The Journal of Foot & Ankle Surgery 57 (2018) 850–853

Contents lists available at ScienceDirect

The Journal of Foot & Ankle Surgery

journal homepage: www.jfas.org

Investigators' Corner

Seeing the Future: A Crystatistical Ball

Daniel C. Jupiter, PhD

Associate Professor, Department of Preventive Medicine and Community Health, University of Texas Medical Branch, Galveston, TX

ARTICLE INFO

Keywords:
population
predictive modeling
risk factor

ABSTRACT

We are used to identifying risk factors, or predisposing factors, for outcomes of interest, whether those outcomes are negative or positive, using the tools of statistical modeling. We are far less accustomed to making predictions of outcomes for specific individuals or testing the accuracy of those predictions. In this commentary, we consider this more difficult task and discuss, in this context, why statistics has such difficulty talking about individuals, even as it easily informs us about populations.

The overarching goal of any biomedical research project is the improvement of patient outcomes and, more broadly, human health. We are trying to discover better treatments—those that are likely to work faster, are more likely to resolve pathology, or are less likely to have side effects. We are trying to better manage patients and their diseases, to build healthcare systems that are more easily navigable and less likely to cause problems for patients or providers. We are trying to understand healthier lifestyles: identifying choices of diets that are less likely to lead to disease or identifying habits that are less likely to have negative effects. Ultimately, we aim to build a healthier society: one that is less likely to have unhealthy citizens and one in which patients and clinicians are likelier to find the healthcare system salutary and user-friendly.

The above is all obvious, yes. Perhaps we do not always so explicitly state our aims or detail them from the individual to the societal and systemic level, but the aims are clear to all of us. However, the way in which the aims are stated above emphasizes their statistical aspects. We are looking for things that are "likely." We are looking at the individual, in terms of their treatment, or their predisposition to particular outcomes. But we are also simultaneously looking at the population, in terms of its overarching trends and behaviors. (Far beyond the scope of this article, of course, is the fact that many of our political debates and decisions concern the desirable balance and tension between these 2 things: the individual and the common good.) The examination of likelihood, and of behaviors at individual and population levels, is exactly the province of statistics.

The topic of this commentary is the contrast between individual and population levels of analysis; it is here that we begin to tie together the threads in our series of 5 commentaries about regression models and model selection. We began by discussing the assumptions of statistical tests and problems that arise when we violate those assumptions (1). Next, we looked at the goals for our models: description, identification of risk factors, or prediction (2). In the last commentary, we examined tools that allow us to see whether our statistical assumptions are met (3). We now, in the current commentary, return to the topic of the second and more fully explore predictive models. Finally, in our last commentary, we will examine how statisticians make model choices: how does one decide what variables are to be included in a model, given the goals of the model? We begin the current commentary by talking about why prediction, which seems so naturally something with which statistics should concern itself, is actually a special and difficult task.

Risk Factors Versus Prediction

As of yet, we have been imprecise about what "prediction" means, beyond a colloquial notion of guessing about future events. This is the unconsciously performed act of prognostication that we do in reading clinical studies: for example, is using a plate or screws more likely to lead to prominent hardware or hardware removal? But since the emphasis in this commentary is on prediction, we should be much more rigorous in our definitions and clearly differentiate between *identification of risk factors* and *making predictions*.

What we are usually doing when we talk about prediction, as in the simple example in the previous paragraph, is better referred to as the identification of risk factors. These risk factors are the variables, or characteristics, that predispose to a particular outcome of interest. They are qualities that make one more likely to experience an outcome; for example, with use of screws, patients are more likely to require

Financial Disclosure: None reported.
Conflict of Interest: None reported.

Address correspondence to: Daniel Jupiter, PhD, University of Texas Medical Branch, Department of Preventive Medicine and Community Health, 301 University Boulevard, 1.134G Ewing Hall, Galveston, TX 77555.
E-mail address: dajupite@utmb.edu (D.C. Jupiter).

1067-2516/$ - see front matter © 2016 by the The American College of Foot and Ankle Surgeons. All rights reserved.
https://doi.org/10.1053/j.jfas.2018.05.001

hardware removal. We are most often interested in *independent risk factors* (4). These are the risk factors that have an effect when we hold all other variables constant. This type of thinking, we recall, is the goal of multivariate regression: to understand how 1 variable affects the outcome of interest, imagining that we are holding all other variables fixed. To give a concrete example, we expand on our screws versus plate example. In such a study, we would likely control for potential confounding factors, such as body mass index (BMI) or sex, in our multivariate analyses. We would conclude, then, from our multivariate analyses that independent of sex and BMI, screws predispose to hardware removal. This means that our result holds regardless of sex: for any 2 men, the man with screws is more likely to have hardware removal than the man with the plate. The same is true for any 2 women, or for any 2 people with the same BMI.

We are being purposefully pedantic in our description of risk factors, as we are attempting to highlight 1 of their salient characteristics. This characteristic is that risk factors speak about populations. Although a risk factor tells us how the likelihood of an outcome changes within an individual, with a change in that individual's having or not having the risk factor, the effect or magnitude of that risk factor is, in some technical statistical sense, averaged across the entire population. This is what independence means.

