Harvard Data Science Review • Issue 2.2, Spring 2020

On the Convergence of Epidemiology, Biostatistics, and Data Science

Neal D. Goldstein¹ Michael LeVasseur¹ Leslie A. McClure¹

¹Department of Epidemiology and Biostatistics, Dornsife School of Public Health, Philadelphia, Pennsylvania, United States of America

Published on: Sep 29, 2020

DOI: <https://doi.org/10.1162/99608f92.9f0215e6>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

Epidemiology, biostatistics, and data science are broad disciplines that incorporate a variety of substantive areas. Common among them is a focus on quantitative approaches for solving intricate problems. When the substantive area is health and health care, the overlap is further cemented. Researchers in these disciplines are fluent in statistics, data management and analysis, and health and medicine, to name but a few competencies. Yet there are important and perhaps mutually exclusive attributes of these fields that warrant a tighter integration. For example, epidemiologists receive substantial training in the science of study design, measurement, and the art of causal inference. Biostatisticians are well versed in the theory and application of methodological techniques, as well as the design and conduct of public health research. Data scientists receive equivalently rigorous training in computational and visualization approaches for high-dimensional data. Compared to data scientists, epidemiologists and biostatisticians may have less expertise in computer science and informatics, while data scientists may benefit from a working knowledge of study design and causal inference. Collaboration and cross-training offer the opportunity to share and learn of the constructs, frameworks, theories, and methods of these fields with the goal of offering fresh and innovative perspectives for tackling challenging problems in health and health care. In this article, we first describe the evolution of these fields focusing on their convergence in the era of electronic health data, notably electronic medical records (EMRs). Next we present how a collaborative team may design, analyze, and implement an EMR-based study. Finally, we review the curricula at leading epidemiology, biostatistics, and data science training programs, identifying gaps and offering suggestions for the fields moving forward.

Keywords: epidemiology, biostatistics, data science, training and education, causal inference, study design, electronic medical records

1. Introduction: A Confluence of Concepts

The fields of epidemiology, biostatistics, and data science, while very distinct in their focus on training, share much in common in that they all rely upon an intersection of various and overlapping concepts. These concepts include statistical methods, research design, and substantive expertise. Rigorous analysis of quantitative data is the common thread among them. When data science is applied to health and medicine for understanding disease etiology, the distinction between the fields becomes blurred. For the data scientist engaging in health-related research, epidemiology and biostatistics provide appropriate complementary knowledge and skillsets through the application of causal inference theory, meticulous study design and measurement, and the development of new statistical methods. Likewise, for the epidemiologist working with massive amounts of health care data, data science provides innovative and robust computational and visualization approaches for high dimensional data that may not be traditionally taught in epidemiology training programs, while biostatistics brings novel statistical methods that could improve inference about the data. For a biostatistician

concerned with developing new methods that lessen bias or reduce variance in a particular field, the epidemiologist can bring topic-matter expertise and data, while the data scientist can play a key role in improving the computational aspects of the approach. In short, there is much to be shared across fields, as well as much contributed from each expert, as exemplified in Figure 1. The epidemiologist and biostatistician may lack computer science skills (labeled as hacking skills and including database design, data management, and informatics), the data scientist may lack research expertise in terms of causal inference, study design, and measurement (envisioned as a third dimension in this figure, intersecting the center), and both the data scientist and epidemiologist may lack the background in statistical theory necessary to improve on the current methodology.

Figure 1. The data science Venn diagram. Reprinted under the Creative Commons license (Conway, 2013).

A brief examination of the history of these fields reveals a natural convergence over time centered around the increasing amount of data available for analyses. Biostatistics surfaced around the mid-1800s for measuring human traits, as well as quantifying morbidity and mortality, but statistical methods applied to health data really took off during the late 1800s with the availability of genetics data (Salsburg, 2001). Meanwhile, epidemiology as a distinct discipline evolved from medicine in response to public health infectious disease crises of the 19th century (Rosen, 1993). As public health research diverged from an infectious disease perspective to a chronic disease perspective, methods were developed to specifically mitigate the effects of bias and confounding resulting from nonexperimental study designs (Greenland, 1987; Susser & Stein, 2009). Biostatistics has provided much of this methodology, and training in biostatistics has emphasized basic programming and data management skills, with this emphasis growing as statistical software has become more readily available. With the rise of electronic medical records (EMRs) in the later part of the 20th century (Henry et al., 2016), as well as the ever-increasing amount of health-related data in other disciplines, the modern-day epidemiologist and biostatistician continue to evolve to better understand disparate data sources.

