10-2017

Not All Evidence is Created Equal: Changes in practice require the highest possible level of statistical testing

Sandy Hunter
Eastern Kentucky University

Recommended Citation

Hunter, S. (2017). Not All Evidence is Created Equal: Changes in practice require the highest possible level of statistical testing. EMS World, 46(10), 46-50.

Not All Evidence Is Created Equal

emsworld.com/node/218772

This is the first in a four-part series on evidence-based practice produced in partnership with the UCLA Prehospital Care Research Forum. Visit www.cpc.mednet.ucla.edu/pcrf.

The practice of medicine has come a long way over the past 150 years. For example, routine use of leeches to remove “bad blood” no longer occurs, and everyone involved in healthcare knows to wear personal protective equipment.

Changes to practice have (at times) been slow in coming. However, we simply can no longer routinely rely on providing care without evidence that it works. Investigators conduct rigorous studies to determine the efficacy of treatments. This philosophy of testing clinical practices using research methods to validate their efficacy and safety is known as *evidence-based medicine.*[1]

To gather the data needed when evaluating a treatment, researchers use a structured approach that utilizes critical thinking. EMS personnel and other clinicians engage in critical thinking as a method of problem-solving every day when deciding upon the best course of action to help a patient. These same skills are used in conducting research. The Center for Critical Thinking’s Linda Elder and Richard Paul wrote (in part) that when engaged in scientific study, we should examine the:

- Purpose of the inquiry;
- Best questions to ask;
- Types of inferences typically drawn;
- Viewpoints in the profession;
- Investigators’ assumptions;
- Implications of the inquiry;
- Types of data (information) to collect.[2]

These dimensions of critical thinking yield different types of data. Authors might differ slightly, but there are generally four accepted types (or levels) of data. These are (from lowest level of rigor to highest) *nominal, ordinal, interval* and *ratio.* Each of these levels of data allows increasingly complex and vigorous statistical testing.

These data are labels or categories that, in themselves, cannot indicate increased or decreased value. For example, two U.S. states have different names. Despite the love one might have for their home state, one of these labeled areas is not more of a state than the other. Nominal data are variables such as sex/gender, race, ethnicity, political affiliation and place of birth. For instance, an investigator might collect information on the numbers of males and females who work each type of shift schedule (e.g., 24-hour and 12-hour).

These labels for sex/gender are constructs; they are titles society has agreed to use. Neither sex (nor gender) is more valuable than the other. These data can only be used for lower-level comparisons. For example, after gathering these data it would be acceptable to report the numbers and distributions of males and females on each type of shift. The number of each sex/gender is an objective measure that can be used for comparisons of the groups.

This would allow for creation of graphics (like line charts) and performance of low-level statistical tests. A researcher could calculate a chi square to find differences between sex/genders. It would be necessary to code sex/gender as a dichotomous (i.e., 0 or 1) variable for this calculation. Keep in mind that this still does not indicate a greater value for one gender/sex over the other.[3–5]

Sometimes an investigator wants to know the order in which things occur. If a researcher were to stand outside of an emergency room and create a log of the order in which ambulances arrived, this list would contain ordinal data (*Table 1*). While it is true that these are more powerful than nominal data, they have a key weakness: If the researcher only focuses upon the order in which the ambulances arrive, they could know which was first to arrive, which was second and so on. They would not know the time needed for the ambulances to reach the emergency room—this would require the additional (higher-level) information included in the table.

While the table indicates the order of ambulance arrival, other data could be collected for better comparisons. You can also see that the en route times (times to drive to the hospital) are not the same for all the ambulances. The researcher can calculate a correlation between the en route times and the number of minutes on scene or the number of minutes out at the hospital (as long as they captured those data). They could also calculate the level of correlation between the order of the ambulances' arrival and the time the ambulances spent on scene or the number of hours the crew had been on duty.[3–5]

Interval

This is the first of the continuous data, meaning you're observing data that have equal distances between points of measurement (as you would see on a measuring tape). They also have the strengths (but not the weaknesses) of nominal and ordinal data.

A characteristic that distinguishes interval data from the highest level of data is that they do not have a (reasonable) zero point on their scale of measure.[4] For example, a typical written exam would allow scores to range from 0–100. If the exam were administered to a group of paramedic students and someone earned a zero because they missed all the questions, it is unlikely the student has absolutely no knowledge about being a paramedic. The score only indicates how the student performed on this exam. Further, the difference between a score of 50 for one student and a score of 100 for another student does not indicate that one has twice as much overall knowledge as the other.

