The Promise of Big Data
Opportunities and Challenges

Harlan M. Krumholz, MD, SM

Medicine is on the cusp of a remarkable transformative change fueled by the prospect of an acceleration and augmentation in knowledge generation and application. Medicine improved as it transitioned from being based on belief to becoming anchored in science. We were limited, however, in our ability to learn from clinical practice, even though natural experiments were occurring every day. Our methods tended to be reductionist, our data crude, our samples constrained, and our inferences speculative. This inability to learn quickly, easily, and meaningfully from everyday clinical practice is yielding to new capabilities and opportunities.

The shift is occurring as a result of advances in computational power and mathematical applications in concert with a rapid growth in the availability of data. Even changes in the last 5 years have produced breathtaking capabilities. We now have the computational heft to handle large, complex, high-dimensional data. We have mathematical applications that enable us to avoid the need to reduce the complexity of the data, maximizing the relevance of the data to patients seen in practice. And we have just experienced a remarkable advance in data availability through transformation of medical data and documentation into digital format.

A precision medicine future, with the generation of personalized estimates of risk of outcomes and response to various therapies, paradoxically depends on being able to analyze data from massive numbers of people. Only with such numbers can the heterogeneity of risk and response be revealed—and can each individual be understood in the context of the experience of people like them who have preceded them in the situation they are in. The knowledge of individual differences—and similarities—will provide opportunities for truly informed choices based on information that is more relevant, accurate, and meaningful.

To fully leverage this opportunity will require several key steps.

Training
Most researchers are not adequately training for a big data era. The basic principles of epidemiology, study design, and biostatistics, which form the core foundation for most clinical researchers, will remain important, but are not sufficient. There are additional areas of importance, including digital data, informatics, and machine learning. Skills in programming will be a core competency to understand the logic of the programs used to extract information.

Application
The production of the new knowledge is not enough to ensure that benefit is derived from it. The application of the knowledge produced from these data sources and methods will need to be optimized to promote benefit and minimize risk. The culture of medicine will need to evolve from a reliance on old methods of risk stratification, for example, to embrace computer-driven estimates that are linked to methods of producing information about net benefits of different strategies.

Replication
A major challenge of data-driven approaches is to ensure that the findings can be replicated. To depend on the estimates will require confidence that, time and again, they can be reproduced in different data sets, particularly for different subpopulations. The possibility of false-positive inferences—or estimates of risk that are not stable across data sets—is substantial.

Dissemination
The dissemination and communication of the knowledge and methods merits consideration. Our journals are ill-prepared to vet advanced analytics applied to massive data resources. In many cases, the approach may be difficult to evaluate by reviewers and editors. Of note, Circulation: Cardiovascular Quality and Outcomes is committed to finding reviewers, including technical experts, to evaluate novel technical approaches. Moreover, the result may be better suited to a dynamic presentation, rather than the static portable document format. In addition, the visualizations may not lend themselves to the 2-dimensional page. The results of research in this area may be better disseminated as a software product or application, with a paper that describes methods. There will be a need to develop platforms that enable this type of dissemination as part of a scientific contribution. Transparency is also important, even as there may be issues of proprietary algorithms and approaches.

Interoperability
Although data increasingly exist in digital format and provide remarkable opportunities for knowledge generation, obstacles

The opinions expressed in this article are not necessarily those of the American Heart Association.

From the Section of Cardiovascular Medicine and the Robert Wood Johnson Foundation Clinical Scholars Program, Department of Internal Medicine, Yale School of Medicine; Department of Health Policy and Management, Yale School of Public Health; Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, CT.

Correspondence to Harlan M. Krumholz, MD, SM, Yale School of Medicine, 1 Church St, Suite 200, New Haven, CT 06510. E-mail harlan.krumholz@yale.edu

© 2016 American Heart Association, Inc.

Circ Cardiovasc Qual Outcomes is available at http://circoutcomes.ahajournals.org
DOI: 10.1161/CIRCOUTCOMES.116.003366

The Promise of Big Data
Opportunities and Challenges

Harlan M. Krumholz, MD, SM

Medicine is on the cusp of a remarkable transformative change fueled by the prospect of an acceleration and augmentation in knowledge generation and application. Medicine improved as it transitioned from being based on belief to becoming anchored in science. We were limited, however, in our ability to learn from clinical practice, even though natural experiments were occurring every day. Our methods tended to be reductionist, our data crude, our samples constrained, and our inferences speculative. This inability to learn quickly, easily, and meaningfully from everyday clinical practice is yielding to new capabilities and opportunities.

The shift is occurring as a result of advances in computational power and mathematical applications in concert with a rapid growth in the availability of data. Even changes in the last 5 years have produced breathtaking capabilities. We now have the computational heft to handle large, complex, high-dimensional data. We have mathematical applications that enable us to avoid the need to reduce the complexity of the data, maximizing the relevance of the data to patients seen in practice. And we have just experienced a remarkable advance in data availability through transformation of medical data and documentation into digital format.

A precision medicine future, with the generation of personalized estimates of risk of outcomes and response to various therapies, paradoxically depends on being able to analyze data from massive numbers of people. Only with such numbers can the heterogeneity of risk and response be revealed—and can each individual be understood in the context of the experience of people like them who have preceded them in the situation they are in. The knowledge of individual differences—and similarities—will provide opportunities for truly informed choices based on information that is more relevant, accurate, and meaningful.

To fully leverage this opportunity will require several key steps.

Training
Most researchers are not adequately training for a big data era. The basic principles of epidemiology, study design, and biostatistics, which form the core foundation for most clinical researchers, will remain important, but are not sufficient. There are additional areas of importance, including digital data, informatics, and machine learning. Skills in programming will be a core competency to understand the logic of the programs used to extract information.

