and complete medical billing. Secondary "core functions" of EHRs, as defined by National Academy of Medicine, include: (1) health information, (2) result management, (3) order entry/management, (4) decision support, (5) communication and connectivity, (6) patient support, (7) administrative processes and reporting and (8) population health management.[1]

The use of de-identified EHR patient health information for research and quality improvement has become more frequent in the last ten years. EHRs' longitudinal encounter structure and extensive laboratory and pharmacy logs make EHRs an attractive data source for potentially impactful and inexpensive clinical outcomes and effectiveness research studies. Additionally, EHRs are also a major source of data used in Comparative Effectiveness Research, an important component of the Patient Protection and Affordable Care Act of 2010.[2,3]

Although there are significant benefits to big retrospective datasets obtained from EHR systems, designing studies that overcome the challenges associated with retrospective cohort and case-control design remain an issue that undermine the generalizability, validity and reliability of results of these otherwise meaningful studies.

The Benefits of Large Retrospective Studies Based on Electronic Health Records and Other Retrospective Data Sources

The benefits and potential impact of data derived from the EHR are clear in some recent, large retrospective studies. Kaelber et al.[4] accumulated EHR data on more than 1.2 million pediatric patients stemming from 196 ambulatory clinics from 27 states across the country. The large dataset obtained from EHRs allowed the researchers to investigate which antihyper-

tensive medications were commonly prescribed within the pediatric population, an analysis that was previously underpowered and simply not possible due to a low frequency of pediatric patients being prescribed this type of medication.

In a retrospective cohort study, Izurieta et al.[5] utilized retrospective data from 2.5 million Medicare beneficiaries ages 65 and older. Influenza vaccination and infection rates were pulled from administrative records, i.e. Healthcare Common Procedure Coding System (HCPCS) or a Current Procedural Terminology (CPT). For the first time and due to the large sample size, researchers were able to show a significant decrease in hospital admissions when patients were given the high-dose influenza vaccination compared to patients given the standard dose. These findings were not shown in previous randomized studies and can ultimately help physicians when recommending influenza vaccinations in senior patients.

Another population-based retrospective cohort study looked at data from US claims and an integrated laboratory database.[6] The study sample included 72,738 newly treated patients with type 2 diabetes who were employed and were commercially insured from all 50 states in the United States. The data was de-identified, and the variables that were pulled from the database encompassed the following: (1) administrative and demographic data (i.e. type of insurance plan, sex, age, dates of eligibility, and income), (2) inpatient and outpatient visits, (3) medical procedures, (4) laboratory tests and results, and (5) pharmacy claims. This was the first large retrospective cohort study to assess the comparative effectiveness and safety of the medication sitagliptin in type 2 diabetic patients. By having thorough clinical data to parse, this study could assist healthcare personnel to provide safer care to patients and more reliable prescribing of medications.[6]

A retrospective observational study looked at stroke/systemic embolism (SE) and major bleeds (MB) in patients with non-valvular atrial fibrillation. This study comprised of 434,046 patients who were matched in six different medicine cohorts. The data was received from the Centers for Medicare and Medicaid Services and 4 US commercial claims databases. This was the largest observational study that compared oral anticoagulants (NOAC) to the oral drug warfarin. After analyzing the data, findings were consistent with previous studies when comparing NOACs and warfarin. This study offers more information regarding the benefits and risks of stroke prevention in nonvalvular atrial fibrillation patients due to greater statistical power and improved generalizability from multiple databases, instead of using a single data source like small randomized control trials (RCTs) have previously done.[7]

A recent retrospective analysis used de-identified WoundExpert EHRs from 242 wound care facilities across the US over a 5-year span. There were a total of 1,498 patients pulled from the EHRs, and data of 1,622 diabetic foot ulcers (DFUs) were analyzed. Variables that were extracted from the EHRs included the following: (1) age, (2) sex, (3) race, (4) body mass index (BMI), (5) wound location, (6) wound size and duration, (7) number of wounds per patient and (8) single/multiple wounds per patient. The researchers found significant differences in frequency and the time of healing when using human fibroblast-derived dermal substitute (HFDS) in patients with diabetic foot ulcers. These findings could imply overall cost savings for medical resources, home health, prescription drugs, physician office visits, emergency department visits and hospitalizations.[8]

