Studies in History and Philosophy of Science 88 (2021) 280–283

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: www.elsevier.com/locate/shpsa

Discussion

Measures of effectiveness in medical research: Reporting both absolute and relative measures

Carl Hoefer [a,b], Alexander Krauss [a,c,*]

[a] Universitat de Barcelona, Gran Via de les Corts Catalanes, 08007 Barcelona, Spain
[b] ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain
[c] CPNSS, London School of Economics, Houghton St, London WC2A 2AE, UK

ARTICLE INFO

Keywords
Clinical trials
Measures of effectiveness
Absolute measures
Relative measures
Base-rate fallacy
Alendronate
Randomised controlled trials
Philosophy of medicine

ABSTRACT

Biomedical research, especially pharmaceutical research, has been criticised for engaging in practices that lead to over-estimations of the effectiveness of medical treatments. A central issue concerns the reporting of absolute and relative measures of medical effectiveness. In this paper we critically examine proposals made by Jacob Stegenga to (a) give priority to the reporting of absolute measures over relative measures, and (b) downgrade the measures of effectiveness (effect sizes) of the treatments tested in clinical trials (Stegenga, 2015a). After exposing significant flaws in a central case study used by Stegenga to bolster his first proposal (a), we go on to argue that neither of these proposals is defensible (a or b). We defend the practice, in line with the *New England Journal of Medicine*, of reporting both absolute and relative measures whenever feasible.

Biomedical research has faced criticisms for engaging in practices that lead to over-estimations of the effectiveness of medical treatments. A central point of discussion in the debate concerns the role of absolute and relative measures of medical effectiveness, with absolute measures being given much greater weight by some philosophers (Stegenga, 2015a, 2018; Sprenger & Stegenga, 2017). This paper critically examines this common view by assessing the work of an influential philosopher of medicine, Jacob Stegenga, in particular focusing on his 2015 paper *Measuring effectiveness,* published in this journal (Stegenga, 2015a).[1] Many of the points that Stegenga makes in the article are well justified. However, some of the central claims in the article are highly problematic, and they do a disservice to medical research and the philosophy of medicine. The purpose of this paper is to address the issue of how measures of effectiveness should be reported by critically responding to several claims made in Stegenga's paper.

Our paper is structured as follows. Section 1 discusses absolute and relative measures of effectiveness, focusing on Stegenga's discussion of the drug Alendronate (Fosamax®). Section 2 discusses the broader context of the debate on these measures of effectiveness. Section 3 discusses Stegenga's proposal of adjusting measures of effectiveness (effect sizes) as a possible strategy for extrapolating results from trial studies to the broader population. Section 4 concludes by drawing implications for the philosophy of medicine and arguing that both absolute and relative measures should generally always be reported in trial studies.

1. Absolute and relative measures of effectiveness: Stegenga on Alendronate

A key criticism advanced by Stegenga on existing measures of medical effectiveness is the following: researchers (and pharmaceutical companies) make the effectiveness of a drug seem more significant than it actually is by highlighting *relative* measures (e.g. of risk reduction or of the beneficial outcome) rather than *absolute* measures. Reporting only relative outcome measures, which are insensitive to the 'base rates' of the given outcomes, he claims, increases the likelihood of people committing something akin to a base rate fallacy and judging treatments to be more effective than they actually are. For this reason, Stegenga argues (2015a, p. 62) that "Effectiveness always should be measured and reported in absolute terms …, and only sometimes should effectiveness also be measured and reported in relative terms". This position is also defended in the paper *Three Arguments for Absolute Outcome Measures* (Sprenger & Stegenga, 2017; cf. Stegenga, 2018, p. 16).

The principal example Stegenga offers to illustrate this issue involves the benefits of the drug alendronate. Stegenga (2015a, p. 67) writes:

[*] Corresponding author. CPNSS, London School of Economics, Houghton St, London WC2A 2AE, UK
E-mail address: a.krauss@lse.ac.uk (A. Krauss).