Data science became formalized during the mid-20th century and is a comparatively new field. Recognizing the need for computer scientists to not only define and develop software and hardware platforms, but to analyze the data captured electronically therein, a cross-disciplinary approach was proposed that incorporated the rigor of various computational approaches with statistics (Cleveland, 2014). Yet it is not solely an applied discipline with a focus on algorithmic development, such as machine learning, or statistics (Meng, 2019). As with epidemiology, it extends beyond methods: formalized theory and frameworks help to define the training and skillset necessary for the data scientist (Borgman, 2019; Floridi & Cowls, 2019). Rooted in common among epidemiology, biostatistics, and data science has been how quantitative—and more recently, qualitative—data can be used to answer research and programmatic questions, including important questions that can be answered with electronic health data. We believe these disciplines have much to learn from and share with each

other, and thus we discuss the education, skills, and competencies that a modern-day researcher who works with electronic health data must possess.

We begin with a motivational research question using data derived from the electronic medical record (EMR), which has become frequent with the near ubiquity of EMRs in medical practice (Henry et al., 2016). We then proceed with a broad overview of health-related research as it applies to etiological questions: did some exposure cause disease? We return to our example of an EMR-based study to demonstrate the complementary roles of epidemiology, biostatistics, and data science in addressing our research question, and then conclude by discussing the state of formalized educational programs in the United States and provide recommendations for cross-training moving forward.

2. A Motivational Example of an EMR-Based Research Question

Suppose a research group is interested in conducting a study on whether the number of occupied beds in an intensive care unit (ICU) is related to risk for infection (Goldstein et al., 2017). The researchers hypothesize that the higher the ICU's occupancy rate, the more likely it is for basic hygiene practices to break down, thus leading to increased exposure to pathogenic organisms such as methicillin-resistant *Staphylococcus aureus*. The researchers expect the patient's admitting diagnosis, comorbidities, and length of ICU stay may also be related to the hypothesis. These data can all be ascertained from the EMR.

Figure 2. Simplified architecture of an electronic medical record system as it relates to our research question: Does the number of occupied beds in an intensive care unit increase risk for infection?

Answering research questions that involve EMRs is inherently cross-disciplinary. EMRs are complex data systems, and require expertise in databases, data linkage, and data abstraction to compile the analytic sample (Figure 2). Meanwhile, understanding risk of infection in a health care environment represents a web of causality: there are many potential factors that could explain the outcome, requiring sophisticated methodologies to unpack. Further, approaches to assessing and mitigating potential biases arising from the data, including from incomplete data, are important to providing the best answer to the researchers' question. In our view, research that utilizes EMR data is beyond the bounds of any single field in isolation, with the best answer to questions such as this arising as a result of team science, including our clinical colleagues. Indeed, a 2018 article exemplifies the potential of data science as part of the team when conducting EMR research: the authors had to mine free text clinical notes in an EMR to derive social risk factors that may otherwise be discrete variables in a prospectively designed epidemiological study (Navathe et al., 2018).

Continuing our hypothetical example, the epidemiologist suggests assembling a retrospective cohort from the EMR records for a one-year period and the data scientist is able to interface with the EMR, retrieve a patient list, and abstract all of the variables necessary for analysis. The biostatistician conducts a rigorous analysis,

including assessing completeness of the data and identifying potential biases in the analysis. The team observes a strong relationship between an increased number of occupied beds and increased risk for infection in the data. Does this reflect some underlying causal relation?

3. Public Health Methodology for the Data Scientist: When Does Correlation Equal Causation?

Public health researchers are trained in the art and science of causal inference—the process of evaluating whether a health-related outcome would have been affected given a change in an exposure. Epidemiologists evaluate causal inference using two separate—but equally important—factors: internal and external validity. Internal validity refers to the ability of a study to correctly ascribe the true underlying relationship within the confines of the study. External validity refers to the ability of a study to correctly ascribe the true causal relationship outside of the confines of the study—that the results are generalizable and transportable. Biostatisticians help ensure studies are designed to maximize both internal and external validity, while also developing statistical methods that better answer the scientific questions posed by public health and clinical researchers. Together, biostatisticians and epidemiologists have developed and adapted numerous methods and study designs to reduce the threats to both internal and external validity and have employed methods for assessing the effect of these threats (Morabia, 2004).

Threats to internal validity include random error, bias (aka systematic error), and confounding. These threats are most often, but not exclusively, found in observational studies, such as our EMR-based example. Random error can be minimized through appropriate sample size and power calculations, although any given study may have the possibility of arriving at an erroneous conclusion on the basis of chance alone. Multiple studies conducted using different study designs in similar settings can increase confidence that the results are not due to chance alone, although researchers need to be aware of effect heterogeneity whereby the same analysis conducted in different samples may produce striking, albeit real, differences (Madigan et al., 2013).

