Received: 12 March 2023 | Revised: 3 May 2023 | Accepted: 4 May 2023

DOI: 10.1111/cts.13545

TUTORIAL

From clinical data management to clinical data science: Time for a new educational model

Richard F. Ittenbach

Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA

Correspondence
Richard F. Ittenbach, Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital, University of Cincinnati College of Medicine, Cincinnati, OH 45229-3039, USA.
Email: richard.ittenbach@cchmc.org

Abstract

The purpose of this article is to propose and provide a blueprint for a graduate-level curriculum in clinical data science, devoted to the measurement, acquisition, care, treatment, and inferencing of clinical research data. The curriculum presented here contains a series of five required core courses, five required research courses, and a list of potential electives. The coursework draws from but does not duplicate content from the foundational areas of biostatistics, clinical medicine, biomedical informatics, and regulatory affairs, and may be reproduced by any institution interested in and capable of offering such a program. This new curriculum in "clinical" data science will prepare students for work in academic, industry, and government research settings as well as offer a unifying knowledge base for the profession.

INTRODUCTION

Clinical data science, like the broader discipline of all data science, integrates tenets of statistics and informatics, but with a specific focus on clinical research. Among the many defining moments in the quest to develop new drugs, devices, and biologics to improve human health, two stand above the others: passage of the Food, Drug, and Cosmetic Act of 1938 and the 1962 Kefauver–Harris Amendment which strengthened the Food and Drug Administration's oversight and enforcement role, requiring manufacturers to provide safety, effectiveness, and reliability data with all new applications.[1–3] The result of this legislation was an increased need for strengthened data management expertise to monitor and manage the flow of data through the research and development pipeline. Hence, the role of clinical data manager and its most recent analog, clinical data scientist, was born.

After 50 years of invaluable service to industry, clinical data managers' roles and responsibilities have expanded to other sectors, including academic, professional, scientific, and financial institutions.[4] Yet, despite the importance of the role to drug and device development over the past half-century, and its contributions to public health more generally, the field has failed to advance beyond that of a technical specialty embedded within a much larger clinical research enterprise. Advances to the field thus far have all come from outside the discipline, namely bioinformatics and biostatistics, the intellectual homes of the newly emerging field. As those fields have matured and added methods, techniques, software, literature, and new professionals to the workforce in a systematic way, clinical data management has not. Without the benefit of formal, degree-granting programs, an established literature base (textbooks, journals, monographs), professionals in upper management positions, and policies and practices that positively influence the other sciences, the field has remained in an expanding but scientifically dormant, service-oriented role. Establishing a foothold in other sectors, most especially the academic and medical communities, will be an important evolutionary step for the field and the professionals who support it. While there

is widespread agreement that the amount, complexity, speed, and sophistication of research data have increased exponentially over the years, the workforce's knowledge base has not evolved like other scientific specialties with which clinical data managers work. Simply stated, the job has changed tremendously but the educational preparation of those who do it has not.

In many ways, the preparation of today's clinical data managers/clinical data scientists looks remarkably similar to that of decades ago. The education and training programs that do exist tend to fall into a discrete number of categories such as rebranded statistics or computer science programs; generic clinical research programs; online learning platforms such as Coursera, edX, and Udemy; short courses offered through professional associations such as the Drug Information Association and Society for Clinical Data Management, two of the strongest advocates for training within the profession; and employer-based training programs. Of the programs mentioned here, the most rigorous ones are the rebranded statistics/computer science programs, yet they generally fail to prioritize the clinical nature of the data. The most typical cases are staff-lead employer-based training programs delivering brief instructional sessions supported with on-the-job mentorship experiences subordinate to the needs of a specific corporation or institution. Missing from this model is a coherent, formalized, and rigorously derived body of knowledge, based on theory and practice, which can serve as the foundation for the knowledge base of this new and emerging field—clinical data science.

