Creating Composite Indices From Continuous Variables for Research: The Geometric Mean

Diabetes Care 2021;44:e85–e86 | https://doi.org/10.2337/dc20-2446

Hertzel C. Gerstein,[1,2,3]
Chinthanie Ramasundarahettige,[1,2]
and Shrikant I. Bangdiwala[1,3]

Clinical research focuses on the relationship between one or more independent variables and some dependent variable or outcome chosen to reflect some underlying process. For categorical variables, the research may either be focused on a specific end point such as myocardial infarction (MI) or on an underlying construct such as vascular disease (with MI as just one exemplar). In the latter instance, a composite index such as major adverse cardiovascular event (MACE), defined as either a nonfatal stroke, nonfatal MI, or cardiovascular death, may be used.

Composite categorical outcomes such as MACE optimize power by ensuring a high event rate, and the results they yield are generalizable to diseases that are consistent with the underlying construct. It is therefore surprising that there is no widely used method to combine continuous variables into composite continuous outcomes. Nevertheless, there is a clear need for such a methodology when the underlying construct cannot be easily captured by one measurement. Glucose control is an example of a construct that can be assessed in many ways, including fasting or postprandial plasma glucose, HbA_{1c}, fructosamine, or "time in target." A composite of two or more of these could provide a better reflection of glucose control than any one alone. Whereas sophisticated statistical techniques such as structural equation modeling (1) can be used to model some underlying construct or latent variable from two or more measurements, a simpler way of combining them into an index that reflects the underlying construct could provide a powerful tool for both researchers and clinicians. Such an approach is described below.

When the same measurements are made using the same scale, the arithmetic mean provides a more precise estimate than any one measurement alone. The challenge arises when different measurements (e.g., heart rate and body temperature) are used to measure some underlying construct (e.g., illness severity) using different scales. In this instance, an arithmetic mean is nonsensical. This is usually managed by converting the measurements into standardized *Z* scores (2) after ensuring that they are normally distributed (or are transformed to a normal distribution) (3). Thus, if a person's heart rate *Z* score is 1.1 and body temperature *Z* score is 0.9, the arithmetic mean of those two *Z* scores may better reflect the construct of illness severity than either *Z* score alone. Clearly, all component measurements must have the same directional relationship with the underlying construct, and any that have inverse relationships need to be reverse-scored before being combined.

A simpler approach is to calculate the geometric mean of the *n* measurements being combined. The geometric mean is simply the *n*th root of the product of the *n* measurements (Supplementary Fig. 1) and can be understood as a function that reflects the multiplicative relationship between the components. Thus, for the geometric mean of measurements "a" and "b," something that increases "a" by some relative amount (e.g., two- or threefold) will yield the same geometric mean as something that increases "b" by the same relative amount. Therefore, in contrast to the arithmetic mean, it can be used to combine measurements from scales with different distributional properties and represents an easy-to-calculate composite index of *n* disparate measures that eliminates the need for standardization before combining them.

The geometric mean can therefore reflect many aspects of some underlying construct, is simple to calculate, and reduces the need for complex multivariable analyses. Limitations (Table 1) include the inability to calculate a geometric mean when either the product of the component variables is a negative number (since the *n*th root of a negative number is an imaginary number) or when the value of any of the components is 0 (because it would return a geometric

[1] *Population Health Research Institute, Hamilton Health Sciences and McMaster University, Hamilton, Ontario, Canada*
[2] *Department of Medicine, McMaster University, Hamilton, Ontario, Canada*
[3] *Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada*

Corresponding author: Hertzel C. Gerstein, gerstein@mcmaster.ca

Received 2 October 2020 and accepted 24 January 2021

This article contains supplementary material online at https://doi.org/10.2337/figshare.13643522.

Table 1—Prerequisites for calculating the geometric mean of two or more variables

- No variable can have a value that is negative or 0, or be missing.
- The component variables should all be different measurements of the same underlying construct.
- Component variables should have near-normal distributions or be transformed so that they follow near-normal distributions.
- Variables to be combined must have the same directional relationship with the underlying construct, such that all are positively or negatively correlated with the underlying construct; when component variables have inverse relationships with the construct, the reciprocal or the complement of all of the positively or negatively correlated variables should be used to calculate the geometric mean.

mean of 0). They also include ensuring that the components of the geometric mean are all either positively or negatively related to the underlying construct and that the measurements are normally distributed or can be transformed to approximate a normal distribution.

