# Data Science in Healthcare
## Implications for Early Career Investigators

Sanjeev P. Bhavnani, MD; Daniel Muñoz, MD, MPA; Akshay Bagai, MD, MHA



## Data Science in Healthcare

The confluence of science, technology, and medicine in our dynamic digital era has spawned new data applications to develop prescriptive analytics, to improve healthcare personalization and precision medicine, and to automate the reporting of health data for clinical decisions.[1] Data science in health care has seen recent and rapid progress along 3 paths: (1) through big data via the aggregation of large and complex data sets including electronic medical records, social media, genomic databases, and digitized physiological data from wireless mobile health devices[2]; (2) through new open-access initiatives that seek to leverage the availability of clinical trial, research, and citizen science data sources for data sharing[3]; and (3) in analytic techniques particularly for big data, including machine learning and artificial intelligence that may enhance the analyses of both structured and unstructured data.[4] As new data sets are created, analyzed, and become increasingly available, several key questions emerge including the following: What is the quality of unstructured data generation? Will the use of nonstandardized methods in data processing with traditional software and hardware lead to data fragmentation and analyses that are nonreproducible? Will healthcare systems incorporate and use big data especially from new publically and patient-generated sources? How will physicians and researchers learn from new open-sourced data and big-data analytics? And ultimately, How can they acquire the skills to create a knowledge translation in data sciences?[5]

## Opportunities and Challenges for the Early Career Investigator

Practicing in an era of continuous payment reform and decline in research funding, early career investigators are challenged to keep up with the accelerating pace of change in medicine, all while being expected to provide meaningful contributions through productive clinical, educational, and research experiences.[6] In this perspective, we aim to highlight how data science can catalyze professional advancement and discuss the implications of big data, open access, and data analytics through 4 main categories for the early career investigator (Figure). These include the following: (1) the evolution and expansion of conventional training programs to incorporate data sciences, (2) changing structure and composition of research teams, (3) new and emerging funding opportunities for data science studies, and (4) academic reward and advancement in the era of open and big data. We aim to provide strategies for how young investigators can maximize benefits and minimize risks through new opportunities afforded by developments in data science.

## Evolution and Expansion of Training Programs

As big data moves into clinical practice, new computer-based predictive analytics such as artificial intelligence and natural language–processing algorithms for precision and personalized health care will invariably change the way clinicians explore, modify, and work with health information. Through big data registries and data analytics, clinicians will need to adapt to understand and rapidly assimilate near real-time health information to support their decision making at the point-of-care. This paradigm shift in our standardized approaches for medical education, clinical exposures, and research methodologies requires a grass roots change in how the current and future generations of healthcare professionals and investigators are educated. Recently, medical schools have started to update their curriculum to incorporate didactic and practice-based modules focused in data science. In this regard, first- and second-year medical students at the New York University are required to participate in Healthcare by the Numbers, a flexible 3-year, individualized, technology-enabled blended curriculum to train and use big data to improve care coordination and quality. In this project, funded in part by a grant from the American Medical Association, students are given access to a database with >5 million deidentified patient records including information on every hospitalized patient in the state for the past 2 years. Through this mandated exposure early in their training, these future clinicians learn to recognize the strength and pitfalls of big clinical databases with the aim to monitor and improve healthcare outcomes.

At present, clinicians trained to understand and assimilate big data analytics are scarce and at a premium. Several strategies exist to extend this analytic skill set to a broader pool of healthcare professionals including didactic training, shadowing or rotating with specialists in clinical informatics and data science, and practice modules created on standardized clinical, genomic, and basic science data sets. Understanding data heterogeneity (accuracy and formatting), data fragmentation (multiple databases and multiple stakeholders), data

From the Division of Cardiology, Scripps Clinic and Research Institute, San Diego, CA (S.P.B.); Division of Cardiology, Vanderbilt University, Nashville, TN (D.M.); and Terrence Donnelly Heart Center, St. Michael's Hospital, University of Toronto, Ontario, Canada (A.B.).