[1] Stegenga's criticisms of biomedical research are also found in another paper of his in this journal (2015b) and especially in his book *Medical Nihilism* (2018).

https://doi.org/10.1016/j.shpsa.2021.06.012
Received 21 January 2021; Received in revised form 12 May 2021; Accepted 27 June 2021
Available online 12 July 2021

> To illustrate the problem that arises when not taking P(Y) [the probability of an individual having a given outcome] into account with relative measures of effectiveness, consider the drug alendronate sodium (Fosamax), claimed to allegedly cause an increase in bone density in women, used with the aim of decreasing the frequency of bone fractures. A large trial compared the drug to placebo over a four year period (Black et al., 1996). The evidence from the trial was touted as showing that the drug reduces the risk of hip fractures by 50%–this was a relative measure of risk reduction (RRR). However, as Moynihan and Cassels (2005) note, only 2% of the women in the control group had hip fractures during the four years of the trial, while only 1% of the women in the experimental group had hip fractures. Thus the RD [risk difference] effect size was a mere 1%–the absolute difference in hip fracture rates between the experimental group and the control group was only 1%–after consuming the drug for four years. Moreover, it was only women at 'high risk' of hip fractures–namely, those who had already had hip fractures–who were included as subjects in the study, and thus the subjects in the study were not representative of the broader target population of patients for whom such an intervention is intended. … In short, alendronate sodium is barely effective, even in the most at-risk patients. The use of a relative outcome measure makes the drug seem more effective than it in fact is.

This example is well-chosen to illustrate the difference between relative measures (in this case, of risk reduction) and absolute measures. The problem is that it is also a mischaracterisation of the cited paper, Black et al. (1996), and of the literature of alendronate studies more generally.

Stegenga's description of the study is misleading because the trial actually aimed to study the effects of alendronate on *vertebral* fractures in women, and one criterion for inclusion in the study was having had one or more previous *vertebral* fractures. Hip fractures were not the main target of the trial, nor was previous hip fracture a selection criterion.[2]

So far, we have here just a misdescription of one important study of alendronate; but a closer look at Black et al. (1996) and at meta-analyses that pool together the results from other large alendronate trials, such as Karpf et al. (1997) and Serrano et al. (2013), paints a different picture than the one Stegenga offers. For example, in Black et al., the *absolute* risk, and the absolute risk reduction, for vertebral fractures, are substantially higher than Stegenga claims – as can be seen from the initial summary of their findings (Black et al., 1996, p. 1535):

> 78 (8.0%) of women in the alendronate group had one or more new morphometric vertebral fractures compared with 145 (15.0%) in the placebo group (relative risk 0.53 95% CI0.41–0.68]). For clinically apparent vertebral fractures, the corresponding numbers were 23 (2.3%) alendronate and 50 (5.0%) placebo (relative hazard 0.45 [0.27–0.72]). The risk of any clinical fracture, the main secondary endpoint, was lower in the alendronate than in the placebo group (139 [13.6%] vs 183 [18.2%]; relative hazard 0.72 [0.58–0.90]). The relative hazards for hip fracture and wrist fracture for alendronate versus placebo were 0.49 (0.23–0.99) and 0.52 (0.31–0.87). There was no significant difference between the groups in numbers of adverse experiences, including upper-gastrointestinal disorders.

In other words, alendronate reduced the relative risk of vertebral fractures by about 50% and reduced the absolute risk from 15% to 8% - a much less trivial result than the incidental results on hip fractures that Stegenga reports. Notice also that the authors present *both* the absolute and the relative numbers – except for those for hip and wrist fractures – up front in the article. In the article, all absolute numbers are later reported clearly, including for hip fractures and other fractures that were not the principal target of the study. Stegenga's presentation of only the latter constitutes a distortion of the study findings of Black et al. (1996).

Meta-analyses of multiple existing studies on alendronate's effectiveness (Karpf et al., 1997; Serrano et al., 2013) provide a positive picture, both in their results and the way they are reported – when considering studies that aimed at measuring bone density changes rather than fracture rates.[3] In trial studies, both absolute numbers and relative numbers tend to be reported. Meta-analyses illustrate that alendronate (and similar bisphosphonate drugs) have significant effects of fracture risk reduction and improvement in bone density in women who are considered to be in at-risk groups for bone fractures. The relative risks of fracture (for treatment group compared to placebo group) tend to be in the range of 50%–80%. Alendronate produces significant reductions in the risk of bone fracture among women in at-risk groups. Stegenga's verdict that "alendronate sodium is barely effective, even in the most at-risk patients" is not justified.

