Special Article

GREENVILLE HEALTH SYSTEM

Meaningful Analysis of Small Data Sets: A Clinician's Guide

Justin Collins, BS; Jordan Brown, BS; Christine Schammel, PhD; Kevin Hutson, PhD; and W. Jeffery Edenfield, MD

From the Institute for Translational Oncology Research, Greenville Health System, Greenville, SC (J.C., W.J.E.); Department of Mathematics, Furman University, Greenville, SC (J.B., K.H.); and Pathology Consultants, Greenville, SC (C.S.)

Abstract

A well-known problem in clinical research is the size of accessible data sets. When analyzing any data set the goal is always to find the actual correlations, quantify these correlations, and assess the likelihood the correlations are not random. Large data will allow for detection of smaller magnitude significant differences between groups, though this statistical significance does not immediately imply clinical relevance. Large research institutions have the resources to curate large databases; however, their patient cohorts differ from patient populations of smaller regional/local medical institutions. Often, smaller hospitals lack the patient volume and resources to maintain large databases. Additionally, access to dedicated medical statisticians may be limited at these places. Thus, these institutions face unique difficulties in statistical analysis. Smaller data sets are not unusable; they merely require a slightly different approach. As smaller institutions outnumber large institutions and provide healthcare locally, the majority of patients will be seen at these places. Properly defining techniques to extract meaningful information from small data sets would allow these hospitals to leverage their patient cohorts, which more closely resemble a random representative sample of the average patient nationally. To date, the methodology of data analysis on small medical data sets has not been expounded on. This paper targets clinicians doing their own statistical analysis and explains the basics of hypothesis testing. Additionally, it sheds light on common problems found in analyzing small medical data sets, provides methods to combat them, and explains why small cohort analysis does not preclude generalizability.

A well-known problem in clinical research is the size of accessible data sets. When analyzing any data set, a differentiation between the signal—the actual pattern that comes from a difference between groups or some explainable connection between variables—and the noise—the random variation in a variable—must be elucidated.[1] This differentiation requires hypothesis testing in which the null hypothesis is often defined as the assumption that the variables have no relationship to each other. Specifically, under the null hypothesis, 2 patient cohorts should look and test as if they come from a single patient population rather than 2 disparate groups. Statistical tests are used to determine the probability that the null hypothesis is true, evaluating the likelihood that 2 (or more) groups in the cohort are from the same sample.

Large data sets allow for more precise attenuation of the signal and, thus, have lower perceived noise, thus allowing a smaller magnitude of difference between groups to be considered statistically significant, and the null hypothesis is more likely to be rejected. Because the power and efficacy of statistical analyses most often used rely on the size of the patient population and data set, conclusions from these large institutions often result in novel medical treatments and add to the generalizable knowledge of the field.

The assumption of these data is that larger samples represent a random sampling of the demographic; however, while large research institutions have the resources to compile and analyze large data sets, large centers operate primarily on referrals and may decline patients based on fidu-

ciary eligibility. Additionally, treatments at a distance from patient homes are often cost prohibitive. As such, these large cohorts may not be truly representative random samples of the population as a whole and the conclusions found may not translate practically into practice at smaller institutions. These smaller institutions, which greatly outnumber large institutions, provide care for the majority of the population. However, smaller centers have neither the patient volume nor the resources to maintain large databases.

As such, an intentional and practical approach to analysis is required to derive meaning from smaller cohorts. Properly using statistical techniques to extract information from small data sets would allow for data more representative of the average cohort to be included in the literature.

In addition, many clinicians have insufficient training regarding statistics to critically analyze statistical analysis in published articles and, as a result, hold the misconception that their own research, made up of small cohorts, has nothing to contribute to the literature as a whole. This fact perpetuates the void in the literature from regional medical institutions. Lacking access to large data sets or statisticians willing to work with small data sets further exacerbates this problem. Elucidating the specifics of statistical hypothesis tests commonly used in medical literature, explaining the considerations of small data sets, and clarifying the generalizations that can or cannot be made from the statistical information will empower clinicians at small regional medical centers to collect, analyze and publish their research, thus filling the previously created void in the body of medical literature.

Background

Overgeneralized statements are one of the biggest obstacles in analyzing and evaluating the results of small data sets. As such, it is imperative to evaluate the statistical tests most commonly used in medical research and understand the effect of small data sets on the results of these tests, the assumptions built into each of the tests, and the actual mathematical question each test is designed to answer. While published methods address the caveats and statistical validity of smaller data sets,[2-5] effective use of these methods is scarce, necessitating a better understanding for the purpose of empowering smaller institutions.

Student's T-test

The student's t-test[6], the most frequently used test in medical statistical analysis, attempts to determine if the difference between the means of 2 data sets arises from random variation. Often, the threshold for statistical significance is arbitrarily defined as $P \leq .05$.

