Journal of Mathematics and System Science 8 (2018) 119-139
doi: 10.17265/2159-5291/2018.05.001

Inverted Simulations Demonstrating Strong Ecological Fallacies in Cohort Studies

Shankar Srinivasan
Resource Tepee LLC, 1 Lane Road, Hillsborough, NJ 08844, USA

Abstract: We start with a description of the statistical inferential framework and the duality between observed data and the true state of nature that underlies it. We demonstrate here that the usual testing of dueling hypotheses and the acceptance of one and the rejection of the other is a framework which can often be faulty when such inferences are applied to individual subjects. This follows from noting that the statistical inferential framework is predominantly based on conclusions drawn for aggregates and noting that what is true in the aggregate frequently does not hold for individuals, an ecological fallacy. Such a fallacy is usually seen as problematic when each data record represents aggregate statistics for counties or districts and not data for individuals. Here we demonstrate strong ecological fallacies even when using subject data. Inverted simulations, of trials rightly sized to detect meaningful differences, yielding a statistically significant p-value of 0.000001 (1 in a million) and associated with clinically meaningful differences between a hypothetical new therapy and a standard therapy, had a proportion of instances of subjects with standard therapy effect better than new therapy effects close to 30%. A "winner take all" choice between two hypotheses may not be supported by statistically significant differences based on stochastic data. We also argue the incorrectness across many individuals of other summaries such as correlations, density estimates, standard deviations and predictions based on machine learning models. Despite artifacts we support the use of prospective clinical trials and careful unbiased model building as necessary first steps. In health care, high touch personalized care based on patient level data will remain relevant even as we adopt more high tech data-intensive personalized therapeutic strategies based on aggregates.

Key words: Ecological fallacies, p-values, cohort studies, case-control studies, inverted simulation, hypothesis testing, aggregate statistics, publication bias, correlation, machine learning, personalized care and therapy.

1. Introduction

Much of statistics is built on a duality between a true state of nature and experiences deriving from it. Such dualities are not uncommon in the sciences. We are always presuming that something meaningful is behind what we experience, observe and measure. Statisticians, refer to this true state of nature as a parameter, typically unknown. Statisticians will go to great lengths distinguishing the average and the mean, the computed standard deviation from an underlying true standard deviation (usually denoted as sigma) and a proportion from a probability. The latter in each of these three tuples is the true state of nature and the former is a computed measure which attempts to get at the true state of nature. We have hypotheses, two, many or from a continuum, about the nature of this truth. These can be invariant truths as in a frequentist statistical framework, or varying truths, having subjective probabilities, when we adopt the Bayesian framework. Data is deemed to devolve from these true states of nature. If we were looking at data on the trajectory over time of a free-falling apple on planet earth, this data derives from and supports Newtonian hypotheses about gravity as the true state behind this experience. The link here between the true state of nature and our trajectory data is deterministic. We come in as statisticians, when this link between what may be true and what we observe is probabilistic.

2. The Statistical Hypothesis Testing Framework

In the classical frequentist framework, we would start with a set of dueling hypotheses. For a new therapy versus a standard therapy for cancer, one

Corresponding author: Shankar Srinivasan, Ph.D., research fields: statistics, observational studies, clinical trials.

would start with the hypotheses that there is no difference between the standard therapy and the new therapy and pit this against the hypothesis of better survival outcome with the new therapy. The former is called the null hypothesis and reflects prevalent opinion, while the latter is called an alternate hypothesis and is something we hope to establish [1]. This framework is consistent with an early philosophy of science framework called falsification [2]—any explanation continues to be relevant till it is rejected by another which is better supported by the data. We spoke earlier about data on a falling object supporting a Newtonian world view as an example of the fact/explanation dichotomy in science and statistics. Let's look at how we moved from that, in order to explain falsification. During World War I, a team of European scientists took a long arduous trip fraught with risk to South America to record data during a solar eclipse. Their goal was to see if light was deflected by gravitational fields, in line with Einstein's theory of relativity or the more strongly supported Newtonian hypotheses of the time. Their data did support the theory of relativity. If the falsification framework is a right perspective, then we can note, that to date, we have been unable to reject Einstein's theory of relativity in favor of another such elegant alternate theory of all things—an alternate explanation as simple as Einstein's explanations using analogies of objects falling from a window of a train, relative to observers on the moving train and relative to those outside. See Einstein's explanations written for the lay person and for the story about the confirmation of his theory in Ref. [3].

There are various data schemas we can use to ascertain the merit of our hypotheses, tested using stochastic data, ranging from retrospective real-world data to controlled prospective designed studies. One may conduct a clinical study, with a sufficiently large number of patients randomly assigned to the two therapies, to assess which of the two hypotheses above are supported. There is usually some quibbling about how conclusions of such studies are expressed. Statisticians are usually taught to express one of two conclusions supported by the data. If the data indicate a lack of a difference between therapies we would say "We were unable to reject the null hypothesis of no difference between therapies" with the odd double negative suggesting that current available evidence does not support differences but perhaps we might with more evidence. If the data support a difference we would say "We reject the null hypothesis of no difference in favor of the alternate hypothesis that the new therapy is superior to standard therapy."

3. The Decision Theoretic Framework

A popular analogy used to explain how we statisticians choose between two hypotheses is based on the criminal justice system. There, we can err when we hold someone guilty when innocent and when we acquit someone when guilty. We are fine when we find someone guilty or innocent when they are. We would like to reduce the error rates. Analogously, in statistical inference, we want to control the error rate of concluding that the alternate hypothesis of effect is true when the null hypothesis of no difference is a better characterization. This is usually called a type I error rate. The error of concluding in favor of no difference when there is one called a type II error. We usually rule in favor of the alternate hypotheses (difference in therapies) when the possibility of being wrong (the type I error rate) is less than 5%. One would then call the result statistically significant. It is necessary to note that this conclusion, very much unlike the legal analogy, pertains to aggregates computed over individuals rather than to each individual. We conclude differences between therapies on aggregate characteristics such as medians and means which may not hold as convincingly, between two individuals with the differing interventions. The fallacy of presuming individual effect based on an effect in aggregates has been criticized for a while by epidemiologists as an "ecological fallacy" [4]. This had led to a movement away from analyses in epidemiology where the data records were aggregates over large units such as counties or other sub-divisions such as median income, disease rates, racial or ethnic composition etc., towards case-control and cohort studies which have data on individual subjects. Here we will be demonstrating that these studies based on individual records (with simulations of Cohort studies) are infected with this fallacy as well.

The general decision theoretic approach described above is not very different using Bayesian approaches applied to clinical trial data. The Bayesian approach has, over time, in the clinical trial context, been forced into a frequentist mold through the use of non-informative prior information (which is the same as not using any prior information) and adaptations of the frequentist decision theoretic framework. The general Bayesian framework may help by holding on to multiple hypotheses, with high or low probabilities, a “this and this” framework rather than a “this or this” framework. Frequently however, a hypothesis with a high Bayesian posterior probability (a revised probability of the hypothesis given observed data) is likely to lead to choices favoring it in just the same manner as those favoring a hypothesis retained by frequentist analysis. As with the frequentist approach we adopt the observation/parameter framework and conclusions tend to be based on aggregates and represent a first step to help in approaches customized to the individual. For instance, Thall et al. [5] support the use of a probability of a true proportion responding to therapy (an aggregate random parameter) in one group exceeding that in another to aid in the choice between therapies. It is important to understand inferences in the aggregate, drawn from statistical analyses, and see why these may not always hold for individuals. Let’s look at a major innovation in statistical theory, used often in frequentist approaches, which drove statisticians into inferences about aggregates—unlike that legal framework discussed earlier which looks at individuals.

4. The Central Limit Theorem

This major result is often called the law of large numbers and features in most inferential analyses. It states that even when the distribution of data on individuals is erratic and non-standard, the distribution of aggregate statistics has a tractable form (usually the symmetric bell-shaped distribution or a related distribution) allowing us to read off probabilities. For instance, we might have a skewed distribution for the reduction in diastolic BP (blood pressure) under therapy for individual patients due to resistance to therapy. This might lead to a distribution with more likelihood of lower values to the left of the peak, reflecting some likelihood of a lack of response, rather than to the right. However, if we looked at the distribution of the average BP reduction of a sufficiently large number of patients, it would tend to have the symmetrical normal distribution. Let’s look at the histograms in Fig. 1. An interactive version of the figure allowing user input is at the webpage in Ref. [6].

Note that this diastolic BP example here and in the rest of the manuscript is pedagogical and illustrative of the points made and likely not reflective of current recommendations for hypertension. In the figure, we have two groups randomized to two therapies capable of reducing blood pressure. We look at the reduction in BP in mmHg for the two groups using the two histograms. The distributions of the individual BP changes are the skewed wide distribution mentioned earlier—it is simulated data and you may not always see the skew in a given simulation at the webpage noted above. The distribution of the average is much skinnier and always peaked and symmetric as shown. The measure of the spread of the distribution of the average, the standard error, is lower than the corresponding measure, the standard deviation, for the parent distribution, by a factor given by the square root of the sample size. Statistical p-values and inferences are drawn based on the separation of the tighter known distributions of the average rather than the wider intractable parent distributions in the histograms. Bayesian formulations in clinical trials, as noted earlier, also rely on a skinny distribution of the aggregate and draw conclusions about the aggregate. Bayesian approaches would look at the distribution of the mean (a random parameter), while the frequentist approach would consider the mean invariant and look at the distribution of the sample average—both are aggregates.