Clinical data science derives from the term data science, a generally amorphous term that means different things to different people. Among the consistent themes, though, is the practice of obtaining meaning from data—most often very large amounts of data—using statistical and informatics-based methods. Clinical data science extends this logic to include a focus on clinical, medical, or biomedical data collected in the interest of advancing healthcare. Yet, because clinical medicine and biomedical research are highly regulated areas heavily influenced by technical and regulatory standards at multiple levels, any attempt to understand and advance the discipline must incorporate regulatory knowledge. Broadly interpreted, clinical data science in the 21st century is not just about a specific science or the discoveries within it, but about bringing together the vastly different worlds of biostatistics, biomedical informatics, clinical medicine (and its clinical operations), and regulatory affairs into a cohesive, scientific whole.

In his landmark paper on clinical and translational science, Zerhouni[5] indicated that scientific progress is often made at the "interface of pre-existing disciplines" and that changes were needed to reduce cultural and administrative barriers to accelerate the advancement of science (p. 1662). Moreover, the knowledge and training for today's translational scientists could no longer be achieved "on the job" as in the past, and what was needed were scientists with a wider range of skills that could reach more broadly and deeply into the behavioral and biomedical problems of the present day. This effort, though, would demand more resources and more methods than before. In essence, the call to action challenged the biomedical research community to create new sciences out of old ones with more rigorous and more interdisciplinary training than ever before. Clinical and translational science is, at base, an integrative science bringing together "knowledge to improve the understanding, efficiency, and effectiveness" of all clinical research.[6] Clinical data science fully embodies this spirit but thus far has lacked the rigorous, interdisciplinary education and training approach needed to realize its potential.

Given the complexity of clinical research today, the expansiveness of the data, clinical data scientists' role in the investigative process, and the dearth of formal education and training programs worldwide, the purpose of this article is to propose and provide a blueprint for a graduate-level curriculum in clinical data science, devoted to the measurement, acquisition, care, treatment, and inferencing of clinical research data. This new curriculum in "clinical" data science will prepare students for work in academic, industry, and government research settings as well as offer a unifying knowledge base for the profession. This curriculum may be reproduced by any institution interested in and capable of offering such a program.

METHODS

Population profile

According to the U.S. Bureau of Labor Statistics,[4] there were approximately 113,300 clinical data managers and clinical data scientists working in professional, scientific, and technical sectors in 2021. Because clinical data managers and clinical data scientists perform highly similar roles across a wide range of settings and given that both occupations carry a single federal job code (U.S. Department of Labor, 15-2051.02), both professions will be referred to using the single unifying term, clinical data scientist, for the remainder of this article.

With an expected annual growth rate of 12%, the profession's future is very bright. Clinical data science is a well-paid profession with an average annual salary of $100,910. Time to maturity in the profession is a mere 2–4 years, with 85% of the workforce having only a bachelor's degree (with 5% holding a 2-year degree). There are no formal requirements to enter the profession, which means staff

come from a wide range of disciplines and backgrounds. With little or no availability of relevant, formal, degree-granting programs dedicated to the scientific treatment of clinical research data, nationally or internationally, education and training for this group of professionals is largely nonexistent or at the discretion of an employer.

The intended audience for this curriculum is broad and diverse. Because of virtual technologies, students may participate from anywhere in the world. For example, as of March 9, 2023, ClinicalTrials.gov reported 444,857 studies worldwide with 31% recruiting in the U.S. only, 53% in non-U.S. countries only, and 5% in both the U.S. and non-U.S. countries (11% locations not provided).[7] Hence, research staff are needed worldwide to care for the data. However, a recent issue of the *Journal of the Society for Clinical Data Management* reported a desperate need for graduate education and training worldwide in academic, government, industry, and non-profit settings.[8–12] As most employers in the field today know, students with average to strong performance in mathematics through advanced algebra, fundamentals of biology/chemistry, and oral and written composition can be successful in the field. The proposed program is a graduate program in a biomedical science, so the ability to think logically, reason systematically, and communicate well with others will be essential for success in the program. While several of the courses are cross-listed at the advanced undergraduate level, a master's degree program was selected as most appropriate, rather than a bachelor's or associate's degree program, because of the cognitive demands of the job, the number of scientists and technical professionals one would work with on a daily basis, and the pragmatics of implementing a self-contained master's degree program. The current salary structure in the field is a much harder sell for many organizations recruiting bachelor's level staff than for master's staff.

