Medicine, statistics, and education: the inextricable link

Katharine Krevans Brieger
Centre Médical Universitaire
1, rue Michel-Servet
1211 Geneva 4, Switzerland
katharine.brieger@unige.ch

Johanna Hardin, PhD
610 N. College Ave.
Pomona College
Claremont, CA 91711
jo.hardin@pomona.edu

INTRODUCTION

Statistics is an integral part of any clinical trial and it is widely recognized that the training of doctors should aim to improve statistical literacy. However, the call for statistical literacy has focused on introductory material and we argue that there is a need for *advanced* statistical literacy in the health sciences (Table 1). In their 2005 article "Statistical Methods in the Journal," Horton and Switzer reviewed all *New England Journal of Medicine* articles from January 2004 – June 2005; they found that more than half of the articles used survival analysis, multiple regression or another advanced technique. A 2007 study appearing in the *Journal of the American Medical Association* looked at residents' understanding of statistics; authors Windish, Huot and Green showed that only 10.5% knew how to interpret the results of a Kaplan-Meier curve and only 37.4% could interpret an adjusted odds ratio from a multiple regression. We provide examples of statistical misinterpretations that have had tangible public health consequences. The need for extended statistical education for our health scientists is greater than previously imagined.

Table 1

Then (1978-1979)	Now (2005)	Soon to come
• t-tests (44%) • contingency tables (27%) • Pearson correlation (12%) • nonparametric tests (11%) • survival methods (11%)	• survival methods (61%) • contingency tables (53%) • multiple regression (51%) • power analysis (39%) • epidemiologic statistics (35%)	• multiple comparisons • analysis of high throughput data • advanced experimental design • Bayesian analysis • Permutation / Bootstrap methods

Top 5 most commonly used statistical methods in *The New England Journal of Medicine* articles across time (first two columns from Switzer and Horton "What your Doctor Should Know about Statistics (but Perhaps Doesn't)," *Chance*, 2007). Percent of papers using the methodology is given in parenthesis. Last column is a prediction of methods which are quickly growing in popularity. Statistical methods used to analyze medical studies are rapidly increasing in sophistication.

Recently, the "dangers" of the vaccine for measles, mumps, and rubella (MMR) made headlines in the United Kingdom, warning parents about the vaccination's associated risk for autism. The alarm was based on an article appearing in *The Lancet*, Wakefield's 1998 study on 12 children who had developmental regression. Given high rates of both MMR vaccination and autism, finding 12 children who had received the MMR vaccine and also had autism is unremarkable. Additionally, the symptoms of autism are often first noticed at the same age as vaccination, so the relationship may have been coincidental, not causal.

In 2004, *The Lancet* retracted Wakefield's 1998 article, but the effect of Wakefield's paper has been particularly damaging to communities where measles, mumps, and rubella had previously been all but eradicated. In the United Kingdom in 2003 – 2004, the MMR vaccination rate reached an all-time low of 80%, much below the herd immunity rate of 95%. In 2008, measles was declared endemic to England and Wales for the first time since 1984.

This example illustrates how misinterpretation of statistical results can have serious public health consequences. Basic knowledge of observational study limitations, variability, and small sample sizes should have prevented the unwarranted conclusion that the MMR vaccine causes autism. However, we would take this example one step further and claim that additional statistical knowledge, on multiplicities for example, could have further reduced the chances of the misinterpretation of the Wakefield paper.

MODERN STATISTICAL CHALLENGES

MULTIPLICITES

We define multiplicities to mean multiple tests of hypotheses on the same data – many variables, multiple endpoints to a study, multiple time points, comparing several treatments, interim analyses, considering an endpoint as continuous or categorical. The prevalence and impact of multiplicities is often underestimated, and there is evidence for a large proportion of false findings in the current literature.

In a controversial 2011 study, Bem tested college students for precognition and premonition abilities, types of extrasensory perception (ESP). Although he obtained highly significant results, some point to Bem's research as an example of the problems with multiple comparisons. Although the majority of Bem's ESP experiments were significant, they represent only a small fraction of all ESP experiments done, most which are not published. Interestingly, the degree of association in the ESP research is much stronger than that in the MMR case; however, the acceptance of the research itself is reversed. The contrast in the reactions to the significance of these studies indicates a need for understanding of the process used to arrive at the results. The

experimental design, statistical analysis, and prior assumptions related to having a true relationship are all components of how likely a research statement is to be true.