Our disagreement here matters. After a (misleading) presentation of results from just *one* article, Stegenga dismisses as "barely effective" a drug that millions of people around the world have taken over the past two decades. His conclusion *may* be right in some sense; the larger issue of whether alendronate's benefits outweigh its costs (economic, and in terms of harmful side-effects) enough to justify its widespread use is not one we assess here. But the conclusion is *not* supported by the study he cites, nor by other studies on alendronate such as Black et al. (2000), Serrano et al. (2013), and Karpf et al. (1997).

The issue is important with regard to at least two classes of potential readers. In the first place, there are individuals who might read Stegenga's verdict on alendronate and factor it into their own personal decision when it comes to weighing what action (if any) to take in case of a diagnosis of osteopenia or osteoporosis. In the second place, as philosophers of science engage more and more with practicing scientists, there is an increasing likelihood of our papers being read by specialists in the relevant fields, i.e., doctors and medical researchers. It is thus crucial that we try to avoid making claims that are likely to be seen as ill-informed, factually mistaken, or misleading by scientific experts. Stegenga's discussion of alendronate is an example of the kind of mistakes that philosophers of science need to avoid committing.

2. Absolute and relative measures of effectiveness: the broader context

After illustrating the way that reporting of relative measures of effectiveness may lead people to perceive treatments as being more effective than they (in absolute terms) really are, Stegenga (2015a, p. 62) makes his central and very strong recommendation: "Effectiveness always should be measured and reported in absolute terms …, and only sometimes should effectiveness also be measured and reported in relative terms" (cf. 2018: 16). Can this be correct? The first question that this recommendation raises is one that Stegenga does not explicitly address: reported *where, and to whom?* There are at least three relevant audiences that one might consider.

In academic journals? We believe it does make sense, generally, to state *both* absolute and relative outcome measures (when feasible) in reporting a study's results. And this is what the world's leading medical

[2] In the quoted passage, Stegenga cites Moynihan and Cassels (2005), *Selling Sickness: How the Drug Companies Are Turning Us All Into Patients,* which discusses osteoporosis and Merck's aggressive marketing of Fosamax. This publication may be the source of Stegenga's idea that Black et al. (1996) studied hip fractures. Although Moynihan and Cassels do not make this assertion, they *do* report the hip fracture results that Stegenga also reports.

[3] In the meta-analysis on alendronate by Karpf et al. (1997), the study authors report positive results, namely that "The estimated cumulative incidence of nonvertebral fractures after 3 years was 12.6% in the placebo group and 9.0% in alendronate group. The relative risk for nonvertebral fracture estimated using the Cox proportional hazards model was 0.71 (95% confidence interval, 0.502-0.997) (*P*=.048)."

journal, the *New England Journal of Medicine*, states in its statistical reporting guidelines for authors to be able to publish clinical trials in the journal: “the editors prefer that absolute event counts or rates be reported before relative risks or hazard ratios. The goal is to provide the reader with *both* the actual event frequency and the relative frequency” (NEJM, 2020; emphasis added). In most trial studies on alendronate, both measures are reported clearly.[4]

In textbooks, pharmaceutical literature, NIH-type guidance documents etc.? Again, it would seem better to report both, when feasible. If medical practitioners are prone to committing base-rate type fallacious inferences, the solution is not to withhold data from them but rather to educate them better about statistics, and make sure that they have the relevant base-rates at their disposal. In order to make the best recommendations to their patients, medical practitioners need to have as much information as possible, *ceteris paribus*. So it is not clear why Stegenga would recommend that relative measures be reported *only sometimes*.

