Perspective

Biomedical informatics meets data science: current state and future directions for interaction

Philip R.O. Payne,1 Elmer V. Bernstam,2 and Justin B. Starren3

1Institute for Informatics, Washington University in St. Louis, School of Medicine, Institute for Informatics, St. Louis, Missouri, USA, 2The University of Texas Health Science Center at Houston, School of Biomedical Informatics, Houston, Texas, USA and 3Northwestern University, Feinberg School of Medicine, Center for Data Science and Informatics, Chicago, Illinois, USA

Corresponding Author: Philip R.O. Payne, PhD, FACMI, Washington University in St. Louis, School of Medicine, 4444 Forest Park Ave, Suite 6318, Campus Box 8102, St. Louis, MO 63110, USA (prpayne@wustl.edu)

ABSTRACT

There are an ever-increasing number of reports and commentaries that describe the challenges and opportunities associated with the use of big data and data science (DS) in the context of biomedical education, research, and practice. These publications argue that there are substantial benefits resulting from the use of data-centric approaches to solve complex biomedical problems, including an acceleration in the rate of scientific discovery, improved clinical decision making, and the ability to promote healthy behaviors at a population level. In addition, there is an aligned and emerging body of literature that describes the ethical, legal, and social issues that must be addressed to responsibly use big data in such contexts. At the same time, there has been growing recognition that the challenges and opportunities being attributed to the expansion in DS often parallel those experienced by the biomedical informatics community. Indeed, many informaticians would consider some of these issues relevant to the core theories and methods incumbent to the field of biomedical informatics science and practice. In response to this topic area, during the 2016 American College of Medical Informatics Winter Symposium, a series of presentations and focus group discussions intended to define the current state and identify future directions for interaction and collaboration between people who identify themselves as working on big data, DS, and biomedical informatics were conducted. We provide a perspective concerning these discussions and the outcomes of that meeting, and also present a set of recommendations that we have generated in response to a thematic analysis of those same outcomes. Ultimately, this report is intended to: (1) summarize the key issues currently being discussed by the biomedical informatics community as it seeks to better understand how to constructively interact with the emerging biomedical big data and DS fields; and (2) propose a framework and agenda that can serve to advance this type of constructive interaction, with mutual benefit accruing to both fields.

Key words: biomedical informatics, data science, big data

INTRODUCTION

Multiple scientific disciplines apply quantitative methods in the biomedical domain. Over the past several years, “data science” (DS) has attracted a great deal of interest. However, the synergies and distinctions between DS and related fields, particularly biomedical informatics (BMI) are not clear. To facilitate an ongoing discussions relevant to such synergies and distinctions, as well as to help organizations develop optimal approaches for both BMI and DS, it is beneficial to develop an overarching model for the relationship between BMI and DS, while also acknowledging the maxim that “all models are wrong but some are useful.” We will present such a model later in this perspective, informed by the following background and analyses.
Current trends in scholarly work concerned with DS and biomedical informatics

As can be seen in Figure 1, web search activity as measured via Google Trends for the terms “data science” and “big data” have grown considerably since 2004, whereas the search activity for “informatics” and “bioinformatics” have fallen over the same period. Comparable increases in PubMed indexed articles related to “data science” and “big data” have occurred over the same period.

The role of ACMI in assessing the current state of interaction between BMI and DS

We engaged a group of domain experts as part of the 2016 Winter Symposium of the American College of Medical Informatics (ACMI, https://www.amia.org/programs/acmi-fellowship). ACMI serves as the college of elected fellows who have made significant and sustained contributions to the field of BMI, and is affiliated with the American Medical Informatics Association (AMIA, http://www.amia.org). ACMI was first incorporated in 1984, and has grown since then to include over 300 elected fellows from diverse geographic and professional backgrounds. During the 2016 ACMI meeting, 49 such fellows were present and participated in the discussions and focus groups that contributed to this perspective (Table 1). Of note, these fellows self-identified their professional settings and geographic location as part of the meeting registration process.

