Journal of Mathematics and System Science 8 (2018) 140-152
doi: 10.17265/2159-5291/2018.05.002

Evaluation of Reported Statistical Inferences

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

Abstract: Someone or the other is always pointing to a published study to justify a point of view or the need for a change in what we do or how we live. There are so many such studies, many reported in top-notch journals, reporting results inconsistent across and often inconsistent within. It is in the interest of increasing the credibility of science, and to safeguard the general public living with its overt and covert influence, to filter good science from bad. Some inferences are good, even when counter-intuitive or seemingly inconsistent, and are likely to withstand scrutiny and some others may represent marginal effects in the aggregate not entirely useful for individual choices or decisions, and are often non-reproducible. The New York Times featured an article in August 2018 debunking some of the reported studies supporting testing for Vitamin D deficiencies and the recommendation of large supplemental doses of Vitamin D. Some of these Vitamin D claims, among other claims, were reported as not holding up on replication in controlled trials [1]. We have noted in Ref. [2] that we need to be wary as individuals about reported signals detected in studies using stochastic data, even when these aggregate signals are of a large magnitude. We demonstrated discordance rates of 30% or higher between subject level assessments of effect and the conclusion drawn in the aggregate. Here we will provide a computation of this discordant proportion as well as post-hoc assessments of aggregate inferences, with emphasis on evaluating studies with time-to-event endpoints such as those in cancer trials. Similar evaluations for continuous, binomial data and correlations are also provided. We also discuss the use of response thresholds.

Keywords: Ecological fallacies, cohort studies, survival analysis, discordant proportions, post-hoc power, correlation, expected p-values, minimally important differences, response thresholds.

1. Introduction

Perhaps you are a journalist covering science, perhaps you are a reviewer with a non-quantitative background evaluating a submitted manuscript, perhaps you are a lead on a research project evaluating computed statistical inferences before publication, or a lay person with a disease or an issue for which an academic publication offers a solution. Lots of us need a sense of whether scientific stochastic data (data associated with noise where signals are hard to discern) reported in the media or on blogs and websites is credible. Whether it amounts to something? Whether it is a call for change in our lives? Whether these reported signals are large enough to warrant societal change? We are getting inundated with scientific claims and it is clear that a lot of these lack merit due to shoddy science and analysis and are often influenced by commercial or partisan considerations. Ideas which appear to have the backing of the "scientific method" have an overly strong influence in our lives.

We argued [2] that current standard inferential presentations of aggregate data exaggerate claims when applied to individuals. We saw a considerable 30% discordance at a subject level even with meaningful effects and a string of 5 zeroes in a p-value from appropriately sized studies. A statement by the American Statistical Association [3] as well expresses caution when using p-values. There can be a push for action based on such aggregate inferences, when clearly it won't work for many and there is a discomfort which needs to be addressed. Perhaps the sheer volume, and the changing, often contradictory claims will negate the near coercive social norms or public policy which might result if the claims persist over time. This high-volume churn and the recent chorus of criticism of science will likely drown out the good as well as the bad. We need to come to the rescue of what works—the ability of quantitative methods, designed to reduce bias, to bring out useful signals from noise. Through this article we hope to

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

provide tools for a measured and proportional evaluation of the merit of reported inferences, thus mitigating any over-reach of science in personal lives. Evaluations of effect, which are not overstated, may paradoxically increase the acceptance of valid scientific data.

2. Post-Hoc Assessment of Reported Study Results

When evaluating reported study results, we examine studies where the investigator went with available data to address a question, rather than planned cohorts of the right size to detect meaningful effects. Often inappropriately sized studies or lopsided unequal cohorts will arise from observational studies. Sometimes, one can have randomized controlled studies with limited resources or those which are over-resourced to address ancillary endpoints such as rare adverse events in clinical trials. The tools we describe will be relevant when there is a degree of discipline exercised when obtaining results—such as statistical analysis plans pre-specifying a path through the data to test a well-defined primary objective. See [2] for a discussion. For uncontrolled studies, comparisons between patient groups receiving a therapy of interest to a suitable control need to be adjusted for inherent difference and biases in such data—see propensity score based methods [4] or methods evolving from this. We had noted earlier, an ecological anomaly leading to as much as a 30% incidence at a subject level of patients in a group which is inferior in the aggregate, being superior on the assessed measure, to members of the group deemed superior in the aggregate. This was in highly statistically significant contexts with relevant meaningful differences. We start our discussion of tools to evaluate reported study results with this flipped statistic, and then move on to standard assessments used for aggregate inferences.

2.1 The Discordant Proportion

We will look at discordance arising in subject records in a study evaluating aggregate differences between some two Groups S and N. For convenience we will label them the Standard and the New Group respectively. The probability of subjects *i* and *j* under Standard and New respectively, having an outcome measure for standard higher than that for the new is given by $P_{ij} = \int_{s=-\infty}^{\infty} F_{jN}(s) dF_{iS}(s)$,where the CDF (cumulative density functions) for the two subjects are $F_{iS}(s)$ and $F_{jN}(n)$ respectively. When the outcome measures have identical and independent distributions for subjects within a group, then we can note that for any randomly chosen pair of subjects chosen from each group, the standard outcome measure will be larger with a probability, given, on dropping the subject subscripts in the CDFs, by $P = \int_{s=-\infty}^{\infty} F_N(s) dF_S(s)$.

