Perspective

Real Life Outcomes vs. Clinical Trial Results

Marco Zarbin, MD, PhD, FACS

Institute of Ophthalmology and Visual Science, Rutgers-New Jersey Medical School, Rutgers University, Newark, New Jersey, USA

ORCID:
Marco Zarbin: https://orcid.org/0000-0002-7811-7132

Abstract

Two questions that clinicians should answer as they attempt to apply the results of clinical trials to clinical practice are: (1) will statistically significant results be reproduced in their clinical practice; and (2) if nothing goes wrong in a clinical trial, is everything alright? Regarding the first question, when considering the results of a randomized, multicenter, prospective, controlled clinical trial, two questions that cannot be addressed by simply by reading the trial results and that only the practicing clinician can answer are: (1) is the study population representative of the patient about to be treated; and (2) is the totality of evidence outside the trial (including the clinician's own extensive experience) consistent with the trial result? Regarding the second question, clinicians are advised to recognize that most studies, even Phase 3 trials, are underpowered to accurately assess the risk of low frequency events.

Keywords: Clinical Practice; Clinical Trial; Confidence Interval; *P* value; Rule of Three; Statistical Significance

J Ophthalmic Vis Res 2019; 14 (1): 88-92

INTRODUCTION

Application of clinical trial results to clinical practice often is not straightforward. Issues such as restrictive enrollment criteria,[1] experimental design limitations,[2] conflicts of interest (both financial and non-financial),[1] publication bias,[3] and biological variability all can underlie the disparity between the outcomes achieved in clinical trials vs. those achieved in clinical practice.[4] Two questions pertaining to this challenge are: 1) will statistically significant trial results be reproduced in one's practice?;[5] and 2) if nothing goes wrong in a trial, is everything alright?[6] Approaches to managing these challenges are explored herein.

Correspondence to:
Marco Zarbin, MD, PhD, FACS. Institute of Ophthalmology and Visual Science, Rutgers-New Jersey Medical School, 90 Bergen Street, Room 6156, Rutgers University, Newark, New Jersey 07103, USA.
E-mail: zarbin@earthlink.net

Received: 20-11-2018 Accepted: 25-12-2018

Access this article online

Quick Response Code:

Website: www.jovr.org

DOI: 10.4103/jovr.jovr_279_18

WILL STATISTICALLY SIGNIFICANT TRIAL RESULTS BE REPRODUCED IN MY PRACTICE?

Reproduce vs. Replicate

Replicability refers to the ability to obtain an identical result when an experiment is repeated under precisely identical conditions.[7] Replicability is likely to be inversely proportional to the number of experimental variables. The stochastic nature of biological processes also decreases the probability of replicating a result even if all the variables in one clinical trial are reproduced in a

How to cite this article: Zarbin M. Real life outcomes vs. clinical trial results. J Ophthalmic Vis Res 2019;14:88-92.

second trial. The patients in the second trial, for example, are not likely to be identical to those in the first trial. Despite similar demographics and disease parameters, the outcomes in the VIEW1 and VIEW2 studies were not identical.[8] Frequentist statistical analysis, which typically is used to analyze clinical trial data, generally assumes replicability of a trial result.

Reproducibility refers to the property that a result will recur even if experimental conditions vary to some degree from experiment to experiment.[7] From the perspective of a practicing clinician, if the findings of a clinical trial depend heavily on replicating the precise experimental conditions of the trial in one's practice, then the result might be considered idiosyncratic and less important compared to a result that can be reproduced by a variety of independent, non-identical approaches. Controlling variables to achieve greater reproducibility of a trial result often will cause the trial to diverge from the heterogeneity of true clinical practice and thus compromise the external validity and importance of the results.[7] Clinicians generally are interested in the reproducibility of a trial result rather than in its replicability.

The importance of reproducibility increases in proportion to the importance of the experimental results, especially if the results challenge established paradigms.[7] This perception arises from the fact that experimental hypotheses cannot be proved, but they can be refuted.[9] Unfortunately, both in clinical science[10,11] and in basic science,[12,13] poor reproducibility of experimental results is an ongoing and important problem.

