

Stability of relative and absolute metrics: empirical evidence from pulmonology

Mátyás Szigeti

Physiological Controls Research Center
Óbuda University
Budapest, Hungary
szigeti.matyas@phd.uni-obuda.hu

Levente Kovács

Physiological Controls Research Center
Óbuda University
Budapest, Hungary
kovacs.levente@nik.uni-obuda.hu

Tamás Ferenci

Physiological Controls Research Center
Óbuda University
Budapest, Hungary
ferenci.tamas@nik.uni-obuda.hu

Abstract—It has been widely argued that absolute treatment effect measurements (such as risk difference) reveal the “clinical benefit” of an intervention. Yet, many previous experience with binary endpoints have shown that they are unlikely to be transportable between populations. As absolute metrics are usually derived from baseline risk and relative metric (such as odds ratio), it seems logical to rather measure relative metrics, assuming they are stable. In the present study, a continuous endpoint was used to assess the stability of both relative and absolute metrics using an empirical data from pulmonology. Results are preliminary due to the low baseline variability, yet, the difference was significantly correlated with the baseline, unlike the ratio, which is in line with previous experience with binary endpoints. Further research is needed to explore the stability with continuous endpoints.

Index Terms—absolute risk, relative risk, treatment effect, clinical benefit

I. INTRODUCTION

Typically, the results of a biomedical investigation can be represented both in relative and in absolute terms [1]. Relative metrics include:

- Relative risk (RR, sometimes also defined as relative risk reduction, RRR)
- Odds ratio (OR)
- Relative rate
- Hazard ratio (HR)
- Percentage change

while the following are absolute metrics:

- Absolute risk reduction, ARR (sometimes also called risk difference, RD)
- Number need to treat (NNT)
- Mean difference (MD)
- Difference in median survival
- Absolute difference in fixed time (e.g. 1 year) survival

The choice of whether to use relative or absolute metrics to measure the outcome is often controversial. Usually absolute metrics are advocated as showing the “clinical benefit” or “clinical advantage” of a treatment or exposure as opposed to relative metrics. As a recent example, Heneghan et al wrote:

The research was supported by the HU-MATHS-IN project number: EFOP 3.6.2-16. This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 679681).

”[r]elative measures can exaggerate findings of modest clinical benefit and can often be uninterpretable, such as if control event rates are not reported” [2].

The rationale of this claim can be particularly clearly illustrated for binary endpoints. Assuming that the treatment reduces the frequency of the endpoint from 2% to 1%, we can say that from 100 treated patients, 98 will be treated unnecessarily, 1 will be treated in vain, and we will achieve a result for 1 patient. Yet, all 100 is exposed to the side-effects, meaning that the rational choice on whether to apply that treatment for a particular patient depends on comparing the consequences of the endpoint to *100 times* the side effects of the therapy. In that sense, clearly the absolute metric (0.02 – 0.01) and not the relative (0.01/0.02) was relevant. Also note that absolute metrics allow us to compare different outcomes: if a treatment reduces an endpoint from 2% to 1%, but increases another from 0.1% to 0.6%, 0.5 and 6 (the relative metrics) cannot be compared, but 0.01 and 0.005 can be compared. (I.e. among 1000 patients, we will prevent 10 from the first endpoint, but cause 5 from the second – in other words, if everything else is neglected, the patient is better off with the treatment if 2 from the first endpoint worth 1 from the second. This is, of course, no longer a statistical question, but the use of absolute metrics helps us to formulate the right question.)

It also worth mentioning that the absolute indicator is usually a product of the relative indicator and the baseline risk, i.e. the risk of the control group. (“Product” is not necessarily understood in the mathematical sense of the word, but simply means “derived from”. However, for binary endpoint, it is literally a product: if p_C denotes the risk in the control group and p_E is the risk in the exposed group, then $ARR = p_C RRR$, if ARR is defined as $p_T - p_C$ and RRR is defined as $p_T/p_C - 1$, as typical.) This might seem trivial, but has an important consequence: it immediately makes the *true* reason clear why statin is likely not given to a healthy 25 year old women, but is likely given to a diabetic 65 year old men with prior myocardial infarction. Not that we assume that the drug is effective in one case but not in the other, but *even the same* effect (i.e. same relative risk reduction) translates to a drastically different absolute difference in the two case (due to different baseline risks), possibly leading to a situation when

the risk outweigh the benefits in one case, but not in the other.

