Editorial

Percent Change Was and Is An Embarrassing Mistake

Predicting the Past

Any fool can predict the future; politicians and investment counselors do it all the time. Modeling past medical and financial events accurately remains a major challenge. Percent change is commonly used in these and other fields of measurement. Quantities in the form of a "percent" are used every day; they are familiar, useful, and generally harmless. But not always.

Proportions as a percent, no problem. "Sixty percent of medical students are female."

Percent as a measure of change: "My tax load has increased by 8% in the past year." No problem so far, but begs the question: 8% of what?

Percent as a measure of relative change: "The group given "NewDrug™" increased bone density by 5%; 50% more than the group given "OldDrug™." This is a gross misrepresentation. "Despite recommendations that bone mineral density precision and followup assessment be based upon absolute measurements (in g/cm^2), the use of relative change (in percent) is still frequently encountered."[1]

Two decades ago, I was 60 and my son 30, half my age. Thirty years later he will be 60, having aged 100%. If I survive, I will age only 33%. Clearly, he is aging 300% faster than I. A grandson, who will be 33 years of age at that time, will age 1000% during that same interval. In time I shall stop aging altogether, so they can catch up.

Are things better in Boston? In the May 6, 2010, issue of the *New England Journal of Medicine*, a group from the Brigham and Women's Hospital report on the use of a technology to reduce medication errors. "...units that did not use [the new technology] reported an 11.5% error rate ...versus [a 6.8% error rate] in those that did... — a 41.4% relative reduction"[2]. (Why not a 4.7% reduction? Or something totally different?)

But surely nothing like this could appear in *The Journal of Rheumatology*! An editorial entitled "When Less Is More"[3] argued: "We could still get some idea by comparing the differences (or deltas) in efficacy between the active and the placebo groups in published studies. The pivotal ASSERT study[4] on IFX using the dose of 5 mg/kg q 6 weeks showed a 61% ASAS20 response in the IFX group compared to 19% response in the placebo group, a delta of 42%."

Ordinal Versus Equal-interval Scales

Two years ago, a colleague and I reported a study on the results of hip surgery entitled "Common Measures and Analytic Techniques Provide Flawed Assessments...."[5] Observed changes in pain severity showed linear fit to the normal distribution, but the same data expressed as percent change had a curved, hyperbolic distribution that invalidated means, standard deviations, and related statistics. A change of 3 units from 3 to 6 (100%) was different from 3-unit change from 6 to 9 (50%). Change measured as percent change is not an equal-interval scale (no measure that plots as a curve can be an interval scale). Further, there was directional bias. Changes (deltas) from 6 to 9 (50%) are different from changes from 9 to 6 (33%). Comparing one outcome expressed as percent change to another was inappropriate at best, outrageous at worst. My grandson will have aged nearly 33 times (3333%) more quickly than me.

Error Functions

All measures are subject to uncertainty. Measurement scatter is commonly calculated using the method of "least squares" and expressed as a standard deviation or related statistic that assumes that the error distribution is "normal" or "Gaussian," or at least symmetrical. This assumption is not valid with percent change. There are many problems, one of which is that the data distribution is often hyperbolic. When we extrapolate in time or space from current data, and need to estimate distant error, this usually increases as the model extends further from the central data. With hyperbolic distributions, error increases as change approaches small values (often the desired target), becoming infinite at zero.

See Informing Response Criteria for PsA II, *page 2559*

The Devil Is in the Denominator

Given that division by a constant does not distort a rectilinear scale, it is not easy to predict problems. Let us take an innocuous example. We are all concerned about fuel efficiency in motor vehicles, commonly measured as miles per gallon or alternatively liters per 100 km. Comparisons among designs is simple with either measure, but see what happens when we compare the measures (Figure 1).

In Figure 1 a perfect fit ($R^2 = 1$) is obtained with reciprocal transformation of values on the x-axis, or alternatively, with log transform of BOTH the x- and y-axis. A single log-transformation of the y-axis fits the data less well. Placing a measure in the denominator dramatically alters the shape of the distribution. Treating this hyperbolic distribution as a single exponential gives an R^2 value of 0.76, and the plotted confidence limits are reassuring, but misleading. If the data ranges in Figure 1 were (quite reasonably) restricted to 100 miles per gallon and 20 liters per 100 km, the hyperbolic relationship might not be perceived. Log transformation of the data in the y-axis would give data that could then be used in further analyses to give wrong answers — as is standard in financial literature, or discussions of climate change, and in "knowledge-based medicine."

You won't find much discussion of hyperbolic distributions in statistics texts, but they are not confined to percent change. A very obvious example is David Sackett's "number-needed-to-treat" (NNT)[6]. The NNT is simply the reciprocal of the absolute risk difference and is widely used and cited. But the answer has a hyperbolic distribution (not specified in the resulting publications). When there is no treatment effect, the absolute risk reduction is zero and the NNT is infinite. Altman has dealt with this problem at length, but avoiding the term "hyperbolic"[7]. To paraphrase: A confidence interval can (must) be quoted for any trial: especially when NNT is used in metaanalysis[7]. With hyperbolic distributions there is a gap around zero where this obligation cannot be filled. This was conceded, and Altman suggested separating NNTB (benefit) from NNTH (harm). Confidence intervals were then calculated from the raw data, before inversion, or after reversing the calculation. Graphically, the central position was occupied by infinity rather than zero.

