Statistics
in Medicine

Commentary

(wileyonlinelibrary.com) DOI: 10.1002/sim.4281 Published online in Wiley Online Library

The end of the beginning: a commentary on 'Evaluation Metrics for Biostatistical and Epidemiological Collaborations'‡

Cathleen Kane[a*†] **and William M. Trochim**[a,b]

The paper 'Evaluation Metrics for Biostatistical and Epidemiological Collaborations' of Rubio *et al.* represents an important initial advance in the evaluation of biostatistics, epidemiology, and research design (BERD). The authors present a sensible three-domain model (collaboration with investigators, application of BERD-related methods, and discovery of new BERD methodologies), rightly acknowledge the importance of team science, and break new ground in illustrating that the Clinical Translational Science Awards can function as a kind of national laboratory for the development and exploration of measures and metrics. Building upon these gains, there are several future considerations worthy of subsequent serious attention: strengthening the connection between BERD evaluation and both the science of team science and the field of evaluation; facing the challenges of operationalization of the conceptual domains; augmenting the work of Rubio *et al.* with standard evaluative models; and anticipating the need for multiplistic mixed methods and experimental and quasi-experimental complements to the proposed BERD metrics. Several common pitfalls will also be important to avoid, including the tendency to conflate the meaning of 'metrics' and 'measures' and the potential for a premature rush to adopt national 'standards' before adequately pilot testing the initial set of methods they have worked so diligently to develop.

Keywords: evaluation; metrics; measures; CTSA; translational research

'Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning.'
–A quote from Winston Churchill's speech at Lord Mayor's Luncheon on November 10, 1942.

Any commentary on 'Evaluation Metrics for Biostatistical and Epidemiological Collaborations' must begin by acknowledging the significant contribution made by the authors in the struggle to learn how to evaluate the complex collaborations inherent in translational research. This work represents a very important preliminary advance in the evaluation of biostatistics, epidemiology, and research design (BERD), and leaves us 'at the end of the beginning' needing to move resolutely ahead in addressing the many other issues that remain in evaluating the Clinical Translational Science Awards (CTSAs), some of which we hope to point out in this paper.

In 'Evaluation Metrics for Biostatistical and Epidemiological Collaborations', the authors' contribution is important to the field, both for its specific applicability to the context of the CTSAs (BERD) and for its implications to translational research and team science more generally. The authors' model is a sensible one that speaks to the critical construct domains and the complex interactions between them. The three outcome domains identified in their Venn diagram cover the territory well (collaboration with investigators, application of BERD-related methods, and discovery of new BERD methodologies). Their assessment of BERD also rightly acknowledges the importance of team science. Furthermore, their work breaks new ground in illustrating that the CTSAs can function as a kind of national laboratory for the development and exploration of measures and metrics. Despite the large size and scale of the CTSAs,

[a] *Weill Cornell Medical College, Clinical and Translational Science Center, 407 E. 61st, 2nd Floor, New York, NY 10021, USA*
[b] *Weill Cornell Medical College, Clinical and Translational Science Center, 407 E. 61st, Room RR-220, New York, NY 10021, USA*
**Correspondence to: Cathleen Kane, Weill Cornell Medical College, Clinical and Translational Science Center, 407 E. 61st, 2nd Floor, New York, NY 10021, USA.*
[†] *E-mail: cmk42@cornell.edu*
[‡] *This publication was made possible by grant award UL1RR024996 from the National Institutes of Health (NCRR CTSA) to Weill Cornell Medical College.*

because the national initiative is still relatively new, multi-center processes with this degree of sophistication have not yet been commonly pursued. Their paper should be characterized as watershed moment in cross-CTSA evaluation and like Churchill's sentiments in the earlier text, a kind of rallying cry for future work.