Broadly speaking, bias can be classified as selection bias or information bias. Selection bias occurs when individuals from an eligible population have a differential probability of being included in a study based on both their exposure and outcome status. Information bias occurs when there is a systematic tendency to erroneously measure the effect, its antecedent cause, or any other covariates that are involved in the exposure to outcome relation. It is important to note that while increasing sample sizes (as the ‘big data’ movement is witness to) can increase the precision of an estimate, it does nothing to mitigate the effects of bias. That is to say, if there is bias present in one’s data, having a larger sample size only means that one has more precisely measured a biased effect.

Confounding occurs when some factor is causally associated with the outcome and noncausally associated with the exposure. This results in a ‘mixing’ of effects that distorts the true effect between exposure and outcome. While some consider confounding to be a form of bias, the key difference between bias and confounding is that

bias is artificially introduced by the researcher whereas confounding exists in nature. Bias and confounding may never be completely removed from a study: the goal is to understand its presence and potential impact on the observed association. Interested readers are referred to the field of quantitative bias analysis (Lash et al., 2009).

Mitigating the effects of bias and confounding occurs at every stage of public health research, including study design (understanding the influence of study design on bias, reducing selection bias in sample selection), data collection (properly measuring variables of interest, reducing the amount of missing data, collecting all data that may be relevant to the evaluation of confounding), data management (properly coding variables, formatting data sets to best answer the research question, summarizing variables in meaningful ways, managing missing data), analysis (using the correct statistical techniques, model building, confounder assessment), and interpretation (appropriate interpretation of the results, sensitivity analyses to assess the influence of bias, error, and assumptions). Even the best designed and executed studies have flaws and may not be externally valid. In fact, randomized controlled trials, which aim to eliminate bias and confounding by randomizing people to treatment, often have strict inclusion criteria that make their inferences nongeneralizable (Rothwell, 2005). Thus, solving one problem often leads to others.

Epidemiologists reflect upon philosophical as much as practical matters when it comes to their approach to science. Before engaging in designing a research study, the epidemiologist, often in collaboration with a biostatistician, will formulate a research question ensuring that it is answerable methodologically. This question ultimately influences the type of study to undertake. Study design has important implications for application of correct analytical procedures and causal inference. Practically, health researchers consider two main categories of study design—observational and experimental—the key difference being the manipulation of the exposure. Experimental studies, such as randomized clinical trials, allow the researcher to manipulate the exposure, whereas observational studies do not. For example, had our motivational research question been, ‘Will altering the ICU admission process and bed location reduce the risk for infection?’ the study design would have been experimental, as the investigators would be directly manipulating the treatment, in this case, the patient admission process. Randomized trials are often considered the gold standard for assessing causality but are infeasible in many research projects. Contrast this to our stated research question, “Does the number of occupied beds in an ICU relate to risk of infection.” This question does not manipulate any exposure effect (we are not moving around patients after all); rather, we simply observe what happens naturally over time in the ICU. This is considered an observational study, the mainstay of epidemiology. Observational studies can further be subdivided into other study types: cross-sectional studies (sometimes termed a prevalence study), case-control studies, and cohort studies, with a variety of hybrid designs possible (Celentano & Szklo, 2018; Rothman et al., 2008; Szklo & Nieto, 2018). A defining factor among different observational study designs is the timing of the exposure and outcome. Cross-sectional studies evaluate both exposure and outcome at a single time point, case-control studies can assess retrospective exposures against an observed health outcome, while cohort studies can either prospectively or retrospectively assess new cases of some outcome given an

exposure. Cohort studies are considered the most flexible, albeit the most time-consuming and can be expensive. Studies performed from a health care system's EMRs, in which a group of patients are followed over time in either the inpatient or outpatient settings, are typically of a (retrospective) cohort nature, as in our example.

Before data are extracted and analyses commence, proper study design can help minimize selection bias and random error. Having a sound theoretical model can help to identify all relevant confounding variables included in the analysis, and equally important, exclude nonrelevant variables. Causal diagrams such as directed acyclic graphs, are conceptual tools that help with variable selection and understanding variable interplay (Rothman et al., 2008). This is especially important given the vast amount of data available in the EMR. During data abstraction and variable operationalization, the research team needs to ensure that all variables have been recorded properly and, to the best of their knowledge, represent the truth, by working with clinical and informatics colleagues. This will hopefully mitigate information bias.