Bias in the Design of Retrospective EHR-Based Research Studies

High-quality observational studies can generate credible evidence of intervention effects, particularly when rich data are already available. Retrospective observational studies are useful, particularly when RCTs are not feasible and too expensive to carry out.[9] One of the major methodological issues and challenges of retrospective observational study design includes cohort selection bias. This bias arises when the study population is not randomly selected from the target population, contains loss of information, including follow-ups, drop outs or deaths and/or an inability to control confounders that might be associated with outcome.[10]

Another major limitation of retrospective study designs is the scope of the already collected data. In research that utilized EHR-derived

data, most of the data were originally collected for other purpose (ex. billing) and not all relevant information is available for analysis, lead to omission of crucial confounders that can introduce bias. For instance, if BMI is an important confounder in a given study and a large percentage of patient data are missing height or weight, the researcher might not be able to use BMI in the analysis. In addition, when conducting case-control or cohort studies, omitted details from the patient's side (ex. specific information collected from patients about treatment episodes that involved multiple treatments) and uncaptured patient characteristics introduce recall and selection biases.[11, 12]

Considering necessary sample size and power to avoid random errors, defining a clear hypothesis, identifying correct populations and treatments with clinically relevant data, maintaining strict inclusion criteria and exclusion criteria, identifying and defining the outcomes that will be used to measure for the study and preparing a suitable study plan in advance will help to minimize risk of research design-related biases.[13]

The Role of Statistical Methods in Reducing Bias in Retrospective Studies

In the previous sections, the many benefits and limitations of using EHRs for retrospective research, such as cohort, case-control and comparative-effectiveness studies, were reviewed. In particular, the many ways bias and error can exist in retrospectively-collected data presents significant challenges in using statistical methods to test trends in the population. These challenges often take the form of missing data, leading to a non-holistic representation of a patient encounter. Additional challenges may arise with incorrect data that was transcribed during the billing process or when patient records are entered into database systems. Furthermore, the design of studies using this data often suffer from a lack of randomization, such that patient populations are not randomly selected or assigned to cohorts. When a lack of randomization exists in the design, confounding and selection bias may be nearly impossible to account for in the data, such as when patient characteristics aren't balanced a priori, particularly when data are missing. In this section, we'll review these issues in detail and propose a selection of statistical analyses that aid in accounting for them.

Statistical Issues in EHRs and Retrospective Studies

Missing data in EHRs can arise from a variety of sources, ranging from a lack of documentation to a mistake in transcription, and can prevent researchers from having a holistic view of the patient encounter.[14, 15] Depending on the nature of the missingness (i.e., the method in which the data are missing), there exists a number of statistical methods that allow us to estimate observations of the patient encounter based on previously observed and complete data.[16] The nature of the missing data can be categorized in three ways: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). MCAR and MAR circumstances are instances when data are missing due to unknown circumstances or when other data can explain why a datum is missing. For example, if a physician forgets to order a certain lab test randomly or a physician does not order one test in lieu of another, these circumstances illustrate MCAR or MAR data respectively. MNAR circumstances are defined as instances when data are missing in a mechanistic and explainable way, such as not ordering any tests because a hospital laboratory is closed for renovations. When data are missing at random, as in the MAR and MCAR circumstances, methods such as imputation and data reduction may be used to generate possible values to replace the missing values. In imputation methods, the goal is to generate "likely" data values via single-value replacement (e.g., substituting the observed mean of the available data for the missing values) or likelihood-based methods (e.g., calculating the most-probable value from predicted values of a regression model). In instances where data are MNAR, such as when patient records are mechanistically not reported, current guidance is to either eliminate observations with the missing values or to remove the variable from consideration. It is worth noting, however, that the removal of observations can often lead to a detrimental drop in power for a given analysis within a study, which removes valuable information relevant to cohort or case-control identities of the patients. Thus, it should only

be used when few patient records are missing the data in question.[17, 18]