Fig. 1 illustrates the integral. Discussion of the case when outcome measures are independent but not identically distributed, with interpretations of the resulting ecological fallacies, are provided in Ref. [2]. We will further evaluate the integral above in the time-to-event context under the proportional hazards assumption. Expressions in the continuous case under normal distributions, the time-to-event case under exponential distributions and the binomial contexts are in Appendix 1. In the time-to-event case, the distributions have support only over positive real numbers. Further, it is customary to work with survival functions instead of CDF's. The integral above is equivalent to $P = \int_{n=0}^{\infty} S_S(n) dS_N(n)$.

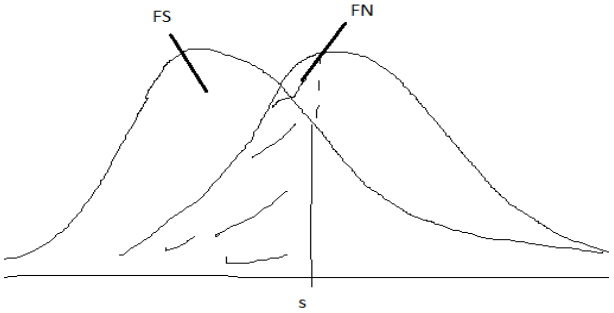

Fig. 1 Densities with CDFs indicated, illustrating the integral to compute the discordant proportion.

Under the proportional hazard assumption, the hazard ratio is independent of time n and we will compute it as the ratio of hazards of New to Standard$\eta = \lambda_N(n)/\lambda_S(n)$. The survival functions are then related as in the following [see 5].

$$[S_S(n)]^{\eta} = S_N(n)$$

Note that

$$dS_N(n)/dn = d[S_S(n)]^{\eta}/dn$$

$$= \eta[S_S(n)]^{\eta-1} * dS_S(n)/dn$$

Hence the integral evaluates as

$$P = \eta \int_{n=0}^{\infty} [S_S(n)]^{\eta} dS_N(n)$$

$$= \eta * \left[\frac{[S_S(n)]^{\eta+1}}{\eta+1} + C\right]_{n=0}^{\infty} = \frac{\eta}{\eta+1}$$

The discordant proportion evaluates as a simple ratio involving reported hazard ratios and is independent of sample sizes and p-values. An identical expression is obtained using the more restrictive exponential assumptions in Appendix 1.In our manuscript on ecological fallacies in cohort studies [2], we had provided a simulated example of a planned study to detect a hazard ratio of 0.7 with 90% power with a two-sided test at a significance level of 0.05 which resulted in a p-value with 5 zeroes after the decimal and estimated hazard ratio of 0.58. When asked what they might expect as the subject level discordance we computed above, many, including statisticians with doctorates guess that this discordant proportion is either 0.000001 or 0.05. Using our expression above the flipped proportion is 0.58/(1+0.58) = 36.7%. It is clear that many may presume individual effect based on aggregate inferences. Fig. 2 illustrates why the discordant proportion is this high despite highly significant effects, statistical and clinical.

Fig. 2 shows the usual stair step survival curves corresponding to a ratio of hazards of about 0.6 and reporting significant effect. The separation of the survival curves is convincing evidence of aggregate effect. Every vertical dip in the survival curves represent follow-up times where events occurred. The dips in the red oval represent short durations to event in the cohort with superior aggregate effect and the dips in the green oval represent long survival durations in the cohort inferior in the aggregate. Clearly this is close to the assessed subject level discordance and not anywhere near the assessed or planned false positive rates corresponding to the test for aggregate effect.

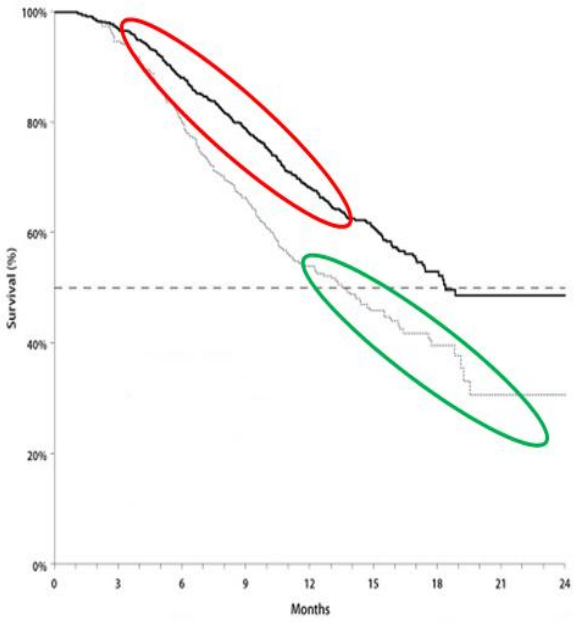

Fig. 2 Graphic depicting high discordance despite well separated aggregate survival curves.