4. Bringing Data Science to Health Research: More Than Just Machine Learning

Data scientists employ a variety of sophisticated methods that noncomputational researchers may not be aware of. Machine learning and artificial intelligence algorithms, one of the many methodological tools of the data scientist, are becoming increasingly utilized in a variety of fields and have advanced causal inference approaches used by epidemiologists and biostatisticians. Various algorithms exist that represent a data-adaptive approach in estimation of causal inference parameters, including targeted minimum loss-based estimation, double/debiased machine learning, and improved construction of propensity scores and their inverse probability weights for predicting exposures (Blakely et al., 2019; Diaz, 2020). So-called Super Learner algorithms may prove useful in the search for candidate risk or protective factors for a given health outcome (Naimi & Balzer, 2018).

At this point, we wish to draw a clear distinction between predictive and causal modeling, and caution against using machine learning and artificial intelligence for the latter, as others have noted (Lin & Ikram, 2019). In a predictive model, one seeks candidate factors from among a larger set that are statistically associated with an outcome. This is useful for hypothesis generation but may lead to the identification of spurious associations. Occasionally a correlation between covariates in the data may be present but not intervenable and possibly irrelevant from a clinical perspective. This can be especially problematic with high-dimensional data where statistical associations may arise but have little meaning (Lin et al., 2013). In causal modeling, a specific exposure is examined to test its causal relation with the outcome, potentially to intervene upon (if harmful) or promote (if protective). Indeed, machine learning and artificial intelligence may not be the panacea to health and health care problems that some have anticipated; without careful scrutiny and regulation, there is the potential for harm (Kaiser Health News, 2019).

Importantly, the data scientist's repertoire extends beyond the more recent innovations of machine learning and artificial intelligence (Meng, 2019). For example, data scientists may be versed in sophisticated approaches to data collection, database system management, novel visualization techniques, complex system modeling, the software development lifecycle, data security, data privacy, and algorithm ethics, among others areas.

Depending upon the data scientists formalized training, even more specialized expertise may be available. For example, data scientists with backgrounds in computer science or software engineering can develop algorithms, program statistical simulations, and optimize existing analyses. Data scientists with expertise in privacy and security can help unpack the complex requirements of sharing and releasing health data inherent in many types of epidemiological research and implement innovative solutions (Goldstein & Sarwate, 2016). Data scientists who are knowledgeable in linguistics can help create discrete variables from free text in the EMR, such as progress notes, through natural language processing, and data scientists who work with high-dimensional data can assist with automated extraction of data from the EMR.

Returning to our hypothetical example, the methods and tools of the data scientist can aid in the investigative process. Suppose the researchers are confident that the observed relation—the number of occupied beds and risk for infection—is not due to chance, bias, or confounding. Attention turns to understanding the mechanism of risk, as well as possible interventions. The data scientist may employ novel visualization techniques to reveal time- and place-based depictions of patients in the ICU, as well as health care workers serving as the pathogen vectors. There may be algorithmic approaches to identifying other salient risk factors in the environment that are intervenable. For example, if the researchers were considering several candidate factors and how they might relate to the infection, one may decide to employ a predictive Super Learner model to generate hypotheses. The data scientist may further be able to collaborate on the development of a complex system simulation of the ICU environment and introduce infection prevention practices to evaluate the potential for staving off pathogen transmission. Theoretically, this simulation model may even reveal the opportune time for an infection prevention practice, such as hand hygiene, to occur.

It is also important to note that the sophisticated techniques we describe are not employed haphazardly. Rather, there is a focus on sound engineering principles, such as testability, maintainability, integrity, reproducibility, and so on. The systems development lifecycle taught in many engineering programs creates a formalized process for planning, creating, testing, and deploying a data system (Figure 3) (Information Management and Security Staff, 2003, Chapter 1). This process can further be decomposed; for example, deploying includes implementation, operations, maintenance, and obsolescence. Continuing with the hypothetical example, the researchers have observed an empiric association in the data between ICU census and risk for infection. The collaboration to develop a complex systems model of the ICU has identified a point of intervention: namely, a hand hygiene reminder at the opportune time. Now, the data scientist, armed with the empiric data obtained from the simulation, can begin the process of deploying such an intervention into the ICU in collaboration with the research team, with careful consideration of testing the algorithm, the appropriate type of implementation (e.g., integrated within the EMR versus a stand-alone application), evaluating the ongoing operation of this

algorithm, including any corresponding maintenance, and planned obsolescence. The epidemiologist and biostatistician can provide expertise in implementation of the intervention and can design an implement an evaluation of its effectiveness, which could in turn result in further refinement by the research team. Truly, this is a cross-collaborative iterative process.