Unfortunately, missingness is not the only issue with retrospectively-collected EHR data. When using these data for studies, such as cohort, case-control or comparative-effectiveness, these designs often suffer from a lack of randomization. Randomization refers to the method in which data was obtained—for example, in RCTs, patients are randomly identified/selected and randomly assigned to a cohort relevant to the study aims. In retrospective studies, however, patients are typically not randomly selected nor are they randomly assigned to cohorts. In fact, the use of retrospective data to these ends typically results in selection or ascertainment bias, as the data are the result of non-random health care processes, such as standards-of-care and extant protocols for illness or disease treatment.[19, 20] This lack of randomization presents several problems: first, many parametric statistical methods rely on randomly-collected data as an assumption for observing trends in the population, thus decreasing the validity and reliability of the tested effects and marring the true relationships inherent in the population; and second, non-randomized data prevents the balancing of nuisance variables or confounders between cohorts, thus, one cohort can suffer from extraneous circumstances over another.

One possible remedy for this lack of randomization is to simulate "randomness in selection" by randomly subsetting the available EHR data into discrete datasets and treating them as separate samples of the population—this provides an opportunity to both satisfy the randomization assumption of parametric statistical methods (thereby increasing the validity and reliability of the test if the same effects are observed across samples), as well as potentially to balance third-variables such that they are contributing equal variances within and between cohorts and across samples. While this method provides opportunities to address the shortcomings that a lack of randomization introduces, this cannot be done when the available data are limited in size and scope (e.g., studying a low-prevalence disease entity). Fortunately, there are statistical methods that can aid in creating balanced cohorts or identifying nuisance variables that may be confounding the results of the analysis.

Selected Statistical Methods for Studies with Retrospective Designs

When random subsetting of a large dataset is not suitable to simulate randomization (e.g., when the sample is small), the goal becomes defining well-balanced cohorts and identifying nuisance variables that can be accounted for between cohorts. Hlatky et al.[21] outlined two classes of analyses that are designed to accomplish these two goals: first, the use of propensity score matching, a technique of determining an individual patient's propensity to receive one treatment over another, to define cohorts; and second, the use of structural modeling and graph-analytic approaches to identify correlational and causal relationships between variables that may indirectly influence relationships between variables pursuant to the study aims. In addition to these methods, data reduction techniques, such as principal components analysis, independent components analysis and multidimensional scaling may be used to extract the relevant observations and variables that maximize explainable differences between cohorts. Both approaches are reviewed below.

Propensity Score Matching

Propensity score matching (PSM) is a method of using available variables, both those relevant to the study aims, as well as possible confounders, to determine a patient's propensity for receiving one form of treatment over another in order to create balanced cohorts.[22, 23] Consider, for example, the examination of treatments for a low-prevalence disease—because few patients are diagnosed and treated for such a disease, they may have wide variability in health traits or symptomology. It would be impossible to compare different treatment options without the influence of these varying traits, and thus our analysis of the data must take into account these traits to determine how they influence the treatment. In its simplest form, PSM allows researchers to balance these nuisance traits across groups by determining how they influence the propensity to receive a particular treatment. Typically, the propensity score is calculated as the prediction of a linear combination of known variables—in most cases,

the treatment cohorts serve as the outcome of a generalized linear model (such as logit or probit models) where known variables serve as predictors. By using a nominal outcome for the logit or probit model, the resulting value is the "likelihood" of an individual patient receiving one treatment over another based on the actual observed treatment method (e.g., Ananthakrishnan et al.[24]). Once these propensities are calculated, a matching algorithm is used to match patients into cohorts based on the propensity and the actual treatment that was given—the most common of these algorithms being nearest-neighbor algorithms, where "near" observations are grouped together.[25] In an ideal setting, this method is used to develop random, balanced samples of patients, where known confounders are considered and relationships between variables of interest can be examined without compromising the assumptions of parametric statistical methods.