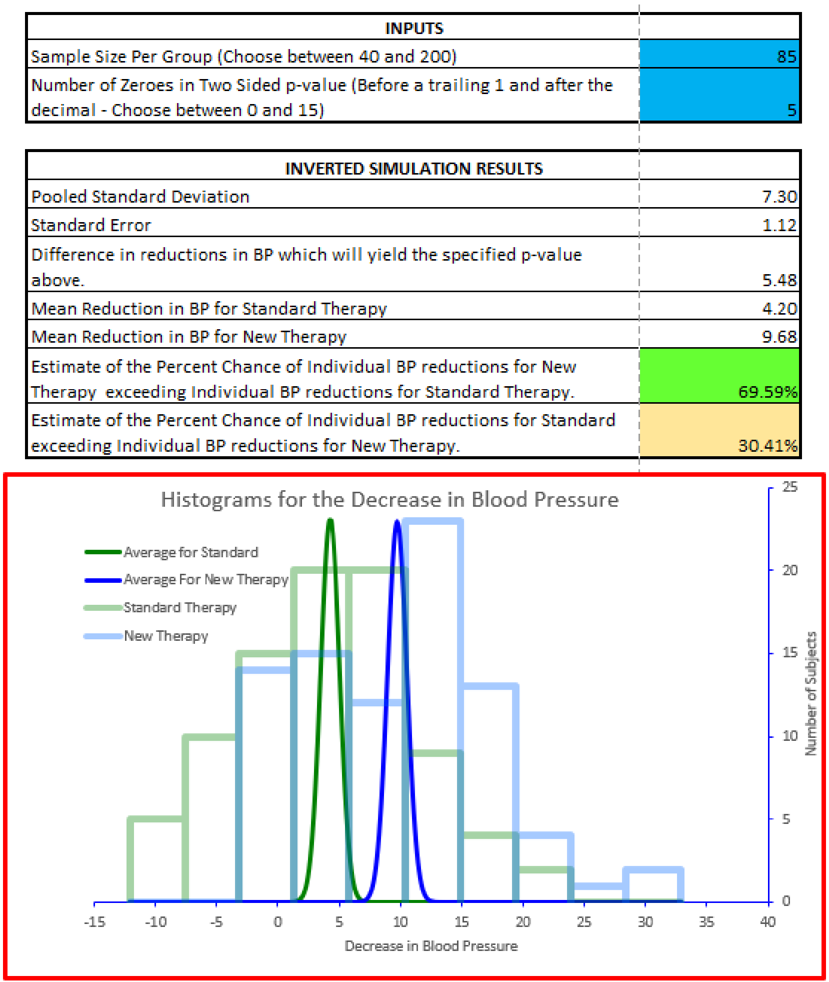
Fig. 1 provides the separation between the distributions, both in the aggregate and in the individual, when you see 5 zeroes in the p-value. This would usually be interpreted as a one in a million chance that the null hypothesis of no effect is supported as opposed to the alternate that the new therapy is superior to the standard therapy. Any statistician would label this "significant" synonymous with something "noteworthy", if you look it up in a dictionary. Further, a clinician would look at the differences in the average reductions in BP across the two therapy groups of much more than half the standard deviation and deem it clinically significant. Half a standard deviation is often used as a threshold to gauge if differences are meaningful [7]. The histograms however indicate considerable overlap in individual BP reductions. The distribution of the average we talk about in this section has no real existence unless we repeat the study 100 times or so, and we usually do it only once. This notional distribution of the average allows us to assess differences in the aggregate and we clearly see fallacies when using these conclusion at the patient level. In our simulation we computed a proportion of the reductions in BP for someone in our "standard" therapy arm exceeding that for someone receiving the "new" therapy at about 30%. The reduction of BP for individuals on new therapy exceed that for standard, despite the highly significant finding, at a rate of about 70%. Note that histograms, a somewhat crude presentation of something called a density estimate, is rarely presented even though considerable research has been conducted in this area. See, for instance, Terrell and Scott [8]. Lo, Mack and Wang [9] look at density estimation in the context of survival data with censoring.

5. The Inverted Simulation and Calculator

We look at three inverted simulations of two-arm clinical trials. The first in Fig. 1 is for continuous data such as the reduction in BP. The second in Fig. 2 has survival data looking at the time to an event such as a death or disease progression. The third in Fig. 3 has binomial data such as the achievement of a response threshold on therapy. The interactive versions of these figures at the web page noted earlier allow changes to the sample size per group (standard therapy or new therapy) and the number of zeroes in the p-value associated with the difference between the two therapies. The default sample sizes in the three interactive graphics of 85, 176 and 230 allow no more than the usual 5% two-sided type I error and a 10% type II errors (errors defined earlier). For details on calculating sample size in these contexts see Desu [10] and Cheng [11].

For the continuous calculator we can detect a meaningful difference of 3.5 (half a standard deviation is often considered meaningful as noted earlier) between a reduction of 5 for standard therapy (something like a diuretic) and a reduction of 8.5 for a new test therapy (something like a diuretic/beta-blocker combination) for a SD of 7 for BP reductions over time. For the other two default sample sizes we use a hazard ratio of 0.7 and a difference in proportion responding of 15%. Further details on the default sample sizes and the inverted simulation are in Appendix 1. Studies with sample sizes larger than these would be called over-powered and would tend to detect trivial differences smaller than those considered meaningful. Smaller studies would be under-powered and would tend to rule a new therapy ineffective even when it has that minimal amount of effectiveness in the aggregate. The default number of zeroes in the p-value is 5—only a 1 in a million-chance supporting the hypothesis of no difference between therapy groups. As noted earlier these default values still leave considerable overlap on the individual data histograms across groups and about an estimated 30% of the standard therapy reductions in BP exceeding that for new therapy—despite that extreme p-value and a difference in reductions between groups of about 50% larger than the clinically meaningful difference of 3.5.

When you increase the sample size for any fixed number of zeroes in the p-value, the skinny notional distributions of the average approach each other and get skinnier. The estimate of the percent chance of individual BP reductions for standard therapy exceeding individual BP reductions for new therapy gets in the 35% to 40% range. So “bigger data” help discriminate smaller differences in the aggregate but do not tell us any more about individual differences—it is likely that big data based “significant” conclusions for stochastic data are no more predictive and often less predictive for the individual. Meta-analyses, combining data from multiple studies, are often considered even better than the constituent complete set of blinded randomized studies testing a hypothesis. They can resolve conclusions about the aggregate when some studies in the mix are considered neutral and some are negative. This however, has the same “bigger data” issue and does not help us any more in evaluating effect in individual subjects. In contrast, if we try the interactive calculator and reduce the sample size below the default value for any fixed number of zeroes in the p-value, you will see that “small” data may actually be more useful.



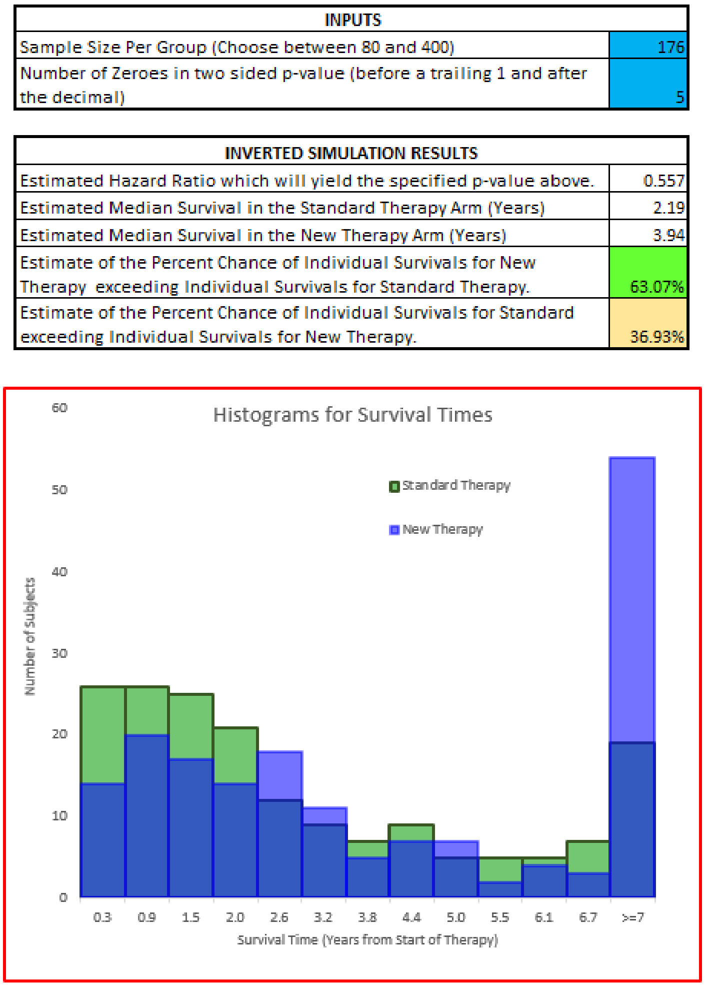
INPUTS	
Sample Size Per Group (Choose between 40 and 200)	85
Number of Zeroes in Two Sided p-value (Before a trailing 1 and after the decimal - Choose between 0 and 15)	5

INVERTED SIMULATION RESULTS	
Pooled Standard Deviation	7.30
Standard Error	1.12
Difference in reductions in BP which will yield the specified p-value above.	5.48
Mean Reduction in BP for Standard Therapy	4.20
Mean Reduction in BP for New Therapy	9.68
Estimate of the Percent Chance of Individual BP reductions for New Therapy exceeding Individual BP reductions for Standard Therapy.	69.59%
Estimate of the Percent Chance of Individual BP reductions for Standard exceeding Individual BP reductions for New Therapy.	30.41%

Fig. 1 Overlap of simulated data distributions corresponding to differences assessed as statistically and clinically significant – continuous data.

