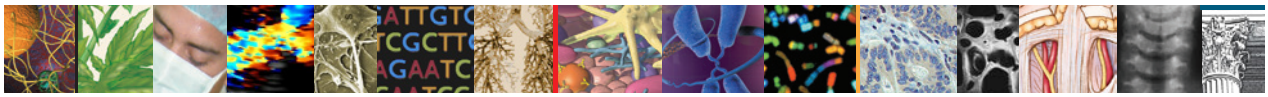

The NEW ENGLAND JOURNAL of MEDICINE

Perspective

Data Sharing and Inductive Learning — Toward Healthy Birth, Growth, and Development

N. L'ntshotsholé Jumbe, Ph.D., Jeffrey C. Murray, M.D., and Steven Kern, Ph.D.

As the international conversation about data sharing shifts from the theoretical ("data sharing is important") to the practical ("this is how data sharing can happen"),[1,2] the Healthy Birth, Growth, and Development–Knowledge Integration (HBGDki) initiative sponsored by the Bill and Melinda Gates Foundation offers one example of how data sharing can be used to improve public health. The initiative aims to develop better interventions for children at risk for faltering growth and neurocognitive deficits. To this end, we have been creating an integrated knowledge base consisting of existing maternal and child health data from 420 clinical and population survey studies in 50 countries, including 137 clinical studies from 26 countries. A team of data scientists has been curating and analyzing the shared data with novel analytic software to explore new questions and develop better strategies to promote healthy birth, growth, and development. The data contributors and maternal and child health experts are collaborating with us to conduct the analyses and interpret the results. Early discoveries from analyses are being validated before publication and are being used to inform decisions about health interventions for children in the communities with the greatest need.

Gathering data for the knowledge base was difficult. More than 90% of the Gates Foundation–funded principal investigators we approached were initially reluctant to share data from studies they performed, citing barriers similar to those that have been described with regard to data from public health agencies — including hurdles related to professional aspirations, economics (the perceived and real costs of data sharing), structural or sovereignty issues (be they political, legal, or ethical), and uncertainty regarding ownership of data-analysis outcomes.[3] We also spoke with principal investigators whose work had not been funded by the foundation. We learned important lessons — about developing a vision for clear and equal reciprocity and addressing concerns regarding data security, quality, and attribution — that helped us build symbiotic, trusting collaborations with many investigators. In addition, we developed a secure analytics platform, stringent data-access protocols, and clear data-use agreements to allay concerns about privacy and security and to facilitate meaningful analyses (see the Perspective article by Merson et al.).

The New England Journal of Medicine
Downloaded from nejm.org on May 24, 2016. For personal use only. No other uses without permission.
Copyright © 2016 Massachusetts Medical Society. All rights reserved.

The purpose of data sharing is not just to amass large numbers of data points. The quality and depth of the data — especially the diversity of covariates — is critically important for making new discoveries in complicated problem areas such as child growth and development. At present, our knowledge base includes some 1700 demographic, clinical, and socioeconomic covariates from more than 8 million children.

When data scientists build statistical models to predict how children may be affected by environmental insults or respond to nutrition interventions, they must distinguish between uncertainty (accuracy of measurements) and heterogeneity (healthy and pathological variability within and among children). Integrated data sets enable us to test correlations and covariation between pertinent variables more effectively than we could with individual data sets, helping us to separate signal from noise, quantify effects at the extremes of distributions, build accurate models, and understand how to give the right intervention to the right child at the right time for the right cost.

We are using this approach to evaluate the relative importance and interaction of multiple and wide-ranging determinants of faltering growth and neurocognitive deficits, including nutrition (both the quantity and quality of foods); infection; gut function; access to clean water, sanitation, and hygiene; and caretaker education level.

Health research that evaluates the risks of the most vulnerable people — women and children in low-income countries — typically focuses on visible threats (above the newsworthy or pandemic-threat threshold), and ignores invisible (below-threshold), developing threats. That limitation is understandable, given that there are so many immediate, above-threshold threats to people's health. However, this focus contributes to our repeatedly being caught flat-footed by diseases that may fester for decades before turning into outbreaks or epidemics.

The process of gathering, sharing, and analyzing high-quality data enables us to make seemingly invisible patterns visible by lowering the threshold for predicting known health threats (e.g., Ebola virus) sooner or responding to expanding threats (e.g., Zika virus) as soon as their clinical impact becomes important. If we were to evaluate disease patterns the same way that meteorologists study weather patterns or air traffic controllers evaluate flight patterns, we could potentially see developing epidemiologic trends in tiny bursts of activity globally. Imagine the impact if we could have predicted the AIDS epidemic of the 1980s with better analytic tools and surveillance capabilities that built on information available as early as the 1930s.

Moral arguments strongly favor data sharing, especially for data generated using philanthropic or public resources. But the practical benefits of data sharing are also compelling. If biomedical researchers continue to share data, the HBGDki and other knowledge bases can become living repositories that advance the field, fulfill the goals of the taxpayers and private funders who enabled the work, and honor the wishes of the participants in the studies. The end result of data sharing, done properly, will be more knowledge that will help all people lead healthy and productive lives.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

From the Healthy Birth, Growth, and Development Core Team, Bill and Melinda Gates Foundation, Seattle.

This article was published on May 11, 2016, at NEJM.org.

1. Longo DL, Drazen JM. Data sharing. N Engl J Med 2016;374:276-7.
2. Taichman DB, Backus J, Baethge C, et al. Sharing clinical trial data — a proposal from the International Committee of Medical Journal Editors. N Engl J Med 2016;374:384-6.
3. van Panhuis WG, Paul P, Emerson C, et al. A systematic review of barriers to data sharing in public health. BMC Public Health 2014;14:1144.

DOI: 10.1056/NEJMp1605441
Copyright © 2016 Massachusetts Medical Society.

The New England Journal of Medicine
Downloaded from nejm.org on May 24, 2016. For personal use only. No other uses without permission.
Copyright © 2016 Massachusetts Medical Society. All rights reserved.