Principal outcome: Curriculum

For purposes of this article, and as a point of reference for readers, a curriculum is defined as a systematically arranged sequence of courses housed within a formal, graduate, degree-granting program. The 'program' then is the administrative housing for the curriculum just as it would be for students wishing to major in accounting, aerospace engineering, or agriculture. Most importantly, while degree programs with the same name may vary from school to school and year to year based on the evolution of knowledge at any given time, the purpose of this curriculum in clinical data science is to advance students' knowledge and skills in the measurement, acquisition, care, treatment, and inferencing of clinical research data. It is the systematically arranged part of the program that gives the curriculum its synergism and allows its return on investment to be much more than the sum of its parts, something ad hoc classes are not likely to achieve.

Procedures

A fundamental requirement of any professional training program is the presence of a theoretical foundation on which to structure the courses. Just as theory helps scientists explain observable phenomena well enough to be acted upon, a theoretical framework helps with the transfer of knowledge from instructor to student as well as within the field itself. It helps instructors organize the material in a logical, coherent, relational manner, and it provides the students with a structure with which to organize the new information. In the case of clinical data science, with such a short occupational history as well as a scientifically complex knowledge base yet little to no foothold in the literature, framing both the knowledge base as well as the coursework is not only prudent and practical but extremely necessary.

Embedded within any graduate program, especially clinical data science, should be the concurrent advancement of both skills and knowledge. Enhanced skills are needed because the profession evolved out of a functionalistic model, one in which a profession's occupational worth was determined by a company's ability to deliver drugs and devices to market more efficiently. Advanced knowledge is needed because the role is now central to a much more sophisticated and technically complex process, and practitioners are expected to interact with professionals from other more technically advanced scientific disciplines. Professionals and thought leaders in the collaborating disciplines now need the scientific guidance of their clinical data science colleagues on multiple fronts. Sadly, the current educational preparation of today's clinical data science professionals not only lags far behind their research colleagues, but now jeopardizes the very process it was intended to help. This new program must offer courses that are as rigorous as they are impactful to all involved. The sequence of steps used in the creation of the proposed curriculum includes development of the following:

1. Theoretical framework of the profession's knowledge base and the disciplines from which it derives, namely biostatistics, biomedical informatics, clinical medicine (and its associated operations), and regulatory affairs.
2. Ordered sequence of core courses characterizing the new profession's knowledge base.
3. Research courses that not only support the new knowledge base, but represent the scientific foundation on which the new discipline rests.

4. Listing of elective courses from which students may pursue specific areas of professional interest.
5. Listing of pervasive skills to be modeled and emphasized alongside the academic content.

Behind each course in the curriculum is a fully developed syllabus including proposed course title, description, textbook, objectives, weekly schedule of material to be covered, support systems for students who may need additional help, outside readings, and evaluation plan. Because this is a graduate curriculum intended to familiarize students with not only the content but process of conducting research, outside readings from the scientific literature are required to supplement the course text. As Tyler[13] has noted, the evaluation component of a course is as crucial as the content itself. Evaluation provides instructors with important information on each student's progress in a given course as well as the program to enable modifications when needed. For this reason, a combination of evaluation strategies will be used across the curriculum, from tests and quizzes to oral/written reports, to instructor observations and portfolios.

Quality assurance

To be sure that the material outlined in the curriculum contains the knowledge and skills needed by professionals in the field, and that the material would pass muster with established professionals at local as well as national and international levels, a two-phase quality assurance plan was used. First, material in the curriculum was mapped to core competencies identified by three influential organizations responsible for professional growth and development in the field: the American Medical Informatics Association,[14] Joint Task Force for Clinical Trial Competency,[15,16] and Society for Clinical Data Management.[17] Second, as courses were developed, the syllabi were shared with two or more local professionals experienced in the area, then two or more regional/national experts in the identified area and, finally, a team of eight international experts from academe, industry, and government, who served as advisors on the development of this curriculum. Material was then modified based on the advisory board's feedback.