Background and working definitions

BMI has struggled to agree upon and communicate a succinct definition that would be useful for comparing and contrasting BMI relative to other fields. In part, efforts to accomplish this are complicated by the existence of a number of similar or complementary scientific domains such as Computer Science, Information Science, Statistics, Mathematics, Cognitive Science, Social Science, and multiple Engineering disciplines with biomedical applications. Thus, a person working in this area has a number of choices regarding how to frame his or her work and professional identity. Further, there is an ongoing debate regarding the differences and commonalities that exist across and between various sub-disciplines of BMI, such as: Bioinformatics, Translational Bioinformatics, Clinical Research Informatics, Imaging Informatics, Medical informatics, Clinical Informatics, Health informatics, and Public Health Informatics.

The published literature uses a variety of definitions for DS, BMI, statistics, and computer science implying a variety of relationships between these fields. These relationships can be summarized as follows:

1. “Biomedical informatics is biomedical data science.”
2. “Biomedical informatics . . . overlaps significantly with biomedical data science, the subfield of data science that is concerned with discoveries using primarily clinical and other health-relevant data.”
3. “Statistics = Data Science”
4. “Data science = Computer Science,” mostly in older literature

No definition will satisfy everyone. However, in order to provide a common frame of reference for this paper, we will adopt the definitions proposed by Bernstam et al. These definition are predicated on a set of underlying definitions related to the branch of philosophy that focuses on computing, wherein:

- Data are observations about the world, some of which are meaningful.
- Information refers to the subset of data that are meaningful.
- Knowledge is “justified true belief.”

| Month | Data Science | Big Data | Bioinformatics | Biomedical Informatics | Informatics |
|-------|--------------|----------|----------------|-----------------------|-------------|
| 2004  |              |          |                |                       |             |
| 2005  |              |          |                |                       |             |
| 2006  |              |          |                |                       |             |
| 2007  |              |          |                |                       |             |
| 2008  |              |          |                |                       |             |
| 2009  |              |          |                |                       |             |
| 2010  |              |          |                |                       |             |
| 2011  |              |          |                |                       |             |
| 2012  |              |          |                |                       |             |
| 2013  |              |          |                |                       |             |
| 2014  |              |          |                |                       |             |
| 2015  |              |          |                |                       |             |
| 2016  |              |          |                |                       |             |
| 2017  |              |          |                |                       |             |
| 2018  |              |          |                |                       |             |
| 2019  |              |          |                |                       |             |
| 2020  |              |          |                |                       |             |
| 2021  |              |          |                |                       |             |
| 2022  |              |          |                |                       |             |

Figure 1. Web search activity as measured via Google Trends for the terms “biomedical informatics”, “big data”, “data science”, “informatics”, and “bioinformatics” (2004–Present).

Thus,

- Biomedical informatics is the science of information applied to, or studied in the context of biomedicine and
- Informaticians study information, its’ usage, and effects; in contrast to focusing on data.

Similarly, will use the following definitions for terms big data and DS:

- “Big data are data whose scale, diversity, and complexity require new architecture, techniques, and analytics to manage and extract value and hidden knowledge from it.”

The term data science is often thought of as the science of reasoning upon Big Data. While the term lacks a clear consensus definition, a common definition is:

- “Data science is the study of the generalizable extraction of knowledge from data.”

Further, the recent National Institutes of Health (NIH) draft strategic plan for DS defined DS as:

- “the interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data.”

While not included in these definitions, DS is both enabled and required by “Big Data”. Therefore, we will collectively refer to the terms big data and DS for the remainder of this report as DS, given our primary emphasis on methodological similarities between such areas and the theoretical and applied methods associated with BMI.

Workshop framing questions

To better understand the current and future relationships that can and should exist between the fields of BMI and DS, and to enumer-
ate future directions for BMI as it seeks to engage with the DS community in a constructive and mutually beneficial manner, a multipart workshop was conducted involving a sub-set of ACMI Fellows as was described above. This workshop was comprised of three stages: (1) a set of five short (between 3 and 5 min) presentations by ACMI Fellows who described DS related activities at their organizations and how those activities intersected or interacted with existing BMI-focused entities and/or programs, (2) a series of moderated break-out sessions where small groups of 5–7 individuals each responded to the following questions:

1. How should the BMI and DS communities engage and communicate with each other so as to coordinate effectively?
2. What are the curricular and workforce development needs incumbent to realizing potential synergies between BMI and DS? What are the distinctions between the two fields in this regard?
3. How does an increasing emphasis on DS in biomedicine impact the types of shared resources and capabilities commonly found in biomedical research enterprises as are regularly overseen by BMI academic or operational units?
4. What critical socio-cultural and policy issues need to be addressed relative to data reuse and open science paradigms as they pertain to providing the “input” for research paradigms that leverage DS approaches?
5. How should DS focused biomedical research be funded and sustained, particularly once initial federal investments (ie, the NIH BD2K program) reach the end of their currently allocated resources?

