When Change Is Not a Good Thing

Kathleen D. Liu and Chi-yuan Hsu*

Over the past 40 years, our understanding of the epidemiology and natural history of a number of common chronic diseases has been markedly enhanced by serial, nationally representative, cross-sectional surveys such as the National Health and Nutrition Examination Survey (NHANES)² (1), longitudinal cohorts such as the Atherosclerosis Risk in Communities (ARIC) study (2), and more disease-specific cohorts including the Chronic Renal Insufficiency Cohort (CRIC) (3) and the Assessment, Serial Evaluation, and Subsequent Sequelae in Acute Kidney Injury (ASSESS-AKI) study (4). Many of these studies have banked biospecimens, which allow tracking of laboratory markers of disease over time. Such stored samples are invaluable resources that can lend insight not only on secular trends in population prevalence of disease but also on how disease appears and evolves over time within a study participant. Both of these require repeated measurements of the same analyte over an extended period of time (years to decades).

This seemingly simple concept—repeated measurement of the same analyte over time—can be challenging to implement well. To begin with, the assay method itself may change. For instance, serum creatinine, the most common test to assess kidney function, can be measured using alkaline picrate (“Jaffe method”) or enzymatic assays (5). Even the Jaffe method has been refined over time to better handle interfering substances (6). In addition, there may be drift in the standards used to calibrate the assay, leading to systematic differences. In the case of serum creatinine, the National Kidney Disease Education Program has led efforts over the last decade to standardize testing, including adoption of isotope dilution mass spectrometry (IDMS)-traceable references (7).

Thus, calibration of assays performed in different eras may be necessary. Laboratory recalibration via repeat measurement of stored specimens at a later (uniform) point in time assumes stability of the analyte and its concentration over time. This fundamental assumption may be violated if degradation of the substance of interest or technical problems such as sample desiccation occur (8).

In the current issue of Clinical Chemistry, Parrinello et al. (9) tested the impact of changes in assays over time in the ARIC study (2). ARIC was conducted in 4 communities in the United States, with 5 study visits to date: 1987–1989, 1990–1992, 1993–1995, 1996–1998, and 2011–2013. The authors identified 200 enrollees who had plasma samples available from all 5 visits. Eight biomarkers were included in the main analysis: creatinine, uric acid, glucose, total cholesterol, HDL cholesterol, LDL cholesterol, triglycerides, and high-sensitivity C-reactive protein. Analytes were remeasured in samples that had been stored at −70 °C since collection.

The authors concluded that analytes remeasured in samples stored for approximately 25 years were highly correlated with original values, but 2 of the 8 analytes (creatinine and uric acid) showed substantial bias at multiple visits. Laboratory recalibration improved reproducibility of test results across visits and resulted in substantial differences in chronic kidney disease (CKD) prevalence. Parrinello et al. should be greatly commended for undertaking this well-executed study, which highlights some important methodological issues in the field. Perhaps as interesting as their results is their approach, which is worth a closer look.

The first decision in undertaking a recalibration study centers on how many samples to remeasure. From both financial cost and sample resource perspectives, it is not practical or wise to remeasure the entire population. Often remeasurement is undertaken among hundreds of study enrollees, as is the case here. But to the best of our knowledge, the full impact of sample size on precision of estimates has not been well described. Formal power calculations are almost never presented (although minimal detectable differences should be possible to estimate). Conceptually, the goal of the analysis is to generate recalibration equations, and the coefficients of those equations would have wider confidence intervals if fewer participants were remeasured and narrower confidence intervals if more participants were remeasured.