CLINICAL DATA SCIENCE CURRICULUM

The creation and transfer of knowledge has the potential to change many things: behaviors, skills, perspectives, and even lives. And so it is with a degree in clinical data science. To deliver on the promise of creating a systematically arranged course of study for students as well as the profession, the following curriculum is divided into five principal components: theoretical framework, required core courses, required supporting research courses, electives, and pervasive skills and evaluation. All are detailed below.

Theoretical framework

Before any courses can be developed, it is important to know what knowledge needs to be conveyed and, as a result, what courses need to be taught, in what order, and in what proportions. For this, a theoretical framework is needed just as a blueprint is needed prior to building a house intended to protect and shelter people. The clinical data scientist curriculum proposed here will be derived from, but not duplicate the knowledge bases of, its four supporting disciplines in equal proportions: biostatistics, biomedical informatics, clinical operations, and regulatory affairs. It is acknowledged that other factors often shape the interpretation and implementation of a discipline's theoretical foundation such as the training requirements of an institution, the expectations of a given professional community, and/or the talent available to teach courses. In the end, however, it all begins with an organizing framework (see Figure 1). Because of the prototypical nature of the program, course numbers and titles are designed to be reflective of the scientific content and sequencing of material and can easily be modified to meet the needs of specific institutions.

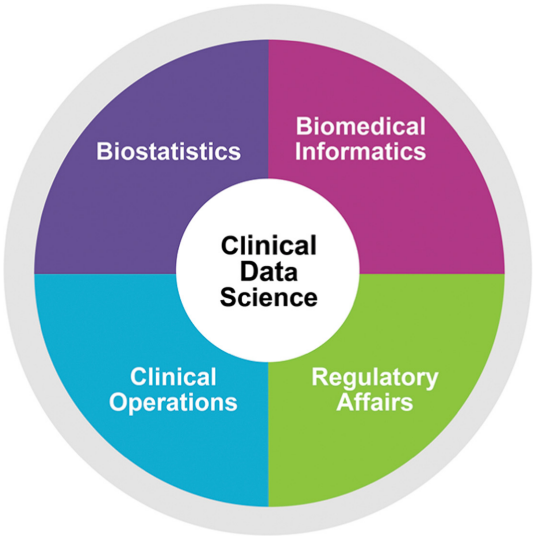

FIGURE 1 Theoretical framework for a clinical data science curriculum.

Core courses

Established academic disciplines generally have a defined knowledge base that not only establishes its content for proficiency, but also defines its boundaries in the broader landscape of education and practice. For a degree in clinical data science, the core sequence proposed here consists of a series of five required, 3-h graduate-level courses. The first course is a foundations class that introduces students to the field of clinical data science, including the history and rationale for the new disciplines, as well as important practices, procedures, supporting areas of science, and standards for good practice (e.g., Good Clinical Practice, Good Clinical Data Management Practices, International Council on Harmonisation), and expectations in performance (CDS 6010, Science of Clinical Data Management [SCDM] I). In this prototype, it will be cross-listed at both the graduate and undergraduate levels and have the capacity to serve as a foundations course for other programs in the institution wishing to have a clinical data science option for their students.

The second course is a deep dive into the content, roles, and responsibilities of the profession beginning with a discussion of data elements, structures, dictionaries, tools, and platforms, moving from study start-up through operations to study close-out, with an introduction to data ownership, stewardship, and security (CDS 6020, SCDM II). This course may also be cross-listed at the advanced undergraduate level. The third course in the sequence would weave together the skills and knowledge from the first two courses by requiring the students to create example documents that they will see and use in their professional practice, including a preliminary research protocol, its corresponding data management plan, electronic case report forms (eCRFs), and a database that conforms to the aforementioned (CDS 6030, SCDM III). The fourth course would be a required, full-time internship at a leading pharmaceutical, clinical research, or academic research organization under the direction of an experienced clinical data manager/scientist (CDS 6040, SCDM IV). In the case of small or developing programs without direct access to major research organizations, or for atypical or hard-to-place students that are surely to enter the program, having back-up field-based placement options at the university or local medical centers offers crucial placement options when needed. The four research courses will be complemented by a required 3-h course in Project Management (CDS 6050) to acquaint the students with the importance of managing people, projects, timelines, and budgets. See Figure 2 for a listing of example courses in this curriculum.