To the public, or patients? What should be told by medical practitioners to their patients is a complex issue that involves the given health problem, the evidence of the effectiveness of the given treatment, the patient's ability to understand the evidence, the patient's interest in knowing the relevant data, and the like. Here, difficult issues of medical ethics arise regarding the physician-patient relationship. But we do not have to enter into those issues in order to see that Stegenga's blanket prescription that absolute measures *always* and relative measures *only sometimes* be presented is not defensible.

Consider the following example. An individual is going to travel to a remote location and sees her doctor about what shots to get before travelling. As it happens, at that time there is a rare tropical disease outbreak in the area she will visit, and her doctor mentions that there is a vaccine she could take, but it only reduces the chance of being infected with the disease by 1.2%. What the doctor neglects to mention is that the baseline incidence of visitors catching the disease is 1.5% and having had the vaccine reduces the probability down to 0.3%. If the vaccine has no, or limited, known negative side-effects, and the disease is one that either kills or debilitates for life, then potential travellers would be well advised to take the vaccine, because when it comes to either dying or being debilitated for life, a 1.5% chance is not something to dismiss. If the doctor in this case, by contrast, in addition indicates the relative risk reduction (80%), the traveller would be much likelier to view the vaccine as a sensible and worthwhile precaution to take. Presenting both measures would be ideal – as is the case in the *New England Journal of Medicine*. This does require that patients have a basic grasp of probability and statistics. But not all patients have such a grasp, and for some patients just hearing that the vaccine only reduces the chance of contracting the disease by 1.2% might be enough to discourage them from taking the vaccine. For such patients, learning also the relative risk reduction can facilitate better decision-making.

Clearly, this case is constructed to maximally highlight the potential danger of reporting only an absolute measure of effectiveness without reporting the relative measure. But with over a million clinical trials conducted (as illustrated in the Cochrane Library), it is likely that some portion of trials have such results, so a blanket statement that absolute measures should *always* and relative measures *only sometimes* be reported to patients is indefensible.[5]

Another question is how newspapers and other media should present clinical trial results to the public. This too is a complex issue, but here as well it is not clear that Stegenga's prescription is defensible. To give another example, suppose that a vaccine against HIV were developed, and clinical trials showed it to be 95% effective (as a relative measure). Is it worth mentioning in public media, that far less than 1% of people (in the general population, or in the given trial) get infected with HIV? That rate may be irrelevant to many people who have reasons to be concerned about exposure to HIV, and may even be dangerous to some of them (depending on their sophistication in matters statistical). So here too, regarding science communication to the public, there are doubts about whether Stegenga's blanket injunction that absolute measures always and relative measures only sometimes be presented is defensible.

Although the issues that arise when considering how medical research results should be reported to the wider public are numerous and complicated, it seems clear that a default prescription of concealing one sort of information in favor of another has little in its favor. Not only is it paternalistic, but it can lead to worse results in some circumstances. For these reasons, we believe a default practice of reporting both absolute and relative measures, where feasible, makes sense in general and not just in journal articles reporting trial results. Arguing for this claim more fully, however, is beyond the scope of this paper.[6]

3. Issues in improving medical research and its reporting

Stegenga (2015a) then explores a number of well-known reasons why it may not be advisable to extrapolate the results about the effectiveness of a treatment, from a clinical trial to the general population. Among the most important difficulties, we can mention (i) *prima facie* relevant differences between the clinical trial group and the potentially-treated population as a whole, and (ii) *publication bias,* namely that clinical trials with negative or inconclusive results are at times left unpublished while trials with positive results are more often published, leading to a bias towards higher treatment effectiveness in the published literature (ClinicalTrials.gov, 2020; Krauss, 2018; Stegenga, 2018). These difficulties with extrapolation, among a number of others, can be significant for a given drug or treatment, and we can rarely assume that the treatment will be as effective in real clinical practice as it was in trial studies.

[4] As mentioned earlier, this is not always the case in existing meta-analyses, which is unsurprising. In meta-analyses, it may be cumbersome to present diverse absolute measures in a non-misleading way. For example, different trials may use different measures, have different end-points, different kinds of patients, and so on. These differences may cause the absolute measures of effectiveness to be extremely diverse, and their significance impossible to assess properly without extensive explanation, whereas concise reporting of the range of relative measures may be informative and non-misleading.