A Clinician's Approach to Assessing Reproducibility

Although policy makers can attempt to avoid the risks associated with generalizing the results of a single study by requiring at least 2 adequate well controlled trials before accepting the result, readers of clinical trial results usually do not have the resources to reproduce the trial in their practice in a controlled manner. One approach to this dilemma of the practicing clinician is to answer 5 questions related to the trial [Table 1].[5] The premise underlying this "5-question test" is that the likelihood that a single research study's findings are "true" (i.e., reproducible) depends on: 1) the prior evidence; 2) the design of the study; and 3) the level of statistical significance.[14,15] Regarding the issue of statistical significance, if the type 1 error rate (α) is set at 0.05 (i.e., in a 2-tailed test of the null hypothesis that there is no difference between two treatments, we are willing to reject the null hypothesis even if it is true provided that the result is likely to occur 2.5% or less of the time), then $P = 0.03$ does not give the same degree of confidence in rejecting the null hypothesis as $P = 0.0001$ although both are "statistically significant". It is important to note that the *P* value indicates how likely one is to observe the given trial data if the null hypothesis is true and given the statistical model of the distribution of outcomes. *The P value does not provide information on the validity of the null hypothesis*.

The "5-question test" poses the following questions:[5]

1. Have steps been taken to minimize bias? [design]
   - Concealed treatment allocation, double masking, randomization
2. Is the result likely due to the treatment? [design]
   - Confounding is best controlled by randomization
3. Is the result unlikely to be due to chance? [level of significance]
   - Are enough patients enrolled for a reliable estimate of the treatment effect?
   - Is the tested hypothesis pre-specified or *post hoc*? ***Post hoc*** analyses are less reliable, as they are subject to bias
   - Is the *P* value << pre-specified type 1 error (α)? $p<<\alpha$ affords greater confidence to reject the null hypothesis (vs. $P \approx \alpha$)
   - Does the 95% confidence interval (CI) clearly exclude the null result, indicating the range of outcomes is unlikely to be due to chance? A large CI indicates the estimate of the treatment effect is not precise. If the confidence interval includes the value of 0 (i.e., no difference between average outcome of the compared cohorts), the difference between compared interventions is not statistically significant (although it may be clinically important[5])
4. Is study population in the trial representative of your patient? [design]
5. Is the totality of evidence consistent with the trial result? [prior evidence]

Table 1. Five question test to determine whether statistically significant results of a single clinical study are likely to be reproduced in routine clinical practice[5]

Question	0/5 "Yes"	1/5 "Yes"	2/5 "Yes"	3/5 "Yes"	4/5 "Yes"	5/5 "Yes"
1. Unbiased result?	No	✓	✓	✓	✓	✓
2. Result likely due to treatment (no confounding)?	No	No	✓	✓	✓	✓
3. Result unlikely due to chance?	No	No	No	✓	✓	✓
4. Study population representative?	No	No	No	No	✓	✓
5. Result consistent with prior objective evidence?	No	No	No	No	No	✓
Likelihood of reproducing results in your practice	Very Low	Very Low	Low	Low	Unclear*	High

- The totality of evidence includes findings from relevant previously published studies as well as extensive personal clinical experience.

Details regarding the concepts underlying the five questions have been discussed elsewhere.[5] It is hypothesized that if the answer to all five questions of the 5-question test is "yes", then the result is likely to be reproduced in one's practice. If not, the likelihood of reproducing the result is low. If the answer is "yes" to all questions except the last (#5, Table 1), then reproducibility is unclear and depends on the strength of prior evidence. If the prior evidence is strong (e.g., multiple randomized, multicenter, controlled trials) and the current trial result is inconsistent with previous studies, then likelihood of reproducing the current result in practice is probably low. If the prior evidence is weak (e.g., uncontrolled case series) and the current trial design is robust (e.g., randomized, multicenter, controlled trial), then the current result is more likely to be reproduced in practice.