As with the continuous calculator you still have, in Fig. 2, individual survival times for the standard therapy better than those for new therapy more than 35% of the time despite the very meaningful estimated ratio of hazards of death of new therapy to standard of less than 0.6, and that one in million p-value supporting the superiority of the new therapy ("in the aggregate"—people sometimes leave that part out). We collect all survivals beyond 7 years into the last bar in the histogram in Fig. 2 revealing a more marked difference. One should consider this and the 65% estimated chance of the new therapy survival being larger than that for standard therapy when making choices. However, with the flip rate of 35% one might consider the standard therapy if it is more tolerable and/or if there are other indicators that one would be in the 35%. The EQ-5D index [12], a quality of life indicator, has a zero score evaluated as a state equivalent to death and a 1 corresponds to good health. Cancer databases often have patients reporting negative indices, presumably a state worse than death.

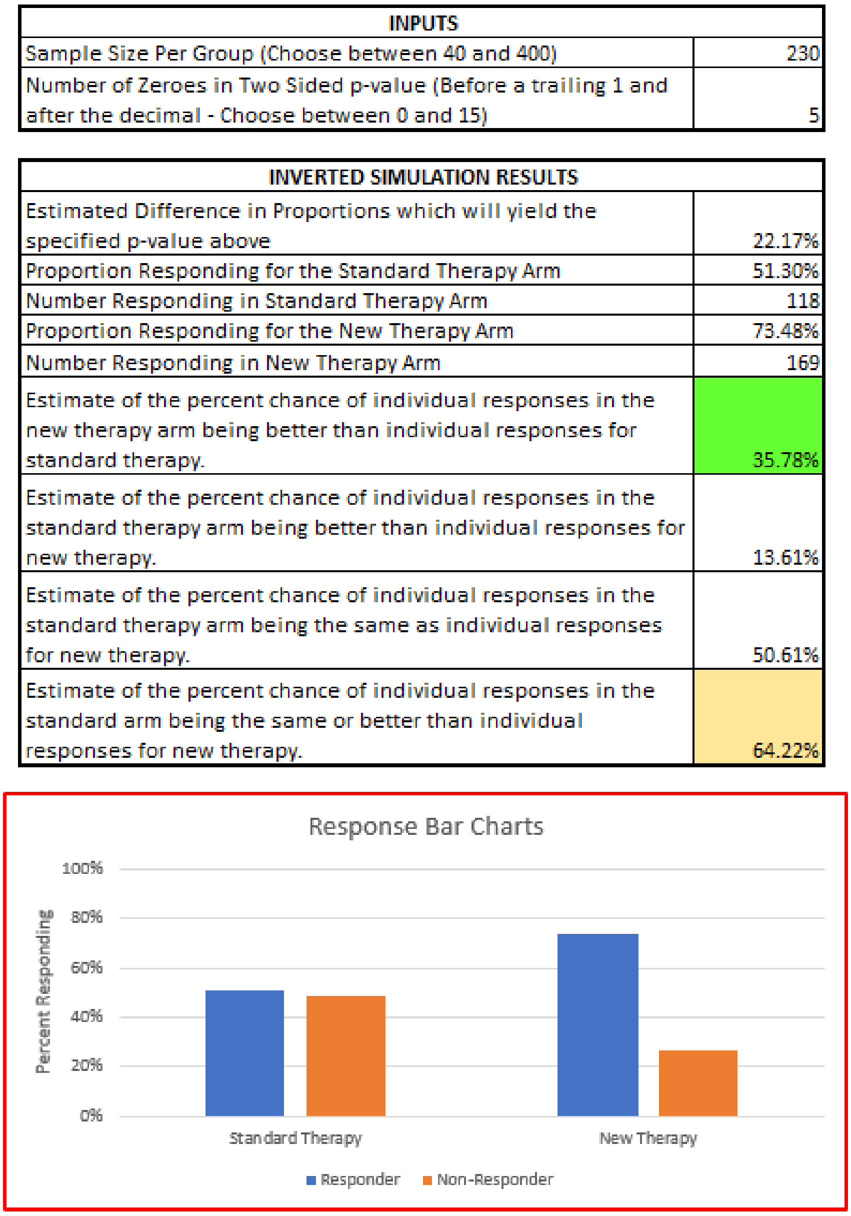
In the binary outcome graphic in Fig. 3, the percent of times we have the standard therapy individuals do as well or better than someone on new therapy higher than 60%. However, it is usually possible to break the "non-responder" and "responder" labels into many ordinal grades and when you do that and reassess how often the standard therapy is as good or better you would get closer to the 30% number.



INPUTS	
Sample Size Per Group (Choose between 80 and 400)	176
Number of Zeroes in two sided p-value (before a trailing 1 and after the decimal)	5

INVERTED SIMULATION RESULTS	
Estimated Hazard Ratio which will yield the specified p-value above.	0.557
Estimated Median Survival in the Standard Therapy Arm (Years)	2.19
Estimated Median Survival in the New Therapy Arm (Years)	3.94
Estimate of the Percent Chance of Individual Survivals for New Therapy exceeding Individual Survivals for Standard Therapy.	63.07%
Estimate of the Percent Chance of Individual Survivals for Standard exceeding Individual Survivals for New Therapy.	36.93%

Fig. 2 Overlap of simulated data distributions corresponding to differences assessed as statistically and clinically significant – survival data.



INPUTS	
Sample Size Per Group (Choose between 40 and 400)	230
Number of Zeroes in Two Sided p-value (Before a trailing 1 and after the decimal - Choose between 0 and 15)	5

INVERTED SIMULATION RESULTS	
Estimated Difference in Proportions which will yield the specified p-value above	22.17%
Proportion Responding for the Standard Therapy Arm	51.30%
Number Responding in Standard Therapy Arm	118
Proportion Responding for the New Therapy Arm	73.48%
Number Responding in New Therapy Arm	169
Estimate of the percent chance of individual responses in the new therapy arm being better than individual responses for standard therapy.	35.78%
Estimate of the percent chance of individual responses in the standard therapy arm being better than individual responses for new therapy.	13.61%
Estimate of the percent chance of individual responses in the standard therapy arm being the same as individual responses for new therapy.	50.61%
Estimate of the percent chance of individual responses in the standard arm being the same or better than individual responses for new therapy.	64.22%

Fig. 3 Overlap of simulated data distributions corresponding to differences assessed as statistically and clinically significant – binary response data.

6. Prospective Clinical Trials

Randomized clinical trials comparing therapies usually involve an adequate amount of follow-up and an adequate number of subjects to uncover unusual and rare adverse events associated with the therapies studied, in addition to providing good estimates of rates for common side-effects. Studies typically have a large number of clinic and hospital sites participating across countries and continents and a large number of contracted organizations to create databases, verify the accuracy of data entered by the site, ensure randomization without bias, and any blinding of patients and site personnel to therapies. These sites and organizations are entities independent of sponsors and are subject to audits by both the regulatory agencies and the sponsor—the data is likely robust and reliable and inferences about aggregate effect are likely accurate.

In Europe, North America and many other countries there is a requirement that sponsors of clinical trials provide details such as the primary and key objectives, hypotheses and endpoints at or before the start of a trial to an online database. In the US such a database is at clinicaltrials.gov. Further regulators require sponsors to provide a study protocol before study starts enrolling and detailed statistical analysis plans shortly after, and before any unblinding and analysis is conducted. Study protocols are also shared with Institutional review Boards (IRBs) before trials are initiated. These commitments reduce publication bias as there is a requirement to report results to clinical trials.gov or a similar online database irrespective of whether results were negative or positive. The prospective statistical plan also helps prevent cherry picking among the many choices when carving out analysis populations, endpoints, data cuts, hypotheses and analysis methods. A statistical plan specifies and selects one from all these choices of data presentation. A statistician will tell you that this controls the type I error rate—the likelihood of falsely concluding a difference between therapies when there is none. A pre-specified analysis has a lot of credibility—it is like calling a shot before making it in billiards. Further, regulatory agencies usually require two successful well controlled studies before the approval of a therapy, thus adding to this credibility of inferences supporting a therapy.

Though we make the case that most statistically significant results often represent some stochastic incrementalism in the aggregate, clinical trials are usually sized right to detect clinically meaningful differences in the aggregate—a chunky incrementalism. A statistician would size a study to detect reasonable improvements in the aggregate reducing the possibility of triggering a signal based on trivial differences. A series of such increments could add up to marked improvements in both efficacy and safety over time. Such an approach is a critical first step even if it may not help entirely in the choice of therapy by and for the individual patient. We will get to a discussion on that shortly. There can be long periods of stasis in drug development where new therapies continue to be compared on efficacy against old standards with little effect, or in trials establishing non-inferiority with a current effective standard. These can get approved based on improved and/or differing safety profiles, quality of life or economic benefit. Conclusions from well conducted clinical trials are likely to be far more reliable than those from data sources I describe below.