Structural Modeling and Graph-Analytic Approaches

Structural modeling techniques, in conjunction with graph-analytic approaches, serve to identify the structure in which variables relate to one another. Particularly, these methods are useful when confounders are not well-specified, whether from a lack of existing literature specifying the relationship or in instances where identifying the confounder is the aim of the study. These methods are generally designed to assess or ascertain the structure of variables by testing linear combinations of known variables to find the optimal relationships between them. For example, if it were unclear whether smoking influences hepatic issues, the structural model would aim to specify whether certain variables, such as BMI or historical diagnoses of digestive organ issues, influence the relationship between smoking and hepatic issues. Similarly, graph-analytic approaches are helpful in visualizing these relationships when the relationship of interest involves a large number of confounders, moderators and mediators (e.g., Williams et al.[26]). When temporal information is available from the EHR, such as the progression of diagnoses in a patient encounter while held inpatient, methods such as marginal-structure models can be used to assess the influence of confounders over time by discounting or accounting for those variables at certain steps of the temporal progression. Thus, these methods may be used to identify and account for nuisance variables when the relationship is unclear, which can rectify the effects of a lack of randomization and allow for a more valid and reliable assessment of the relationship between variables of interest.

Principal Component Analysis and Multidimensional Scaling

The final method to be reviewed in service to correcting the lack of randomization is data reduction techniques comprised of methods such as component analysis (i.e., principal or independent components analysis; PCA and ICA, respectively) or multidimensional scaling (MS). Generally, the goal of these techniques is to transform datasets, where many variables are to be considered, into the most informative and succinct subset of data.[27] This method is ideal in situations where biomarkers or health characteristics are not well-defined—if it were the case where we wanted to find specific lab test values or comorbidities that indicated septic shock, among a swathe of hundreds of variables, data reduction techniques allow us to separate relevant, variance-explaining data from uninformative data. In particular, component analyses such as PCA or ICA are used to identify optimal "components" of the data that maximize the variances that can be accounted for across the entire dataset. Similarly, MS methods are used to find the optimal dimensions (often 2 or 3 dimensions) that can successfully capture the variability between data points without sacrificing the meaningfulness of the differences between observations. These methods are particularly useful in situations where the effect of a set of certain variables are not known and must be established before the relationships of interest are examined—although, the resulting dimensions identified in these analyses can be difficult to interpret when there is no clinical or observable realistic relationship to be determined. Thus, dimensionality reduction should be reserved for instances when confounders are known but their relationship between variables of interest are not known and rather are suspected via clinical or care-related knowledge.

The Role of Effect Size in Studies with Large Sample Sizes

Effect size, which measures magnitude rather than how rare a statistical difference is, has been deemed more accurate for gaging statistical significance of a between group difference than p-values in studies with large sample sizes.[28] It has also been adopted in clinical literature as a more useful to approximate clinical significance.[29] For instance, in a hypothetical study, a complex, robotically-assisted cardiac surgery technique is developed to improve upon a traditional surgical method. The 30-day mortality of these 2 techniques are compared retrospectively via a large EHR database using a chi-squared test. A significant p-value is obtained, and a manuscript recommending one technique over the other is written.

What this hypothetical analysis leaves out is the odds ratio for the comparison was only 0.96, which is a very small difference of only 4% from the traditional technique. The authors of the hypothetical analysis would likely not recommend the use of the new surgical technique given that the investment in teaching the new technique and buying the machinery might not be worth a 4% reduction in 30-day mortality. The use of p-value in hypothesis testing suffers from a fatal flaw. As sample size increases, p-values become more likely to be significant. Repeating a 100 patient study with 100,000 patients will result in higher chances of producing significant p-values, even if the between group difference is exactly the same. This would happen because p-value is an indication of how likely a result is to have come from chance alone. The larger the sample size, the less likely a result is to be chance, and thus, lower p-values. That does not mean the result is any more clinically significant, just that it is unlikely to be formed by random chance alone.

The use of EHRs commonly results in very large sample sizes. With the ability to harvest tens of thousands of patients in some of the larger hospital network EHRs, traditional statistical power becomes trivial, and p-values become far less trustworthy to be the sole determining factor of significance alone. The p-value should only be part of the determination of the value of a result in these cases, and researchers should always be wary of large sample studies reporting only a p-value and not the adjoining effect size.