Building upon these gains, there are several future considerations worthy of serious attention as the authors and their BERD colleagues move forward, namely, further strengthening the connection between BERD evaluation and the science of team science; facing the challenges of operationalization within the CTSAs; augmenting the work of Rubio *et al.* with standard evaluative models; and anticipating the need for multiplistic mixed methods as well as an experimental and quasi-experimental complement to the proposed BERD metrics. This commentary also seeks to illuminate several common pitfalls our colleagues in BERD would be wise to avoid as they organize their evaluative approach(es) at the national level, primarily, avoiding the conflation of metrics and measures and a premature rush to adopt national 'standards' before adequately pilot testing the initial set of methods they have worked so diligently to develop. The remainder of this commentary makes the assumption that Rubio *et al.* is a groundbreaking evaluative advancement worthy of a series of serious recommendations written from the perspective of professional CTSA evaluators.

Biostatistics, Epidemiology, and Research Design Evaluation and the Science of Team Science. Despite the single acronym, from the onset, Rubio *et al.* pointed out that BERD is not itself a unitary entity. Although some institutions might attempt to organize faculty and professionals into a single unit, at most institutions, the relevant expertise will reside across a number of traditional disciplines and departments. The authors correctly noted that'BERD units need to understand how to leverage the various strengths of their BERD practitioners.' Future assessments of the role of BERD in translational research should be informed by the emerging field of team science [1–3] as both areas share many common issues and concerns.

The work of Rubio *et al.* suggests several key questions pertinent to the 'team science' agenda. For instance, how often and to what extent do BERD professionals, whether they are consultants or research faculty, become members of the team of researchers that they are assisting? What factors affect how well those teams function? How do different team members feel about their role on the team? How is credit apportioned in publications, and to what degree were consultants included in subsequent work? These kinds of questions should be considered a part of a broader agenda of team science in BERD and added to the evaluation agenda.

Biostatistics, Epidemiology, and Research Design and Evaluation. The endeavor the authors set out to address is inherently a self-reflective one. To evaluate BERD, one would most logically turn to professionals like BERD experts themselves. The profession of evaluation is itself a multi-discipline with its own professional organizations (e.g., the American Evaluation Association), professional journals (The American Journal of Evaluation, Evaluation Review, Evaluation and Program Planning, and New Directions in Evaluation), graduate training programs, history, sets of traditions and norms, and others. Like BERD professionals, evaluators are also located differently at different universities and at many, have little or only tangential interaction with the traditional BERD practitioners. Evaluation is a broad endeavor[4–6] that addresses both formative efforts including needs assessment[7], structured conceptualization[8], evaluability assessment[9], implementation evaluation and process evaluation[10], and summative issues, including outcome and impact evaluation[11], cost-effectiveness and cost-benefit analysis[12], secondary analysis, and meta-evaluation[13]. Evaluation integrates into its work both (bio) statistics and research design, including the design and measurement issues commonly associated with epidemiology. We should give serious thought to working on integrating these allied disciplines more effectively, if not through formal structural mechanisms, at least through social and professional interdisciplinary networks and teams. BERD evaluation itself seems like the perfect opportunity for doing so.

The Challenges of Operationalization. At first glance, the authors' proposed metrics appear fairly straightforward and easily measured (e.g., number of consultations, number of investigators). However, as BERD practitioners begin to apply some of the proposed metrics and measures to their work, several vexing definitional and operational issues are bound to arise. For instance, what exactly is a 'consultation'? If in one context, a single BERD professional holds a short meeting of less than an hour with a researcher, whereas in another context, several BERD professionals take on a lengthy and distinct subproject within a research protocol, are we to consider each of these a single 'consultation'? BERD practitioners do not yet benefit from industry-wide standards for measuring their contributions with the same uniform 'consultative unit' such as an attorney's billable hours. Thorny definitional issues such

as this are not limited to the consultative example mentioned earlier. Operationalizing BERD's use of the metric 'number of investigators' poses just as many ontological challenges. Will the 'investigators' associated with BERD services be limited only to the researchers they have directly consulted with or to all of the investigators involved with each associated protocol? In addition, across various institutions, fields, and disciplines, there are at present, wildly varying norms for defining co-investigators and co-principal investigators on research protocols. How many variations of 'investigator' should be counted in association with BERD work on any given project? Until these critical metrics (and others) are fully defined and operationalized, any attempt to compare the work of BERD professionals across the sites or settings (e.g., CTSAs) cannot be accurately interpreted.