Application
The production of the new knowledge is not enough to ensure that benefit is derived from it. The application of the knowledge produced from these data sources and methods will need to be optimized to promote benefit and minimize risk. The culture of medicine will need to evolve from a reliance on old methods of risk stratification, for example, to embrace computer-driven estimates that are linked to methods of producing information about net benefits of different strategies.

Replication
A major challenge of data-driven approaches is to ensure that the findings can be replicated. To depend on the estimates will require confidence that, time and again, they can be reproduced in different data sets, particularly for different subpopulations. The possibility of false-positive inferences—or estimates of risk that are not stable across data sets—is substantial.

Dissemination
The dissemination and communication of the knowledge and methods merits consideration. Our journals are ill-prepared to vet advanced analytics applied to massive data resources. In many cases, the approach may be difficult to evaluate by reviewers and editors. Of note, Circulation: Cardiovascular Quality and Outcomes is committed to finding reviewers, including technical experts, to evaluate novel technical approaches. Moreover, the result may be better suited to a dynamic presentation, rather than the static portable document format. In addition, the visualizations may not lend themselves to the 2-dimensional page. The results of research in this area may be better disseminated as a software product or application, with a paper that describes methods. There will be a need to develop platforms that enable this type of dissemination as part of a scientific contribution. Transparency is also important, even as there may be issues of proprietary algorithms and approaches.

Interoperability
Although data increasingly exist in digital format and provide remarkable opportunities for knowledge generation, obstacles

The opinions expressed in this article are not necessarily those of the American Heart Association.

From the Section of Cardiovascular Medicine and the Robert Wood Johnson Foundation Clinical Scholars Program, Department of Internal Medicine, Yale School of Medicine; Department of Health Policy and Management, Yale School of Public Health; Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, CT.

Correspondence to Harlan M. Krumholz, MD, SM, Yale School of Medicine, 1 Church St, Suite 200, New Haven, CT 06510. E-mail harlan.krumholz@yale.edu

© 2016 American Heart Association, Inc.

Circ Cardiovasc Qual Outcomes is available at http://circoutcomes.ahajournals.org
DOI: 10.1161/CIRCOUTCOMES.116.003366
still abound in coalescing data from different sources. Although there are common data standards for many data types, adherence to the standards vary. Moreover, data holders, independent of concerns for privacy, have often seen data as assets to be hoarded rather than shared, even among those participating in research consortia and public health initiatives. There will be a need for continuing efforts to combine data, with attention to privacy and security, to fulfill their promise to produce knowledge that will help the next person who needs assistance.

**Funding**

Funders are not accustomed to applications that focus on big data as a means of knowledge generation. These projects require infrastructure to make data available, method development to produce tools customized to the particular challenges in medicine, and support for empirical work that may not be hypothesis driven, but is just as capable of producing valuable insights. In addition, funders need reviewers who transcend traditional medical research paradigms if this work is to get the support it needs.

**Collaboration**

Finally, and perhaps most importantly, this work will require new collaboration for clinical research. Advances in medicine are likely to derive from work with mathematicians, statisticians, and computer scientists. Such teams will bring together data scientists and clinical scientists. The recognition that medicine is an information science will begin to spread. These teams will need to invest in learning each other’s culture, language, and capabilities to forge collaborative activities where the whole is truly larger than the sum of its parts.

Medicine already lags behind so many fields in the use of digital data. Our research and research funding is largely anchored in a model that has been stable for decades. Our meetings and journals largely remain bastions of traditional research and their presentation. Our practices on the wards are moving slowly toward the adoption of truly learning health system approaches.

And yet, we are at one of those amazing junctures in medicine that will define a new future as profoundly different from today as the pre-microbiology era was from what preceded it. We may have the chance to see what was formerly undecipherable. But the hazards of change are also great, and our ability to tackle these challenges is far from guaranteed. We will need ways to prove the value of these new approaches and tools—and they should be subjected to the same scrutiny as any new diagnostic or therapeutic intervention.

In this issue, we feature an array of insightful articles that address different facets and applications with respect to big data methods.1–11

The studies are not just about data size, but about approaching the data in new ways. The publication of these contributions signals a receptivity by the data in new ways. The publication of these contributions addresses different facets and applications with respect to big data as a means of knowledge generation. These projects require infrastructure to make data available, method development to produce tools customized to the particular challenges in medicine, and support for empirical work that may not be hypothesis driven, but is just as capable of producing valuable insights. In addition, funders need reviewers who transcend traditional medical research paradigms if this work is to get the support it needs.

**Acknowledgments**

I thank Sharon-Lise Normand for her partnership in coediting this issue.

**Disclosures**

Dr Krumholz is a recipient of research agreements from Medtronic and from Johnson & Johnson (Janssen), through Yale University, to develop methods of clinical trial data sharing; is the recipient of a grant from the Food and Drug Administration and Medtronic to develop methods for postmarket surveillance of medical devices; works under contract with the Centers for Medicare & Medicaid Services to develop and maintain performance measures; chairs a cardiac scientific advisory board for UnitedHealth; and is the founder of Hugo, a personal health information platform.

**References**


The Promise of Big Data: Opportunities and Challenges
Harlan M. Krumholz

Circ Cardiovasc Qual Outcomes. published online November 8, 2016;
Circulation: Cardiovascular Quality and Outcomes is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2016 American Heart Association, Inc. All rights reserved.
Print ISSN: 1941-7705. Online ISSN: 1941-7713

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circoutcomes.ahajournals.org/content/early/2016/11/08/CIRCOUTCOMES.116.003366.citation

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation: Cardiovascular Quality and Outcomes can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at: http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation: Cardiovascular Quality and Outcomes is online at: http://circoutcomes.ahajournals.org//subscriptions/