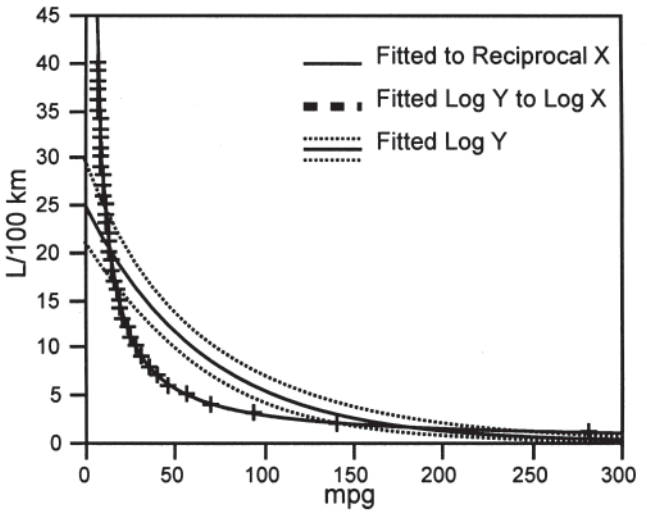

Figure 1. Hyperbolic curve and fitting strategies based on data from our study[5].

David Sackett, Douglas Altman, and the authors of an article in this issue[8] have been heroes of mine, and remain so. Controversy freely entered should cement friendly relations.

The Devils Are in the Denominators

But I am not finished. Percent change and NNT are not the only commonly used analytic techniques that place experimental variables in the denominator, yielding hyperbolic distributions that are usually unrecognized and ignored in subsequent statistical analyses. These often include modeling for time, and often omitted most relative effect sizes, most ratios (e.g., relative risk, likelihood ratios, and derivatives such as ROC curves) and any calculation arising from 2 by 2 tables. Are flu vaccines effective? The published evidence relies on ratio statistics with perhaps flawed estimates of error.

Let us look at these dilemmas in another way. Liters per 100 km and miles per gallon are simple statistics and usually graph as straight lines on the y-axis against scales where the values in the horizontal axis are without error. There is nothing wrong with the measures. The problem illustrated in Figure 1 is having variables measured with error in both axes, a problem treated with respect (if not clarity) by many generations of statisticians. These topics require exploration beyond that relevant to this review.

Psoriatic Arthritis Response Criteria (PsARC)

The authors in this issue[8] have used strategies that are sound. They have experimental data derived from double-blind, controlled trials, and have used them to create a training set; from which new response criteria are derived, and independent testing sets. They used a number of inventive strategies to evaluate the hypotheses that emerged, all sound. The difficulty (if that is the right word) is that they were too diligent. They measured changes in raw data (joint counts, pain, etc., and percent changes in the same data) — a belt and suspenders approach. They could not avoid this because the American College of Rheumatology has used percent changes as standards since 1994[9,10]. They model multiple strategies to predict drug versus placebo use in previous controlled trials in terms of the benefit achieved, with considerable success. But how would the measures fare if the patients worsened? Changes of 2 units in visual analog scale pain or joint count would simply reverse direction. The difference between 2 and 4 would be the same as the difference between 6 and 8, as in an equal-interval scale.

Examination of distributions would indicate symmetrical form, or indicate the need for data transformations. But percent change is demolished by direction reversal. Change from 2 to 4 (100% worsening) has to be different from 6 to 8 (33% worsening). Not only that, but the difference from 6 to 8 is different from the earlier 8 to 6 (25%).

It gets worse. In a population with mixed outcomes (some worse and some better), the distributions of the new hybrid scale are or will be undescribable. Means will be meaningless, and with no mean there can be no standard error or analysis of variance. These effects are modeled in Figure 2 (all data from Smythe and Bogoch[5]).

Comparing Means

The same clinical data from our study[5] were used in the x- and y-axes of the 2 graphs in Figure 2. For the raw data (left plot), the means had equal numerical values with opposite signs, and standard deviations were equal. All standard parametric tests could be employed. This is not true with analyses related to the right plot. Calculated averages and deviations are different from the raw data, and vary greatly, reflecting change of direction. The data are the same, but the means are meaningless.

The data in Figure 2 were drawn from patients who improved. Not plotted were data on patients who were unchanged or became worse; or from outcome measures that rose with improvement. These complexities present no difficulties when the raw data are analyzed, but are impossible with analysis of mixed outcomes expressed as percent change.

It is laborious, and I believe not necessary, to show how calculations from fundamentally flawed data will flow through to flawed estimates of sensitivities and specificities, and from these to appropriate evidence-based therapeutic targets.