, and (3) a discussion by all participants concerning findings from each break-out group. The key concepts and arguments presented during stage (3) were recorded by the authors in the form of field notes for subsequent thematic analysis. Subsequently, the authors reviewed and thematically summarized the field notes to generate a synthesized set of findings that could inform a set of recommendations. This thematic analysis was performed using a grounded-theory based approach.

While there have been many discussions regarding the relationship between BMI and DS, we believe this workshop represents a first-of-its-kind mixed-methods effort to develop a multi-expert position concerning the intersection of BMI and DS. As such, while it may recapitulate ideas and shared concepts understood by the BMI and DS communities in an anecdotal manner, it still serves to verify and validate such perspectives in a more robust and defensible manner as is needed when considering a topic with a strategic impact on the advancement of science, policy, and practice.

FINDINGS FROM THE 2016 ACMI WINTER SYMPOSIUM

Conceptual model of BMI and DS

An early version of the conceptual model show in Figure 2 was presented to the workshop participants by the authors in order to stimulate discussion. Based on workshop discussion, and the thematic analysis of the field notes collected during the workshop, we have revised the model to visually convey the position of the participating ACMI Fellows that the fields of BMI and DS overlap but are not identical. In addition, our thematic analysis has led the authors to make the following two observations:

1. Every system, whether biological, social, or digital, produces data. This may be through observation or instrumentation in the case of biological systems. It may also be a byproduct of normal functioning in the case of digital systems. DS and the disciplinary areas that it subsumes focus on making sense of this data to “extract value and hidden knowledge.” This process creates new knowledge and insight. This type of conversion of data into information and ultimately knowledge is immediately familiar to any Biomedical Informatician. A subtle, but useful, distinction between the fields of DS and BMI can be found by comparing the verb “extract” in the International Medical Informatics Association definition of Big Data and the Dhar definition, with the verb “use” in the AMIA definition of BMI. This will become more clear in below. Building on this recognition and drawing an analogy to neurology, the preceding phenomena (eg, the differentiation of “extract” vs “use”) could be viewed as an afferent loop, in the same way that sensory neurons bring information from the world into the brain.

2. However, BMI does not stop with the discovery of new knowledge. An important function of BMI is to leverage this knowledge to create, implement, and evaluate new tools that impact the original source system. Continuing the analogy described in observation 1, this action on the original system could be viewed as an efferent loop, analogous to motor neurons that cause the body to interact with the world. Many in the BMI community extend the data to knowledge flow as data to information to knowledge to action. Of note, BMI training programs often use the phrase “tool builder vs tool user” to describe the distinction between a BMI graduate student and other biomedical domain graduate students who adopt or adapt such tools. One example of this is to utilize new knowledge about pharmacogenomics interactions to create better decision support systems for EHRs. In the case of human
systems, this requires leveraging contributions from fields rarely associated with DS, such as: human factors, organizational theory, implementation science, management, cognitive science, and sociology.13

This model can provide a basis for working towards a common understanding spanning the complementary fields of BMI and DS, which ultimately is needed to ensure that advances resulting from the use of DS in biomedicine are cumulative, and not a recapitulation of said experiences and outcomes generated by the BMI field over the past several decades. Further, by positioning one’s work within this model, it is possible to better understand alignment with one or both of these areas. Overall, we believe that such a model is critical to the fields of BMI and DS, as not achieving such a shared understanding has a potential to reduce the potential impact, efficacy, and efficiency of efforts that synergize between these complementary fields.

Specific findings and recommendations
Workshop responses to the individual framing questions are summarized in Table 2.