Fortunately, the answer to questions #1-3 above is almost always "yes" in prospective, randomized, controlled, multicenter clinical trials. In this circumstance, it remains for practicing clinicians to focus their attention primarily on questions #4 and 5. Therefore, let us consider two examples that illustrate the importance of the latter questions.

Totality of Evidence

The Diabetic Vitrectomy Study (DRVS) assessed the value of early vitrectomy for severe non-clearing vitreous hemorrhage in patients with diabetic retinopathy and best corrected vision ≤5/200 for at least one month.[16] The results indicated that at year-2, early vitrectomy was better than deferral (P = 0.0001) only for patients with type 1 diabetes mellitus, not for those with type 2 disease. The DRVS was unbiased, randomized, multicentered, and tested a large number of patients (~500) typically found in clinical practice, but the result was not consistent with the totality of evidence. Evidence from practice strongly suggested that the complication rate (e.g., 20% no light perception rate) would be lower in practice than was observed in the study. As a result, most surgeons did not withhold early vitrectomy for severe non-clearing vitreous hemorrhage for patients with type II diabetes mellitus. It is important to note that this prior evidence did not come from additional randomized clinical trials, but from extensive observations in clinical practice. Current practice outcomes justify this approach. The totality of evidence is quite important.

Composition of the Study Population

Protocol S was a multicenter, randomized, prospective clinical trial to evaluate whether intravitreal ranibizumab was non-inferior to panretinal photocoagulation (PRP) for visual acuity outcomes in patients with proliferative diabetic retinopathy (PDR).[17] Patients randomized to ranibizumab treatment had a lower risk of retinal detachment vs. those randomized to PRP (12 (6%)/191 eyes vs. 21 (10%)/203, respectively, adjusted difference (95%CI) −4 (−9 to + 1), P = 0.08).[17] The ranibizumab cohort also had a lower risk of needing vitrectomy vs. the PRP cohort (8 (4%)/191 eyes vs. 30 (15%)/203, respectively, adjusted difference (95%CI) -9 (-15% to -4%), P < 0.001).[17] This result was somewhat surprising, as intravitreal anti-vascular endothelial growth factor (anti-VEGF) injections are known to increase the risk of traction retinal detachment in patients with PDR.[18] Patients typically at risk for this phenomenon have a ring of fibrovascular tissue, extensive retinal capillary non-perfusion, and no previous laser treatment [Figure 1].

Among the patients enrolled in Protocol S, 69 eyes (37%) had high risk PDR (ETDRS severity score 71-75 [Figure 2][19]) in the ranibizumab cohort, and 73 (37%) had high risk PDR in the PRP cohort. In the ranibizumab cohort, only 2 (1%) patients had advanced PDR with the macular center attached, and 1 (<1%) had advanced PDR with the macular center detached. One patient in the PRP cohort had advanced PDR with the macula detached. Thus, only 4 (1%) of 394 eyes in Protocol S had a clinically important risk of anti-VEGF induced retinal detachment or exacerbation thereof. As a result, is not entirely clear that in a real world practice setting, patients receiving anti-VEGF therapy will have a lower risk of retinal detachment or need for vitrectomy vs. patients receiving PRP. It is critically important to be sure that the study population is representative of the patient one is treating as one assesses the likelihood of reproducing of a study result in one's practice.

IF NOTHING GOES WRONG IN A TRIAL, IS EVERYTHING ALRIGHT?[6]

Consider, as an example, a hypothetical Phase 3 clinical trial involving 33 patients who receive subcutaneous injections of a new anti-VEGF agent to treat diabetic

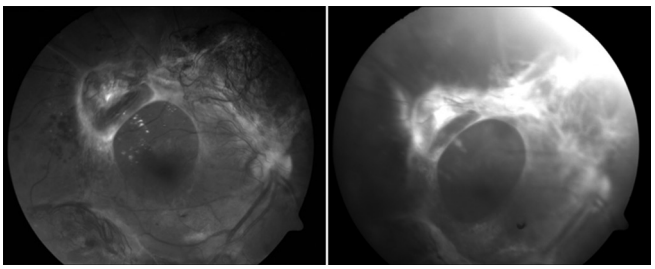

Figure 1. Preinjection photograph showing C-shaped fibrovascular proliferation along the temporal arcades and no laser treatment. One week after intravitreal injection of 1.25 mg bevacizumab. Note evolution of traction retinal detachment along the superotemporal arcade as well as mild vitreous hemorrhage creating mild image blur. Courtesy of Robert Avery, MD.