In contrast, however, when we speak of *prediction* proper, we do not want to do this type of averaging. Rather than averaging the effect of a risk factor across the possible values of the other variables, rather than thinking about independence of variables, we want to account for the values of the other variables. In other words, we want to explicitly understand the absolute risk for each given combination of values of risk factors. We want to know how a man with a given BMI proceeds through his care. We are not interested in how changing his BMI would change his course, keeping his sex fixed; we are interested in this particular individual and his particular progress.

Another way to say this is: given that you have a particular combination of factors, what is the likelihood that you have the outcome of interest? This is a very different question than whether, at a population level, having a particular risk factor predisposes an individual to a given outcome.

Interestingly, the tools to examine risk factors and predictions are essentially the same; but how we use them, how we understand our results, and how we view populations, are different.

The Power of Statistics, the Weakness of Statistics

These 2 views of prediction, broadly viewed, highlight an important duality in statistics that it is worth taking some time to explore. (And there may be no other appropriate place to do so in this series.) The strength of statistics is easily seen, even more so than in biomedical studies, in the fact that statisticians are usually able to make remarkably accurate predictions about presidential elections in the United States (both popular and electoral college votes) based on samples of only a few thousand potential voters. That is a generalization to 100 million voters from a tiny percentage of that population. The use of statistical tools, then, allows us to infer from tiny samples to much larger populations. And it allows us to do so with great precision and even with a notion of how uncertain the inferences are. (This uncertainty is quantified in things like confidence intervals.)

However, those inferences are only about average behavior. We can speak about populations. We can speak about gestalt trends. But we can say nothing about individuals and little about subpopulations. We would be hard-pressed to use the same sample on which we based our national-level election prediction to predict, say, how women of a given age group within a given ethnic group will vote. We are stuck with looking at overall populations. This is exactly what we observe with biomedical studies. Above, we noted that in our population we can predict that, on average, screws lead to more frequent hardware removal, and that, on average in the whole population, male sex predisposes to more frequent hardware removal. Again, we said nothing about individuals or subpopulations.

Thus, the great strength of statistics is, in some sense, its great weakness. In looking broadly, and aiming for generality, we lose sight of the particular. In a way this is obvious, or at least suggested by all those assumptions we discussed in our earlier commentary (1). Why indeed do we care about the normal curve, with its nicely describable central tendency? Why are we so often focused on long-term behaviors of averages in large populations when we think about statistical theorems such as the central limit theorem? Because in large populations, average behaviors predominate, and outlier behaviors vanish into the background. But who exhibits outlier behaviors? Individuals. They vanish into the background when we study populations.

To make this contrast even clearer, let us delve deeper into our screws versus plate example. We probably should consider sex in our study, as sex affects outcomes, and we are interested in the potential need for different treatment in men and women. Sex gives us 2 subpopulations. We should also consider diabetes status, because diabetes may interfere with bone healing, and that may, in turn, predispose to further complication. Diabetes status gives us 2 subpopulations. Put together, that's 4 subpopulations. If we add in another variable, say ethnicity, with 3 or 4 possible values, we now have 12 or 16 subpopulations. If instead we examine BMI, we should have a reasonably sized subpopulation at each BMI level. We see here how quickly the number of subpopulations increases with the number of risk factors or predictors of interest. To make inferences about the population as a whole, we need reasonable numbers of subjects within each subpopulation, say 3 or 4 with each treatment type; to make inferences about each subpopulation on its own, we need even larger numbers of subjects within each subpopulation, say 15 or 20.

As the subpopulations become more specific, the size of the overall sample must increase to ensure good inference about each subpopulation. Putting this another way, to study an individual's behavior, you need a population of that individual. It does not suffice to simply look at the entire population containing 1 instance of that individual.

What Is a Population?

Although it is not entirely germane to our overarching discussion, it is worthwhile to digress slightly to talk about populations and sampling frames. We are happy to generalize from our sample of several thousand voters to the population of all voters in the U.S. presidential election. However, we should be very careful in our polling not to survey foreign nationals living in the United States or those not eligible to vote. In other words, to infer to a particular population, our sample must be entirely from that specific population. We must also carefully choose a representative sample of voters, not exclusively polling only 1 sex, 1 ethnic group, or 1 socioeconomic stratum. This is all in service of making valid inferences about a predefined population of interest.

In a biomedical study, the problem is somewhat reversed. We are usually constrained in the samples we can look at it, especially in observational studies. For example, we can only examine screw and plate fixation in those who have already had such fixation. We cannot examine these treatments in those who might possibly have them at some point in the future. Nor can we ensure that we have sampled in a representative fashion, manifesting in the fact that there is no way to directly account, in our sampling, for things such as selection bias. Even in a randomized trial, we are limited to who presents to our clinic or hospital and who consents to participate in the trial. We are also limited or affected by the catchment area, and