Some clinically relevant potential composite indices that could be constructed by calculating geometric means of continuous variables are noted in Supplementary Table 1.

Data from the Outcome Reduction With Initial Glargine Intervention (ORIGIN) trial (4) were used to estimate indices reflecting three disease constructs. These included the fasting plasma glucose and HbA_{1c} levels for the degree of dysglycemia (available for 12,345 people), the urine albumin-to-creatinine ratio (ACR) and the estimated glomerular filtration rate (eGFR) for renal disease (available for 12,187 people), and the Mini-Mental State Examination and the Digit Symbol Substitution Test scores for cognitive status (available for 3,676 people). Because the ACR and eGFR have opposite relationships to renal disease, the reciprocal of eGFR was used. All variables were natural log (ln)-transformed. The ACR and reciprocal of eGFR were both multiplied by 1,000 prior to this transformation to avoid ln-transformed values that were either 0 or negative. Each participant's ln-transformed value for each variable was then converted to a *Z* score, and the mean of the two *Z* scores was calculated and compared with each participant's geometric mean using a quantile-quantile plot. All analyses were done using SAS (version 9.4), and figures were drawn using R (version 4.00).

Mean values of the ln-transformed measurements of each of these variables, the arithmetic means of the index derived from the two *Z* scores, and the index derived from the geometric mean of the ln-transformed variables are listed in Supplementary Tables 2 and 3. The quantile-quantile plots of these two indices (Supplementary Fig. 2) demonstrate that higher quantiles of the geometric mean reflect higher quantiles of the mean of the two *Z* scores.

These examples illustrate that the geometric mean of different measurements of an underlying disease construct has the same ordinal relationship as the arithmetic mean of the standardized scores of these measurements. The facts that the examples were based on different types of measurements reflecting different health-related states and that they used data collected as part of a large trial suggest that the findings are robust and generalizable. This implies that when more than one measurement is available that reflects some underlying health state, and when the measurements all satisfy the conditions in Table 1, they can be combined into an index by computing their geometric mean. This index can then be used as either a dependent variable (i.e., outcome) or independent variable (risk factor or covariate) in subsequent analyses.

This approach reduces the number of variables that need to be included in analyses and combines information from different aspects of an underlying health state into one single index. The performance characteristics of each index created in this way, as well as any thresholds for classifying people with or without disease, would clearly need to be evaluated in a variety of databases. Nevertheless, the geometric mean represents a novel and simple way of creating composite indices from continuous data that is likely to promote new insights and identify new health-related outcomes and risk factors.

Duality of Interest. H.C.G. holds the McMaster-Sanofi Population Health Institute Chair in Diabetes Research and Care. He reports research grants from Eli Lilly, AstraZeneca, Merck, Novo Nordisk, and Sanofi; honoraria for speaking from AstraZeneca, Boehringer Ingelheim, Eli Lilly, Novo Nordisk, and Sanofi; and consulting fees from Abbott, AstraZeneca, Boehringer Ingelheim, Eli Lilly, Merck, Novo Nordisk, Janssen, Sanofi, and Kowa. S.I.B. reports research-related grants to his institution from Bristol-Myers Squibb and Sanofi.

Author Contributions. H.C.G. and S.I.B. designed the analytic plan. H.C.G. wrote the first draft. C.R. performed the statistical analyses. All authors reviewed and edited the manuscript and contributed to interpretation of the data. H.C.G. is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

References

1. Tomarken AJ, Waller NG. Structural equation modeling: strengths, limitations, and misconceptions. Annu Rev Clin Psychol 2005;1: 31–65

2. Bancks MP, Carnethon MR, Jacobs DR Jr, et al. Fasting glucose variability in young adulthood and cognitive function in middle age: the Coronary Artery Risk Development in Young Adults (CARDIA) Study. Diabetes Care 2018;41:2579–2585

3. Altman DG, Bland JM. Statistics notes: the normal distribution. BMJ 1995;310:298

4. Gerstein HC, Bosch J, Dagenais GR, et al.; ORIGIN Trial Investigators. Basal insulin and cardiovascular and other outcomes in dysglycemia. N Engl J Med 2012;367:319–328