Sidebar

In a provocative 2005 article (with a provocative title, "Why most published research findings are false"), Ioannidis demonstrates that under certain typical conditions, most published research findings are false. In order to understand the findings, we first consider the set-up of a typical experiment in a particular field of science. Of course there is not any static field of science, but we imagine a given set of all hypothesis tests done with that field. Using the notation from Ioannidis, let *c* be the total number of tests in the field, and let R represent the ratio of the number of "true relationships" to "null relationships" among those tested in the field. Note that R is typically unknowable, and we differentiate between "true" relationships and "significant" relationships; R considers the former. Additionally, assume the standard notation for the probability of making a type I error (rejecting a null relationship, α) and the probability of making a type II error (not rejecting a true relationship, β).

1. Using the identity: # true relationships = R*(# null relationships), we can calculate the proportion of true relationships.

$$proportion\ of\ all\ tests\ which\ are\ true\ relationships = R/(R+1)$$

2. The number of true and null relationships is simply the appropriate proportions of the total, *c*, number of tests.

$$\#\ of\ true\ relationships = c\frac{R}{R+1}$$

$$\#\ of\ null\ relationships = c\frac{1}{R+1}$$

3. Using type I and type II error rates, we can find the number of significant relationships out of those that are true and the number of significant relationships out of those that are null.

$$\#\ of\ relationships\ that\ are\ significant\ and\ true = c\frac{R}{R+1}\ (1-\beta)$$

$$\#\ of\ relationships\ that\ are\ significant\ and\ null = c\frac{1}{R+1}\ \alpha$$

4. All significant relationships are given by the sum of those from the true and null groups.

$$\#\ of\ significant\ relationships = \frac{c}{R+1}\ (\ R(1-\beta) +\ \alpha)$$

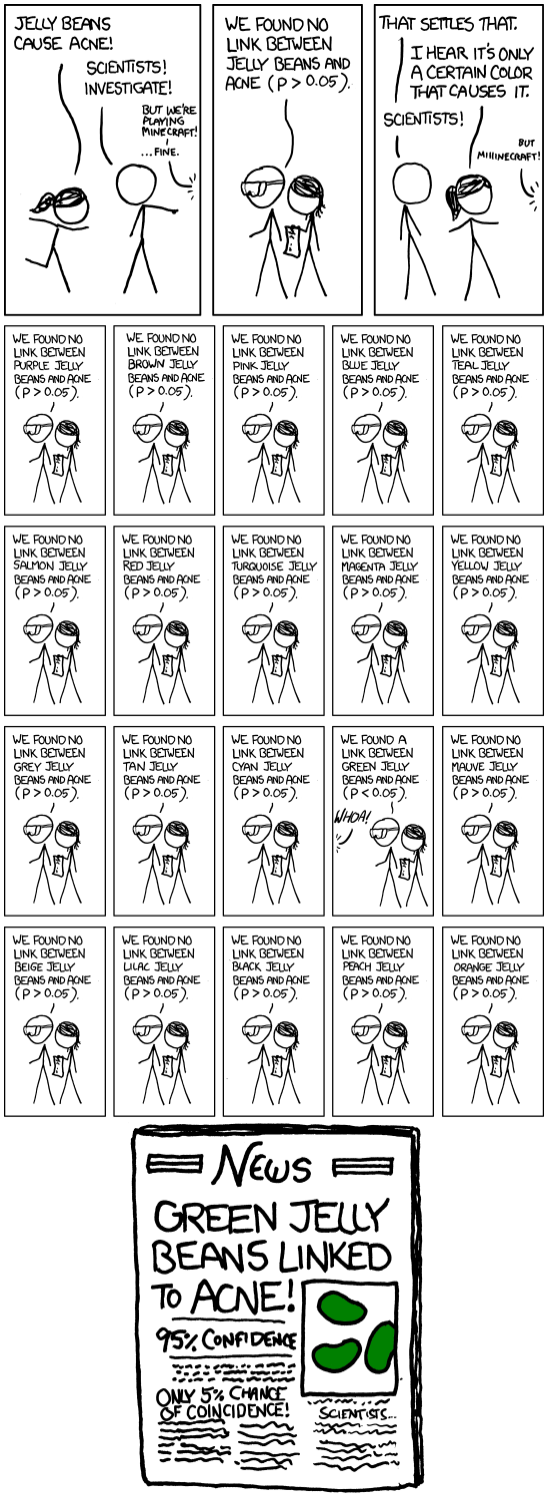
5. Finally, the positive predicted value (PPV) is calculated as the proportion of true studies out of those which are reported to be significant.