2.2 Evaluating Aggregate Inferences

As noted, often scientists work with available cohorts in observational or randomized settings which may or may not be of an appropriate size to detect clinically meaningful differences. Oversized studies can trigger trivial effects and undersized studies may reflect large effects despite a result which is not flagged as statistically significant. Here we will provide, based on standard sources [6, 7], expressions to obtain likely p-values to help reproduce, approximately, the p-values the scientists obtained. Further, expressions are provided for calculating the post-hoc power given the cohort sizes available to the scientist using [8-10].

These together with an assessment of the size of the effect through standardized statistics can help assess if a reported result is meaningful. These formulations are provided for continuous data and binomial data in Appendix 2. For survival data, under the proportional hazard assumptions, the p-value can be obtained using an estimate $\hat{\eta}$ of the ratio of hazards λ_N/λ_S, the total number of events and the proportion in first group r_S through the following expression

$$p = 2 * \left\{1 - \Phi\left[|LN(\hat{\eta})|\sqrt{(E_S + E_N)r_S(1 - r_S)}\right]\right\},$$

where, Φ is the cumulative density function of the standard normal distribution. The post-hoc power for other hazard ratios η considered more relevant or worth detecting obtains as

$$\Phi\left\{|LN(\eta)|\sqrt{(E_S + E_N)r_S(1 - r_S)} - Z_{\alpha/2}\right\}$$

The interpretation of the subject level and aggregate statistics discussed in the last two sections are provided in the next section.

2.3 Interpretive Examples

We use an online calculator [11] to evaluate hypothetical scenarios involving time-to-event data. A screen shot from the calculator is in Fig. 3. At the online calculator [11], the reader can enter data specific to the study being evaluated in the blue cells. The calculator allows for the post-hoc assessment of inferences for continuous data, time-to-event data, binomial data and for correlations in four tabs of the online spreadsheet. In Fig. 3, we have a screen shot of evaluations for three scenarios involving time-to-event data involving events such as disease progressions or death.

Scenario 1 has a hypothetical study having one group with 217 events in 431 subjects and another group with 191 in 407. It is adequately powered (82.10% for a hazard ratio of 0.75—see bottom box, 80 to 90% is often the norm for planned studies) given the number of events observed and reports results which are not statistically significant (p-value about 0.1583 in the second box). The study's reported hazard ratio of 0.87 would not be considered of a relevant magnitude. Further, we have computed a proportion of subject pairs selected, one from each group, with the group inferior in the aggregate (Group A) having a larger event free survival period in these pairs at 46.51%. All indicative of very little difference, in this hypothetical context, between the two groups.

In the second scenario, we have one group with 50 events in 112 subjects and another group with 26 in 106 with a meaningful reported ratio of hazards of 0.65. Further our flipped proportion is lower at 39.53%. These reflect, relative to scenario 1,the possibility of a reasonable effect despite the non-significant p-value of 0.0639. One might give some credence to this result despite the lack of statistical significance when the study report indicates a lack of bias and a disciplined path through the data. The non-significant p-value is likely due to the smaller number of events, less than a fourth of those in scenario 1 and thus having a computed power to detect a ratio of hazards of 0.6, at a low 60.82%. Scenario 3 has an unimpressive reported hazard ratio of 0.85. However, the event rates are more than twice those in scenario 1 resulting in a statistically significant p-value of 0.0219, which is not surprising as there is a high 97.66% power to detect a hazard ratio of 0.75. The discordant proportion is 45.87%. All indicative of a marginal difference between groups.

The remaining three tabs of the spreadsheet provide a similar set of three scenarios for survival, dichotomized binomial response endpoints and correlations. As in the continuous case, the first scenario presents an inconclusive result which we find is adequately powered, the second presents a result which we find is inadequately powered to detect differences but demonstrates the likelihood of effect despite the non-significant p-value, and the third scenario presents marginal differences triggering a statistically significant effect due to large cohort sizes leading to an overly sensitive discernment of effect.

INPUTS: DESCRIPTIVE STUDY DATA			
	Scenarios		
	1	2	3
Number of Events in Group A (With the Anticipated Larger Hazard of an Event)	217	50	407
Number of Events in Group B (With the Anticipated Smaller Hazard)	191	26	360
Reported Hazard Ratio (Group A to Group B)	1.15	1.53	1.18
Sample Size for Group A	431	112	804
Sample Size for Group B	407	106	822
Hazard Ratio Computed as Group B to Group A	0.87	0.65	0.85
Proportion of Subjects in Group A	51.43%	51.38%	49.45%

APPROXIMATE OBSERVED STUDY RESULTS			
	Scenarios		
	1	2	3
P-value for Two-Sided Test	0.1583	0.0639	0.0219
Estimated Proportion of Pairs Selected, one from each Group, with the Group A Subject having the Larger Time-to-Event.	46.51%	39.53%	45.87%
Estimated Proportion of Pairs Selected, one from each Group, with the Group B Subject having the Larger Time-to-Event.	53.49%	60.47%	54.13%