Conclusion

Despite the numerous limitations outlined in this paper, the use of EHRs for retrospective studies presents a valuable opportunity to explore novel research questions that are generally underpowered and unable to be answered in a prospective research setting. Powerful statistical techniques exist that aid in correcting these issues related to missing data and lack of randomization, making EHRs a well-suited mechanism to evaluate the efficacy of treatments and outcomes in the healthcare setting and much more.

Conflicts of Interest

The authors declare they have no conflicts of interest.

The authors are employees of HCA Healthcare Graduate Medical Education, an organization affiliated with the journal's publisher.

This research was supported (in whole or in part) by HCA Healthcare and/or an HCA Healthcare affiliated entity. The views expressed in this publication represent those of the author(s) and do not necessarily represent the official views of HCA Healthcare or any of its affiliated entities.

Author Affiliations

1. HCA Healthcare Mountain Division
2. HCA Healthcare Mid-America Division
3. HCA Healthcare Continental Division

References

1. Kim E, Rubinstein S, Nead K, Wojcieszynski A, Gabriel P, Warner J. The Evolving Use of Electronic Health Records (EHR) for Research. *Semin Radiat Oncol*. 2019;29(4):354-361. https://doi.org/10.1016/j.semradonc.2019.05.010
2. Masica A, Collinsworth A. Leveraging Electronic Health Records in Comparative Effectiveness Research. *Prescriptions for Excellence in Health Care Newsletter Supplement*. 2012:1(14). http://jdc.jefferson.edu/cgi/viewcontent.cgi?article=1105&context=pehc
3. Thorpe JH. Comparative effectiveness research and health reform: implications for public health policy and practice. *Public*