For every apparently simple metric or measure for BERD, there exists a serious (though certainly not impossible) knot of definitional issues in need of operational untangling. These issues in BERD evaluation will become even more complex as we take on more qualitative and judgmental metrics such as 'peer evaluations', 'success stories', or 'leadership in professional societies.' But the danger here is not that these terms cannot be defined or standardized; the real risk associated with the authors' seemingly simple list of BERD measures or metrics is that they will give readers the false impression that that they are somehow immediately applicable in evaluation. Within each proposed measure or metric, there are numerous challenges of definition and operationalization that will require future serious consideration from the authors and others.

Alternative and Enhanced Models. Early on in their argument, the authors referenced the systematic logic model of the National Institute of Environmental Health Sciences as an example of how 'new approaches to evaluation link research activities to spatially and temporally remote outcomes.' Although the authors' three-domain model is an important contribution to our understanding, Rubio *et al.* have not yet attempted to develop a similar conceptual model for BERD evaluation outright. Even a generic logic model for BERD, especially in its causal pathway form[14], could encompass their three outcome domains and show how they are generally related to the types of activities BERD professionals engage in. Further still, the development of a process model that details the steps in the (again generic) BERD process might illuminate some of the operational issues mentioned earlier and clarify and standardize how different BERD activities relate to other parts of the biomedical research process (Institutional review board and contracts review, clinical research management, publication preparation, dissemination, etc.). Augmenting the work of Rubio *et al.* with these two powerful types of evaluative models (logic and pathway models) would also help identify potential barriers to and opportunities for successful BERD integration in research efforts.

The Need for Multiplistic Mixed Methods. The authors stated that 'Evaluating performance in interdisciplinary biomedical research is highly complex, requiring a pluralistic approach that extends beyond conventional metrics.' That is certainly the case. But perhaps what is more important is that such evaluation goes beyond metrics alone, whether they are conventional or otherwise capture the complexity involved. What is called for is multiplism [15] and mixed methods[16], concepts already familiar to the evaluation community. Although metrics are an obviously useful start, even seemingly simple or 'conventional' metrics typically create more questions than they answer. For example, when you determine the number of publications associated with BERD efforts across multiple CTSA centers, this does not so much answer a question about the effect of BERD on scientific productivity as it raises a whole host of questions about what the numbers mean. Is this a lot or a few publications? Compared with what? What causes variation in rates within and across centers? How is the rate affected by the fields or disciplines published in? Further questions arise about what actions might be taken in response to such a number. Can publication rates be increased? Should they? If so, what factors related to BERD might be potential causal ones? The metrics suggested by the authors are for the most part actually quite conventional, but their simplicity will prove to be advantageous as a foundation for future multiplistic mixed methods that probe the meaning behind the numbers and suggest how they might be interpreted. BERD evaluators would be wise not to frame their short list of metrics and measures suggested in this article as cutting edge or even as the end point of evaluative inquiry and instead count them as 'the end of the beginning' a la Churchill's rallying cry.

Complementary Methods for Metrics. Rubio *et al.* expertly illuminated the wide variety of BERD team dynamics and types of substantive contributions to translational research. The complexity they described would greatly benefit from the kind of evaluative designs that could isolate and control for the most influential BERD contributions within a complex array of services and teams. More to the point, although metrics themselves have utility as a feedback mechanism, they have their greatest value when combined with an active program of intervention and evaluation. For instance, to collect metrics on the

discovery of novel BERD-related methodologies like numbers of proposals or grants submitted does not tell us much, unless we examine quasi-experimentally whether these might be related to structural factors like number and locations of BERD personnel, methods of BERD collaborations, and others. Even more prospectively, all of their proposed metrics would have greater utility if they were to conduct systematic studies explicitly designed to improve the BERD consulting process, such as an experimental or quasi-experimental test of a novel BERD approach to consulting on proposal and grant development.