7. Big and Easy Data

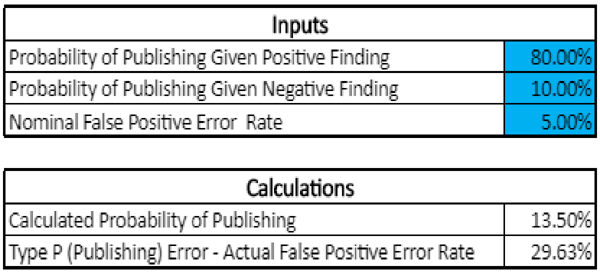
Obamacare accelerated trends towards the use of electronic records and other uncontrolled prospective and retrospective data. Acquiring such data can cost a tenth or lesser than running a controlled clinical trial to obtain the data. Very large datasets can be obtained with the downside mentioned earlier of the triggering of trivial results as noteworthy. Further there can be substantive biases due to the uncontrolled nature of the data. There exist methods, using propensity scores, as described in Rubin and Rosenbaum [13], which can control for these biases when all likely confounding variables are available. It is noted that while statisticians can try to pre-specify analyses for such data, one is not required to publish either the analysis plan or any obtained negative results.

Further the analyses can be overly managed at the institution conducting the research, resulting in multiple inferential analyses, population carve-outs, endpoints, hypotheses and analysis methods and subsequent choices amongst these on results worth publishing. Other reasons why results do not see the light of day are that they are negative or neutral, not in the best interest of the organization, inconsistent with other published results emanating from the institution and contrary to opinions of influential external opinion leaders. Conclusions in such contexts should note that results are exploratory or hypothesis generating and that multiple analyses in addition to the reported results were conducted without adjustments for multiple testing to control the overall false positive rate. Such acknowledgements are necessary, unless all conducted analyses agreed substantially, as data presentation for such data often patterns presentations for prospective controlled clinical trials.

8. Publication Bias Calculator

In the last section we looked at a large number of reasons why results did not get published. We will look at the consequence of just two of the reasons I mentioned, negative or neutral results, on the validity of results that did get published. The calculator uses Bayes' Theorem and mathematical details are in Appendix 2. In a second calculator at our webpage mentioned earlier [6] you can enter information as shown in Fig. 4.

Fig. 4 has the probabilities that analyses conducted at an institution will be published given a positive finding (a statistically significant result) and that for publishing given a negative finding. The results for values of 80% and 10% respectively are shown in Fig. 4. The third entry is the nominal false positive rate used in the analysis—usually a two-sided 5%. The actual false positive rate corresponding to the nominal 5% is actually close to 30%. We are still talking aggregates here and predictability in individuals is likely much lesser. Even with our p-values with 5 zeroes, which needed no adjustment upwards, the patient level estimated rate of flipped efficacy was near 30%. Further adjustments for any of the multiple analyses mentioned above would make the results even less credible. Young and Karr [14] looked at 52 claims from uncontrolled studies with significant results which were published in reputed journals like NEJM, JAMA and JNCI, and noted that none of these significant findings held up in randomized clinical trials—5 were supported in the opposite direction.



Inputs	
Probability of Publishing Given Positive Finding	80.00%
Probability of Publishing Given Negative Finding	10.00%
Nominal False Positive Error Rate	5.00%

Calculations	
Calculated Probability of Publishing	13.50%
Type P (Publishing) Error - Actual False Positive Error Rate	29.63%

Fig. 4 Publication bias estimator.

9. Correlation Is Not Much Association Either

A typical example on correlation starts with data on the number of sand castles built at a beach plotted on the *Y*-axis against ice-cream sales at the beach on the *X*-axis. Such a scatter plot would have a narrow cloud of data aligned with the lower end of the cloud at lower values of both the number of sand castles and the ice-cream sales and with upper end having both high. One may be lead to think, through this graphic, that the high ice-cream sales lead to increases in the number of sand castles built. This notion is then rejected by the presenter who will let you know that a third factor, the number of people at the beach, likely lead to highs or lows on both simultaneously. The conclusion drawn is that correlation is merely association and not necessarily causation. A p-value with the 5 zeroes when rejecting a null hypothesis of a zero correlation and 50 records of data would correspond to a calculated correlation coefficient of 0.613. A correlation coefficient between two measures is roughly a measure of the tendency of one thing measured, to be above or below its average, when the other measure is also above or below its average and goes from -1 for a perfect negative correlation and +1 for a perfect positive. It uses the two averages, something we have a little trouble with earlier, and like the average it is an aggregate statistic. It can be shown, that for a close to ellipsoidal scatter plot of data, a good 29.9% of data points will show an association in the opposite direction when the correlation coefficient is the 0.613 above—in 29.9% of the days at the beach, ice-cream sales will be below the average ice-cream sale with the number of sand castles built being above average and vice-versa. Details on calculations are in Appendix 3. The discordant proportion evaluates to the simple expression $COS^{-1}(\rho)/\pi$. The correlation of 0.613 above is not much association for a good 29.0% of the individual data points—they associate in an opposite direction to that indicated by the correlation coefficient. The discordant percentages are 33.3%, 25.3% and 20.5% for correlations of 0.5, 0.7 and 0.8 respectively. All three correlations would be considered large. In Cohen [15] correlations of 0.1 to 0.3 are small, 0.3 to 0.5 are medium and greater than 0.5 are deemed large. Scientists working in the social sciences, psychology, quality of life, mental health, drug abuse and addiction, criminal justice and other similar disciplines rely considerably on statistical tools using correlations. Many of their conclusions likely have ecological fallacies making them inappropriate for prescriptive recommendations at an individual level.

10. Fallacies in the Histogram and the Standard Deviation as Well

First recall how we spoke earlier about the skinny notional distribution of the average with its spread characterized by the narrow standard error and the wider distribution for individuals with its spread characterized by the standard deviation. That "distribution" for individual data can be a fallacy as well unless every individual's measurement has an identical distribution—otherwise what are we talking about when we look at the histogram? This assumption is also made in Bayesian approaches when obtaining the likelihood used to update the prior distribution of an underlying aggregate to obtain the updated posterior distribution of the aggregate as discussed in Spiegelhalter et al. [16]. These measurements should be independent of each other, otherwise, for instance for a positive dependence, if we get some measures on one side of the histogram most others would tend to be on that side skewing the histogram away from the true distribution. These assumptions drive the classical central limit theorem used to derive the notional distribution of the average and as noted, the Bayesian approaches as well.

We have frequently referred to an alternate additional statistic reflecting individual variation through the proportion of times a patient on the standard had a better reduction in BP than a patient on a new therapy. We can try to move from this data driven fact to a state of nature true for all patients that underlies it. We might infer that this proportion is an estimate of the probability of "any" patient responding better to the standard therapy than to the new therapy in similar patients elsewhere outside the confines of the clinical trial. The assumptions in both the frequentist and Bayesian setting that we are making about "identical" distributions across patients and independence across patients would support the previous statement. Returning to our BP example, this would mean that all patients under standard and new therapy would tend to hit reductions in BP of say 5 and 8.5 respectively, give or take about a standard deviation of 7. This however, would not be true, if, for instance, a patient had a skinnier or wider distribution under the two therapies than that reflected by our wide histograms and if the patient distributions are centered at different BP reduction levels than those for the wide histograms. It can be shown that drawing independent identical observations from a distribution obtained as the average of individual non-identical distributions is equivalent to drawing an observation from each of the independent non-identical individual distributions. The grouped averages and the histograms could be estimating either something interpretable as the true underlying mean and true underlying distribution of identically distributed individual measures or the average of the differing underlying means and distributions of individuals with differing distributions. See Appendix 4 for details demonstrating these observations. The grouped variance obtains as an average of the individual variances plus the mean of the squared differences of the individual means from the average of the means. At one extreme this identity allows differing individual means with all variances of zero—these distributions have singularities at the mean values. At the other extreme all patients have the same underlying mean with identical or differing variances. One could argue fixed or singular data at the subject level or an inherent volatility. Likely we have differing underlying variances and means for individuals and an argument that they have identical distributions is likely specious. Computed statistics such as the estimated aggregate treatment differences with confidence intervals, and p-values will be identical—inferences that are likely to be masquerading as applicable equally in all individuals.

The data derived proportion of times patients on standard have a larger BP reduction under standard is likely estimating the mean probability over all patients of patients responding better on standard, rather than the probability for any and every patient. An estimated proportion of 30% could mean some patients have a 20% or less probability of doing better and some have a probability of 40% or more. So, the statistic we used to demonstrate an ecological fallacy, itself has an ecological fallacy. The "underlying" mean, variance and distribution, used often in the previous paragraph, presumed to be different or the same across patients was used to make the scientific distinction that I had made earlier between an experience and an explanation. We are encouraged to look at an ordering of the underlying means from a clinical trial and presume that a patient is always innately better off on one therapy and not look at the experience which tells us that a good number of patients ended up with a better response on standard than on new therapy. We are trained to value an explanation or a rule more than experience/phenomenon, as all experience should fall in place when we have the rule. A little explanation in physics will tell us that we will come out screaming, but alive and unhurt, from every roller-coaster ride. But this context here is somehow different. When evaluating whether someone will respond to a therapy we should be referencing both the mean measures on therapies, any established ordering of means for the patient (differing or identical distributions across patients) and observed prior data such as the flipped proportion estimate. In some social contexts, when someone is getting fixed up for good, the prospective bride (and sometimes the groom) is usually told that we know their family, we know their community, they are good people! I am reminded about a little ditty the girls used in my hometown when a boy misbehaved, which goes "handsome is as handsome does"! There is a strong move now with modernization, towards allowing this assessment adequately. We need that and it is still helpful to reference the "good" if it is not based on some prejudice.