Health Rep. 2010;125(6):909-912. https://doi.org/10.1177/003335491012500619
4. Kaelber DC, Liu W, Ross M, et al. Diagnosis and Medication Treatment of Pediatric Hypertension: A Retrospective Cohort Study. *Pediatrics*. 2016;138(6). https://doi.org/10.1542/peds.2016-2195
5. Izurieta HS, Thadani N, Shay DK, et al. Comparative effectiveness of high-dose versus standard-dose influenza vaccines in US residents aged 65 years and older from 2012 to 2013 using Medicare data: a retrospective cohort analysis. *The Lancet Infectious Diseases*. 2015;15(3):293-300. https://doi.org/10.1016/S1473-3099(14)71087-4
6. Eurich DT, Simpson S, Senthilselvan A, Asche CV, Sandhu-Minhas JK, Mcalister FA. Comparative safety and effectiveness of sitagliptin in patients with type 2 diabetes: retrospective population based cohort study. *BMJ*. 2013;346(apr25). https://doi.org/10.1136/bmj.f2267
7. Lip GY, Keshishian A, Li X, et al. Effectiveness and Safety of Oral Anticoagulants Among Nonvalvular Atrial Fibrillation Patients. *Stroke*. 2018;49(12):2933-2944. https://doi.org/10.1161/STROKEAHA.118.020232
8. Sabolinski ML, Capotorto JV. Comparative effectiveness of a human fibroblast-derived dermal substitute and a viable cryopreserved placental membrane for the treatment of diabetic foot ulcers. *Journal of Comparative Effectiveness Research*. 2019;8(14):1229-1238. https://doi.org/10.2217/cer-2019-0001
9. Lu CY. Observational studies: a review of study designs, challenges and strategies to reduce confounding. *International Journal of Clinical Practice*. 2009;63(5):691-697. https://doi.org/10.1111/j.1742-1241.2009.02056.x
10. Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM. *Developing a Protocol for Observational Comparative Effectiveness Research: A Users Guide*. Agency for Healthcare Research and Quality; 2013.
11. Galvan RF, Barranco V, Galvan JC, Batlle S, Garcia FS. Limitations and Biases in Cohort Studies. *Intech*. 2016;13. https://doi.org/10.5772/57353
12. Roche N, Reddel H, Martin R, et al. Quality Standards for Real-World Research. Focus on Observational Database Studies of Comparative Effectiveness. *Annals of the American Thoracic Society*. 2014;11(Supplement 2). https://doi.org/10.1513/AnnalsATS.201309-300RM
13. Marko NF, Weil RJ. The Role of Observational Investigations in Comparative Effectiveness Research. *Value in Health*. 2010;13(8):989-997. https://doi.org/10.1111/j.1524-4733.2010.00786.x
14. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. *Medical Care*. 2013;51. https://doi.org/10.1097/MLR.0b013e31829b1dbd
15. Raman SR, Curtis LH, Temple R, et al. Leveraging electronic health records for clinical research. *American Heart Journal*. 2018; 202:13-9. https://doi.org/10.1016/j.ahj.2018.04.015
16. Hegde H, Shimpi N, Panny A, Glurich I, Christie P, Acharya A. MICE vs PPCA: Missing data imputation in healthcare. *Informatics in Medicine Unlocked*. 2019;17:100275. https://doi.org/10.1016/j.imu.2019.100275
17. Manly CA, Wells RS. Reporting the Use of Multiple Imputation for Missing Data in Higher Education Research. *Research in Higher Education*. 2015;56(4):397-409. https://doi.org/10.1007/s11162-014-9344-9
18. Masconi KL, Matsha TE, Erasmus RT, Kengne AP. Effects of Different Missing Data Imputation Techniques on the Performance of Undiagnosed Diabetes Risk Prediction Models in a Mixed-Ancestry Population of South Africa. *Plos One*. 2015;10(9). https://doi.org/10.1371/journal.pone.0139210
19. Haneuse S, Daniels M. A General Framework for Considering Selection Bias in EHR-Based Studies: What Data are Observed and Why? *eGEMs (Generating Evidence & Methods to improve patient outcomes)*. 2016;4(1):16. https://doi.org/10.13063/2327-9214.1203
20. Mcculloch C, Neuhaus J. Statistical Methods for Reducing Bias in Comparative Effectiveness Research When Using Patient Data from Doctor Visits. *PCORI*. 2019. https://doi.org/10.25302/6.2019.ME.130601466
21. Hlatky MA, Winkelmayer WC, Setoguchi S. Epidemiologic and Statistical Methods for Comparative Effectiveness Research. *Heart Failure Clinics*. 2013;9(1):29-36. https://doi.org/10.1016/j.hfc.2012.09.007
22. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55. https://doi.org/10.1093/biomet/70.1.41
23. Rosenbaum PR. Model-Based Direct Adjustment. *Journal of the American Statistical Association*. 1987;82:387-394. https://doi.org/10.1080/01621459.1987.10478441
24. Ananthakrishnan AN, Cagan A, Cai T, et al. Comparative Effectiveness of Infliximab and Adalimumab in Crohn's Disease and Ulcerative Colitis. *Inflammatory Bowel Diseases*. 2016;22(4):880-885. https://doi.org/10.1097/MIB.0000000000000754
25. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*. 2014;33(6):1057-1069. https://doi.org/10.1002/sim.6004
26. Williams TC, Bach CC, Matthiesen NB, Henriksen TB, Gagliardi L. Directed acyclic graphs: a tool for causal studies in paediatrics. *Pediatric Research*. 2018;84(4):487-493. https://doi.org/10.1038/s41390-018-0071-3
27. Kalankesh L, Weatherall J, Ba-Dhfari T, Buchan

I, Brass A. Taming EHR Data: Using Semantic Similarity to Reduce Dimensionality. *Studies in Health Technology and Informatics*. 2013;192:52-56.
28. Halsey LG. The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? *Biology Letters*. 2019;15(5). https://doi.org/10.1098/rsbl.2019.0174
29. Sullivan GM, Feinn R. Using Effect Size—or Why the P Value Is Not Enough. *Journal of Graduate Medical Education*. 2012;4(3):279-282. https://doi.org/10.4300/JGME-D-12-00156.1