For example, in medical research, t-tests are usually used to compare 2 different treatment groups based on continuous outcome variables such as BMI. Each of the 2 groups will be treated by some changing parameter (drug type, surgery type, etc.), and the outcomes will be compared. The value returned from the t-test is compared to the z-distribution and yields a probability score defining the likelihood that the difference in means is the result of sampling error. This probability can be given from a 1- or 2-tailed version of the test. If the directionality of the relationship of the variables is known, a 1-tailed t-test is appropriate; however, because effects of individual variables on specific outcomes are typically ambiguous, a 2-tailed t-test is appropriate if the relationship between variables is unknown or under investigation. In general, very specific information about variable relationships must exist for a 1-tailed t-test to be appropriate.

Assumptions for performing these tests are as follows: samples in each group are independent, the 2 populations being compared are normally distributed, and the 2 samples should have equal variance. If these conditions are not met, or good reason exists to doubt whether they are fulfilled, a non-parametric test would be more appropriate. For example, survival in days is almost never normally distributed. Additionally, the variance in each sample compared to the other can be unequal. In this case, the student's t-tests would be appropriately replaced by a Welch's t-test.

Chi-square Test

The Chi-square test allows for hypothesis testing for categorical variables and categorical outcomes. In medical research, rare frequencies of certain outcomes can make tests challenging to use routinely; however, when data can be classified into a contingency table, a Chi-square test is an appropriate statistical option.

For Chi-square analysis, events must be mutually exclusive and have a total probability of one. Contraindications for Chi-square analysis include events with a very low probability or low frequency; for these scenarios, a Fisher's exact test (if the data are in a 2x2 contingency table), binomial test, or a g-test would be more accurate.[7] When the test statistic is computed and then transformed into a probability score, the *P* value returned gives the likelihood the groups are from the same distribution, meaning the probability that the test groups' difference is the result of random sampling error.

For example, Chi-square tests are used when you have 2 or more groups based on treatment with a categorical outcome such as mortality.

ANOVA

An analysis of variance (ANOVA) is optimally used when data will have more than 2 groups and consist of continuous outcomes. Like the t-test, ANOVA attempts to determine if the difference between the means of multiple data sets (3 or more groups) arises from random variation. While technically this difference could be determined by running pairs of t-tests with all possible groups, the likelihood of incurring type-I errors is high, causing the null hypothesis to be rejected and a statistically significant false positive to be found.[8]

Type-I errors, also known as false positives, are when the null hypothesis is mistakenly rejected. Type-II errors, on the other hand, are also called false negatives. These occur when the null hypothesis is mistakenly accepted.

In ANOVA, observations among groups must be independent, the groups must each have a normal distribution, and variances among groups must be equal to evaluate the null hypothesis. In medicine, this situation requires control over the 3 or more groups for all possible variables when investigating whether a single factor affects the means of the groups. The probability, calculated from the F-statistic, gives the likelihood that the difference in the groups' means is the result of random sampling error. While this test is widely used in medicine, sometimes there are not large samples for each of the different groups. This scenario will cause ANOVA to produce type-II errors, meaning the null hypothesis will fail to be rejected even when it should be rejected, producing a false negative. In analysis of medical data, ANOVA analysis is used just like a t-test but for more than 2 treatment groups.

Tests That Are Not Parametric

Measures exist to determine if data meet the appropriate assumptions for analysis using parametric tests like t-tests, Chi-square tests, and ANOVA. For example, the Shapiro-Wilk test, a test for normality, can determine if the samples from a population have a normal distribution, thus indicating the ability to use a t-test/ANOVA. If the Shapiro-Wilk test returns a P value below the specified alpha level (usually $P < .05$), the null hypothesis that the data come from a normal distribution would be rejected; analysis must commence using non-parametric tests.[9]

Non-parametric equivalents of almost all parametric statistical tests exist. In each case, they make far fewer assumptions but sacrifice sensitivity. As an example, some non-parametric tests transform the data into rank order, allowing for comparison of the medians instead of the means. This method eliminates the need for data to be normally distributed and also lowers the variability, thus eliminating the effect of outliers.

The drawback of non-parametric tests is a reduction of power when compared to its parametric counterpart, which then requires larger cohorts to show a significant difference of equal magnitude. While too many non-parametric analogs of parametric tests exist to list here, some examples include the Kruskal-Wallis (One-way ANOVA), Mann-Whitney U-test (t-test), and the Wilcoxon signed-rank test (paired t-test). In medicine, age is often compared between groups; however, because age is related to many disease processes and outcomes, it is rarely normally distributed and, thus, is often analyzed with non-parametric tests.

Potential Problems

Too Few Patients, Too Many Groups

In medical research at smaller institutions with a limited sample size, groups may exist that contain only a few patients. In these cases, many hypothesis tests are unlikely to reveal any significant group differences, necessitating another analytic approach.