POST-HOC EVALUATION OF THE STUDY			
	Scenarios		
	1	2	3
Ratio of Hazards Considered Worth Detecting (Group A to B)	1.33	1.67	1.33
Two Sided Significance Level of the Test for a Hazard Ratio ≠ 1.0	5.00%	5.00%	5.00%
Ratio of Hazards Above Computed as Group B to A	0.75	0.60	0.75
Power Associated with Difference Considered Worth Detecting	82.10%	60.82%	97.66%
Likely Proportion of Pairs Selected, one from each Group, with the Group A Subject having the Larger Time-to-Event (for the Ratio of Hazards Considered Worth Detecting)	42.92%	37.45%	42.92%
Likely Proportion of Pairs Selected, one from each Group, with the Group B Subject having the Larger Time-to-Event (for the Ratio of Hazards Considered Worth Detecting)	57.08%	62.55%	57.08%

Fig. 3 Post-hoc assessment of reported results in time-to-event studies.

2.4 Additional Note on Correlations

A Pearson 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. In Cohen [12] correlations of 0.1 to 0.3 are small, 0.3 to 0.5 are medium and greater than 0.5 are deemed large. In our calculator [11] we have analogously used these intervals to classify the difference between correlation coefficients under the null and alternate hypotheses. It can be shown, that for a close to ellipsoidal scatter plot of data, a good percent of subjects will show an association in the opposite direction from that indicated by the aggregate correlation coefficient. For instance, for a correlation coefficient of 0.55, which would be considered "large", there would be a 31.46% discordance rate. 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. As discussed at a greater length in Refs. [2, 13], many conclusions are

likely to have ecological fallacies making them inappropriate for prescriptive recommendations at an individual level.

3. Expected p-Values: Below Analyst Expectations

When quarterly results for public corporations are announced, the impact of the reported earnings per share, revenue growth and other such key financial summaries on movement in share prices depend more on how these compare to analyst expectations rather than absolute results. One often sees results which appear to be quite impressive and yet there is often a drop in stock prices when results are released. The market had moved to a larger price level based on some kind of a consensus analyst valuation, leading to a correction, as the reported results, though impressive, were below these expectations. A similar expectation can be derived for planned studies.

Planned studies are sized based on limits to the false positive rate, typically to a two-sided 5%, and on the power or the ability to detect aggregate effect, which usually varies from 80 to 90%. Power is the chance of seeing a reported p-value less than 0.05, indicative of a statistically significant effect. So, with 50% power one would tend to hit 0.05 and be lower or higher 50% of the time—an expected p-value of 0.05. With 80% or higher power we would expect to hit a much lower value as we now could have p-values greater than 0.05 only about 20% of the time with smaller values 80% of the time—an expected p-value much lower than 0.05. The expected p-value is obtained as $p = k * \left[1 - \Phi\left(Z_{\alpha/k} + Z_{\beta}\right)\right]$, where $k = 1$ or 2 depends on whether the test is one or two-sided. $Z_{\alpha/k}$ and Z_{β} are values of the standard normal distribution associated with probabilities equal to the subscript in the right tail of the distribution. α and $(1-\beta)$ are the false positive rate and power at the planning stages. For a discussion of expected p-values see Ref. [14]. This expectation at the planning of a study can be used to roughly gauge a study's eventual aggregate estimated effect relative to the anticipated effect. The screen shot of a calculator in Ref. [15] displayed in Fig. 4, considers a study planned to provide 85% power to detect aggregate effect using a two sided test at a 5% significance level. The expected p-value of 0.00273 is less than a tenth of what is adequate to claim a statistically significant aggregate signal and is an approximate benchmark against which we can evaluate reported p-values. This kind of an assessment holds for group sequential trials having interim analyses with futility assessments and no stopping for efficacy. Further, it can be used to gauge expected effect in trials allowing for early termination for effect if the trial does not terminate early. For reported trails which were terminated earlier one can instead gauge aggregate and individual effect using the online calculator in Ref. [11] noted earlier.

Input: Statistical Design Parameters	
Significance Level	0.050
Two-sided or 1-Sided Test	2
Power	85.0%

Output	
Expected P-Value	0.00273

Fig. 4 Expected p-values.

4. Assessment of Dichotomous Response

Often one sees evaluations of effect using improvements by more than a threshold value at a specific time-point or at any time over a schedule of assessments. We will see that these endpoints can be problematic even when the thresholds represent an MID (minimally important differences). We use half a standard deviation for the MID, often seen to be relevant [16]. Flagging subject response based on the crossing of a threshold at any time over a large number of assessments is particularly susceptible to overstatements of effect. This is like evaluating if a subject is happier or healthier at any time over 5 or 10 visits—likely a lot of people will be.

4.1 Calculator for Assessment of Response Classifications

We evaluate response criteria based on improvement by various threshold values on measures from study baseline to post-baseline at this online calculator [17]. The estimated proportion of subjects likely to be deemed to have responded based on such thresholds is provided for user specified aggregate mean improvements post-baseline (in MID units). We look at subject level improvement thresholds in multiples M of the MID on the measure of interest and evaluate the proportions of subjects crossing these improvement thresholds given aggregate effect. Details on computations are in Appendix 3.A screen shot of the calculator is in Fig. 5.