$$PPV = \#\ of\ true\ studies\ |\ significance = \frac{R(1-\beta)}{R(1-\beta) +\ \alpha}$$

The positive predicted value (PPV) gives a measure of what is desired in science: the probability of a true relationship given a significant result. Note that PPV > 0.5 (a significant result is more likely true than false) only if (1-β)R > α.

Using the results from Ioannidis in the sidebar, we can assess the MMR and ESP examples. Keep in mind that, ideally, we would be able to evaluate our work using the probability of a true relationship given a significant result, as opposed to the standard hypothesis testing structure

where we constrain the probability of a significant result given a null relationship. The positive predictive value (PPV) measures the proportion of true relationships out of those found significant, we should hope for a large PPV for all of our work.



It is worth considering the values of R (the ratio of true relationships to null relationships of all those tested in the field) and β (type II error rate) in each of the studies. One might have a prior belief that with ESP, R is particularly low. Additionally, with the small sample size in the MMR study, β would be particularly high. Low R and high β give rise to a small PPV and thus the results of such studies should be considered carefully. Ioannidis extends his analysis to include situations with bias and multiple independent research teams working on the same problem. A standard type of bias is to use some form of multiplicity – variable selection, change in endpoint, etc. Each statistical analysis that is pushed toward significance via some form of multiplicities induces a lower PPV.

The issues of multiplicities can be ameliorated with validation studies and a more thorough understanding of variability across the system at hand. Purely chance findings are often published and mistakenly considered important simply because positive results are more likely to be published. We need to use our statistical knowledge not only to assess the data, but also to evaluate the merit of published results. Medical education would greatly benefit from a second course in statistics that better prepared doctors to interpret results for themselves.

LARGE DATA SETS

High-throughput data, as from microarray, proteomic, and next generation sequencing, have become ubiquitous and invaluable in medical research. However, unlike many statistical techniques applied in the medical literature, methods used to analyze high-throughput data are sophisticated and not taught in standard statistical curricula. Additionally, high-throughput data are prone to measurement and preprocessing errors.

We present the 2009 work of Baggerly and Coombes, published in *Annals of Applied Statistics*, in what they term "forensic bioinformatics." In 2007, Duke University initiated a series of clinical trials based on analyses from microarray studies which modeled genetic characteristics predictive of sensitivity to certain drugs. The initial work by Potti et al., published in *Nature Medicine* in 2006, was retracted after patients were enrolled in three different clinical trials that were subsequently suspended in June of 2010.

The underlying problem with high-throughput data is lack of intuition about how certain markers or tests should behave, particularly diagnostic procedures based on tens or hundreds of genetic markers. It is difficult to make relevant predictions based on data that is so large that it cannot be

conceptualized. Because analysis of high-throughput data is typically exploratory, simple mistakes are easy to make but difficult to discover. In the Duke research, simple mistakes led to the complete invalidity of their results. The experimental conditions were confounded with the date of the experiments making any differences in conditions unknowable. A gene label mix-up created a model which included genes *that had not been measured* to predict sensitivity to chemotherapy agents. The outcome labels were switched on the microarrays for a group of patients, thus creating a model predicting exactly the *opposite* treatment than that which should have been applied.