Correspondence to Sanjeev P. Bhavnani, MD, Division of Cardiology, Scripps Clinic and Research Institute, 9888 Genesee Ave, San Diego, CA 92037. E-mail bhavnani.sanjeev@scrippshealth.org

**(*Circ Cardiovasc Qual Outcomes.* 2016;9:683-687. DOI: 10.1161/CIRCOUTCOMES.116.003081.)**

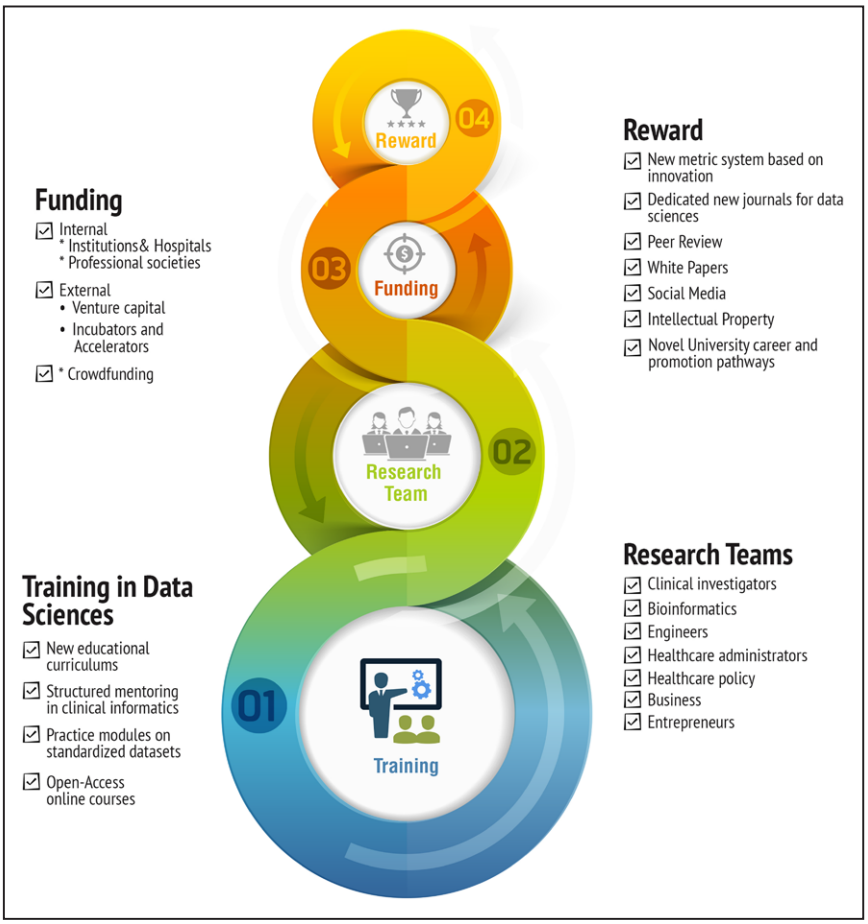


**Figure.** Core components for early career investigator advancement in data sciences.

availability and handling (management, access, querying, and sharing), data privacy and integrity (prevention of corruption and hacking), and data conceptualization (ontologies) are necessary and important skills as clinical investigators navigate health information technology, patient care, research, and administration. Formal education in clinical informatics, computational biology, and data visualization are among the tools that will further position the early investigator for success to ultimately design effective analytic plans using existing databases or from new open-access sources. Several online and open curriculum-based training and certification courses are now available at IBM's Big Data University (www.bigdatauniversity.com) and Cloudera University (www.cloudera.com). These provide a hands-on and practitioner's approach to the techniques and tools required for big data analytics and for the various concepts pertaining to multivendor and multitechnology utilization. Supported by real-world use cases, newly developed didactic and online curricula may leverage industry and academic best practices for their translation to research and for patient care.

## Changing Structure and Composition of Research Teams

If open-access and big data analytics are considered external innovations—that is developed outside of conventional medical educational and clinical arenas—for them to be successful, fundamental changes to the internal structure and composition of clinical and research teams are necessary. Some have proposed standardization and an approval process for access to open-sourced data to ensure that research teams possess the necessary skills to manage, analyze, interpret, and report results from open-access data sets.[7] To be effective, research teams need to not only include clinical investigators but also incorporate individuals with expertise in big data analytics, bioinformatics, technology, engineering, healthcare administration, business and entrepreneurship, and healthcare policy. Similar to the objectives of established data sources such as census and public health data sets, or standardized patient registries such as the National Cardiovascular Data Registry where data are structured and aggregated to monitor population trends, develop guideline-based care, and infer changes to healthcare policy, new citizen science and crowdsourcing initiatives aim to leverage public and patient participation to collect health data and vital statistics through new massive, open, and online data repositories.[8]