retinopathy. In this trial, 0 of 33 patients experience a myocardial infarction (MI). What can we conclude about the risk of MI with this new agent? The answer is given by the "rule-of-3", which states that if a certain event did not occur in a sample with n subjects, the interval from 0 to $3/n$ is a 95% CI for the rate of occurrences of the event in the population, assuming the outcomes are independent events with a binomial distribution.[6] The formula for this result is: (1-maximum risk of event) ^{n} = chance of not observing event in n subjects. In other words, if 0 of n patients experiences a complication, one can guess the true chance of the complication occurring is $3/n$ or less, and this guess will be correct approximately 95% of the time. In a Phase 1 trial, for example, if 0 of 15 patients develops an MI, the upper limit of the 95% CI for the rate of MI with this treatment is 20% (3/15), which means the risk of MI could actually be 20%, yet the trial would yield an observed 0 MI rate 5% of the time!

This caveat applies to large scale as well as small studies. Consider for example the results of the ANCHOR and PIER trials, both of which were multicenter, randomized, prospective Phase 3 clinical trials involving the treatment of neovascular complications of age-related macular degeneration with ranibizumab. The event rate for stroke excluding transient ischemic attack (TIA) was 0 in both trials [Table 2]. Application of the rule-of-3, however, yields an estimate of the upper limit of the 95% CI for the ANCHOR study of 2% and for the PIER study 5% [Table 2]. A patient-based meta-analysis of six randomized clinical trials, however, yielded a rate of stroke excluding TIA of 1.6%,[20] consistent with both estimates and close to the estimate of the ANCHOR trial, which enrolled 2.3 times more patients than the PIER study.

The clinical significance of the rule-of-3 is that *zero observed complications does not equal zero risk*. It is difficult to conclude that a treatment is safe and effective if one has a low tolerance for catastrophic outcomes (e.g., stroke, death) unless the sample size is very large. If, for example, one observes 0 MIs in a cohort receiving treatment A, one would require $n \geq 600$ patients in that treatment arm to conclude the true risk of MI with treatment A is $\leq$0.5% (i.e., upper limit of the 95% CI = 0.5%).

Figure 2. Fundus photograph of level 75 disease with high risk proliferative diabetic retinopathy. New vessels at the disk occupy more than one-third of the disc area.

Table 2. Rule-of-3 applied to a pooled safety analysis of ranibizumab in neovascular AMD[20]

Trial	Event	Event Rate (n/N/PY) 0.5mg RBZ	Upper Limit of 95% CI (3/n)
ANCHOR (phase 3)	Stroke excluding TIA	0/140/260	2%
PIER (phase 3b)	Stroke excluding TIA	0/61/107	5%

n, number of subjects with event; N, number subjects; PY, patient-years follow-up

CONCLUSIONS

Two questions that clinicians should answer as they attempt to apply the results of clinical trials to clinical practice are: (1) will statistically significant results be reproduced in their clinical practice?; and (2) if nothing goes wrong in a clinical trial, is everything alright? Regarding the first question, when considering the results of a randomized, multicenter, prospective, controlled clinical trial, two questions that cannot be addressed by simply by reading the trial results and that only the practicing clinician can answer are: (1) is the study population representative of the patient about to be treated?; and (2) is the totality of evidence outside the trial (including the clinicians own extensive experience) consistent with the trial result? Regarding the second question, clinicians are advised to recognize that most studies, even Phase 3 trials, are underpowered to accurately assess the risk of low frequency events.

Acknowledgements

Supported in part by the Joseph J and Marguerite DiSepio Retina Research Fund, the Eng Family Foundation, and the New Jersey Lions Eye Research Foundation.