Research courses

The supporting research courses are crucial in any curriculum because they serve to support and strengthen the knowledge gained in the core courses. They take on an additionally important role in this degree program because of their historical significance in giving rise to the profession.[11,12] The research courses here consist of five, 3-h required classes, the first four of which map directly onto one of the four supporting areas of the core knowledge base: Introduction to Biostatistics (BSTAT 7000), Introduction to Biomedical Informatics (BMIN 7050), Introduction to Clinical Operations (CDS 7020), and Regulatory Science I (CDS 7010). While introductory in nature, students who have taken them at any school will quickly attest to the sophistication and demand of the material. In this case, it is presumed that the courses will be taught in their home department where possible and appropriate. For those universities without such programs, syllabi can be provided. Offering these four courses, complemented by a 3-h Clinical Research Ethics course (CDS 7030), is essential given that today's professionals must often confront and respond to unanticipated moral and ethical dilemmas that arise from their day-to-day work.[18]

Consistent with the quality assurance protocol noted above, all core and research courses were mapped onto the competencies and foundational domains for the three most prominent professional associations influencing the field today: the American Medical Informatics Association,[14] Joint Task Force for Clinical Trial Competencies,[15,16] and Society for Clinical Data Management.[17] Astute readers will quickly recognize important practice-related differences within the foundational areas that must be acknowledged. In the case of biomedical informatics, some students will undoubtedly be more interested in clinical (care delivery) as opposed to research, or even molecular as opposed to nonmolecular. While the emphasis of this curriculum is actually on clinical "research" informatics, courses should offer students enough of a working knowledge across the subspecialty areas to allow them to subspecialize where appropriate. The same would be true for students wishing to gravitate toward statistics, either biomedical, social, or otherwise. And, importantly to the field itself, within the area of clinical operations, some will prefer to work within medicine, others nursing, others the allied health sciences, and still others laboratory or imaging medicine. Again, this curriculum was developed with the general case in mind, and allows institutions the freedom to adjust their coursework accordingly. See Figure S1 for a subset of the cross-mapping of courses to professional competencies.

RESEARCH AREAS

Biomedical Informatics

Biostatistics

Clinical Operations

Regulatory Affairs

Master of Science Curriculum in Clinical Data Science

COURSE	COURSE TITLE	HOURS
Core Courses (15 Hours required)		
CDS 6010	Science of Clinical Data Management I: Overview *	3 Credits
CDS 6020	Science of Clinical Data Management II: Roles/Responsibilities *	3 Credits
CDS 6030	Science of Clinical Data Management III: Programming	3 Credits
CDS 6040	Science of Clinical Data Management IV: Internship	3 Credits
CDS 6050	Project Management in Clinical Research *	3 Credits
Research Courses (15 Hours required)		
BMIN 7050	Introduction to Biomedical Informatics	3 Credits
BSTAT 7000	Introduction to Biostatistics	3 Credits
CDS 7010	Regulatory Science I *	3 Credits
CDS 7020	Introduction to Clinical Operations	3 Credits
CDS 7030	Clinical Research Ethics	3 Credits
Electives (6 Hours Required - select two)		
BMIN 7070	Data Science for Biomedical Research	3 Credits
BMIN 7080	Database Management Systems	3 Credits
BMIN 7090	Economics and Cost Analysis *	3 Credits
BSTAT 7060	Principles of Clinical Trials	3 Credits
BSTAT 7080	Introduction to Epidemology	3 Credits
BSTAT 7090	Regression Analysis	3 Credits
BSTAT 8080	Successful Scientific Writing	3 Credits
CDS 7012	Innovation and Regulatory Science II	3 Credits
CDS 7022	Seminar in Clinical Operations	3 Credits
PHAR 8010	Global Regulatory Science	3 Credits
PHAR 8040	Drug and Medical Device Development	3 Credits
CDS 8000	Thesis in Clinical Data Science	3 Credits

Note. * denotes a class that is cross-listed at the advanced undergraduate level. Thesis option may be elected after the 6-hour requirement is satisfied.