DISCUSSION AND FUTURE DIRECTIONS

A synopsis of the outcomes of the 2016 ACMI Winter Symposium

The working definitions, models, and outcomes generated by the 2016 ACMI Winter Symposium serve to clarify that BMI and DS are not subsumed by one another, but rather, exist as synergistic disciplines as part of a broader data, information, and knowledge lifecycle. Just as Big Data is often informally defined as “data that is too big for the hardware I have,” BMI, but not necessarily DS, addresses the situation when: “Big Data meets the human condition.” In this way, we believe that questions regarding whether DS subsumes BMI or vice versa, which have been raised in both formal and informal venues over the past few years, are misinformed. It makes no more sense to ask this question than to ask whether the field of statistics subsumes BMI. Rather, DS (like computer science, information science, statistics, social science, etc.) is a contributing discipline to BMI. If we accept this position, then it can consequently be argued that:

1. Biomedical Informaticians must be trained in DS, although not as deeply as dedicated Data Scientists;
2. Biomedical Informaticians have and continue to make significant contributions to the field of DS, and vice versa;
3. Not all BMI is DS and not all DS is BMI; and
4. In this interdisciplinary world, it is to be expected that individuals and organizations will cross-fertilize activities and organizational structures between the fields of BMI and DS.

These arguments can lead to a conclusion that in the future, any well-trained biomedical researcher will need a basic understanding of both DS and BMI. At the same, our efforts to define the synergies and distinctions between BMI and DS, as are reflected in this manuscript, also demonstrate that substantial work is needed to ensure a broad and shared understanding of the critical issues noted in this report.

Limitations and future directions

Our findings are limited by the nature of the convenience sample of domain experts engaged in the formulation of our report as well as the nature and scope of individuals who participated in the 2016 ACMI Winter Symposium. Further, the stratification of backgrounds for such domain experts was limited to the self-identification of professional and geographic settings. However, we believe that these perspectives remain generalizable and representative of a key cross-section of the BMI and DS communities due to the intrinsic composition of the ACMI membership. That being said,
Question 2: What are the curricular and workforce development needs incumbent to realizing potential synergies between BMI and DS? What are the distinctions between the two fields in this regard?

Finding 3: It is important for the BMI and DS communities to design and deliver effective curricular frameworks that ensure mastery of critical competencies incumbent to both domains.
- Achieving this goal will require the creation of better pipeline for trainees who can enter the two field; and
- It will be important to engage potential employers to ensure that such curricula are harmonized with workforce development needs.

Finding 4: It will also be important articulat in an accessible manner what unique capabilities BMI and DS training provides to individuals, and how differentiation therein will lead to impact and outcomes in “real world” settings depending on educational and competency requirements as are needed to achieve such results.

Question 3: How does an increasing emphasis on DS in biomedicine impact the types of shared resources and capabilities commonly found in biomedical research enterprises as are regularly overseen by BMI academic or operational units?

Finding 5: Addressing this question will require the initial creation of inventories that align such services and capabilities with end user needs and requirements. Such services and capabilities can include (but are not limited to):
- Data storage
- Data “wrangling”
- Computational methods
- Quantitative methods
- Visualization

Finding 6: Our community must be mindful of existing cores and services and determine how to leverage to infrastructures, rather than re-invent them, and therefore, reduce economies of scale.

Finding 7: Funding models for shared BMI and DS infrastructure and services must be rationalized with demand and operational environments. Critical issues to be addressed in this regard include:
- Who is responsible for data capture and management and who serves as the owner/steward/champion for such data assets;
- How do we adequately address multi-disciplinary interactions around said services/capabilities and ensuing stakeholder engagement; and
- Who pays for such services and capabilities and how are they rendered sustainable, particularly as they become an essential and necessary substrate for modern biomedical research, education, and practice.

Finding 8: Recent publications in the biomedical domain concerning the re-use of primary data for secondary research purposes have made arguments that such approaches should be discouraged. This perspective is very much at odds with the substantial benefits afforded by the responsible and rigorous secondary use of data as have been shown in many other scientific fields. Such approaches should be encouraged and supported, rather than discouraged.

Finding 9: Achieving the goals of Finding 1 relative to this question will require the BMI and DS communities further develop:
- Case studies that show the explicit and implicit value of data reuse (reproducibility, cumulative open science, meta-analyses, etc.);
- Metadata standards that facilitate data reuse;
- Collaborative platforms for team-oriented open innovation;
- Mechanisms and incentive structures to empower data generators as part of derivative projects/teams;
- Inventories of best practices from other countries to inform US-based efforts in this regard; and
- Mechanisms to provide credit to data generators when their data is used in other contexts/projects.