[5] Stegenga (2015a, p. 68) seems to recognize that in some cases knowing just the absolute measures may induce poor decisions by patients. Here is his discussion of this: “Schwartz and Meslin (2008) suggest that the use of absolute measures could cause patients to make irrational decisions (say, to forgo treatment in cases similar to those above, in which the absolute effect sizes are tiny), and for at least some cases they seem to suggest that this is an argument in favor of the use of relative measures. Their argument is: for a patient to make an autonomous medical decision they must be informed about the extent to which a particular medical intervention is effective; since people display a low degree of numeracy, absolute outcome measures might hinder patients’ understanding of effectiveness; thus, employ relative measures. I hope to have shown that such a comparison between people’s comparative understanding of relative versus absolute outcome measures is dubious. Relative measures, by promoting the base rate fallacy, fundamentally mislead patients into overestimating effectiveness.” Ordinary citizens may be prone to making poor judgments at times when given only relative measures, and also prone to making poor judgments at times when given only absolute measures.

[6] There is a further point about what should be reported regarding medical treatments that we wish to raise, one with which we expect Stegenga would agree. In addition to quantitative outcome measures (both absolute and relative), there are further facts about the effects of medical interventions that patients and practitioners need to take into account. There are non-quantifiable effects of a treatment on patients’ level of pain and quality of life that are, by their very nature, not directly amenable to quantitative analysis. In some trials, for example, the primary outcome is an increase in survival rates; but those treated can also be more likely to suffer from highly adverse side-effects. Only providing quantitative outcome measures in studies, and not also collecting and reporting such qualitative information, can lead to the omission of important information concerning whether patients who live longer due to a treatment may also suffer more intensely and for longer periods of time (Krauss, 2018).

The difficulties with simple extrapolation that Stegenga highlights are important and well known (ibid.). The complex question is: How should we attempt to overcome these problems? Stegenga (2015a, p. 69) recommends a practice of adjusting measures of effectiveness downward: "a method of extrapolation could take publication bias into account by decreasing estimates of effectiveness as measured in published studies when predicting the effectiveness of the medical intervention in a target population". But he does not specify when or how this should be done, nor does he distinguish between efficacy trials (conducted in highly controlled conditions) and larger effectiveness trials (conducted in real world conditions that generally produce smaller effect sizes). Stegenga also does not specify *who* should lower expected effectiveness: The researchers who have carried out the trial study? Doctors considering prescribing the treatment to patients? National or international drug regulation agencies like the FDA? Journal editors considering the publication of a study?

Let us suppose that Stegenga suggests that it is drug regulation agencies that should implement the adjusted reduction in expected effectiveness, and ensure that the "corrected" numbers are made available to the relevant actors. Even if these proposed downplaying measures were nuanced and well-designed, neither medical researchers nor the pharmaceutical industry would be likely to agree with them. And when it comes to individual cases, they would often be right: a general measure of downgrading, no matter how nuanced, cannot be defended on the grounds of bringing greater accuracy in every case. In many cases, the measure of effectiveness reported in a given trial would be more accurate than such a downgraded measure.[7]

Medical researchers and journals are however already in the process of taking more serious measures to exclude the kinds of biases and errors that have been identified as the causes of over-estimation of effectiveness. Stegenga does not, in his (2015a) paper, mention the measures that *have been* taken to improve medical research, such as the requirements of pre-registration of trials and pre-publication of protocols, requirements of making trial outcome data publicly available even if negative, registered reports in which a study protocol is peer reviewed and provisionally guaranteed publication before the study is conducted, regardless of the results (to counteract publication bias), and so on (ClinicalTrials.gov, 2020).

Both epistemically and socially, it would be better overall for philosophers of medicine to recommend that medical research be conducted in ways that prevent error rather than trying to push for the adoption of measures that bring about compensatory bias or errors that may at times be unjustified, and even harmful in specific cases.