INPUTS			
	Scenarios		
	1	2	3
Aggregate Mean Improvement post-baseline in MID units (**Delta**)	0	0.5	-0.5
Correlation between baseline and post-baseline measurements	0.7	0.7	0.7
Number of Periodic Measurements (**k**) post-baseline	6	6	6
Number **M** of MID units to be used to assess a change as a 'Response'	1	1	1
Measurements from Baseline to Endpoint (in MID units)	2.020	2.020	2.020

Assessments: Immediate Step Change by Delta Post-Baseline			
Estimated Proportion with an Improvement at time k larger than M MIDs	31.03%	40.22%	22.89%
Estimated Proportion with an Improvement anytime at or before time k larger than M	89.23%	95.44%	78.97%
Estimated Proportion with a Deterioration at time k less than M MIDs	31.03%	22.89%	40.22%
Estimated Proportion with a Deterioration anytime at or before time k less than M MIDs	89.23%	78.97%	95.44%

Assessments: Linear Change to Delta at k^{th} Post-Baseline Measure			
Estimated Proportion with an Improvement at time k larger than M MIDs	31.03%	40.22%	22.89%
Estimated Proportion with an Improvement anytime at or before time k larger than M	89.23%	93.37%	83.86%
Estimated Proportion with a Deterioration at time k less than M MIDs	31.03%	22.89%	40.22%
Estimated Proportion with a Deterioration anytime at or before time k less than M MIDs	89.23%	83.86%	93.37%

Fig. 5 Assessment of dichotomized response criteria.

The first box of the online calculator [17] allows user input of data. The second box assesses subject level response rates given an immediate step change post-baseline due to the intervention being studied. The third box assesses subject level response rates given a gradual linear change post-baseline due to the intervention. We examine three scenarios. In the first scenario, there is no net effect due to the intervention.

The aggregate mean improvement is 0. The second row allows input of a correlation between the baseline and the post-baseline measures. We entered 0.7 as a convenience, as this results in a standard deviation of the change post-baseline of the same magnitude as that for the baseline measure—about 2 MID units. We

enter 6 periodic measures post-baseline and a 1 MID subject level threshold for change, to assess the subject as having a response. This results in an estimated 31.03% response rate "at the last visit" despite the null net effect.

This brings back our little paradox with the large proportion discordant to an aggregate effect. If a mean and distribution framework were "real", then one would discount the 31.03% response rate as unreal and arising out of random variation around the mean. A scientist having data with close to a null aggregate effect, will, if he pores through individual subject records, see changes for about 30% of his subjects which are of a relevant and important magnitude and hence "real". An 89.23% estimated response rate, with this null net effect, is likely if we flag response based on the crossing of the threshold "at any time". Now, this might be something that the scientist will likely see as unreal, as those flagged as having response may not have an effect that persists over the periodic assessments. A requirement of persistence of response is similar to requiring multiple thresholds to be met concurrently to be deemed a responder. We will discuss more stringent thresholds and multiple thresholds in the next section. Deterioration statistics "at last visit" or "at any visit", which one would not typically see reported, have the same estimated proportions under null net effect.

In scenario 2 we consider an aggregate effect of 0.5 MID. The proportion having a response "at the last visit" goes up to 40.22%. This is a little deceptive in contexts without a parallel control. In such contexts one should perhaps compare against a putative control rate of 31.03%—the rate we obtained in scenario 1 when there is no net effect post-baseline. In scenario 3, with an aggregate worsening by 0.5 MID, we see a 22.89% response rate "at last visit" and a 78.97% response rate when we look at response "at any time". The numbers in the third box of the calculator are somewhat lower for scenario 2 and higher for scenario 3 for the "at any time" assessment as the change occurs gradually. Note that high within subject pre versus post correlations on a measure help, and lower correlation results in increases in subjects crossing the threshold. You can enter a value of 0.2 instead of the 0.7 for the correlation at the online calculator [17] to see this effect.

4.2 Some Observations on Constructing a Response Classification

If you increase M to a larger number than the default 1 MID, you make the threshold more conservative, resulting in a lower "zero error" proportion when there is null net effect. However, it is likely that there will be loss of information with too conservative a threshold in certain contexts where interventions are only moderately effective. Cut-off thresholds can be obtained using a "gold" standard on what constitutes response, through assessments of tradeoffs between specificity and sensitivity as we vary thresholds [18]. Even here, if the obtained thresholds do not reflect marked effects, one could convert the threshold to MID units and evaluate the null effect proportions and perhaps the symmetric "deterioration" assessment. Stronger response assessment can also be constructed by using a composite requiring a threshold on multiple measures.