HUGH A. SMYTHE, MD, FRCP,
Associate Editor,
The Journal of Rheumatology,
Toronto, Canada

Address correspondence to Dr. H.A. Smythe, 2 Heathbridge Park, Toronto, Canada M4G 2Y6. E-mail: hasmythe@rogers.com

REFERENCES

1. Leslie W. The importance of spectrum bias on bone density monitoring in clinical practice. Bone 2006;39:361-8.
2. Poon EG, Keohane CA, Yoon CS, Ditmore M, Bane A, Levtzion-Korach O, et al. Effect of bar-code technology on the safety of medication administration. N Engl J Med 2010;362:1698-707.
3. Deodhar A. When less is more. J Rheumatol 2010;37:1089-90.
4. van der Heijde D, Dijkmans B, Geusens P, Sieper J, DeWoody K, Williamson P, et al; Ankylosing Spondylitis Study for the Evaluation of Recombinant Infliximab Therapy Study Group. Efficacy and safety of infliximab in patients with ankylosing spondylitis: results of a randomized, placebo-controlled trial (ASSERT). Arthritis Rheum 2005;52:582-91.
5. Smythe HA, Bogoch ER. Common measures and analytic techniques provide flawed assessments of pain: modeled data, and hip replacement study. J Rheumatol 2008;35:2400-5.
6. Sackett DL, Cook RJ. Understanding clinical trials. What measures of efficacy should journals provide busy clinicians. BMJ 1994;309:755-6.

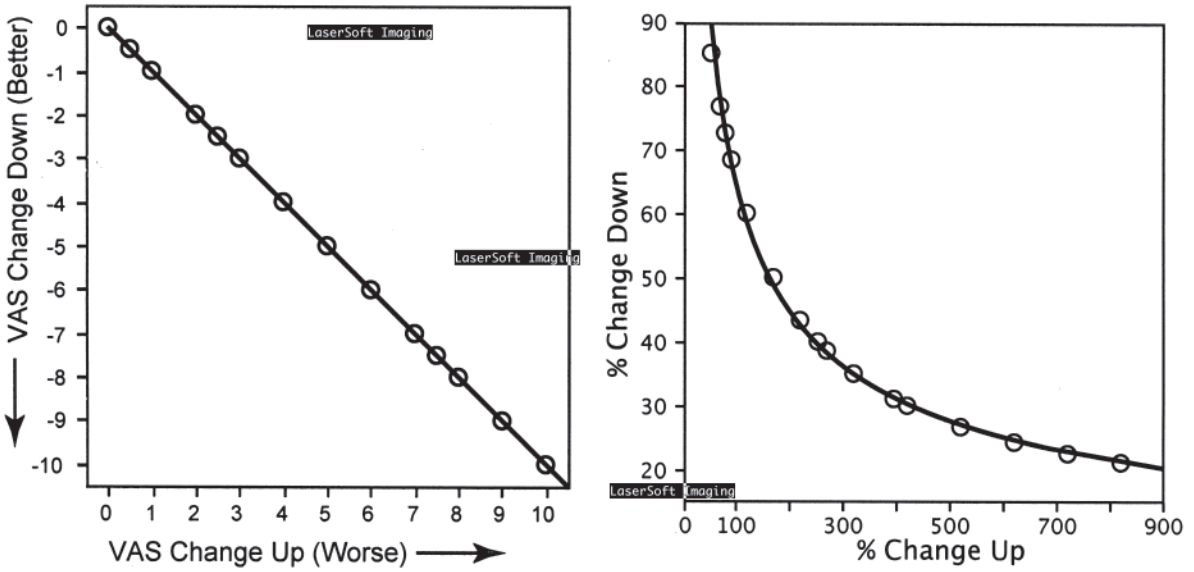

Figure 2. Percentage change and comparison of means modeled based on data from our study[5]. Change in VAS pain (left, as recorded) from patients who improved, against the same data recalculated as if time was reversed, so that preoperative pain became the outcome. The line of fit is rectilinear, and the intervals comparable. On the right, the same contrasts are applied to the same data calculated as percent change. The fit to a log/log plot (a power curve) is excellent, but percent change up is greatly magnified as paired with percent change down. The fitting and the fit of the curve are more important than the mathematical form. Analysis as a hyperbola gives similar but slightly different results. Once fitted, the form can be expressed in terms of 2 parameters and an asymmetric and non-uniform "confidence" plot, but NOT as a mean and standard deviation.

7. Altman DG. Confidence intervals for the number needed to treat. BMJ 1998;317:1309-12.
8. Gladman DD, Tom BDM, Mease PJ, Farewell VT. Informing response criteria for psoriatic arthritis. II: Further considerations and a proposal — The PsA joint activity index. J Rheumatol 2010;37:2559-65.
9. Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, et al, and the Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. Arthritis Rheum 1993;36:729-40.
10. Boers M, Tugwell P, Felson DT, van Riel PL, Kirwan JR, Edmonds JP, et al. World Health Organization and International League of Associations for Rheumatology core endpoints for symptom modifying antirheumatic drugs in rheumatoid arthritis clinical trials. J Rheumatol 1994;21 Suppl 41:86-9.

J Rheumatol 2010;37:2448–51; doi:3899/jrheum.100943