FIGURE 2 Proposed clinical data science curriculum. *Denotes a class that is cross-listed at the advanced undergraduate level. Thesis option may be elected after the 6-h requirement is satisfied.

Elective courses

Two 3-h elective courses are required for completion of the degree. Similar to the research courses, all electives map to one of the four foundational research areas of biostatistics, biomedical informatics, clinical operations, and regulatory affairs. The two elective courses may be taken within a single area or from two different areas. For example, within the biostatistics area, students must first take the required Introduction to Biostatistics course (BSTAT 7000) and then, if the student so wishes, may elect to take one or two of the more advanced biostatistics courses, such as Principles of Clinical Trials (BSTAT 7060), Introduction to Epidemiology (BSTAT 7080), or Regression Analysis (BSTAT 7090). Similarly, for students inclined toward informatics (i.e., bioinformatics, medical informatics, or biomedical informatics), advanced options are also available here: Data Science for Biomedical Research (BMIN 7070), Database Management Systems (BMIN 7080), or Economics and Cost Analysis (BMIN 7090). For those entering the program with an interest in regulatory science, advanced options include Innovation and Regulatory Science II (CDS 7012), Global Regulatory Science (PHAR 8010), or Drug and Medical Device Development (PHAR 8040). Students wanting more coursework in clinical operations may choose from a series of Seminars in Clinical Operations containing advanced content in such areas as clinical medicine, nursing, the allied health sciences, or laboratory medicine, as possible examples (CDS 7022) (see Figure 2).

Pervasive (soft) skills and evaluation strategies

Employers can quickly identify the knowledge and job-related skills needed to be successful in a specific job.

To that extent, one need only ask a job candidate what courses were taken in school or during previous employment, examine a transcript, review a syllabus, or have the applicant submit a work example. Extremely important but far less prioritized are the pervasive skills (also known as soft- or people-skills) that transcend the formal curriculum and connect the technical content with day-to-day practice. Pervasive skills include critical thinking, communication, leadership, and social skills—those skills that allow one to work well with others even on a hard day. In a world where timelines and competing obligations encourage shortcuts, it is the pervasive skills that can make the difference between a smoothly running, professional, and collegial environment, and one that is not. These types of skills should not only be taught but intentionally addressed, modeled, and discussed as the content allows during the course of the semester.

In many college courses, midterms and final examinations are the default means of evaluation. For other courses it may be a series of weekly quizzes or a final paper and presentation. In addition to the very traditional quizzes, tests, and final examinations, this evaluation plan also includes instructor evaluations of students' interactions with others as professionals, particularly in the project management and internship classes. Finally, the curriculum evaluation plan also includes student portfolios that can house the students' relevant work examples prior to graduation. The portfolio can include everything from tests to reports to presentations to program code and instructor feedback. In this curriculum, using a combination of evaluation strategies will be emphasized to include tests, quizzes, and opportunities for oral and written reports representative of the work done by clinical data scientists. Not only will some excel in one form of learning over another, but each means of evaluation offers unique information to instructors and program administrators on which to base a student's progress in a class as well as the program. See Figure 3 for a listing of recommended pervasive skills and evaluation strategies mapped to the core and research courses.

DISCUSSION

With the size of the digital universe reportedly doubling every 2 years[19] and the complexity of the clinical data universe expanding in a comparable fashion, training today's data scientists for the world of tomorrow remains a daunting task. From paper case report forms of 20 years ago, to electronic data capture systems of 10 years ago, to the wireless, continuous data streams of intensive care unit patients yielding as many as 2000 data points per second today, the world of clinical data science has changed remarkably over the years.[20] Sadly, the educational training of clinical data scientists has not.