Understanding Measures and Metrics. The language of 'metrics' tends to come from the tradition of business performance management, and the language of 'measures' has its historical roots in fields of research. In recent years, in a number of fields, there has been a disturbing trend of treating the two terms interchangeably and assuming they are synonymous. The danger here is in the possibility that quantitative *metrics* will be conflated with qualitative *measures*. In their paper, Rubio *et al.* took a very broad view of what a 'metric' is, at least if this is to be inferred from their tables. For instance, they included in their list of metrics things like student course evaluations, success stories, and BERD contributions to activities. Although there may be some quantitative component to these, there would almost necessarily need to be a qualitative one as well. So, to what extent are these definable or describable as 'metrics'? This comes dangerously close to perpetuating the growing general confusion about what a metric means. In its most traditional and narrow definitional sense, a metric is a *quantitative* measure typically used to monitor some process, although not all quantitative measures are necessarily metrics.

From an evaluator's perspective, it is true that performance metrics can be considered a type of evaluation, but it is only a subset. But here, as in many other contemporary settings, this original sense has been broadened to mean virtually anything that is measured with respect to a construct, whether qualitative or quantitative. We would argue that, given the more expansive approach taken by the authors, the paper should have been titled 'Evaluation Measures for Biostatistical and Epidemiological Collaborations' rather than using the narrower 'Evaluation Metrics' terminology, even though we empathize with the impulse to appeal to the current informal lexicon by relying on the term 'metric'. Still, in regard to the measures/metrics identified in this work, as the authors advance their agenda and move to execute their recommended evaluation strategies for BERD, they should maintain a conscious distinction between the meaning of the terms metrics and measures. They would also be wise to recall that a qualitative measure can be converted into quantitative metric, but the reverse is not true. Put more simply, there is an artificial rigor and deceptive simplicity in dubbing anything of an evaluative nature a 'metric'. That is why the term has become so popular in common language. Being sensitive to the distinction will help assure that the illusion of rigor does not substitute for the hard work still needed to define and operationalize what we are observing.

National 'Standards' for BERD Evaluation. The authors suggest that their three-domain model be adopted as a standard, at least across the CTSAs: 'While flexibility in the details of the metrics is essential to their broad implementation, each BERD unit should use the three-domain framework and definitions provided here.' (emphasis added). Although we heartily support the development of national CTSA BERD standards, it is important at the onset to note several crucial caveats. Recall our questions surrounding operationalization and the utility of logic and/or process models. If multiple CTSAs begin applying the metrics within the three domains and the myriad subdomains without more clarity on definitional issues while still subject to the widespread heterogeneity of BERD teams and consultative roles, a national effort would quickly founder on the grounds that they were 'comparing apples and oranges'. Further still, without some effort to develop measures of the different structural configurations of BERD functions across institutions, we would not be able to determine what factors contribute to institutional variations in metrics. In short, if the metrics are prematurely fixed before BERD professionals account for the structural differences, this important work may be dismissed, and BERD practitioners may prematurely settle on a standard of measurement that does not reflect important situational differences.

Rubio *et al.* did an excellent job in pointing out the structural differences across BERDs and in using the CTSA as a natural evaluative laboratory. In lieu of prematurely adopting their short list of measures as a rigid set of national standards, we would encourage them to take the intermediate step of expanding their initial work towards the design of a national pilot of their proposed measures with the explicit aim of evaluating their proposed evaluative methodology. Why rush the call to standardize? The work of Rubio *et al.* deserves expansion to a larger sample of settings so that they can look empirically at how their measures perform and especially how they might be interpreted and used. Then, we would be in a better position to talk about standardization.

Happily, BERD practitioners are in a unique position to fully understand the future recommendations enumerated in this commentary. In their everyday work as research consultants, BERD professionals

and scholars are called upon to create solid research designs, and they also understand the importance of replication, large data sets, and the slow forward progress of phased hypothesis testing. The evaluation of BERD should be implemented with the same high standards and disciplined professional best practices used in their work with biomedical researchers. In this case, 'good design' for BERD evaluation will require the added use of clear models and definitions for BERD. The importance of replication will demand clarity of operational definitions, and the structured use and analysis of large data sets will necessitate the maintenance of a clear distinction between metrics and measures. Finally, the forward progress of phased hypotheses will be best served via a disciplined focus on evaluative outcomes (what happened) as well as process (why did it happen) questions. This will in turn call for the eventual adoption of experimental and quasi-experimental methods to isolate causal variables to establish links between the work of BERD professionals and the more traditional outcomes of research productivity (e.g., presentations, publications, and grants). Of course, all this will only be possible after such time as the metrics and measures proposed by Rubio *et al.* have been fully piloted with a much wider variety of BERD colleagues across the CTSAs and elsewhere.