Figure 3. The traditional systems development lifecycle. Adapted from Information Management and Security Staff, 2003, Chapter 1.

5. Opportunities for Training: Brick and Mortar Barriers to Collaboration

Given the importance of a collaborative model in health research, the question as to whether students are afforded an opportunity to cross-train arises. To assess the current state of formalized training in epidemiology, biostatistics, and data science, we undertook a review of curricula as of the Fall 2019 academic year at the top 20–ranked U.S. News and World Report public health programs (U.S. News and World Report, 2019). For each program, we evaluated the curriculum for each master’s level epidemiology, biostatistics, and data science degree-granting program to assess three factors: 1) the program offering the degree, 2) whether an epidemiology or statistics course (for data science) or data science course (for epidemiology or biostatistics) was required, and 3) if not required, whether these courses were available as an elective. We chose to use as the basis of our review public health program rankings in the United States as opposed to data science program rankings for several reasons. First, to our knowledge, no equivalent list exists ranking the top data science programs. Second, data science is inherently a cross-discipline field and can be housed in schools of engineering, computer science, business, and others. Thus, any ranking system specific to these broader disciplines would be incomplete or include unrelated degrees. Third, as one of our aims was to assess whether an epidemiology component was included in the data science programs, having access to the appropriate faculty would likely necessitate formalized public health degrees at the institution. Therefore, this review can be viewed as the top 20 public health programs in the United States, and whether these universities also offer master’s-level degrees in data science.

There are several other qualifiers to our review we wish to highlight at the outset. Our interest in master’s programs is because they represent a degree that is most likely to be sought by those doing applied work, as opposed to the more academic-focused goals of a doctorate. When assessing whether an epidemiology or biostatistics training program included coursework in data science, we considered courses with a primary focus in computational science (aside from statistical computing), informatics, or data management to be sufficient to label as data science. Likewise, when assessing whether a data science training program included coursework in epidemiology, we considered courses with a primary focus in study and experimental design or causal inference to be sufficient to be labeled as epidemiology, even if not taught by faculty in public health.

Occasionally, a university offered competing or similar programs of study out of different schools. In this case, we focused on the program labeled as or most directly aligned with data science (e.g., as opposed to a degree in health informatics or business analytics). Additionally, if a program offered multiple master's degrees, we evaluated the research-focused degree (e.g., a master's in science superseded a master's in public health). When evaluating electives, unless explicitly indicated in the curriculum, we included only courses within the school/college offering the degree.

Figure 4. Results of the curriculum review for inclusion of data science in epidemiology (A) and biostatistics (B) training programs; and inclusion of epidemiology (C) and statistics (D) in data science training programs.

Among the 22 reviewed programs (top 20 plus ties; Figure 4 and Supplement 1), all conferred epidemiology- and biostatistics-related degrees as part of a master's in public health or a master's in science, and 18 (82%) conferred a data science–related degree most often as a master's of science, although several nonthesis and engineering degrees were available. Data science degrees were offered out of a variety of schools and colleges, indicating the cross-disciplinary nature of the field (Supplement 1). Most commonly, these data science programs were found in Schools and Colleges of Arts and Science ($n = 5$, 28%) and Engineering ($n = 6$, 33%). In some cases, the program was housed in an interdisciplinary institute such as Brown University's Data Science Initiative and University of Washington's eScience Institute.

A data science component was more often required in a biostatistics training program ($n = 6$, 27%) than in an epidemiology training program ($n = 4$, 18%), and it was frequently available as an elective for both ($n = 14$, 64% for biostatistics; $n = 16$, 73% for epidemiology). Most often, data science coursework was available through a biostatistics course designation, suggesting an alignment of data science with biostatistics. Contrast this with an epidemiology component in a data science program, where it was required in a similar proportion ($n = 4$, 21%) but less frequently available as an elective ($n = 2$, 11%). Statistical coursework in a data science program, being a core component of the discipline, was offered more frequently ($n = 16$, 89%), though it was not ubiquitous.