4. Conclusions

Philosophy of science should engage with actual scientific practices and play a normative and important role in understanding and improving research. Medicine is one of the most fruitful areas where such engagement can occur. Stegenga (2015a) does present a range of important critical observations concerning measures of effectiveness of clinical trial research. Unfortunately, his paper's central claim and recommendation about the importance of absolute over relative outcome measures, derived from a misleading analysis of alendronate research, is highly problematic. Trial studies should (as a rule) always report both absolute and relative outcome measures. Stegenga's proposal, by contrast, is intrinsically not defensible, and is unlikely to be taken seriously by the greater medical community. Likewise, Stegenga's (2018) general thesis in *Medical Nihilism*, namely that we should have little confidence in the effectiveness of most medical treatments, goes against the simple historical fact that medical researchers have developed effective treatments for a wide range of diseases. Medical research has enabled us to cure illnesses, develop vaccines to prevent others and has contributed to expanding our life expectancy. Given the stakes involved in carrying out and applying medical research — namely, our health and lives — philosophers should take greater care when formulating such criticisms of research practice, and making general recommendations.

Funding

Research for this paper by Hoefer was supported by grants from the Catalan agency AGAUR (2017-SGR-0063) and from Spanish ministries (MINECO FFI2016-81858-REDC, MICINN FFI2016-76799-P). Research for this paper by Krauss was supported by grants from the Catalan agency AGAUR (Marie-Curie/Beatriu de Pinos grant 2019 BP 00004, and 2017-SGR-0063).

Acknowledgments

The authors would like to thank Nancy Cartwright, Genoveva Martí, and both anonymous reviewers for helpful comments on an earlier version of this paper.

References

Black, D., Cummings, S., Karpf, D., Cauley, J., Thompson, D., Nevitt, M., … Ensrud, K. (1996). Randomised trial of effect of alendronate on risk of fracture in women with existing vertebral fractures. Fracture Intervention Trial Research Group. *Lancet, 348*(9041), 1535–1541.

Black, D., Thompson, D., Bauer, D., Ensrud, K., Musliner, T., Hochberg, M., … Cummings, S. (2000). Fracture risk reduction with alendronate in women with osteoporosis: The fracture intervention trial. *Journal of Clinical Endocrinology & Metabolism, 85*(11), 4118–4124. https://doi.org/10.1210/jcem.85.11.6953

ClinicalTrials.gov. (2020). *ClinicalTrials.gov*. U.S. National Library of Medicine. https://clinicaltrials.gov/ct2/about-site/history.

Karpf, D., Shapiro, D., Seeman, E., Ensrud, K., Johnston, C., Adami, S., … Thompson, D. (1997). Prevention of nonvertebral fractures by alendronate: A meta-analysis. *JAMA, 277*(14), 1159–1164. https://doi.org/10.1001/jama.1997.03540380073035

Krauss, Alexander (2018). Why all randomised controlled trials produce biased results. *Annals of Medicine, 50*(4), 312–322.

New England Journal of Medicine (NEJM). (2020). Submitting to NEJM: Statistical reporting guidelines. *New England Journal of Medicine*. www.nejm.org/author-center/new-manuscripts.

Serrano, A., Begoña, L., Anitua, E., Cobos, R., & Orive, G. (2013). Systematic review and meta-analysis of the efficacy and safety of alendronate and zoledronate for the treatment of postmenopausal osteoporosis. *Gynecological Endocrinology, 29*(12), 1005–1014. https://doi.org/10.3109/09513590.2013.813468

Sprenger, J., & Stegenga, J. (2017). Three arguments for absolute outcome measures. *Philosophy of Science, 84*(5), 840–852.

Stegenga, J. (2015a). Measuring effectiveness. *Studies in History and Philosophy of Biological and Biomedical Sciences, 54*, 62–71.

Stegenga, J. (2015b). Effectiveness of medical interventions. *Studies in History and Philosophy of Biological and Biomedical Sciences, 54*, 34–44.

Stegenga, J. (2018). *Medical Nihilism*. Oxford: Oxford University Press.

[7] By "more accurate" we mean closer to the eventual effectiveness (in terms of effect sizes), defined in the same way as in the clinical trial, observable (ideally, if not practically) in the full treated population.