Widely available crowdsourcing programs such as PatientsLikeMe (www.patientslikeme.com) have amassed participation from >400 000 patients across 2500 disease conditions who actively share health-related data on an open and online platform that tracks and collects important patient-reported outcomes.[9] The United Kingdom's BioBank is a large-scale biomedical data set containing detailed phenotypic, genotypic, and multimodal imaging findings to determine the genetic and nongenetic determinants of health and

disease in a contemporary cohort of >500 000 participants. Available through open access, research collaborations have advanced our knowledge in the risk prediction of cardiovascular, psychiatric, and cerebrovascular diseases and have identified important anthropometric and genetic traits of metabolic health including diabetes mellitus and obesity.[10] These citizen science and open-access initiatives are creating new data sets that see clinicians, researchers, and patients operating in digital networks and is democratizing the scientific process by transforming research from a purely investigator-centered focus to a publically-participated one.[11]

Apple's Research Kit is a high-profile example of a public–academic–industry collaboration and the creation of an app-based clinical trial coordination platform that has seen tens of thousands of individuals downloading and participating in population-based digital health trials.[12] As a true open-access initiative, the app and software are available on GitHub (www.github.com), a widely popular, public, open-access, and code-sharing platform. With an immediately available and functional trial platform in place, young investigators may see their ideas move rapidly from protocol to execution. Although universal access is a fundamental tenet of such open initiatives, a multidisciplinary team with a diverse expertise may be best suited to generate meaningful insights and research findings. Such teams will be well positioned to effectively tackle different analytic plans resulting from various open-sourced data sets in basic, clinical, and translational science and will be sufficiently skilled to formulate competitive proposals for funding, publication, and subsequent trial designs.

## Conventional and Unconventional Funding Opportunities

In evaluating proposals, highly competitive funding agencies traditionally rely on preliminary data and a proven track record of investigator productivity. Although evolving, established agencies may not have sufficient funding allocated specifically for data sciences or to accommodate a large number of open-access proposals.[13] In pursuit of funding opportunities, early career investigators invariably face concerns stemming from scientific value, preliminary data, and competition. Perhaps, the most productive route to funding the young investigator is to view data science opportunities as a stepwise process that begins with recognizing that meaningful contributions often occur in small increments. Research with new biomedical innovations begin with pilots, efficacy investigations, proof-of-concept, and first-in-man studies.[14] Thus, funding opportunities must parallel the proposed research designs, whether resulting from new data sources or those ideas generated through open access.

We propose 2 mechanisms for early career investigators seeking funding for data science projects. The first is internal and the primary responsibility of the investigator's institution and professional society in which a mutual agenda exists to advance knowledge with new innovations and those resulting from big data and open access. As outlined above, institutions and societies must acknowledge the potential obstacles, risks, and unknowns, and decide on a mutually beneficial funding pathway, administrative support, and provide access to the necessary teams and resources, whether available internally or acquired externally. We agree with Majmudar et al[5] and Dittrich[15] that this may not be applicable everywhere; however, all institutions with a mandate to improve healthcare quality, education, and training are obligated to build an ecosystem that empowers young investigators to succeed along these pathways.

The second mechanism builds on the first and a pathway that sees funding similar to the growth process commonly undertaken by startups companies. In contrast to the long time horizon required to secure RO1 and career pathway funding, startup funding while equally challenging to secure is often of shorter duration especially in the initial get-off-the-ground phase. Early-stage companies commonly go through a phased approach to growth that begins with self and public funding (our example of institutional funding), acquiring seed investments from venture capital and outside funders, and subsequent expansion that scales the innovation from an idea to a deliverable. For the early career investigator, potential funding sources include industry, venture capital, and regional incubators and accelerators. On one hand, academic institutions may view these as unconventional. On the other, they may be increasingly viable as young investigators seek productive pathways and to work with a new generation of companies and organizations focused on healthcare innovations. An attractive and potentially risk-averse pathway is crowdfunding and online campaigns such as www.indigogo.com, www.kickstarter.com, and www.experiment.com. These platforms organize financing for new ideas through public funding in both regional and global settings.[16] Although not peer-reviewed in the established form, or substantiated with academic validation, garnishing sufficient public funding may be associated with a successful vetting process of those ideas that may have the most merit and sustainability. Leveraged by initial seed funding, these ideas may be scalable to the next phase of research and trial design.