Drastically changing landscape

If it were simply about the amount and speed of the data, the solutions would be more easily imagined. However, the entire research landscape in which today's clinical and biomedical research is conducted has changed. According to the global clinical research and healthcare analytics provider IQVIA and their *Global Trends in R&D*, clinical trial starts increased by 39% in the past 10 years, with phase I, II, and III studies increasing by 14% alone in 2021. And, while Food and Drug Administration (FDA), National Institutes of Health (NIH), and Patient Centered Outcomes Research Institutes (PCORI) budgets have all continued to rise over the years, spending tells only part of the story. One may consider PCORI's change in research commitments between 2020 and 2022 as a case in point. PCORI's research commitments increased from US$187M (approved) in 2020 to US$390M (approved) in 2021 to US$500M (planned) in 2022.[21,22] As such, today's clinical data scientists are now finding themselves managing larger and more complex portfolios, with projects that are correspondingly larger and more analytically complex. These studies extend well beyond the traditional and compartmentalized phase I, II, and III studies of the

		Pervasive Skills				Evaluation Methods			
Core Courses		Critical Thinking	Communication	Leadership	Social Skills	Tests	Report Writing	Observation	Portfolio
CDS 6010	Science of CDM I: Overview	●				●	●		
CDS 6020	Science of CDM II: Roles/Responsibilities	●	●	●	●	●	●	●	●
CDS 6030	Science of CDM III: Programming	●	●	●		●	●		
CDS 6040	Science of CDM IV: Internship	●	●	●	●		●	●	●
CDS 6050	Project Management in Clinical Research	●	●	●	●	●		●	●
Research Courses									
BMIN 7050	Introduction to Biomedical Informatics	●	●			●			
BSTAT 7000	Introduction to Biostatistics	●	●			●			
CDS 7010	Regulatory Science I	●	●	●	●	●	●	●	●
CDS 7020	Introduction to Clinical Operations	●				●		●	
CDS 7030	Clinical Research Ethics	●	●	●	●		●	●	

FIGURE 3 Pervasive (soft) skills and evaluation methods by course for core and research courses.

past, to now include everything from basic science to action research to program projects with multiple protocols embedded within them. Adaptive clinical trials are displacing randomized controlled trials while umbrella trials now simultaneously contain phase I, II, and II studies all within a single protocol. With little to no training in study design, today's clinical data scientists have no real way of interpreting how the data, the designs, or even the studies relate to one another.

When it comes to interacting with the other, related disciplines, the rules of the game have changed as well. For example, scaling the data from a classical two-arm clinical trial is vastly different from that of a larger basket or umbrella trial, thus requiring a deeper ability to communicate and interact with biostatisticians when it comes to designing, capturing, pulling, authenticating, and validating the data. And getting the standards appropriately aligned between the study documentation and data operations (e.g., Study Protocol, eCase Report Forms, Statistical Analysis Plan, Data Management Plan, Data related Standard Operating Procedures), whether one is adhering to CDISC (Clinical Data Interchange Standards Consortium), CDASH (Clinical Data Acquisition Standards Harmonization), or FHIR (Fast Healthcare Interoperability Resources) standards, requires a much deeper ability to understand and communicate with both the informatics team and the regulatory team to assure the team is in compliance with the highest ethical and scientific standards of good clinical practice particularly when submitting applications to the European Medicines Agency or the FDA. Similar examples can be found with respect to the clinical components of the role whether in medicine, nursing, or the allied health sciences.

While the technical components of the role are well documented, the professional expectations are much more subtle, placing the new professionals under more pressure and more scientific scrutiny than ever before. With bigger portfolios come larger study teams with which to interact and manage, but there is also increased pressure from sponsors and advocacy groups to certify the data and produce results before they are analytically ready.[23] For clinical data scientists, and especially for those with more advanced responsibilities, there is an increased need to monitor the literature, publish results, and contribute to other submissions—the list can be endless. It is a vastly different world from that of only a few decades ago, with clinical data science colleagues being forced into roles for which they have not been formally trained or are simply not ready. The logical consequence for many will be failure, burnout, or endless job-hopping to find the right fit in an ill-fitting occupation. When those most entrusted to monitor and care for the data quality cannot keep up or lack the scientific knowledge of their colleagues on study teams, the entire study become vulnerable to mistakes, omissions, and misinterpretations of great clinical and scientific significance.