We see in Fig. 2 patients on standard therapy surviving 4 years when a patient on new therapy survives 3 years. Would that standard therapy patient have survived 5 years if he had taken the new therapy instead—hard to tell—most of us have only one life to live. We need to be able to assess the distribution for measures of interest, within subject and under both therapies before we can make a statement about a personalized probability of one therapy being better than the other for the patient. Sounds very much like the thing with the average—we usually have just one observation on a patient and that too on just one therapy. Perhaps the notion of a distribution underlying that one thing, identical or different across patients, is a convenient bit of fiction as well. More so for survival data and for diseases where we have only one shot at therapy. Cross-over and N-of-1 designs, where we can switch between short term therapies (as the disease recurs on stopping therapy), can allow us to use a notion of a patient "distribution" and help crudely assess that distribution. Even for survival data some kind of a distribution could be assessed using response on some quick leading indicators of future survival.

11. Tea Tasting Experiment

We hear a lot about randomized, blinded designs and it all started round about when a famous statistician, Sir Ronald Fisher, had a little tea with a lady. The table was set. There were walnut scones, jumbleberry jam, biscuits perhaps, and tea—lots of that, 8 cups—all for the lady. Let me tell you why. Earlier, the lady, lady Muriel Bristol, had claimed that she could tell if the tea or the milk was added to the cup first. Now our statistician, Sir Ronald Fisher, likely a bit of a skeptic, said “Let us find out if you can?”. The rest is part of our folklore. Sir Ronald Fisher had 4 cups made with the milk first and 4 with the tea first and then randomly ordered them on the table with the lady blinded to the ordering. The lady picked all four cups out correctly. Sir Ronald Fisher figured a p-value of 1 in 70 or 0.0143. For the tea tasting design and details on experimental design see Hinkelmann and Kempthorne [17].

This experiment, you should notice, is very much an N of 1 trial. There was N = 1 Lady, two interventions occurring 4 times each and the outcome was a correct guess. The probability that this individual, Lady Muriel Bristol, was guessing at random is just 1.4% so we reject the null and conclude that she very likely was not. The conclusion was not about an aggregate—we are not making the much weaker conclusion that people in the aggregate tend to be able to tell if the tea was added first. We could do a study asking a large number of people if they can tell if a cup of tea had the milk or the tea added first. A 65% correct rate could be statistically significantly different (with those 5 zeroes in the p-value) than guessing (null of 50%), if we had the right number of subjects. A 15% difference—some may even call it meaningfully significant – but 35% of the individual subjects got it wrong!

12. Getting to Personalized Medicine

The standard deviation (SD) of diastolic blood pressure measures (usually about 8 mmHg) is a little larger than that for a measure of the reduction in blood pressure (about 7 mmHg) and this is due to the correlation of before and after measures for a subject. A correlation of 0.7 would have given close to the same SD for the two measures. A cross-over design would consider a similar difference of effects of two or more therapies given to the same subject and can be more efficient when correlations between effects are greater than zero. This would lead to a smaller sample size requirement than a parallel arm study. This would also get at the within subject distribution we referred to earlier as something which would help us overcome issues with aggregates. However, the cross-over analyses reported will additionally provide between subject aggregates of within subject differences between therapies and are hence subject to similar fallacies to those described earlier when used at a patient level. Menard et al. [18] report results of a hypertension trial involving cross-overs between a beta-blocker alone, a combination of two diuretics, and the two diuretics in combination with the Beta-blocker. This manuscript focused on design of hypertensive trials and did not report adverse event rates for the therapies used. Information on safety is available on the product labels [19-21]. The mean reductions from prior placebo wash over after adjusting for carryover effects in the Menard report were 3.2, 4.0 and 6.5 respectively. These differences were not demonstrated to be statistically significant in this cross-over study with 24 patients. After patients went through the three therapies, they continued on the therapy with the largest reduction in Diastolic BP. An N-of 1 trial would have been similar and would have had just one patient with more repetition of therapies to gauge effect and variance for that one subject [22]. Like the Menard trial this would allow for a more informed patient choice should the patient provide informed consent to such an approach. There would be the additional burden to the patient of switching back and forth through therapy and wash-out periods. And there are scores of options, for hypertension and other indications. Other issues with cross-over designs include difficulty with drop-outs, inappropriateness for many conditions, carry over of previous treatment effects and some difficulty in analysis and in confidence in findings given carry-over and period effects [23]. In the Menard trial, 3 patients (12.5%) had insufficient response on any therapy. Seven (29.2%) had their largest reduction in BP on Beta-blockers—the therapy with the lowest aggregate response. Four (16.7%) had their largest reduction in BP on the diuretics and 10 (41.7%) on the Diuretics in combination with Beta-blockers—the therapy with the largest aggregate response. Had it been a larger study yielding statistically significant results (The difference between the largest average reduction and the smallest in the Menard study was close to half the standard deviation), a winner takes all treatment strategy would have put all subjects on the diuretics/beta-blocker combination. This result supports earlier discussions on the considerable percent of subjects who buck aggregate data supporting a therapy and do well on a therapy deemed inferior.

In Figs. 1-3 we were simulating completely randomized designs or designs where there was blocking just to ensure that the randomization did not give us an imbalance in the size of the two groups. Often, we have randomized block designs where subjects are randomized within blocks—typically a set of 4 patients within strata (for two treatment groups), with 2 patients randomly assigned to each group. One might stratify a hypertension trial by gender and age (< 50, >= 50 years). Computing our statistic on the flipped proportion within each of the 4 {gender × age} strata should lead to a smaller proportion when there is increased homogeneity of measures within strata for both therapy groups. This allows some degree of personalization of effect for patients. Dynamic randomization which balances group membership within a larger number of strata as well as propensity score matching in the non-randomized setting could help tease out an effect better for more complex patient profiles. Supervised machine learning methods used in AI (artificial intelligence) are built on a decision theoretic framework reducing error around an expected mean given a patient profile [24]. One must note that predictions based on machine learning are also based on aggregate data though they are trained to a larger degree by data closer to a profile. The ideas discussed earlier about the skinny distribution for aggregates and the wide distribution for the patient still applies. These distributions now associate with a complex patient profile including characteristics and therapeutic interventions rather than those we saw earlier standing on one feature (a particular therapy—new or standard). Both the skinny and the wide distributions will be tighter due to homogeneity when looking at a narrow patient profile. There are a number of machine learning methods to choose from which will give somewhat different widths, locations and shapes for these distributions. Hastie et al. [24] provide expressions, for some models, for the much wider prediction intervals for individuals as well as the confidence interval of the aggregate prediction. Clear expressions of this distinction between patient and aggregate predictions are provided in standard books on multivariate regression [25]. The patient prediction is a lot more useful for a physician treating a patient—indicative that some patients with an adverse profile might do well and some others with a good profile might do poorly, calling for a more personalized look at patient data.

Another good statistic to look at is the concordance probability—it is the probability that a randomly selected pair of patients, one with a poorer outcome than another, will be correctly identified based on inputting the two patient profiles in the model [26]. This statistic inspires the flip proportion statistic in this commentary and like that statistic it tends to be about the same when the size of the data set changes—it is not an artifact of big and bigger data. It tends not to be too large. The historically popular Framingham Heart Study Model (2002 version) had a concordance probability of a little more than 70% [27]. Likely other more predictive disease contexts and a good choice of the model and relevant predictors could provide larger concordance probabilities, while still leaving 10% to 35% of patients characterized incorrectly. Patient factors and populations not considered during training of the model could lead to biased estimates through an AI model. Even randomized and blinded studies lead to strong expectancy effects with data moving towards patient and physician biases. The data itself is "trained" towards these biases and the AI model trained on the data may quantify these and perpetuate them. In a different context, a predictive model used by Chicago law enforcement, resulted in inappropriate profiling not very different from human racial profiling [28], possibly due to inherent biases in databases. Often the model is based on correlations without a strong theoretically linked basis. For instance, a model may use the bill payment history of subjects and a poor history may signal a cognitive disability requiring intervention. Clearly there are other contexts such as poverty or unemployment which could lead to such a poor history.

13. Implications for Patients

Significant stochastic propensities for efficacy, small or large, are likely to be associated with ecological fallacies when we look at individual patients. We computed proportions of patients on one therapy responding favorably despite aggregate data significantly (and even clinically significantly) supporting its inferiority to an alternate, as a statistic demonstrating this fallacy. Discomfort with p-values is reflected in the official statement by the American Statistical Association [29] which discounts some interpretations of p-values which were more acceptable historically, likely due to increasing perceptions of subject level anomalies. The relevance of p-values may continue to hold for aggregate inferences. We had some difficulty ascribing meaning to our discordant statistic given ecological fallacies in its construction as well. This flipped proportion was based on a single core endpoint and not on the complete profile of effects of a new therapy including safety, financial costs, convenience and some assessment of short and long term adverse effects. When an informed patient and physician choice for therapy occurs using such complete holistic effects surrounding therapy, the flipped proportion is likely even less predictive of the eventual intervention and outcome.