Another example is body temperature. When a thermometer is used, the heat of the body is calculated on a scale that has consistent markings (or digital increments). The body might reach a temperature of zero on commonly used scales, but it is unlikely the patient would reach a temperature of absolute zero, at which there is no molecular movement.

Interval data allow for powerful calculations, including comparing one group with another and looking for significant differences, while controlling for variables that can affect your results, known as extraneous variables.[3–5] Most readers are familiar with research results that note a study controlled for specific things; this is the type of variable best used here. For example, an investigator might want to know whether length of shift (e.g., 12-hour vs. 24-hour) plays a role in the number of traffic accidents involving ambulances. It is possible that the types of ambulances involved also play some unknown role in this phenomenon. However, since length of shift is the focus, the researcher will control for ambulance type by including it at a specific point in the statistical calculation.

Ratio

These are similar to interval data. They are so similar that they can be mistaken for (and used in place of) each other. Ratio data are also measured on a scale that has consistent intervals. The key difference is that ratio data allow for the possibility of a true zero on the scale used to measure a variable. Two examples are speed and hours on duty.

Two vehicles traveling at 30 mph and 40 mph have the same separation between their speeds as two other vehicles traveling at 55 mph and 65 mph. A vehicle traveling at 100 mph is traveling twice as fast as one traveling at 50 mph. Both of these objects can be completely still. Similarly, the number of hours on duty can be expressed along a scale of consistent intervals and have a zero point.[3–5]

Each variable used in a study must be evaluated for its strength if you want to use it as grounds for making a change in clinical practice. Authors led by AHRQ’s David Atkins suggest we evaluate evidence collected to decide whether a new or different clinical approach is needed using the GRADE (*grading* of *recommendations assessment, development* and *evaluation*) system.[6,7]

Within this system, evidence is graded as high, moderate, low or very low.[8] High-quality evidence (e.g., a randomly controlled trial [RCT]) leads to a conclusion, and gathering more research would probably not influence the decision(s) being made. Moderate-quality evidence leads to a conclusion, but more research is likely to influence the decision(s) being made. With low-quality evidence, more research is very likely to lead to different outcome. If the evidence is very low quality, any decisions being made based upon it should be suspect.

An example of data that would be initially seen as high quality could come from an RCT. These data could be demoted to a weaker status if a review of the study finds problems such as weak internal validity (e.g., lack of randomization or blinding) or weak external validity (e.g., small sample size or a sample not representative of the population).

Strengths of the GRADE system include growing and general acceptance of the model and ease of use.[8] It does contain subjective elements that might be an issue (e.g., at what point does a researcher decide data from an RCT should be downgraded?). To lower the perceived strength of evidence requires both at least a general understanding of the research model being used and an understanding of the subject being investigated.

This can be seen in a research study on the effect of a new prehospital respiratory medication. If the study were carried out in a laboratory setting, testing a drug and a placebo with neither the patient nor the treatment administrator knowing which was used, this evidence would be valued as high. If the trial were carried out by asking a few paramedics to administer the new medication to some patients during a short period of time in the field, then comparing these to patients who received nothing, the evidence would be weaker.

EMS Example

When studying a clinical topic (such as hypertension or hemorrhage control), an investigator needs to use the highest level(s) of data available to apply rigorous and appropriate statistical tests. These tests reduce the chance that results are found by chance. An example would be an agency that wanted to know whether using 24-hour shifts vs. 12-hour shifts would affect patient outcomes. This is a broad question, and a primary step would be to narrow the focus. So here, it will be limited to: Does having two-person paramedic units responding to nonarrest cardiac calls during 24-hour shifts result in higher patient morbidity and or mortality?

A directional hypothesis for this example is: 12-hour crews will have statistically significantly higher average patient morbidity and mortality rates than 24-hour crews; and 12-hour crews will have statistically significantly worse patient condition outcomes at discharge than 24-hour crews (note: frame hypotheses as the opposite of what you think is true). Some of the data to be collected will be: shift length, crew demographics (e.g., age, experience, etc.), type of cardiac complaint, time of the call, en route (elapsed) time, patient condition at the emergency room, and patient final condition at discharge.