Disclosure

Dr. Zarbin has served as a consultant for Cell Cure, Chengdu Kanghong Biotechnology Co., Coherus Biosciences, Daiichi Sankyo, Frequency Therapeutics, Genentech/Roche, Healios KK Inc., Iridex, Isarna Therapeutics, Makindus, Novartis Pharma AG, Ophthotech, and Percept Corp.

Financial Support and Sponsorship

Nil.

Conflicts of Interest

There are no conflicts of interest.

REFERENCES

1. Smith GA. Review of patient inclusion exclusion criteria. Available from: 2013: http://www.fda.gov/downloads/Drugs/NewsEvents/UCM341154.pdf.
2. Macleod M, Lawson McLean A, Kyriakopoulou A, Serghiou S, de Wilde A, Sherratt N, et al. Risk of bias in reports of *in vivo* research: A focus for improvement. *PLoS Biol* 2015;13:e1002273.
3. Kyzas PA, Denaxa-Kyza D, Ioannidis JP. Almost all articles on cancer prognostic markers report statistically significant results. *Eur J Cancer* 2007;43:2559-2579.
4. Holz FG, Tadayoni R, Beatty S, Berger A, Cereda MG, Cortez R, et al. Multi-country real-life experience of anti-vascular endothelial growth factor therapy for wet age-related macular degeneration. *Br J Ophthalmol* 2015;99:220-226.
5. Zarbin MA. Challenges in applying the results of clinical trials to clinical practice. *JAMA Ophthalmol* 2016;134:928-933.
6. Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? Interpreting zero numerators. *JAMA* 1983;249:1743-1745.
7. Casadevall A, Fang FC. Reproducible science. *Infect Immun* 2010;78:4972-4975.
8. Heier JS, Brown DM, Chong V, Korobelnik JF, Kaiser PK, Nguyen QD, et al. Intravitreal aflibercept (VEGF trap-eye) in wet age-related macular degeneration. *Ophthalmology* 2012;119:2537-2548.
9. Popper KR. The Logic of Scientific Discovery. New York, Basic Books; 1959.
10. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 2005;294:218-228.
11. Open Science C. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* 2015;349:aac4716.
12. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. *Nature* 2012;483:531-533.
13. Prinz F, Schlange T, Asadullah K. Believe it or not: How much can we rely on published data on potential drug targets? *Nat Rev* 2011;10:712.
14. Goodman SN. Toward evidence-based medical statistics. 1: The *P* value fallacy. *Ann Intern Med* 1999;130:995-1004.
15. Goodman SN. Toward evidence-based medical statistics. 2. The Bayes factor. *Ann Intern Med* 1999;130:1005-1013.
16. Early vitrectomy for severe vitreous hemorrhage in diabetic retinopathy. Two-year results of a randomized trial. Diabetic Retinopathy Vitrectomy Study report 2. The Diabetic Retinopathy Vitrectomy Study Research Group. *Arch Ophthalmol* 1985;103:1644-1652.
17. Writing Committee for the Diabetic Retinopathy Clinical Research Network, Gross JG, Glassman AR, Jampol LM, Inusah S, Aiello LP, et al. Panretinal photocoagulation vs intravitreous ranibizumab for proliferative diabetic retinopathy: A randomized clinical trial. *JAMA* 2015;314:2137-2146.
18. Arevalo JF, Maia M, Flynn HW Jr, Saravia M, Avery RL, Wu L, et al. Tractional retinal detachment following intravitreal bevacizumab (Avastin) in patients with severe proliferative diabetic retinopathy. *Br J Ophthalmol* 2008;92:213-216.
19. Grading diabetic retinopathy from stereoscopic color fundus photographs--an extension of the modified Airlie House classification. ETDRS report number 10. Early Treatment Diabetic Retinopathy Study Research Group. *Ophthalmology* 1991;98:786-806.
20. Zarbin MA, Francom S, Grzeschik S, Tuomi L, Haskova Z, Macfadden W, et al. Systemic safety in ranibizumab-treated patients with neovascular age-related macular degeneration: A patient-level pooled analysis. *Ophthalmol Retina* 2018; 2:1087-1096.