We discussed emerging trends in personalized medicine. Models with a high degree of complexity incorporating methods for reducing bias and variability are likely needed to help make individualized predictions and choices. Personalized medicine in many contexts is currently too rudimentary to justify a normative therapeutic recommendation based on a patient's genetic or disease profile, particularly when there are larger variances associated with patient specific predictions. We have to hold on to a large number of choices for patients. Randomized controlled clinical trials are a good critical first step in demonstrating the validity each of these choices. Uncontrolled studies can support a therapeutic intervention if there are a large number of other independent analyses supporting the intervention. In a parallel publication [30], we provide a set of tools to evaluate the subject level and aggregate utility of cohort comparisons in randomized or uncontrolled settings.

This rather complex landscape, without clear answers for all patients, necessarily requires consent from patients after alternatives are discussed adequately. When faced with a set of equally unfavorable choices, patients should have the right to forego any therapy. It is necessary to factor in the patient's subjective quality of life assessments under therapy and consider changes in therapy which improves it. One may consider available care options, which are infrequently used due to less commercial backing, such as therapies which are off patent or those for which intellectual property rights do not apply. In many cases, evidence is available about these therapies from randomized controlled trials. See for instance, a review about Hibiscus *sabdariffa* extract for Hypertension by Hopkins et al. [31]. One sometimes hears of elderly patients having as many as a dozen pills a day. Perhaps there is an overuse and the physician can monitor closely and do nothing or consider preventive care, especially when evidence supporting action is weak. The physician's intuition derived from training, skill, and experience, bedside communications, patient reports of health, sickness and pain and patient specific data continue to be critical. Aggregate data and model-based predictions are helpful but are not likely to entirely replace personalized care.

References

[1] Neyman, J., and Pearson, E. S. 1933. "IX. On the Problem of the Most Efficient Tests of Statistical Hypotheses." *Phil. Trans. R. Soc. Lond. A.* 231 (694-706): 289-337. doi:10.1098/rsta.1933.0009. ISSN: 0264-3952.

[2] Popper, K. 1959. *The Logic of Scientific Discovery* (2002 pbk; 2005 ebook ed.). Routledge. ISBN: 978-0-415-27844-7.

[3] Einstein, A. 1920. *Relativity: The Special and General Theory*. Methuen.

[4] Rothman, and Greenland. 1998. *Modern Epidemiology*. Lippincott, Williams, and Wilkins.

[5] Thall, P. F., Simon, R. M, and Estey, E. H. 1995. "Bayesian Sequential Monitoring Designs for Single-Arm Clinical Trials with Multiple Outcomes." *Stat Med* 14 (4): 357-79. PMID: 7746977.

[6] http://resourcetepee.com/commentaries/alternative-truths-the-slag-and-smoke-of-statistics/.

[7] Norman, G. R., Sloan, J. A., and Wyrwich, K. W. 2003. "Interpretation of Changes in Health-Related Quality of Life: The Remarkable Universality of Half a Standard Deviation." *Medical Care* 41 (5): 582-92.

[8] Terrell, G. R., and Scott, D. W. 1985. "Over smoothed Nonparametric Density Estimates." *Journal of the American Statistical Association* 80: 209-14.

[9] Lo, S. H., Mack, Y. P., and Wang, J. L. 1989. "Density and Hazard Rate Estimation for Censored Data Via Strong Representation of the Kaplan-Meier Estimator." *Probability and Related Fields* 80: 461-73.

[10] Desu, M. M., and Raghavarao, D. 1990. *Sample Size Methodology*. Boston: Academic Press.

[11] Chow, S.-C., Wang, H., and Shao, J. 2007. *Sample Size Calculations in Clinical Research* (Second Edition). Boca Raton, Florida: Chapman & Hall/CRC.

[12] The Euro Qol Group. 1990. "Euro Qol—A New Facility for the Measurement of Health-Related Quality of Life." *Health Policy* 16 (3): 199-208.

[13] Rosenbaum, P. R., and Rubin, D. B. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41-55.

[14] Young, S. S., and Karr, A. 2011. "Deming, Data and Observational Studies: A Process Out of Control and Needing Fixing." *Significance* 8 (3): 116-20.

[15] Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

[16] Spiegelhalter, A., and Myles, B. 2004. *Approaches to Clinical Trials and Health-Care Evaluation*. Chichester, England: John Wiley.

[17] Hinkelmann, K., and Kempthorne, O. 1994. *Design and Analysis of Experiments*. Vol. 1. New York: John Wiley.

[18] Menard et al. 1988. "Cross-Over Designs to Test Anti-Hypertensive Drugs with Self-Recorded Blood Pressure." *Hypertension* 11 (2): 153.

[19] Oxprenolol Hydrochloride Patient Leaflet. https://www.medicines.org.uk/emc/files/pil.1039.pdf.

[20] Chlorthalidone FDA Package Insert. https://www.iodine.com/drug/chlorthalidone/fda-package-insert.

[21] Triamterene FDA Package Insert. https://www.iodine.com/drug/triamterene/fda-package-insert.

[22] Kravitz, R. L., Duan, N., and the DEcIDE Methods Center N-of-1 Guidance Panel (Duan, N., Eslick, I., Gabler, N. B., Kaplan, H. C., Kravitz, R. L., Larson, E. B., Pace, W. D., Schmid, C. H., Sim, I., Vohra, S.). 2014. *Design and Implementation of N-of-1 Trials: A User's Guide. AHRQ Publication No. 13(14)-EHC122-EF. Rockville, MD: Agency for Healthcare Research and Quality*. http://www.effectivehealthcare.ahrq.gov/N-1-Trials.cfm.

[23] Senn, and Barnett. 2002. *Cross-Over Trials in Clinical Research*. Chichester, England: John Wiley.

[24] Hastie, T., Tibshirani, R., and Friedman, J. 2008. *The Elements of Statistical Learning* (2nd ed.). Springer.

[25] Myers, R. H. 1989. *Classical and Modern Regression with Applications*. Boston, MA: PWS-KENT.

[26] Harrell, F. E., 2001. *Regression Modelling Strategies*. New York: Springer-Verlag.

[27] Pencina, M. J., and D'Agostino, R. B. 2004. "Overall C as a Measure of Discrimination in Survival Analysis: Model Specific Population Value and Confidence Interval Estimation." *Statistics in Medicine* 23 (13): 2109-23.

[28] Saunders, J., Hunt, P., and Hollywood, J. S. 2016. "Predictions Put into Practice: A Quasi-Experimental Evaluation of Chicago's Predictive Policing Pilot." *J. Exp. Criminol* 12: 347. https://doi.org/10.1007/s11292-016-9272-0.

[29] Wasserstein, R. L., and Lazar, N. A. 2016. "The ASA's Statement on p-Values: Context, Process, and Purpose." *The American Statistician* 70 (2): 129-33. Doi: 10.1080/00031305.2016.1154108.

[30] Srinivasan, S., 2018. "Evaluation of Reported Statistical Inferences." *Journal of Mathematics and System Science* 8 (5): 140-52.

[31] Hopkins, A. L., Lamm, M. G., Funk, J., and Ritenbaugh, C. 2013. "Hibiscus Sabdariffa L. in the Treatment of Hypertension and Hyperlipidemia: A Comprehensive Review of Animal and Human Studies." *Fitoterapia* 85: 84-94. doi:10.1016/j.fitote.2013.01.003.

[32] University of Michigan. *Bivariate Normal Probability Calculator*. http://socr.umich.edu/HTML5/BivariateNormal/.

[33] Cambanis, S., Huang, S., and Simon, S. 1981. "On the Theory of Elliptically Contoured Distributions." *Journal of Multivariate Analysis* 11: 368-85.

[34] Jensen, D. R., and Srinivasan, S. S. 2004. "Matrix Equivalence Classes with Applications." *Linear Algebra and Its Applications* 388: 249-60.

[35] Srinivasan, S. 1995. "Pitman Estimation for Ensembles and Mixtures." V.TechWorks Home, Virginia Tech. https://vtechworks.lib.vt.edu/handle/10919/39155.

Appendix 1: Details on the Inverted Simulations

For all three cases we start by obtaining the Wald Statistic Z which maps to the k of zeroes in the two-sided p-value (between the decimal and a 1) using the expression

$$Z = \varphi^{-1}[1 - 0.5 * 10\hat{}(-k - 1)],$$

where φ^{-1} is the inverse of the cumulative distribution function of the standard normal.