The case for absolute metrics as the good indicator of clinical benefit seems to be straightforward from the previous example, but in practice, several questions complicate the issue. It suffices to mention two here.

First, in their not well-known, but highly interesting paper, Snapinn and Jiang have shown [3] that absolute metrics can contradict each other: they have demonstrated that for time-to-event endpoints, under certain conditions (such as exponential survival time distribution) the difference in median survival and the difference in fixed (e.g. 1 year) survival change in opposite direction with changing hazard ratio, all else equal. This result directly contradicts the above perception of absolute metrics shown "the" clinical benefit of the treatment. Furthermore, they have also shown that the "cost"-effectiveness of the treatment (where cost is understood in its most general sense: a side effect is also a cost, and not only the financial cost) will depend only on the hazard ratio – i.e. the relative metric. In other words, neither fixed time survival change, nor median survival change had anything to do with whether the patient is better off with the treatment; it depended solely on the hazard ratio.

The second problem is even more directly pertinent to our investigation. As an illustration, consider the highly controversial paper of Diamond and Ravnskov from 2015, titled "How statistical deception created the appearance that statins are safe and effective in primary and secondary prevention of cardiovascular disease" [4]. The authors claim – among others – that the results of statin trials are presented in relative form to intentionally mislead the readers, as the benefit would be minimal when presented in absolute – and therefore clinically relevant – format (e.g. a reduction from 0.76% fatal and non-fatal heart attack rate to 0.35% presented as -54% instead of -0.41 percentage points). As it was subsequently pointed out by one of the authors of the present paper [5] this claim is unfounded – even if we accept that absolute metrics capture the clinical benefit! The problem is that the length of follow-up in the studies cited by the paper are typically in the range of a few years (in the concrete example presented previously, it was 1.9 years in median [6]), while the usual duration of treatment with such drugs, the time during which cardiovascular risk builds up is usually measured in *decades*. This is clear from the previous numbers as well, which show a sub-percentage risk even in the control group (when cardiovascular causes represent 30% of all deaths globally [7], half of which is due to heart attack). If we assume that the relative risk is stable, i.e. is the same when measured from the first few years, then the absolute risk reduction will be immediately much more dramatic when extrapolated to the practically relevant time horizon.

The point here is more general. *If* we assume that the relative metric is stable, then it can be used to "transport" the results to other populations. This is much more general question than extrapolating in time: it also means that the results apply even if the patients had different baseline risk (for instance because their risk profile is different than those

who are in the trial, e.g. they're older or have more comorbidities). That is an important question: trial populations are almost always different from the general population of the patients, sometimes drastically [8]. This is often raised as a criticism against trials, and sometimes used to advocate real-world evidence, i.e. observational studies. However, if our assumption on stable relative metrics is sound, we can report those from the trial and then "put it into context" (or "translate" it, as we previously phrased) when actually deciding on the treatment of a concrete patient by applying the relative metric to his/her baseline risk [9], [10]. In other words, this doesn't refute the notion that absolute metric is the clinically relevant, but says that in a trial nevertheless the relative metrics should be measured, as it is the property of the drug; it is a separate issue to later transform it to the – clinically relevant – absolute measure. Or, as noted statistician Stephen Senn once wrote: " 'Additive at the point of study, relevant at the point of application', ought to be our motto" [11]. Also consider our previous discussion: if the absolute metric is a derived indicator, it is logical that a trial should measure not the composite (affected by factors unrelated to the drug's effectiveness) but rather the component that is the property of the drug, i.e. the relative metric. But again, saying that the relative measure is the property of the drug assumes that it is stable.

The importance of this issue is clear. Should a trial be "representative"? Perhaps not in a sense that the distribution of the variables is the same as in the population, but many would argue that we should "stretch the covariate space" (i.e. include diverse age groups, patients with all possible combinations of comorbidities etc.). Unquestionably this is the safest, because we will then need no assumption on the stability of the metrics, but it is often unfeasible. On the other hand, if this assumption holds then we can use trials, even those enrolling atypical subjects, to infer on the general patient population of the clinical practice, by using relative measures.

But are relative metrics indeed stable?

To our best knowledge, relatively few studies addressed this question empirically. Subgroup analyses, quite often supplementing results in today's trials provide certain evidence (as far as they show consistent effect). Long-term follow-up studies, where results from different time horizons can be compared, are instructive [12], but can only shed light on the time extrapolation issue. A way to directly investigate the stability issue is to compare the relative/absolute effectiveness of the same drug (or drug class) observed in different trials. An important paper from Smeeth et al [13] directly investigated the question by comparing 5 statins from populations with vastly different background risk, finding that the relative metrics are remarkably stable. (And thus, of course, absolute metrics are substantially different.) Several further authors discussed the question using much more comprehensive selection of studies.