Oftentimes, patients can be clinically reclassified to reduce the degrees of freedom, yielding more patients per group and, thus, a more balanced analysis. This situation requires a biological understanding of the underlying question asked of the data. For example, a data set of patients that received various chemotherapy regimens for multiple myeloma containing 77 patients across 14 different groups could be analyzed via ANOVA. However, to strengthen the power of the analysis, an evaluation of the mechanism of action of the drugs in each regimen was completed to reduce the number of treatment regimens in this cohort. Patients were then reclassified by combined mechanism of action of the drugs in each respective regimen. This reclassification yielded only 5 biologically and clinically different groups instead of the previous 14, allowing for a more sensitive analysis.

The importance of understanding the medicine and biology behind the data cannot be overstated. Many times, when working with small medical data sets, a familiarity with the underlying science is necessary to solve problems that are not immediately straightforward. This scenario is what makes working with small medical data sets simultaneously difficult and rewarding.

Categorical Data With No Expected Values

Binary categorical outcome variables can be recoded as 0 and 1, respectively, and then analyzed via t-test. This method would be applicable when the assumptions for a Chi-square or Fisher's exact test cannot be met; however, while the data will likely follow a binary distribution, it may approximate a normal distribution. When using this method, if the output variable is non-binary, a more sophisticated layering of binary variables is required (see effects coding or contrast coding for more information).

Small Data Sets Are Generalizable

The most pronounced problem in analyzing small data sets is the perceived deficiency in the generalizable nature of the conclusions. Typically, hypothesis tests consider the sample size being analyzed such that a statistically significant difference in a small cohort is just as valid as those found in larger cohorts. The difference between small and large data sets lies in the magnitude of difference needed to reject the null hypothesis with an alpha level of .05.

For example, if an actual difference of 25% between groups could be detected by a certain size data set, a larger cohort of patients would be required to detect a 10% difference. Thus, smaller data sets avoid false positives by requiring a larger difference between cohorts. This means that statistically significant findings from a small data set with a particular alpha level (compared to the same finding in a larger data set at the same alpha level) are likely of equal or greater magnitude than that of the larger data set without a loss of confidence. Thus, statistical significance in small data sets occurs much less frequently and only when the difference between the groups is large; however, there is difficulty in contributing to the generalizable knowledge when the null hypothesis must be rejected to achieve a "statistically significant difference" between groups.

Conclusion

In exploring some of the possible problems and solutions in analyzing small data sets, the goal is to demonstrate the logic and intentionality required to find meaning in this information. While research at smaller centers is currently under-represented in medical literature, this review illustrates some common problems and the solutions to facilitate the inclusion of the data generated at these centers into the body of generalizable medical knowledge. Additionally, it can serve as a call-to-arms of sorts for small medical institutions to begin meaningful research while simultaneously giving them the tools necessary to analyze their data. Specifically, this work can help clinicians who want, or need, to analyze their own data know where to start. The key is to understand the statistical tests, what they mean, how data must be appropriately coded, meaning of the output produced by a particular analysis, and power/limitations in applicability. The most important rule in these analyses is to assure all transformations of the data do not affect the information stored therein. This fact requires knowledge of the underlying science and how the statistical test used deals with the data on a computational level.

As medicine moves toward evidence-based practice, the findings generated by these tools at smaller institutions will serve the patient population well, allowing clinicians to modify practice that more accurately reflect their demographic. Additionally, the generalizability of statistically significant hypothesis testing from small cohorts should not be dismissed outright but, rather, critically evaluated for what the statistical testing actually shows.

Correspondence

Address to:
W. Jeffery Edenfield, MD, Greenville Health System, Institute for Translational Oncology Research, 900 W Faris Rd, Greenville SC 29605 (jedenfield@ghs.org)

References

1. Silver N. *The signal and the noise: why so many predictions fail—but some don't.* New York: Penguin Press; 2012.
2. Bacchetti P, Wolf LE, Segal MR, McCulloch CE. Ethics and sample size. *Am J Epidemiol.* 2005;161:105-10.
3. Biau DJ, Kerneis S, Porcher R. Statistics in brief: the importance of sample size in the planning and interpretation of medical research. *Clin Orthop Relat Res.* 2008;466:2282-8.
4. van de Schoot R, Broere JJ, Perryck KH, Zonderman-Zwijnenburg M, van Loey NE. Analyzing small data sets using Bayesian estimation: the case of post-traumatic stress symptoms following mechanical ventilation in burn survivors. *Eur J Psychotraumatology.* 2015;6:25216.
5. *Small Clinical Trials. Issues and Challenges.* Evans CH, Ildstad ST, eds. Washington DC: National Academy Press; 2001.
6. Du Prel J, Röhrig B, Hommel G, Blettner M. Choosing statistical tests: part 12 of a series on evaluation of scientific publications. *Dtsch Arztebl Int.* 2010;107:343-8.
7. Ludbrook J, Dudley H. Issues in biomedical statistics: analyzing 2 x 2 tables of frequencies. *Aust N Z J Surg.* 1994;64:780-7.
8. Laerd statistics. One-way ANOVA in SPSS statistics. https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php. Accessed June 18, 2016.
9. Whitley E, Ball J. Statistics review 6: nonparametric methods. *Crit. Care.* 2002; 6:509-13.