Several programs warrant specific comments or highlighting. Harvard University offers both a master's of science in data science through the School of Engineering and Applied Sciences and a master's of science in health data science through the School of Public Health. This latter degree explicitly included an epidemiology requirement, whereas the former did not. The University of North Carolina (epidemiology and biostatistics), University of Washington (biostatistics), and University of Pittsburgh (biostatistics) offered a data science–

specific track in their public health programs. The University of Michigan at Ann Arbor School of Information offered a health data science concentration in their master's of applied data science program, which included coursework in experimental design and analysis. The University of Pittsburgh offered data science tracks within two programs: a master's in health informatics through their School of Health and Rehabilitation Sciences and a master's in information science through their School of Computing and Information, with both emphasizing the requirements of analyzing large data sets regardless of discipline, common in data science. Yale University's Department of Statistics and Data Science offered a postgraduate certificate in data science without a formal degree program—a nondegree option we suspect is available elsewhere—while the University of Iowa College of Liberal Arts and Sciences offered an undergraduate degree in data science. The only program to offer a specific course in causal inference in a data science program was the University of California at Berkeley School of Information, albeit as an elective. More often than not, the epidemiology component was either satisfied by taking a course directly through the public health program or an independent course in study and experimental design.

Despite the importance of cross-training students to prepare them for collaboration, there are physical 'brick and mortar' barriers to doing so. Our review uncovered that epidemiology and biostatistics are traditionally taught in Schools and Colleges of Public Health (clinical epidemiology is also offered in many medical schools, although we did not explicitly evaluate this subdiscipline), whereas data science was more likely to fall under Schools and Colleges of Engineering, Arts and Science, or Information. In cases where data science was housed in Schools or Colleges of Engineering or Computing and Information, we observed statistics was less often a required course, compared to Schools or Colleges of Arts and Science or Public Health offering data science degrees. Data science appeared to be a hybrid program more often than epidemiology, meaning the training drew on expertise across departments and programs more often. Yet the fields are still siloed: aside from the programs specific to health data science, we did not observe any data science program that included coursework from an epidemiology training program, whereas most programs do include a statistics or biostatistics course. Unfortunately, being physically in separate spaces may hamper collaboration, as some training programs stipulated that coursework may not be derived from outside of the school or college. This separation may also translate into research barriers among established faculty, not just students. We believe a primary reason for the close relationship between epidemiology and biostatistics is due to the fact that they are located in the same school or college, if not the same department.

6. Discussion: What Does the Future Hold?

Data literacy underscores our themes in this article. Data are inextricably embedded in everything we do as researchers; we all struggle with issues of data quality, measurement error, bias, and missing data. Training students to understand the possibilities, and more importantly, the limitations of data is paramount. As was argued in the first issue of HDSR, the approach to training data scientists can be tiered, with different levels of theoretical and methodological expertise depending on the type of student (Garber, 2019). This is also true for

epidemiologists and biostatisticians: while they do not necessarily need to become experts in machine learning, artificial intelligence, database systems, and other data science approaches, they do need a foundational knowledge to enable them to communicate across members of the interdisciplinary teams required to answer important scientific questions today. This has been recognized by recent efforts to bring a related data-driven discipline into the public health training program: informatics (Dixon et al., 2015).

In our view, specialized health data science degrees and data science concentrations or tracks in epidemiology and biostatistics programs represent an appropriate paradigm for training the future generation of public health researchers, given access to the expertise in these areas in a School or College of Public Health. Additionally, training in the art and science of causal inference, study design, statistical methods, and measurement should be brought into pure data science programs, to guard against resources being invested in spurious or confounded associations. These methods need not always be taught in the context of public health, but may come from fields such as economics, psychology, and others. Economics, for example, has also developed rigorous causal inference techniques that, while often similar to those employed in public health research, represent a convergent evolution of methods (Angrist & Pischke, 2008, Chapter 5.2). The more we cross-train in other disciplines, the more we appropriately blur the distinction in the fields (Figure 1). The relative weightings of the constructs in Figure 1 can be aligned with a researcher's—or student's—interests and skills, and while an equal balance is likely unachievable, and perhaps even unnecessary, within an individual the overlap or lack thereof can emphasize a researcher's specific skillset and draw attention to the collaborators needed for a given project. The selection of potential collaborators and mentors can play to the researcher's respective strengths: epidemiologists are well versed in study design and causal inference, biostatisticians have an arsenal of analytic methods and the theoretical knowledge to develop new methods as needed, and data scientists understand data provenance and visualization. Further, while the availability of online computational resources is nearly endless, the appropriate use of these tools demands content-area knowledge that can only come from experience, training, and collaboration.