These papers, however, all investigated binary endpoints. The present study aims to deepen our knowledge by investigating a continuous endpoint: forced expiratory volume in 1

second (FEV1) in patients suffering from chronic obstructive pulmonary disease (COPD). COPD is a prominent disease associated with a downward spiral of progressive breathlessness and functional decline. It is present in 10% of the population [14], predominantly in smokers. Mortality remains high and high 90-day-mortality and readmission rates are seen following COPD admissions.

II. MATERIALS AND METHODS

Forty-three studies were randomly selected from the systematic review and meta-analysis of Aziz et al [15], and were read in full-text to assess eligibility. Studies were included if they reported the investigated drug, measurement interval(s) and both baseline and follow-up measurements of FEV1. In particular, studies that only reported difference (without explicitly giving the baseline value) were excluded. This resulted in a final sample size of 24.

Numerical values for these studies were extracted from the text and both absolute and relative changes were calculated. Results were visualized separately for different follow-up durations.

Calculations were performed using the R statistical program package version 3.5.1 [16], with `lattice` version 0.20-35 [17].

III. RESULTS

The distribution of the baseline and follow-up FEV1 values are shown on Figure 1.

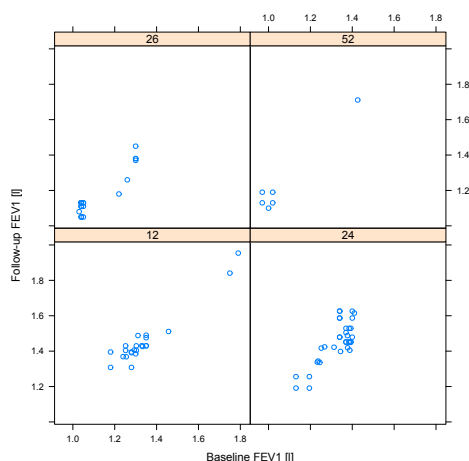


Fig. 1. Baseline and follow-up FEV1 values stratified according to follow-up time.

Both metrics (relative and absolute change) together with the baseline values are shown on Figure 2.

IV. DISCUSSION

Previous research in this field mostly concentrated on binary endpoints. (I.e. comparing relative risk or odds ratio to absolute risk reduction.) The paper from Deeks [18] is particularly interesting as it analyzes a vast amount of data, namely 551 meta-analysis (each consisting of at least five trials). Both

relative and absolute metrics were calculated for all trials, and then heterogeneity was assessed at meta-analysis level for both metrics using the Q -statistic with inverse variance weighting. Median Q was 5.36 for OR, 4.99 for RR, but 7.08 for RD.

Schmid et al [19] likewise performed a meta-analysis level investigation using three metrics (RD, log RR and OR) and found that correlation between the control group event rate was much more likely to be significant with RD than with RR or OR, again indicating that RD was less stable. Engels et al [20] noted (again using a same methodology, with 135 meta-analyses) that decisions were qualitatively not different with absolute and relative metrics, but again found less heterogeneity with the relative metrics. Furukawa et al [21] calculated concordance rate – proportion of non-significant difference – in similar arrangement (55 meta-analysis), and again found that it is substantially lower for RD than for RR or OR.

Whether there is a theoretical reason for this phenomenon, or how wide it is supposed to be is a matter of debate (see [22] for discussion).

Our study was somewhat underpowered, as the variability in the baseline was limited. Yet, the correlation between baseline and measurement (with all follow-up durations lumped together) was 0.239 for the difference (95% CI: 0.0286 – 0.0429, $p = 0.02668$), but 0.0316 for the ratio (95% CI: -0.181 – 0.214, $p = 0.7726$), confirming that ratio was able to capture information that was less dependent on baseline.

This results is in line with previous findings, but extends it to continuous endpoints.

V. CONCLUSION

These results present only the first step to better understand how the stability of absolute and relative metrics behave for continuous endpoints. The next step can be the extension to a more comprehensive selection of studies, incorporation of drug (which is more practical on a larger sample size), and the application of a heterogeneity metric, such as the Q statistic. Nevertheless, these preliminary results are in line with previous experience with binary endpoints.