The data listed above include some nominal variables (e.g., shift schedule), some extraneous variables to be controlled for (e.g., crew demographics) and some interval data (e.g., en route time). Condition at the emergency room could be coded so that it is interval data: You would create a scale (e.g., from 1–10, with 1 being dead and 10 being asymptomatic). Each patient's medical record would be reviewed and placed into the appropriate category.

Coding for the scale based upon a predetermined group of symptoms and signs is appropriate because there is an *a priori* (deduced) argument that being dead is much worse than being asymptomatic and happy, and you could determine what would constitute the other levels based upon the patient's condition. Some might argue that data on condition and disposition are ratio-level data; that is a reasonable postulate. Here that is acceptable.

Moving forward in this study, all the nonarrest cardiac calls run over a predetermined time would be reviewed. An investigator would need two groups from which to collect data. These could be two (or more) sets of paramedics working at the same time (e.g., 12-hour and 24-hour shifts over 6 months) or one set working over a longer period (e.g., 12-hour shifts for 6 months and then 24-hour shifts for the next 6 months). The former allows you to better control for extraneous variables such as changes in seasons or pay or updates to protocols.

Reviewing data on the two groups of paramedics allows reporting of descriptive statistics. This would include numerical and graphical representations of the mean, median, mode and standard deviation of each of the nominal, ordinal, interval and ratio variables. Nominal and ordinal data are important but should not be used to make critical decisions. It is the higher levels of data that allow for the most powerful testing if the data are still high-value (using the GRADE system).

A researcher could use interval and ratio data to compare the averages for the patients' conditions upon arrival at the emergency room. These data allow one to determine whether there is a statistically significant difference between the groups (12-hour vs. 24-hour shifts). If there is a significant difference, an agency needs to consider the real-world impact that difference represents.

This decision-making will be aided by a combination of classic critical thinking and the use of GRADE. For example, an agency might find that after 400 cardiac calls (n=200 of the 24-hour shift and 200 of the 12-hour shift), software indicates there is a statistically significant difference between groups in patients' conditions at the emergency room. However, the average for one group could be 8 and the other 8.4 (on a scale of 1–10), or the data might be gathered from a study with low internal validity. This significant difference might not be large enough or trustworthy enough to disrupt the agency's practices.

Summary

Each level of evidence is useful. Nominal data allow for solid descriptive reporting. Ordinal data can allow for stronger tests (e.g., correlation) and be controlled for in complex testing. However, to perform the types of in-depth statistical tests needed before changing a clinical practice, an investigator should strive to gather the highest levels of data available. Interval and ratio data allow a researcher to compare groups with precision and confidence.

Therefore, as an investigator plans a research project, it is incumbent upon her or him to think about what types of questions are being asked, collect the appropriate level(s) of data and evaluate the strength of the specific variables being used. Changes in practice affect lives, and decisions related to how medicine is practiced require the strongest possible evidence.

References

1. McAlister FA, Straus SE, Sackett DL. Why we need large, simple studies of the clinical examination: the problem and a proposed solution. *Lancet,* 1999 Nov 13; 354(9,191): 1,721–4.

2. Elder L, Paul R. *The Thinker's Guide to Analytic Thinking: How to Take Thinking Apart and What to Look For When You Do.* Dillon Beach, CA: Foundation for Critical Thinking, 2007.

3. Forister JG, Blessing JD. *Introduction to Research and Medical Literature for Health Professionals,* 4th ed. Burlington, MA: Jones & Bartlett, 2016.

4. Gay LR, Mills GE, Airasian PW. *Educational Research: Competencies for Analysis and Applications,* 10th ed. Boston: PEAR, 2012.

5. Salkind NJ. *Statistics for People Who (Think They) Hate Statistics: Excel 2010 Edition,* 3rd ed. Thousand Oaks, CA: SAGE Publications, 2013.

6. Atkins D, Eccles M, Flottorp S, et al. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches, The GRADE Working Group. *BMC Health Serv Res,* 2004 Dec 22; 4(1): 38.

7. GRADE Working Group. GRADE, http://www.gradeworkinggroup.org/.

8. American Thoracic Society. The GRADE Approach (Part 2 of 12), https://www.youtube.com/watch?v=IjxZ_-HI8BE.

Sandy Hunter, PhD, NRP, is a professor with the paramedic program at Eastern Kentucky University and a graduate of the doctoral program in educational psychology at the University of Kentucky. He holds a master's in health education and an undergraduate degree in emergency medical care. His research interests include diversity, self-efficacy, learning theories and EMS safety.