Lastly, we recommend that regardless of the discipline or field of study, epidemiologists, biostatisticians, and data scientists embrace transparent and open science (Hamra et al., 2019). There has been a recent push in public health and medicine toward releasing both data and code, as the description of methods alone may be insufficient for reproducibility, and we call upon our data science colleagues to do the same (Goldstein, 2018; Goldstein et al., 2019). Even though the programming languages of choice may differ, having data and code publicly available may help guard against erroneous findings and promote insight into complex methodologies as scientists adapt each others' code (Piwowar et al., 2007; Stodden et al., 2013). An example of this comes from the COVID-19 pandemic of 2019–2020, where a flurry of mathematical models were used to inform difficult policy decisions and the analytic codes to many of these models were released in the public domain (Wynants et al., 2020). This is a positive first step, but publishing a model is not an end in and of itself, especially if it has not been peer reviewed. Rather, this should be the beginning of a dialogue between the

modelers, epidemiologists, and public health policymakers to ensure that assumptions entered into the model are valid and policy recommendations are in line with other considerations.

In summary, we are excited about the evolution of these fields, as we seek to answer more difficult public health questions with increasingly more complicated data sources. They are converging at an opportune time around the use of electronic health data. Studies of health phenomena are complex endeavors requiring large teams with expertise along all the steps of the research continuum. Having collaborators with complementary skills as part of the research team can provide insight and direction in the ongoing quest for better ways to prevent and treat disease, and we can only accomplish this through synergies in training epidemiologists, biostatisticians, and data scientists.

Disclosure Statement

Research reported in this publication was supported by the National Institute Of Allergy And Infectious Diseases of the National Institutes of Health under Award Number K01AI143356 (to NDG). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

- Angrist, J., & Pischke, J. S. (2008). *Mostly harmless econometrics*. Princeton University Press.
- Blakely, T., Lynch, J., Simons, K., Bentley, R., & Rose, S. (2019). Reflection on modern methods: When worlds collide-prediction, machine learning and causal inference. *International Journal of Epidemiology*, 49(6), 2058–2064. <https://doi.org/10.1093/ije/dyz132>
- Borgman, C. L. (2019). The lives and after lives of data. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.9a36bdb6>
- Celentano, D. D., & Szklo, M. (2018). *Gordis epidemiology*. Elsevier.
- Cleveland, W. S. (2014). Data science: An action plan for expanding the technical areas of the field of statistics. *Statistical Analysis and Data Mining*, 7(6), 414–417. <https://doi.org/10.1111/j.1751-5823.2001.tb00477.x>
- Conway, D. (2013). *The data science Venn diagram*. <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>
- Díaz, I. (2020). Machine learning in the estimation of causal effects: Targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*, 21(2), 353–358.

<https://doi.org/10.1093/biostatistics/kxz042>

Dixon, B. E., Kharrazi, H., & Lehmann, H. P. (2015). Public health and epidemiology informatics: Recent research and trends in the United States. *Yearbook of Medical Informatics*, 24(1), 199–206.

<https://doi.org/10.15265/IY-2015-012>

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.8cd550d1>

Galea, S. (2013). An argument for a consequentialist epidemiology. *American Journal of Epidemiology*, 178(8), 1185–1191. <https://doi.org/10.1093/aje/kwt172>

Garber, A. M. (2019). Data science: What the educated citizen needs to know. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.88ba42cb>

Goldstein, N. D. (2018). Toward open source epidemiology. *Epidemiology*, 29(2), 161–164. <https://doi.org/10.1097/EDE.0000000000000782>

Goldstein, N. D., Hamra, G. B., & Harper, S. (2019). Are descriptions of methods alone sufficient for study reproducibility? *Epidemiology*, 31(2), 184–188. <https://doi.org/10.1097/EDE.0000000000001149>

Goldstein, N. D., Ingraham, B. C., Eppes, S. C., Drees, M., & Paul, D. A. (2017). Assessing occupancy and its relation to healthcare-associated infections. *Infection Control & Hospital Epidemiology*, 38(1), 112–114. <https://doi.org/10.1017/ice.2016.239>

Goldstein, N. D., & Sarwate, A. D. (2016). Privacy, security and the epidemiologist in the era of electronic health record research. *Online Journal of Public Health Informatics*, 8(3), Article e207. <https://doi.org/10.5210/ojphi.v8i3.7251>

Greenland, S. (ed.). (1987). *Evolution of epidemiologic ideas: Annotated readings on concepts and methods*. Epidemiology Resources Inc.

Hamra, G. B., Goldstein, N. D., & Harper, S. (2019). Resource sharing to improve research quality. *Journal of the American Heart Association*, 8(15), Article e012292. <https://doi.org/10.1161/JAHA.119.012292>

Henry, J., Pylypchuk, Y., Searcy, T., & Patel, V. (2016). Adoption of electronic health record systems among U.S. non-federal acute care hospitals: 2008–2015. In *ONC Data Brief* 35.