REFERENCES

- [1] L. Citrome, "Relative vs. absolute measures of benefit and risk: what's the difference?" *Acta Psychiatrica Scandinavica*, vol. 121, no. 2, pp. 94–102, 2010. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0447.2009.01449.x>
- [2] C. Heneghan, B. Goldacre, and K. R. Mahtani, "Why clinical trial outcomes fail to translate into benefits for patients," *Trials*, vol. 18, no. 1, p. 122, Mar 2017. [Online]. Available: <https://trialsjournal.biomedcentral.com/articles/10.1186/s13063-017-1870-2>
- [3] S. Snapinn and Q. Jiang, "On the clinical meaningfulness of a treatment's effect on a time-to-event variable," *Statistics in Medicine*, vol. 30, no. 19, pp. 2341–2348, 2012. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4256>
- [4] D. M. Diamond and U. Ravnskov, "How statistical deception created the appearance that statins are safe and effective in primary and secondary prevention of cardiovascular disease," *Expert Review of Clinical Pharmacology*, vol. 8, no. 2, pp. 201–210, 2015. [Online]. Available: <https://www.tandfonline.com/doi/full/10.1586/17512433.2015.1012494>

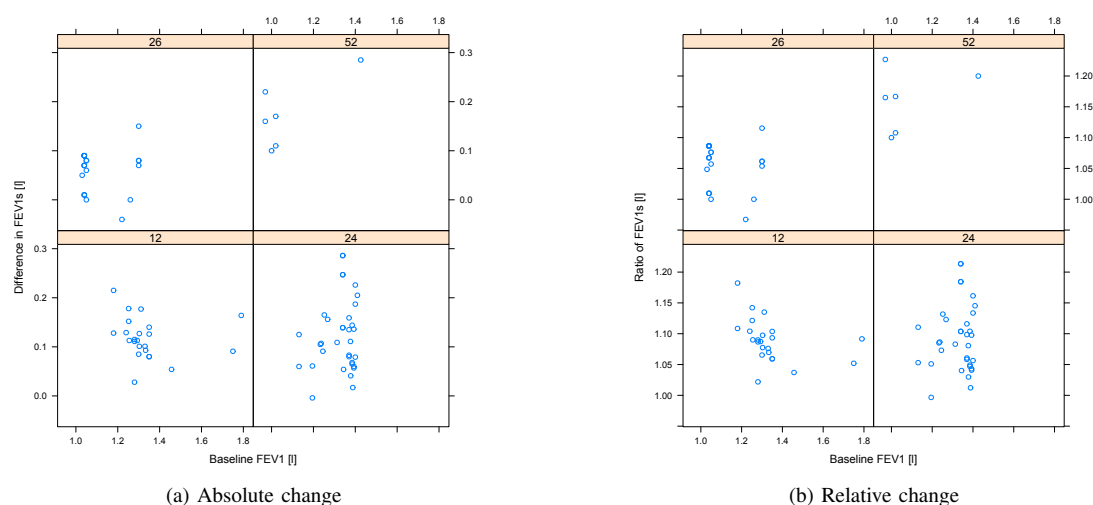


Fig. 2. Simulation results for the network.