The ACR (American College of Rheumatology) 20, 50 and 70% improvement measures [19] require improvements on tender joint counts and swollen joint counts and three of five other measures, before a patient with rheumatoid arthritis is deemed to have had a response. What strengthens the presentation of response in this context is the use of increasing thresholds for improvements from 20 to 70% as well as the requirement that the thresholds be met on multiple measures. Notice the "and" in the assessment —you do not want to see "or" as that will result in a weaker threshold than the use of the constituent measures. Note that the constituent measures should ideally provide additional independent information. There are moderate correlations across the ACR

measures making it likely that there is some residual response rate triggered even with stable disease post-baseline, especially with the ACR composite with the lower 20% threshold. A number of studies, which allow the use of a less effective older standard such as methotrexate in the placebo group, report placebo response rates on ACR20 of about 15% [20]. Likely one may see ACR(Minus20), defined analogously, looking at deterioration instead, reporting a similar 15% rate in these control groups if methotrexate lacks net effect.

5. Discussion

We present subject level as well as aggregate criteria for the assessment of reported results. Strong results in the aggregate, supported by extreme p-values, are often an artifact of large cohort sizes detecting small effects. Even when these results reflect meaningful effect in the aggregate, they often associate with a very large discordant effect when evaluated at a subject level. Note that one of the earlier exponents of the use of the p-value, Sir Ronald Fisher [21], when describing it through his tea tasting experiment, evaluated it in the context of a single subject. Our difficulties seem to arise with the use of this notion in inferences about aggregates over subjects.

When evaluating survival data we noted that the discordant proportion obtains as a simple ratio using the ratio of Hazards HR, as HR/(1+HR), under the usual proportional hazards model and is independent of the sample size or obtained p-values. We had noted a subject level discordance with an aggregate Hazard ratio of 0.58 supporting effect, of 0.58/(1+0.58) = 36.7%.One could attribute a patient's lower survival despite being given the therapy superior in the aggregate to the argument that we may be looking at sicker patients in the lower tail of the distribution of survival times for the superior cohort. This presumes that those in the lower tail of the distribution of the inferior cohort are not sicker. Further, Forrest plots looking at effects in numerous subsets, including sick subsets, often obtained using risk criteria derived from a composite of assessments of susceptibility to the disease (see Ref. [22] for the IMWG risk criteria in Multiple Myeloma), result in hazard ratios in the same ball-park as the hazard ratio computed across all subjects, or worse. Such subset hazard ratios, if different from those for the entire cohort, can often be artifacts of the multiple assessments across multiple subsets [23].

We could argue bad luck in the roll of the die and always advocate the test therapy. A variability argument based on assessed aggregate distributional parameters and the ordering of a superior test group to a control. Here we may be looking at marginal instead of a more predictive conditional distributions using all information in patient profiles. A similar discordance rate assessed as a complement of the concordance index [24], which inspires our discordance proportion in our simpler contexts here involving cohort comparisons, reduces discordance somewhat but persists in large magnitude in complex predictive models using such complete patient information. The Framingham Heart Study Model (2002 version) had a concordance probability of about 70% [25]. We will often be left supporting the conjecture that the large discordant percentages reflect real discordance and there is a lack of effect, often reversed, to the "better" option for a number of patients. We need evaluations of patient profiles, as well as synergistic and antagonistic interactions of these patient characteristics with therapeutic options.

Similar subject level discordance was demonstrated in the use of the correlation coefficient and for continuous and discrete data. In all cases the discordance rate was independent of p-values and sample sizes, both of which are seen as problematic when evaluating statistical inferences [3]. We hope to help the reader assess if a reported scientific finding based on stochastic data represents adequate effect in the aggregate, and when it does, to evaluate the extent

to which an option flagged as the better option in the aggregate will work for the reader or for someone to whom the reader might refer the finding. Whether it amounts to something? Whether it is a call for change in our lives? Whether these reported signals are large enough to warrant societal change? We might often conclude, that any social rule, action or guidance, based on reported scientific results using stochastic data, should be governed by the presumption of exception. When there is a need to act, one should usually pivot to addressing and accommodating exceptions, rather than an overly rigid adherence to the rule.

References

[1] 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.

[2] Srinivasan, S. 2018. "Inverted Simulations Demonstrating Strong Ecological Fallacies in Cohort Studies." *Journal of Mathematics and System Science*8 (5): 119-39.

[3] Ronald, L. W., and Nicole, A. L. 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.

[4] 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.

[5] Allison, P. D. 2010. *Survival Analysis Using SAS: A Practical Guide* (2nd Edition). SAS® Institute, Cary, NC.

[6] Elisa, T. L., and John, W. W. 2003. *Statistical Methods for Survival Data Analysis* (3rd Edition). Wiley-Inter Science.

[7] Ott, L.1984. *An Introduction to Statistical Methods and Data Analysis.* Boston, Massachusetts: PWS Publishers.

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

[9] Shein, C. C., Hansheng, W., and Jun, S. 2007. *Sample Size Calculations in Clinical Research* (2nd Edition). Boca Raton, Florida: Chapman & Hall/CRC.

[10] *Cytel EAST 5 User Manual.* 2010.

[11] https://resourcetepee.com/evaluation-of-reported-statistical-inferences/post-hoc-assessment-of-reported-study-results/.

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

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

[14] Proschan, M. A., Lan, G. K., and Wittes, J. T. 2006. *Statistical Monitoring of Clinical Trials—A Unified Approach*. New York, NY: Springer Science + Business Media, LLC.