Hersh, W. (2015). What is the difference (if any) between informatics and data science? *Informatics Professor*. <https://informaticsprofessor.blogspot.com/2015/07/what-is-difference-if-any-between.html>

- Information Management and Security Staff. (2003). *DOJ systems development lifecycle guidance*. <https://www.justice.gov/archive/jmd/irm/lifecycle/ch1.htm>
- Kaiser Health News. (2019). *A reality check on artificial intelligence: Are health care claims overblown?* <https://khn.org/news/a-reality-check-on-artificial-intelligence-are-health-care-claims-overblown/>
- Lash, T. L., Fox, M. P., & Fink, A. K. (2009). *Applying quantitative bias analysis to epidemiologic data*. Springer-Verlag. <https://doi.org/10.1007/978-0-387-87959-8>
- Lin, M., Lucas, H. C., & Shmueli, G. (2013). Research commentary—too big to fail: Large samples and the p-value problem. *Information Systems Research*, 24(4), 906–917. <https://doi.org/10.1287/isre.2013.0480>
- Lin, S. H., & Ikram, M. A. (2019). On the relationship of machine learning with causal inference. *European Journal of Epidemiology*, 35(2), 183–185. <https://doi.org/10.1007/s10654-019-00564-9>
- Madigan, D., Ryan, P. B., Schuemie, M., Stang, P. E., Overhage, J. M., Hartzema, A. G., Suchard, M. A., DuMouchel, W., & Berlin, J. A. (2013). Evaluating the impact of database heterogeneity on observational study results. *American Journal of Epidemiology*, 178(4), 645–651. <https://doi.org/10.1093/aje/kwt010>
- Meng, X.-L. (2019). Data science: An artificial ecosystem. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.ba20f892>
- Morabia, A. (Ed.). (2004). *A history of epidemiologic methods and concepts*. Birkhauser-Verlag.
- Naimi, A. I., & Balzer, L. B. (2018). Stacked generalization: An introduction to super learning. *European Journal of Epidemiology*, 33(5), 459–464. <https://doi.org/10.1007/s10654-018-0390-z>
- Navathe, A. S., Zhong, F., Lei, V. J., Chang, F. Y., Sordo, M., Topaz, M., Navathe, S. B., Rocha, R. A., & Zhou, L. (2018). Hospital readmission and social risk factors identified from physician notes. *Health Services Research*, 53(2), 1110–1136. <https://doi.org/10.1111/1475-6773.12670>
- Piwovar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS One*, 2(3), Article e308. <https://doi.org/10.1371/journal.pone.0000308>
- Rosen, G. (1993). *A history of public health*. Johns Hopkins University Press.
- Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology*. Lippincott Williams & Wilkins.
- Rothwell, P. M. (2005). External validity of randomised controlled trials: "To whom do the results of this trial apply?" *Lancet*, 365(9453), 82–93. [https://doi.org/10.1016/S0140-6736\(04\)17670-8](https://doi.org/10.1016/S0140-6736(04)17670-8)
- Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. Henry Holt and Company.

Stodden, V., Guo, P., & Ma, Z. (2013). Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PLoS One*, 8(6), Article e67111. [https://doi.org/10.1016/S0140-6736\(04\)17670-8](https://doi.org/10.1016/S0140-6736(04)17670-8)

Susser, M., & Stein, Z. (2009). *Eras in epidemiology: The evolution of ideas*. Oxford University Press.

Szklo, M., & Nieto, J. (2018). *Epidemiology: Beyond the basics*. Jones & Bartlett Learning.

U.S. News and World Report. (2019). *Best public health schools 2019*. <https://www.usnews.com/best-graduate-schools/top-health-schools/public-health-rankings>

Wynants, L., Van Calster, B., Bonten, M. M. J., Collins, G. S., Debray, T. P. A., De Vos, M., Haller, M. C., Heinze, G., Moons, K. G. M., Riley, R. D., Schuit, E., Smits, L., Snell, K. I. E., Steyerberg, E. W., Wallisch, C., & van Smeden, M. (2020). Systematic review and critical appraisal of prediction models for diagnosis and prognosis of COVID-19 infection. *medRxiv*. <https://doi.org/10.1101/2020.03.24.20041020>

Supplementary File

[Supplement 1 - Degree program review.xlsx](#)

37 KB

©2020 Neal D. Goldstein, Michael T. LeVasseur, and Leslie A. McClure. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the article.

Preview image is “The Data Science Venn Diagram” by Drew Conway, available under [Creative Commons Attribution-NonCommercial \(CC BY-NC 3.0\) license](#) at <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>.