## Academic Reward and Advancement

How do academic institutions credibly view new developments in data sciences? Should research findings resulting in intellectual property and commercialization lead toward academic promotion and career advancement? And, should promotion committees continue in a culture of research and publication or evolve into a hybrid that also recognizes the contributions resulting from creating new open-access and crowdsourced data sets or new analytic methodologies? These questions are particularly germane to the development of early investigators seeking a productive and reward-driven pathway that are commonly dependent on the clinical translation of new discoveries and findings in data sciences.[17]

Similar to most disciplines, the advances in data science in health care are not entirely new. Established big data sources including electronic medical records, heath insurance claim databases, and the digitization of radiographic images have used a variety of analytical methods ranging from decision trees, computer-assisted diagnostics, and ridge regression to produce useful learning models for disease classification and prediction. In terms of data sources, what is new are recent

initiatives that aim to provide open access to postpublication clinical trial data sets,[18] the development of new digital infrastructures to search, download, and analyze shared biomedical research such as OpenImmport (www.immport.org) and the Yale Open Data Access (www.yoda.yale.edu),[19] and new crowdsourced and patient-generated data repositories. In terms of data analytics, new analytical methods including machine-learning, artificial intelligence, and cloud-based analytics are being translated from nonhealthcare setting to medicine for clinical decision support, predictive modeling, and therapeutic personalization. Although attractive, several unknowns exist including the validation of new data analytics to conventional diagnostic, risk, and therapeutic approaches, their impact on outcomes, and ultimately the adoption of new data models by academic organizations and healthcare systems. The latter is of particular importance and commonly requires a long time duration that may see proposed analytic methodologies surpassed by newer techniques.

Academic evaluative mechanisms need to be developed for research methodologies with new data sets, open-sourced findings, and new data analytics. For example, granting agencies and promotion committees may view research findings generated from open-sourced databases as not original or credible because data sets were not curated by early career investigators. The sharing of new programming, software, and analytic algorithms may be considered exploratory and hypothesis generating or lead to concerns regarding the replicability of proposed research plans.[20] Early career investigators may see their efforts disseminate in the form of white papers, social media, and open-access and online publications as expanding digital infrastructures for communication and information sharing gain momentum in the medical and scientific arenas. Contributions in these venues may be additive to traditional peer reviewed journals and conference presentations; however, require validation as appropriate metrics for academic productivity. In this context, evaluative systems for data sciences are not required to overcome these concerns but rather to create merit-based pathways that recognize the importance of innovation, technology transfer, and leadership that are outside conventional training environments. Such a system can be complimentary and in parallel to established academic reward and promotions pathways, and one that positions early investigators, directors, and deans along a mutual trajectory toward scholarly achievement and scientific contribution.

## A Path Forward

Data science offers the early career investigator new and promising opportunities to forge a pioneering niche in big data, generate study results from open-access trials, and expand on a multitude of skills that lead to personal and professional growth. As with most new innovations, enthusiasm is curbed by risks. Early career investigators and clinical innovators must acknowledge that failures may, more often than not, outnumber successes especially in a new and rapidly changing discipline that does not have a regulatory or a standardized framework. To achieve clinical and academic productivity requires emersion in new training and educational programs, access for funding from established and unconventional pathways, the creation of research teams to harness multidisciplinary collaborations, and academic advancement that may initially track along hybrid promotion pathways. In the aggregate, these are the functionalities that need to be brought together in new integrated biomedical–computing–research environments. Success will not be measured by our ability to take risks but rather in our preparation for the obstacles and challenges inherent to change. As such, current and future generations of early career investigators may be best poised to move new healthcare innovations in data science from the bench and ultimately to the bedside.

## Acknowledgments

We are deeply indebted to Gary V. Heller, MD, PhD, and Paul Teirstein, MD, for their insights, thoughtful review, and critique of our article.

## Disclosures

S.P. Bhavnani reports receiving an educational and research grant from the Qualcomm Foundation to Scripps Health, is a consultant for Proteus Digital, and is an advisory board member for iVEDIX and Wellseek. A. Bagai is an advisory board member for Astra Zeneca. The other author reports no conflicts.