The difficulties with large and high-throughput datasets are not all simple and technical. In order to quantify genetic activity from microarray data, image processing techniques are applied to a scan of a biological sample labeled with florescent dye and hybridized to a chip. Additional algorithms adjust the numerical values to account for systematic changes – e.g., dye biases, edge effects, heteroskedasticity. The choice of image processing or pre-processing algorithms and parameters necessarily affect the numerical value outcomes that represent genetic activity. The statistical training required to understand microarray analyses is much more thorough than a conventional introductory curriculum.

EXPERIMENTAL DESIGN

Not every clinical trial is straightforward with 1:1 randomization between two treatments; researchers use historical controls, inverse sampling, and adaptive randomization. Though there are benefits to each design, trials using advanced experimental designs are difficult to analyze. From the outset, it is imperative to state the explicit design structure in order to accurately

control the type I error rate. Additionally, later statistical analyses must take into account any unusual data structures.

One alternative to 1:1 randomization is to use adaptive randomization, randomizing more patients to the treatment arm which shows the best prognosis. In the 1980s, a series of trials was conducted to ascertain whether extracorporeal membrane oxygenation (ECMO) was effective for treatment of persistent pulmonary hypertension of the newborn (PPHN). At the time, there was mounting evidence that ECMO was superior to conventional therapy, and there were ethical concerns over implementing a standard randomized design in order to demonstrate ECMO's efficacy. Bartlett et al. conducted an initial adaptive design experiment in 1985; only one patient was given conventional treatment, and that patient died. Though statistically significant, the small sample size was unconvincing to the medical community, warranting further study. In 1989, Ware then designed a two-stage clinical trial where treatments were selected using permuted-block randomization. The study eventually enrolled 30 patients with the ECMO treatment, 25 of whom survived; all nine patients enrolled on the conventional treatment died. According to Ware's *Statistical Science* piece "Investigating therapies of potentially great benefit: ECMO," as compared to standard 1:1 randomization, the creative study design subjected fewer patients to the less efficacious conventional treatment.

The decision to implement an adaptive design is not always straightforward. Often the motivation for an adaptive design is to create a setting where fewer patients are subjected to the treatment arm which is substantially worse. In 2011, though, Korn and Freidlin demonstrate through simulations that the sample sizes needed to power an adaptive study lead to more

patients on the "worse" arm of treatment than standard randomization would, although they are a smaller proportion of total subjects. However, in their article "Outcome-Adaptive Randomization: Is It Useful?" Korn and Freidlin neglect to mention the patients who are not enrolled in the study but do have the disease. Those patients must also be treated, so the effective number of patients given the "worse" treatment is likely still less in the adaptive setting.

Though adaptive designs may be well-suited for situations with large differences (as with ECMO) or multi-arm trials, they also bring with them a host of logistical complications that can undermine the actual adaptation. Collecting data in a timely manner from multiple sites is not trivial, and we should be wary to think that the theoretical justification trumps other concerns in designing studies that will communicate maximal information to the medical community.

CONCLUSION

In 1937, an article in *The Lancet* criticized physicians' "blind spot" in laboratory and clinical medicine to be simple statistical methods and a decade later, in 1948, the British Medical Association recommended that statistics be included in medical education. Yet, for example, it was not until 1975 that statistics became mandatory at the University of London School of Medicine. In 2009, the Association of American Medical Colleges listed among its recommended competencies for medical school graduates the ability to "apply quantitative knowledge and reasoning – including integration of data, modeling, computation, and analysis – and informatics tools to diagnostic and therapeutic clinical decision making."

Increasingly, many of the important statistical considerations in medical studies are far beyond the knowledge gained in introductory statistics; we must remember that just as medicine is a

dynamic and fast-moving field, so is statistics. As we keep abreast in both fields simultaneously, we enhance our ability to expand upon knowledge of human health.

Further Reading

Altman, DG, and JM Bland. 1991. Improving doctors' understanding of statistics. *J R Statist Soc A* 154: 223-267.

Berry, Donald A. 2011. Adaptive Clinical Trials: The Promise and the Caution. *Journal of Clinical Oncology* 29(6): 606-609.

Mathematics and Medicine. 1937. Editorial, *The Lancet* 31(1).