Finding 10: The BMI and DS communities need to overcome stumbling blocks for funding generated by NIH administrative “overhead” so as to ensure the timely and impactful implication of basic and applied research effort therein.

Finding 11: The NLM should be the home for both BMI and DS within the NIH. The NIH must significantly increase the scale, scope, funding, and broad recognition of the NLM in such a capacity.

Finding 12: There remain substantial challenges surrounding cross-fertilization and understanding throughout NIH relative to the necessary and appropriate roles of BMI and DS and practice.

Finding 13: The BMI and DS communities need to better understand and engage a broader portfolio funding mechanisms beyond the NIH (such as NSF, PCORI, AHRQ, Foundations, Corporations, etc.), to minimize the impact of funding decisions in any single program or source on the progress of the broader field.

BMI: biomedical informatics; DS: data science; NLM: National Library of Medicine; NIH: National Institutes of Health; NSF: National Science Foundation; PCORI: Patient Centered Outcomes Research Institute; AHRQ: Agency for Healthcare Research and Quality.

it is our hope to extend and expand such information gathering via future qualitative and survey-based research involving a broader cross-section of domain experts and informants. Such future work should serve to bolster and contextualize the important findings and lessons learned that are reflected in this perspective.
CONCLUSIONS

Ultimately, addressing the current interactions between DS and BMI, and future directions for collaboration between those fields, needs to be pursued in a timely and transparent manner. Such dialog can and should lead to sustained and constructive interaction amongst all key constituencies involved in the BMI and DS communities, with demonstrable benefits in terms of the speed, efficiency, and impact of ensuing efforts as they collectively seek to improve the human condition using data, information, and knowledge.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the contributions to this manuscript made by the participants and discussants at the 2016 ACMI Winter Symposium as well as the AMIA staff who coordinated the meeting.

CONTRIBUTORS

The authors (P.P., E.B., and J.S.) contributed equally to the conception of this manuscript; the acquisition, analysis, or interpretation of relevant qualitative and quantitative data for the work; the preparation of the manuscript content, and the final approval of the version to be published.

Conflict of interest statement. None declared.

REFERENCES

1. Box GE, Draper NR. Empirical Model-Building and Response Surfaces. New York, NY: Wiley; 1987.
2. Cimino JJ, Sholtliffe EH. Biomedical Informatics: Computer Applications in Health Care and Biomedicine (Health Informatics). Springer-Verlag New York, Inc.; 2006.
3. Kulikowski CA, Sholtliffie EH, Carrie LM, et al. AMIA Board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline. J Am Med Inform Assoc 2012; 19 (6): 931–8.
4. Ohno-Machado L. Data science and informatics: when it comes to biomedical data, is there a real distinction? J Am Med Inform Assoc 2013; 20 (6): 1009.
5. Brennan PF, Chiang MF, Ohno-Machado L. Biomedical informatics and data science: evolving fields with significant overlap. J Am Med Inform Assoc 2018; 25(1):2–3.
6. Computer Science. Wikipedia; 2018. https://en.wikipedia.org/wiki/Data_science. Accessed March 27, 2018.
7. Bernstam EV, Smith JW, Johnson TR. What is biomedical informatics? J Biomed Inform 2010; 43 (1): 104–10.
8. Luciano F. Semantic Conceptions of Information. Stanford University; 2017. https://plato.stanford.edu/entries/information-semantic/. Accessed March 27, 2018.
9. Bellazzi R. Big data and biomedical informatics: a challenging opportunity. Yearb Med Inform 2014; 9 (1): 8.
10. Dhar V. Data science and prediction. Commun ACM 2013; 56 (12): 64–73.
11. NIH Strategic Plan for Data Science. Bethesda, MD: NIH; 2018.
12. Friedman CP. A “fundamental theorem” of biomedical informatics. J Am Med Inform Assoc 2009; 16 (2): 169–70.
13. Payne PRO, Pressler TR, Sarkar IN, Lussier Y. People, organizational, and leadership factors impacting informatics support for clinical and translational research. BMC Med Inform Decis Mak 2013; 13 (1): 1.