[15] https://resourcetepee.com/evaluation-of-reported-statistical-inferences/expected-p-value-calculator/.

[16] 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.

[17] https://resourcetepee.com/evaluation-of-reported-statistical-inferences/assessment-of-dichotomous-response/.

[18] Mithat, G. 2007. *Receiver Operating Characteristic Curves with SAS®* . SAS® Institute, Cary, NC.

[19] Felson, D. T., Anderson, J. J., Boers, M., Bombardier, C., Furst, D., and Goldsmith, C.1995. "American College of Rheumatology. Preliminary Definition of Improvement in Rheumatoid Arthritis."*Arthritis Rheum*38 (6): 727-35.

[20] van Vollenhoven, R. F. et al2011. "ACR Hybrid Analysis of Certolizumab Pegol Plus Methotrexate in Patients with Active Rheumatoid Arthritis: Data from the RAPID 1 Trial." *Arthritis Care Res (Hoboken)* 63 (1): 128-34.

[21] Fisher, R. A. 1971. *The Design of Experiments* (9th Edition). Macmillan.

[22] Chng, W. J. et al 2014. "IMWG Consensus on Risk Stratification in Multiple Myeloma." *Leukemia* 28: 269-77.

[23] Rothwell, P. M. 2005. "Subgroup Analysis in Randomized Controlled Trials: Importance, Indications, and Interpretation." *The Lancet* 365 (9454): 176-86.

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

[25] 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.

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

Appendix 1: Computations of the Discordant Proportion

The probability of subjects *i* and *j* in Group S (Standard) and Group N (New) respectively, having a higher measured outcome for Group S despite an aggregate trend favoring Group N is given by

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

where the CDF (cumulative density functions) for the two subjects are $F_{iS}(s)$ and $F_{jN}(n)$ respectively. When the measured outcome has identical and independent distributions for subjects within a group then we can note that for any randomly chosen pair of outcomes, the Group A measure will be larger with a probability given, on dropping the subject subscripts in the CDFs, by

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

This flipped proportion can be computed based on actual subject level data from a study by tallying the number of subjects in Group A having a lower outcome than each subject in Group B and dividing the sum of these tallies across Group B subjects by the product of the Group A and Group B sample sizes. The likely proportion can be estimated when subject level data is not available post-hoc, by forcing distributional assumptions, or using resampling from estimated distributions, separated appropriately, if other historical data is available.

Continuous Data

For continuous data we use independent identical within group normal distributions to compute the flipped proportion in the calculator. For subjects *i* and *j* in Group S and Group N respectively, the distribution of the difference in measures is given through the Group means and variances by

$$X_i - X_j \sim N\left\{\mu_S - \mu_N, \sqrt{(\sigma_S^2 + \sigma_N^2)}\right\}$$

The estimated probability of the flipped proportion can be obtained as

$$P(X_i - X_j > 0) = \Phi\left\{(\mu_S - \mu_N)/\sqrt{(\sigma_S^2 + \sigma_N^2)}\right\}$$

where, Φ is the CDF of the standard normal distribution. When computing this estimate for observed data we can use the sample means and the sample standard deviations S_S and S_N. Forthe post-hoc evaluation we use the sample standard deviations and the mean differences worth testing instead of the sample means.

Survival Data (Exponential Case)

For survival data we use independent identical within group exponential distributions to compute the flipped proportion in the calculator. The text of the manuscript derives the proportion under proportional hazards assumptions. For subjects *i* and *j* in Group A (Standard) and Group B (New) respectively, the distribution of the differences $Y = X_i - X_j$ in the times to event is given by the Laplace distribution whose CDF is given by

$$F_Y(y) = [\lambda_S/(\lambda_S + \lambda_N)]EXP(y * \lambda_N) \text{ for } y \leq 0 = [\lambda_N/(\lambda_S + \lambda_N)]EXP(y * \lambda_S) \text{ for } y > 0$$

For a hazard ratio $\eta = \lambda_N/\lambda_S$,

$$P(Y = X_i - X_j > 0) = \lambda_N/(\lambda_S + \lambda_N) = \lambda_S * \eta/(\eta * \lambda_S + \lambda_S) = \eta/(\eta + 1)$$

When computing this estimate for observed data we can use the reported hazard ratio. For the post-hoc evaluation we use the hazard ratios worth testing instead of the reported hazard ratios. Note that the estimate is identical to that in the text of the document obtained using the less restrictive proportional hazards assumption.

Binomial Data

The estimate of the percent chance of individual responses in Group S being better than individual responses for Group N is obtained

from the proportions responding as $P_S(1-P_N)$. Similarly, $P_N(1-P_S)$ for Group N subjects better than S. The estimate of the percent chance of individual responses in the two groups to be the same is

$$P_S P_N + (1-P_S)(1-P_N)$$

Finally, the estimate of the percent chance of individual responses in the Group S (inferior in the aggregate) being the same or better than individual responses for Group N is the sum of the first and the third estimates above. When computing this estimate for observed data we use the reported proportions. For the post-hoc evaluation we use proportions reflecting a difference worth testing.