Continuous Case

For the continuous case we consider a null distribution of the reduction in BP with a mean reduction in BP of 5 in the standard therapy group and a standard deviation sigma of 7. A somewhat skewed distribution is obtained as a 25%/75% mixture of two normal distributions with means of 4 and 5.333. Then for the given *M* subjects per group two data sets are simulated under this null distribution for 'standard therapy' and for the 'new therapy'. The standard deviations *SS* and *SN* and the averages $\overline{XS}$ and $\overline{XN}_0$ were computed for these null distribution-based standard and new therapy data sets. The difference in sample means that would yield the p-value with the *k* zeroes was obtained as

$$Difference = \overline{XN}_1 - \overline{XS} = Z * SQRT\left[\frac{SS^2}{M} + \frac{SN^2}{M}\right]$$

Where $\overline{XN}_1$ is the sample mean we need to have when the p-value has those *k* zeroes. We now add the following to each observation in the null based new therapy data so that it now has the appropriate difference in sample means corresponding to the p-value.

$$\overline{XN}_0 - \overline{XS} + Difference$$

The two datasets are then plotted in a histogram with the number of bins computed by rounding up

$$1 + 2 * M\hat{}(1/3)$$

The distribution of the averages are overlaid on the plots and are normal distributions centered around group averages $\overline{XS}$ and $\overline{XN}_1$ with scale parameters $SS/\sqrt{M}$ and $SN/\sqrt{M}$. The estimate of the percent chance of individual BP reductions for new therapy exceeding individual BP reductions for standard therapy is computed by counting the number of reductions in the standard group which are lower than that for each simulated subject in the new therapy group and then adding these tallies across all new therapy subjects. This sum is then divided by the total number of such comparisons given by *M^2*. The estimate of the percent chance of individual BP reductions for standard exceeding individual BP reductions for new therapy is computed in a similar manner.

The default value 85 for the *M* per group in the calculator would be required to detect a difference of 3.5 between mean reductions in BP of 5 for standard therapy and 8.5 for new therapy with 90% power using a two-sided 5% level test using the standard Wald Statistic based on the asymptotic normal distribution of the difference in averages. The difference 3.5 is half the standard deviation of 7 – having a difference of at least that fraction of the standard deviation is often considered meaningful.

Survival Case

For the survival case we consider a null exponential distribution of survival times with a median of 2 years. This corresponds to a hazard of 0.3466. Then for the given *M* subjects per group two data sets are simulated under this null distribution for 'standard therapy' and for 'new therapy' with a censoring at 7 years. The estimated value of the ratio of hazards of standard to new HR_1, that would yield the p-value with the *k* zeroes was obtained using the approximation to the variance of the natural logarithm of the hazard ratio given by the reciprocal of a fourth of the total number of events across the two groups [11]. The expression used was as follows

$$HR_1 = EXP[Z/0.5 * SQRT(ES + EN)]$$

Initial estimates *ES* an *EN* were obtained as the expected number of events within 7 years under medians of 2 years for standard and 2.857 for the new therapy (Hazard ratio = 0.7). The hazards under the null simulations for the standard and new therapy of *H1* and $H2_0$ were obtained as dividing the number of events *ES* and *EN* by the sum of the event and censor durations in the two groups obtained by the null simulation. The hazard ratio under the null simulation was obtained as HR_0= $H1/H2_0$ and a correction was computed as HR_1/HR_0. This correction is the amount you would have to multiply the survival times generated under the null for the new therapy to get data consistent with the estimated HR_1computed earlier.

This initial computation of HR_1 used expectations for ES and EN. We can now use the ratio transformed new therapy data to update the number of events EN, recompute HR_1 with this EN and the number of events ES in the null simulation of the standard therapy group and recompute EN. We repeat this process twice to get stable values for EN and HR_1. The reported hazard ratio in the calculator is the reciprocal of $HR_1$and represents the ratio of hazards of new therapy to standard. This inverted ratio is the one usually presented.

The two estimates on ordering of individual survivals on the two therapies are derived in a similar manner to that for the continuous data. We use the simulations of survival times before the censoring is applied to compute these estimates.

The default value of 176 for the M per group in the calculator would be required to detect a reduction of 30% in the hazard of an event in the new therapy compared to standard therapy (a hazard ratio of 0.7) with 90% power using a two-sided 5% level test using the log rank statistic in fixed 7 year follow-up context. Hazard ratios of at most 0.7 or less are usually considered worth testing.

Binomial Case

For the binomial case we consider a null distribution of the proportion P responding of 0.5. Then for the given M subjects per group the number of responders for standard therapy XS was obtained as random outcome from a binomial distribution with M Bernoulli trials with a probability of 0.5 of responding. Then an initial estimate of the value of the ratio of the odds of responding to new therapy to the odds of responding to standard therapy, corresponding to the p-value with the k zeroes, is obtained using the standard error SE of the natural logarithm of the odds ratio as $OR_1 = EXP[Z * SE]$. The initial computation uses a crude estimate of the standard error given by

$$SE = SQRT[2/(M * P * (1 - P))] = SQRT[2/(M * 0.25)]$$

Then we calculate the sample proportion responding $\widehat{PS} = XS/M$ for standard therapy and compute the odds of response for standard therapy as $OS = \widehat{PS} * (1 - \widehat{PS})$. Then an initial estimate of the odds of responding for new therapy is given by $ON = OR_1 * OS$. This can then be used to compute the sample proportion responding to the new therapy as $\widehat{PN} = ON/(1 + ON)$. This can then be used in a more accurate estimate of the standard error given by

$$SE = SQRT\left[1/\left(M * \widehat{PS} * (1 - \widehat{PS})\right) + 1/\left(M * \widehat{PN} * (1 - \widehat{PN})\right)\right]$$

This is followed by an update of the odds ratio $OR_1 = EXP[Z * SE]$, the odds of responding to new therapy ON, the proportion $\widehat{PN}$ responding to new therapy, the standard error SE and odds ratio corresponding to the p-value. We repeat this process twice to get stable values for $\widehat{PN}$and OR_1. The number responding to new therapy XN is obtained by multiplying the proportion $\widehat{PN}$ by M and rounding. The calculator provides two proportions and a difference in proportions corresponding to a p-value.

The estimate of the percent chance of individual responses in the new therapy arm being better than individual responses for standard therapy is obtained as *[XN*(M-XS)]/M^2*. The estimate of the percent chance of individual responses in the standard therapy arm being better than individual responses for new therapy is *[XS*(M-XN)]/M^2*. The estimate of the percent chance of individual responses in the standard therapy arm being the same as individual responses for new therapy is

*[XS*XN + (M-XS)*(M-XN)]/ M^2*

Finally, the estimate of the percent chance of individual responses in the standard arm being the same or better than individual responses for new therapy is the sum of the last two estimates above.

The default value of 230 for the M per group in the calculator would be required to detect a difference of 15% between the proportion responding to standard therapy of 50% versus a proportion of 65% responding to new therapy (an odds ratio of 1.86) with 90% power using a two-sided 5% level test using the Wald Statistic based on the asymptotic normal distribution of the log odds ratio. Differences in proportions of at least 15% or more are usually considered worth testing.

Appendix 2: Details on the Publication Bias Error Calculator

Denote by $Prob(R)$, the probability of an erroneous rejection of a null hypothesis i.e. a rejection when the null is true. Then $Prob(R^c)$, the probability of not rejecting a null when it is true is $1 - Prob(R)$.

For the publication bias related calculator, let $Prob(P|R)$ be the probability of publishing when the null hypothesis is rejected. Let $Prob(P|R^c)$ be the probability of publishing when the null hypothesis is not rejected. Then by Bayes theorem the probability of an erroneous rejection of a null hypothesis given that the result is published is given by

$$Prob(R|P) = \frac{Prob(P|R) * Prob(R)}{Prob(P)}$$

where, $Prob(P)$ is the probability of publishing given by

$$Prob(P) = Prob(P \; and \; R) + Prob(P \; and \; R^C)$$
$$= Prob(P|R) * Prob(R) + Prob(P|R^C) * Prob(R^C)$$

Appendix 3: Correlation Related Calculations

Inverted Estimated Value of Sample Correlation Coefficient

The sample correlation coefficient r associated with the two-sided k zero p-value, when rejecting a null correlation of ρ_0, is obtained by using the approximate normality for large sample sizes (n>=25) of the statistic below using [25].

$$(1/2)Ln\frac{1+r}{1-r} \sim Normal\left\{Mean = \left(\frac{1}{2}\right)Ln\frac{1+\rho_0}{1-\rho_0}, Variance = 1/(n-3)\right\}$$

Then to reject a null correlation of zero we would compute the Wald Statistic

$$Z = 0.5 * SQRT(n-3) * Ln\frac{1+r}{1-r}, which\ implies\ r = \frac{Exp[(2Z)/SQRT(n-3)] - 1}{Exp[(2Z)/SQRT(n-3)] + 1}$$

Using $Z = \varphi^{-1}[1 - 0.5 * 10\hat{}(-k-1)]$, with φ^{-1} being the inverse of the cumulative distribution function of the standard normal, k = 5 and n =50 we get a correlation coefficient r = 0.613 in the example.