As in Churchill's rallying cry, the work of Rubio *et al.* has done a great deal to establish evaluation within the field of BERD evaluation by bringing us to 'the end of the beginning' of the search for appropriate measures within BERD. The authors have succeeded in providing a real contribution to the field by addressing the lack of consensus or usage of common metrics and measures and in using the CTSA as a 'natural laboratory' for this inquiry. Their work is a good model for other collaborative groups and cross-CTSA communities in areas like education, administration and community engagement as we collectively struggle with how to measure and evaluate critical components that contribute to the larger endeavor of translational research.

References

1. Falk-Krzesinski HJ, Börner K, Contractor N, Fiore SM, Hall KL, Keyton J, Spring B, Stokols D, Trochim W, Uzzi B. Advancing the science of team science. *Clinical and Translational Sciences* 2010; **3**:263–6.
2. Börner K, Contractor N, Falk-Krzesinski HJ, Fiore SM, Hall KL, Keyton J, Spring B, Stokols D, Trochim W, Uzzi B. A multi-level systems perspective for the science of team science. *Science Translational Medicine* 2010; **2**:cm24.
3. Falk-Krzesinski HJ, Contractor N, Fiore SM, Hall KL, Kane C, Keyton J, Klein J, Spring B, Stokols D, Trochim W. Mapping a research agenda for the science of team science. *Research Evaluation*. in press.
4. Scriven M. The methodology of evaluation. In *Perspectives on Curriculum Evaluation, AERA Monograph Series on Curriculum Evaluation*, Tyler R, Gagne R, Scriven M (eds). Rand McNally: Skokie, IL, 1967; 38–83.
5. Trochim W, Donnelly JP. *The Research Methods Knowledge Base*, 3rd ed. Thomson (Atomic Dog) Publishing: Cincinnati, OH, 2006.
6. Mathison S. *Encyclopedia of Evaluation*. Sage: Thousand Oaks, CA, 2005.
7. Witkin BR, Altschuld JW. *Planning and Conducting Needs Assessments: A Practical Guide*. Sage: Thousand Oaks, CA, 1995.
8. Trochim W. An introduction to concept mapping for planning and evaluation. *Evaluation and Program Planning* 1989; **12**:1–16.
9. Wholey JS. Evaluability assessment: developing agreement on goals, objectives and strategies for improving performance. In *Organizational Excellence: Stimulating Quality and Communicating Value*, Wholey JS (ed.). Heath: Washington, DC, 1987.
10. Stufflebeam DL, Madaus GF, Kellaghan T. *Evaluation Models: Viewpoints on Educational and Human Services Evaluation*, 2nd ed. Kluwer Academic Publishers: Boston, 2000.
11. Cook TD, Campbell DT. *Quasi-Experimentation: Design and Analysis for Field Settings*. Houghton Mifflin Company: Boston, 1979.
12. Levin HM, McEwan PJ. *Cost-Effectiveness Analysis: Methods and Applications*, 2nd ed. Sage: Thousand Oaks, CA, 2001.
13. Boruch RF, Wortman PM, Cordray DS. *Reanalyzing Program Evaluations*. Jossey-Bass: San Francisco, CA, 1981.
14. Urban JB, Trochim W. The role of evaluation in research-practice integration: working toward the "Golden Spike". *American Journal of Evaluation* 2009; **30**:538–53.
15. Shadish WR. Planned critical multiplism: some elaborations. *Behavioral Assessment* 1986; **8**:75–103.
16. Greene JC, Caracelli VJ. Advances in mixed-method evaluation: the challenges and benefits of integrating new paradigms. *New Directions for Evaluation* 1997; **74**.