## References

1. Sengupta PP. Intelligent platforms for disease assessment: novel approaches in functional echocardiography. *JACC Cardiovasc Imaging*. 2013;6:1206–1211. doi: 10.1016/j.jcmg.2013.09.003.
2. Bhavnani SP, Narula J, Sengupta PP. Mobile technology and the digitization of healthcare. *Eur Heart J*. 2016;37:1428–1438. doi: 10.1093/eurheartj/ehv770.
3. Krumholz HM, Gross CP, Blount KL, Ritchie JD, Hodshon B, Lehman R, Ross JS. Sea change in open science and data sharing: leadership by industry. *Circ Cardiovasc Qual Outcomes*. 2014;7:499–504. doi: 10.1161/CIRCOUTCOMES.114.001166.
4. Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. *Nat Rev Cardiol*. 2016;13:350–359. doi: 10.1038/nrcardio.2016.42.
5. Majmudar MD, Harrington RA, Brown NJ, Graham G, McConnell MV. Clinician innovator: a novel career path in academic medicine: a presidentially commissioned article from the American Heart Association. *J Am Heart Assoc*. 2015;4:e001990. doi: 10.1161/JAHA.115.001990.
6. Bagai A, Udell JA. Academic practice plans for early career clinician investigators: the fourth pillar of success. *J Am Coll Cardiol*. 2015;66:1839–1840; discussion 1841. doi: 10.1016/j.jacc.2015.08.864.
7. Strom BL, Buyse M, Hughes J, Knoppers BM. Data sharing, year 1–access to data from industry-sponsored clinical trials. *N Engl J Med*. 2014;371:2052–2054. doi: 10.1056/NEJMp1411794.
8. Topol E. The Patient Will See You Now: The Future of Medicine Is In Your Hands. New York: Basic Books; 2015.
9. Chiauzzi E, Rodarte C, DasMahapatra P. Patient-centered activity monitoring in the self-management of chronic health conditions. *BMC Med*. 2015;13:77. doi: 10.1186/s12916-015-0319-2.
10. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12:e1001779. doi: 10.1371/journal.pmed.1001779.
11. Haug CJ. From patient to patient–sharing the data from clinical trials. *N Engl J Med*. 2016;374:2409–2411. doi: 10.1056/NEJMp1605378.
12. Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, Doerr M, Pratap A, Wilbanks J, Dorsey ER, Friend SH, Trister AD. The mPower study, Parkinson disease mobile data collected using ResearchKit. *Sci Data*. 2016;3:160011. doi: 10.1038/sdata.2016.11.
13. Collins FS. Exceptional opportunities in medical science: a view from the National Institutes of Health. *JAMA*. 2015;313:131–132. doi: 10.1001/jama.2014.16736.
14. Bhavnani SP, Srivastava A, Meyer D, Kuo R, Nowaczyk J, Wolman L, Bardarian S, Heywood T. Digitizing medication adherence monitoring

with a novel ingestible nanosensor - first in man application among recipients of continuous flow left ventricular assist devices. *J Heart Lung Transplant.* 2016;35:S330.
15. Dittrich HC. Cultivating the clinician innovator: is there pay dirt in academic medicine? *J Am Heart Assoc*. 2015;4:e002489. doi: 10.1161/JAHA.115.002489.
16. Dahlhausen K, Krebs BL, Watters JV, Ganz HH. Crowdfunding campaigns help researchers launch projects and generate outreach. *J Microbiol Biol Educ*. 2016;17:32–37. doi: 10.1128/jmbe.v17i1.1051.
17. Sanberg PR, Gharib M, Harker PT, Kaler EW, Marchase RB, Sands TD, Arshadi N, Sarkar S. Changing the academic culture: valuing patents and commercialization toward tenure and career advancement. *Proc Natl Acad Sci USA*. 2014;111:6542–6547. doi: 10.1073/pnas.1404094111.
18. Taichman DB, Backus J, Baethge C, Bauchner H, de Leeuw PW, Drazen JM, Fletcher J, Frizelle FA, Groves T, Haileamlak A, James A, Laine C, Peiperl L, Pinborg A, Sahni P, Wu S. Sharing clinical trial data–a proposal from the international committee of medical journal editors. *N Engl J Med*. 2016;374:384–386. doi: 10.1056/NEJMe1515172.
19. Krumholz HM, Waldstreicher J. The Yale Open Data Access (YODA) project–a mechanism for data sharing. *N Engl J Med*. 2016;375:403–405. doi: 10.1056/NEJMp1607342.
20. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. *Nature*. 2014;505:612–613.

KEY WORDS: access to information ■ automatic data processing ■ diffusion of innovation ■ education ■ electronic health records ■ precision medicine ■ research

# Data Science in Healthcare: Implications for Early Career Investigators

Sanjeev P. Bhavnani, Daniel Muñoz and Akshay Bagai

*Circ Cardiovasc Qual Outcomes.* 2016;9:683-687; originally published online November 8, 2016;
doi: 10.1161/CIRCOUTCOMES.116.003081
*Circulation: Cardiovascular Quality and Outcomes* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 1941-7705. Online ISSN: 1941-7713