Estimating Individual Discordance with Aggregated Sample Correlation Coefficient

The estimated discordance rate in individuals can be estimated for a given correlation coefficient by using a probability calculator for the bivariate normal at a University of Michigan website [32].For two variable X and Y with means μ_X and μ_Y the discordant probability for a positive correlation is given by the sum of the probabilities

$$Prob[(-\infty < X < \mu_X) \; and \; (\mu_Y < Y < \infty)] + Prob[(\mu_X < X < \infty) \; and \; (-\infty < Y < \mu_Y)]$$

The correlation coefficient of these two variables as well as the probability above is invariant under translation and scaling of X and Y. Hence the required estimated discordant probability can be obtained using a standard normal marginal (mean zero and variance 1) for X and Y and varying the correlation coefficient. The bounds to X and Y need to be entered twice in the calculator at the University of Michigan site for each correlation coefficient to get the two probabilities in the expression above. In general, there is no closed form expressions to compute probabilities over regions defined through intersections of intervals on two bivariate normal variables X and Y. The U MICH site provides a numerical computation. We derive an expression for the probability in the limited context of the interval intersections above, and used the U MICH calculator to check our result. This analytical result holds for the class of elliptically symmetric distributions, which includes the multivariate normal distribution. We had noted, rather loosely, that the result about the discordant probability holds for any “ellipsoidal” cloud of data – and for that I must say - aye! Here is the $R\boldsymbol{UB}$!

An elliptical distribution can be constructed as the distribution L given by

$$L(\boldsymbol{W}) = L(\boldsymbol{\mu} + R\boldsymbol{UB})$$

where $\boldsymbol{W}$ and $\boldsymbol{\mu}$ are k dimensional row vectors corresponding to the variables and their means, $\boldsymbol{U}$ is a k dimensional random row vector from a uniform distribution on a k-dimensional unit sphere, R is some radial distribution on [0, ∞) and $\boldsymbol{B}$ is a rank factorization of a matrix of scale parameters $\boldsymbol{\Sigma}$ [33]. Different radial measures produce different classes of elliptically symmetric distributions. For the normal distribution the density of the radial measure R is given by

$$f(r) = [2/\mathrm{SQRT}(2\pi)]*\mathrm{EXP}(-0.5*r\hat{}2)$$

The matrix $\boldsymbol{\Sigma}$ has an interpretation as a variance-covariance matrix when second order moments exist. If $\boldsymbol{E}$ is the matrix containing the eigen vectors of $\boldsymbol{\Sigma}$ as it columns and $\boldsymbol{D}$ is the diagonal matrix containing the square root of the corresponding eigen values, then $\boldsymbol{B} = \boldsymbol{ED}$. Since the two components of our discordant probabilities are defined relative to the means, we can consider the random row vector $\boldsymbol{Z} = \boldsymbol{W} - \boldsymbol{\mu}$ with null mean $\mathbf{0}$. Then $\boldsymbol{Z}$ is related to a spherically symmetric distribution $\boldsymbol{S}$ as in $\boldsymbol{Z}$=$\boldsymbol{SB}$ =$\boldsymbol{SED}$. The column vectors of $\boldsymbol{B}$ project each of the elements of $\boldsymbol{Z}$ onto the co-ordinates of a transformed metric where the distribution is spherical. The discordant probabilities are probabilities over quadrants about $\boldsymbol{0}$ in a two-dimensional sub-space defined by two elements in the $\boldsymbol{Z}$ metric (or the entire space if k =2). The elements of $\boldsymbol{Z}$ are perpendicular to each other in this metric. Reading probabilities is much easier in the transformed metric as it has a characterization through just the uniform distribution $\boldsymbol{U}$ on the unit sphere. The radial measure is not relevant in our context as the probabilities are over infinite wedges radiating from $\boldsymbol{0}$ and the measure integrates out over each ray in the wedge. When we project $\boldsymbol{Z}$ onto the transformed metric, the elements of $\boldsymbol{Z}$ have wedges between them in the transformed metric which are no longer perpendicular. The angles in radians between two such elements can be found as the inverse of the cosine of the ratio of the dot product of the corresponding row vectors of $\boldsymbol{B}$ divided by the product of the lengths of these vectors. Probabilities in the quadrants in the bivariate marginal distributions of two elements of $\boldsymbol{Z}$ are obtaining by dividing the angle by $2*\pi$ (360 degrees). Probabilities over orthants or other wedges in the $\boldsymbol{Z}$ metric can be worked out in similar manners.

Note that for this translation and scale distributional context, one can get a correlation matrix with 1's for diagonals and the correlations as off diagonal values by subtracting out the vector means and scaling each marginal by the marginal scale parameter. This removes all parameters except the correlation coefficient. When we pick two elements of $\boldsymbol{Z}$, say X and Y, having a correlation ρ, then the corresponding correlation matrix has eigen values (1+ ρ) and (1 - ρ) correspond to the eigen vectors of $[1/\sqrt{2},\ 1/\sqrt{2}]$ and $[1/\sqrt{2}$, -$1/\sqrt{2}]$. For characterization of equicorrelated and other similar matricessee [34,35]. The matrix $\boldsymbol{B}$ can be obtained as

$$\boldsymbol{B} = \boldsymbol{ED} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2}) & -1/\sqrt{2} \end{bmatrix} * \begin{bmatrix} \sqrt{(1+\rho)} & 0 \\ 0 & \sqrt{(1-\rho)} \end{bmatrix} = \begin{bmatrix} \sqrt{0.5*(1+\rho)} & \sqrt{0.5*(1-\rho)} \\ \sqrt{0.5*(1+\rho)} & -\sqrt{0.5*(1-\rho)} \end{bmatrix}$$

Denote as $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ the two row vectors of $\boldsymbol{B}$. Then $|\boldsymbol{b}_1|$= $|\boldsymbol{b}_2|$ = 1 and dot product $\boldsymbol{b}_1.\boldsymbol{b}_2 = \rho$. As the discordant probability is computed over two quadrants in the $\boldsymbol{Z}$ metric the total probability is computed as twice the fraction of the circle spanned by the angle between two elements $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ as in

$$\frac{2*COS^{-1}[\boldsymbol{b}_1.\boldsymbol{b}_2/(|\boldsymbol{b}_1|.|\boldsymbol{b}_2|)]}{2*\pi} = \frac{COS^{-1}(\rho)}{\pi}$$

Thus, our discordant probabilities depend solely on the correlation coefficient across the class of elliptically symmetric distributions. Computed numbers based on this expression agree with those from the University of Michigan calculator.

Appendix 4: Theoretical Statistical Justification for Observations about Distributions

The probability of subjects i and j under Standard and New therapy respectively, having a reduction in BP for standard higher than that for the new therapy is given by

$$P_{ij} = \int_{s=-\infty}^{\infty} F_{jN}(s)dF_{iS}(s)$$

Where the cumulative density functions (CDF) for the two subjects are $F_{iS}(s)$ and $F_{jN}(n)$ respectively. Note that we make references to a BP measure and a standard and new therapy to remain consistent with the text but the expressions provided hold more generally. When the reductions in BP have identical and independent distributions for subjects within a group then we can note that for any randomly chosen pair of reductions in BP, the standard reduction will be larger with a probability given, on dropping the subject subscripts in the CDFs, by

$$P = \int_{s=-\infty}^{\infty} F_N(s)dF_S(s)$$

When the reductions in BP are independent but not identically distributed for subjects within a group, then we can note that over all chosen pairs of reductions in BP the standard reduction will be larger with a mean probability given by the following with expressions for averaged distributions $\bar{F}(s)$ below it.

$$\bar{P} = \frac{\sum_{i=1}^{M_S}\sum_{j=1}^{M_N}\int_{s=-\infty}^{\infty} F_{jN}(s)dF_{iS}(s)}{M_S * M_N} = \int_{s=-\infty}^{\infty}\int_{n=s}^{\infty} \bar{F}_N(s)d\bar{F}_S(s)$$

$$\bar{F}_N(s) = {}^1/_{M_N}\sum_{j=1}^{M_N} F_{jN}(s) \ and \ \bar{F}_S(s) = {}^1/_{M_S}\sum_{i=1}^{M_S} F_{iS}(s)$$

It is noted that drawing *M* independent identically distributed observations from the $\bar{F}(s)$ distributions will be indistinguishable to drawing 1observation each from *M* independent non-identically distributions. If we were to believe that the observations are not identically distributed, then the expected value of a function *g(s)* under the $\bar{F}(s)$ distributions is the average of the expectation over the distributions $F_{iS}(s)$ and $F_{jN}(n)$. These two statements are supported in the following

$$\int_a^b d\bar{F}_N(s) = \int_a^b {}^1/_{M_N}\sum_{j=1}^{M_N} dF_{jN}(s) = {}^1/_{M_N}\sum_{j=1}^{M_N}\int_a^b dF_{jN}(s), \text{ and}$$

$$\int_{-\infty}^{\infty} g(s)d\bar{F}_N(s) = \int_{-\infty}^{\infty} g(s)\,{}^1/_{M_N}\sum_{j=1}^{M_N} dF_{jN}(s) = {}^1/_{M_N}\sum_{j=1}^{M_N}\int_{-\infty}^{\infty} g(s)dF_{jN}(s)$$

Setting *g(s) = s* and *g(s) =s^2*, we have $E_{\bar{F}}(S) = {}^1/_M \sum_{j=1}^{M} E_{F_j}(S)$ and

$$VAR_{\bar{F}}(S) = E_{\bar{F}}(S^2) - \left(E_{\bar{F}}(S)\right)^2 = {}^1/_M\sum_{j=1}^{M} E_{F_j}(S^2) - \left[{}^1/_M\sum_{j=1}^{M} E_{F_j}(S)\right]^2$$

$$= {}^1/_M\sum_{j=1}^{M} VAR_{F_j}(S^2) + {}^1/_M\sum_{j=1}^{M}\left[E_{F_j}(S) - E_{\bar{F}}(S)\right]^2$$