- [5] T. Ferenci, "Absolute risk reduction may depend on the duration of the follow-up," *Expert Review of Clinical Pharmacology*, vol. 10, no. 12, pp. 1409–1410, 2017. [Online]. Available: <https://www.tandfonline.com/doi/abs/10.1586/17512433.2015.1102008>
- [6] P. M. Ridker, E. Danielson, F. A. Fonseca, J. Genest, A. M. Gotto, J. J. Kastelein, W. Koenig, P. Libby, A. J. Lorenzatti, J. G. MacFadyen, B. G. Nordestgaard, J. Shepherd, J. T. Willerson, and R. J. Glynn, "Rosuvastatin to prevent vascular events in men and women with elevated c-reactive protein," *New England Journal of Medicine*, vol. 359, no. 21, pp. 2195–2207, 2008. [Online]. Available: <https://www.nejm.org/doi/full/10.1056/NEJMoa0807646>
- [7] P. Joseph, D. Leong, M. McKee, S. S. Anand, J.-D. Schwalm, K. Teo, A. Mente, and S. Yusuf, "Reducing the global burden of cardiovascular disease, part 1," *Circulation Research*, vol. 121, no. 6, pp. 677–694, 2017. [Online]. Available: <https://www.ahajournals.org/doi/abs/10.1161/CIRCRESAHA.117.308903>
- [8] P. M. Rothwell, "External validity of randomised controlled trials: "to whom do the results of this trial apply?," *The Lancet*, vol. 365, no. 9453, pp. 82–93, 2005.
- [9] S. Senn, "Controversies concerning randomization and additivity in clinical trials," *Statistics in Medicine*, vol. 23, no. 24, pp. 3729–3753, 2004. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2074>
- [10] P. P. Glasziou and L. M. Irwig, "An evidence based approach to individualising treatment," *BMJ*, vol. 311, no. 7016, pp. 1356–1359, 1995. [Online]. Available: <https://www.bmj.com/content/311/7016/1356>
- [11] S. Senn. (1999, 6) At odds with reality. [Online]. Available: <https://www.bmj.com/rapid-response/2011/10/27/odds-reality>
- [12] P. S. Sever, C. L. Chang, A. K. Gupta, A. Whitehouse, N. R. Poulter, and on behalf of the ASCOT Investigators, "The Anglo-Scandinavian Cardiac Outcomes Trial: 11-year mortality follow-up of the lipid-lowering arm in the UK," *European Heart Journal*, vol. 32, no. 20, pp. 2525–2532, 2011. [Online]. Available: <https://academic.oup.com/eurheartj/article/32/20/2525/487760>
- [13] L. Smeeth, A. Haines, and S. Ebrahim, "Numbers needed to treat derived from meta-analyses—sometimes informative, usually misleading," *BMJ*, vol. 318, no. 7197, pp. 1548–1551, 1999. [Online]. Available: <https://www.bmj.com/content/318/7197/1548>
- [14] A. S. Buist, M. A. McBurnie, W. M. Vollmer, S. Gillespie, P. Burney, D. M. Mannino, A. M. Menezes, S. D. Sullivan, T. A. Lee, K. B. Weiss, R. L. Jensen, G. B. Marks, A. Gulsvik, and E. Nizankowska-Mogilnicka, "International variation in the prevalence of copd (the bold study): a population-based prevalence study," *The Lancet*, vol. 370, no. 9589, pp. 741 – 750, 2007. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0140673607613774>
- [15] M. I. A. Aziz, L. E. Tan, D. B.-C. Wu, F. Pearce, G. S. W. Chua, L. Lin, P.-T. Tan, and K. Ng, "Comparative efficacy of inhaled medications (ICS/LABA, LAMA, LAMA/LABA and SAMA) for COPD: a systematic review and network meta-analysis," *International Journal of Chronic Obstructive Pulmonary Disease*, vol. 13, p. 3203, 2018.
- [16] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018. [Online]. Available: <https://www.R-project.org/>
- [17] D. Sarkar, *Lattice: Multivariate Data Visualization with R*. New York: Springer, 2008, ISBN 978-0-387-75968-5. [Online]. Available: <http://lmdvr.r-forge.r-project.org>
- [18] J. J. Deeks, "Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes," *Statistics in Medicine*, vol. 21, no. 11, pp. 1575–1600, 2002. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1188>
- [19] C. H. Schmid, J. Lau, M. W. McIntosh, and J. C. Cappelleri, "An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials," *Statistics in Medicine*, vol. 17, no. 17, pp. 1923–1942, 1998. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0258%2819980915%2917%3A17%3C1923%3A%3AAID-SIM874%3E3.0.CO%3B2-6>
- [20] E. A. Engels, C. H. Schmid, N. Terrin, I. Olkin, and J. Lau, "Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses," *Statistics in Medicine*, vol. 19, no. 13, pp. 1707–1728, 2000. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/1097-0258%2820000715%2919%3A13%3C1707%3A%3AAID-SIM491%3E3.0.CO%3B2-P>
- [21] T. A. Furukawa, G. H. Guyatt, and L. E. Griffith, "Can we individualize the 'number needed to treat'? an empirical study of summary effect measures in meta-analyses," *International Journal of Epidemiology*, vol. 31, no. 1, pp. 72–76, 2002. [Online]. Available: <https://academic.oup.com/ije/article/31/1/72/655925>
- [22] F. A. McAlister, "Commentary: Relative treatment effects are consistent across the spectrum of underlying risks ... usually," *International Journal of Epidemiology*, vol. 31, no. 1, pp. 76–77, 2002. [Online]. Available: <https://academic.oup.com/ije/article/31/1/76/655927>