A second, and closely related, decision is which participants to sample. In this analysis, Parrinello et al. stratified the cohort by age, sex, and race to ensure representative sampling of the original population. Because the focus was on a panel of analytes, this is a reasonable approach. However, if a single marker is of predominant
interest, or if a marker is associated with a specific disease state that is relatively rare, it may make sense to oversample those more likely to have the disease of interest to ensure more robust representation in the range of biological values of interest. Similarly, undersampling those in whom marker levels are likely to be below the limits of detection would give efficiency gains. It is worth pointing out that disease is often defined by a threshold at the far end of a bell-shape distribution (e.g., CKD is defined by an estimated glomerular filtration rate (eGFR) of <60 mL·min⁻¹·(1.73 m²)⁻¹, and the prevalence of CKD in the overall ARIC cohort is low), so small changes in the absolute value [e.g., eGFR in mL·min⁻¹·(1.73 m²)⁻¹] as a result of recalibration can translate into large differences in the area under the curve (10). In presenting the final conclusions, a confidence interval around the prevalence estimate after calibration would appear to be desirable to convey the uncertainty introduced by the calibration.

After measurements are completed, further choices have to be made regarding analysis. For example, outliers can greatly influence the coefficients of regression equations. As the authors point out, outliers may arise from processes irrelevant to recalibration, such as isolated sample degradation or labeling/data entry error. The Parrinello et al. study (9) used an iterative approach to eliminating outliers. Observations >3 SDs away from the mean difference were removed, and the process was repeated until no more outliers were identified. These steps were also taken in prior studies by the same research group (11). Although we are unaware of a formal analysis of the impact of modifying the outlier boundary, this approach seems reasonable, since eliminating measurements that are >3 SDs should represent only 0.3% of measurements based on a normal distribution. Overall, 4.5% of paired measurements were removed in the current study (9) (there was no comment on whether this proportion was similar across the 8 analytes or how many iterations were required). The substantial impact of removing these outliers can be illustrated by how it affects the final estimated prevalence of CKD: using the recalibrated creatinine values, the prevalence of CKD was 1.3%, 2.2%, and 6.4% at ARIC visits 1, 2, and 4, respectively. In contrast, when the outliers were retained, the prevalence of CKD was 8.5%, 3.4%, and 6.5%, respectively. These empiric data support removing outliers, because when this is not done, the anticipated increase in CKD prevalence with aging is not observed. Going by the premise that it would be ideal to have confidence bounds around these prevalence estimates, the influence of outlier removal on these parameters warrants close scrutiny in future studies in this area.

Other analytic considerations include choice of regression methodology. The authors used Deming regression (9), which is appropriate as it accounts for errors in both the x- and y-axis observations (vs standard linear regression, which accounts only for error in the dependent variable). Decisions also have to be made regarding whether the regression analysis should be done using log-transformed data. The typical default approach, and the one used by Parrinello et al. (9), is the linear scale, which assumes that more flexible models would not improve calibration, an assumption that is almost never tested. More complex approaches such as equipercentile standardization (12) can be considered if there are additional shifts in measurement resulting in changes in the SD of the analyte distribution in the population. However, this approach typically requires measurements in a relatively large number of young, healthy participants to convert one scale to the other. Finally, usually only the β-coefficients for the intercept and slope of the recalibration equations are carried into the next step (e.g., see Table 2, rightmost column, in Parrinello et al. (9)). However, as discussed above, these β-coefficients actually have confidence intervals, which could inform our interpretation of final study conclusions.

To close, as growing numbers of longitudinal cohorts with biobanking accrue more person-years of follow-up, the issues addressed by this article (9) will garner increasing attention in the future. Parrinello et al. are commended for doing leading-edge work in this important area. We believe that further refinement is possible for future investigators, and we have highlighted 1 such potential area related to statistical methodology. To facilitate the interpretation of study findings, it would be an advance if the imprecision stemming from the recalibration process—based on factors mentioned above, such as choice of sample size/framework and regression method, elimination of outliers—can be somehow captured from beginning to end and lead to the generation of confidence bounds around the final derived parameters (such as revised disease prevalence estimates in a population or a revised rate of disease progression within a study participant).

Author Contributions: All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.

Authors’ Disclosures of Potential Conflicts of Interest: Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:

Employment or Leadership: None declared.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: K. Liu, NIH (R01 DK98233); C-y. Hsu, NIH (R01 DK98233).
Expert Testimony: None declared.
Patents: None declared.
Acknowledgments: We thank Dr. David Glidden for helpful discussions.

References