Correlation Data

The proportion of subjects with data discordant with aggregate correlation coefficient is given by $cos^{-1}(\rho)/\pi$. This derivation is provided in Appendix 3 in Ref. [2].

Appendix 2: p-Value and Post-Hoc Power Calculations

This appendix contains details on computing likely approximate p-values for inferential comparisons across groups as a check on reported p-values for the continuous, binomial and correlation contexts. The expression for time-to-event data is in the text of the manuscript. Expressions for the post-hoc power to detect clinically meaningful differences is also provided, given cohorts that were available in studies for which we have reported results.

Continuous Data

For continuous data we compute the pooled standard deviation (first box of calculator) using the sample sizes and standard deviations in Group S (Standard) and Group N (New) as

$$S_P = \sqrt{\{(N_S S_S^2 + N_N S_N^2)/(N_S + N_N)\}}.$$

Under approximate normality and independence, the student's-t statistic can be computed using the difference in observed sample means as

$$t = (\bar{X}_S - \bar{X}_N)/\big(S_P(1/N_S + 1/N_N)\big) \sim t_{N_A+N_B-2}$$

The two-sided p-value then obtains as $2 * \big(1 - \Phi(|t|)\big)$, where Φ is the cumulative density function for the student's-t distribution above. We use half the pooled standard deviation as a measure of the MID (minimal important difference). The power is obtained by using differences in means considered worth detecting using the expression below. For the true pooled standard deviation σ_P, one can still use S_P.

$$\Phi\Big\{|(\mu_S - \mu_N)|/\big(\sigma_P(1/N_S + 1/N_N)\big) - t^{\alpha/2}_{N_S+N_N-2}\Big\}$$

Binomial Data

For binomial data the p-value can be obtained using the estimated proportions $\hat{P}$ and the post-hoc power through differences in proportions *P* worth detecting using

$$p = 2 * \left\{1 - \Phi\left[|(\hat{P}_S - \hat{P}_N)|/\sqrt{\hat{P}_S(1-\hat{P}_S)/N_S + \hat{P}_N(1-\hat{P}_N)/N_N}\right]\right\} \text{ and}$$

$$\text{power} = \Phi\Big\{\big[|(P_S - P_N)|/\sqrt{P_S(1-P_S)/N_S + P_N(1-P_N)/N_N}\big] - Z_{\alpha/2}\Big\}.$$

Correlation Data

The p-value associated with a reported sample correlation coefficient *r*, when rejecting a null correlation of ρ_0, is obtained by using the approximate normality for large sample sizes ($n \geq 25$) of the statistic above[26].

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

Then to reject a null correlation of ρ_0 we would compute the Wald Statistic:

$$Z = 0.5 * SQRT(n-3) * \left[Ln\frac{1+r}{1-r} - Ln\frac{1+\rho_0}{1-\rho_0}\right]$$

The p-value can then be obtained as using $p = 2 * [1 - \Phi(|Z|)]$, with $\Phi(.)$ being the cumulative distribution function of the standard normal. The power associated with a two-sided α level test of a correlation coefficient of ρ_1 considered meaningfully different from ρ_0 is given by

$$\Phi\left\{\left[0.5 * SQRT(n-3) * \left|Ln\frac{1+\rho_1}{1-\rho_1} - Ln\frac{1+\rho_0}{1-\rho_0}\right|\right] - \Phi(1-\alpha/2)\right\}$$

Appendix 3: Details on the Dichotomized Response Assessment Calculator

In this calculator we evaluate responder criteria based on improvement by various threshold values on measures from study baseline to post-baseline. The estimated proportion of subjects likely to be deemed responders based on such thresholds is provided based on the aggregate mean improvement post-baseline. We look at improvement thresholds in multiples M of the minimally important differences (MID) on the measure of interest. The MID is often seen to be about half a standard deviation of the measure [17]. Hence if we standardize the data to MID units we obtain a standard deviation of the measure σ_m = 2 in MID units. The standard deviation of the improvement σ_δ is obtained using the correlation between baseline and post-baseline as

$$\sigma_\delta = 2 * \sqrt{(2 - 2\rho^2)}$$

We will consider probabilities of improvements crossing the threshold at different periods assuming: (1) An immediate aggregate mean improvement μ_δ post-baseline, which stays stable and; (2) A linear change to the specified aggregate mean improvement μ_δ at the last post-baseline visit.

In the first context, the probability of an improvement crossing the threshold at the i^{th} periodic post-baseline measurement is obtained approximately through the cumulative distribution function Φ of a normal distribution with mean μ_δ and standard deviation σ_δ as

$$p_i = 1 - \Phi\{M\}$$

In the second context, for the i^{th} of k post-baseline periodic assessments this probability is obtained in a similar manner for mean $(i * \mu_\delta)/k$ and standard deviation σ_δ. In both cases we can compute probabilities of deteriorations of the same multiple M of the MID as

$$p_i = \Phi\{-M\},$$

and the probability of crossing the threshold, for the improvement or deterioration context using the correspondingp_i at any time in k assessments by

$$p = 1 - \prod_{i=1}^{k}(1 - p_i).$$