As the research landscape increases in complexity, speed, and multidimensionality, the scientists and institutional support systems must strive to keep up.[24] Clinical data science is no exception—except that clinical data science is operating without a consistent and reproducible knowledge base. Under the current educational model, clinical data scientists are underprepared for today's research and will soon be vastly unprepared for the science of tomorrow. Their training simply does not compare to that of the biostatisticians, informaticians, medical/clinical staff, and regulatory professionals with whom they routinely interact. We must prepare today's researchers, and our clinical data scientists more specifically, for the clinical and translational world of tomorrow.[6,25] But that preparation must go beyond knowledge of the data in its current form; it must also include data's role in science, the scientific method, and how the other sciences rely on data to solve the problems of today and hope to solve the problems of tomorrow. Clinical data scientists must be able to see the limitations as well as the possibilities of their data and see them through the lenses of their colleagues, the biostatisticians, informaticians, and clinical scientists. For this reason, clinical data scientists will be reacquainted with the scientific method in lecture 1 of course 1 on day 1 of the program and then build from there.

Strength of a rigorous but flexible curriculum

Well-thought-out curricula can achieve outcomes that poorly planned curricula and ad hoc courses cannot. A key feature of a strong educational curriculum is that it provides enough guidance for administrators to have a blueprint for success but does not overregulate what goes on in the classroom.[26] First, a curriculum is structured to build a student's knowledge about a subject from the ground up, from foundational principles to more complex concepts and ideas. The curriculum and the courses on which it is built generally have well-defined learning objectives and prespecified content that allows the students to meet those objectives, with stipulated priorities for evaluation for students to know when, how, and to what standards they will be held when evaluated. The content of the curriculum should be as complete as possible without being redundant or excessive, when preparing students for the next phase of their education or life.

Given the competitive nature of programs today, and students' access to virtual options worldwide, keeping the program accessible and financially feasible is a must. The

curriculum presented here is a full 36 h; however, administrators may opt for a reduced 30-h curriculum by dropping one of the electives, and opting for three 2-h research courses in research ethics (CDS 7030), clinical operations (CDS 7020), regulatory science I (CDS 7010), where necessary. Yet, whatever form the curriculum takes, the content must be rigorous enough to be recognized as appropriate by established professionals in the field, yet nimble enough to meet the objectives of the learner. Finally, and most importantly, a curriculum must be premised upon three main factors: the needs of society, the attributes of the learner, and a specialized knowledge base worthy of transmission to others.[13] In short, a strong curriculum should meet the needs of key stakeholders simultaneously, all of which make the educational experience possible.

Unifying knowledge base

The clinical data science core courses will draw from but not duplicate information from its four foundational disciplines: biostatistics, biomedical informatics, clinical operations, and regulatory affairs. For example, in their core courses the students will begin applying information on sampling and probability theory from their biostatistics course with respect to sample selection, recruiting, retention, error rates, and auditing of records. From their informatics class they will learn the rationale and principles of user mapping of external sources and the importance of interoperability of resources in healthcare. From the clinical operations course they will be introduced to basic terminology in medicine, nursing, and the applied health sciences needed for eCRFs, database builds, and adverse event reporting. And from the regulatory science course the students will learn about relationships between data, risks, informed consent, and the regulations and ethical guidelines that influence their day-to-day work. The courses in this curriculum should be as integrative as the program itself. The better prepared the workforce is, the better able they will be to understand the causes and consequences of everyone's actions, and the better able they will be to focus on what matters most when it comes to the quality of the data.

Sequencing of courses

Because the courses represent an integration of four very complex disciplines, it is extremely important to sequence them in a way that maximizes learning for the students (see Figure 4). This begins with a concept known as vertical articulation in which courses are intentionally connected through the teaching and reinforcement of specific skills and content at subsequent points in the curriculum.[26] This would be the case in teaching the roles and responsibilities of the practicing clinical data scientist in Science of Clinical Data Management I (CDS 6010) through Science of Clinical Data Management IV (CDS

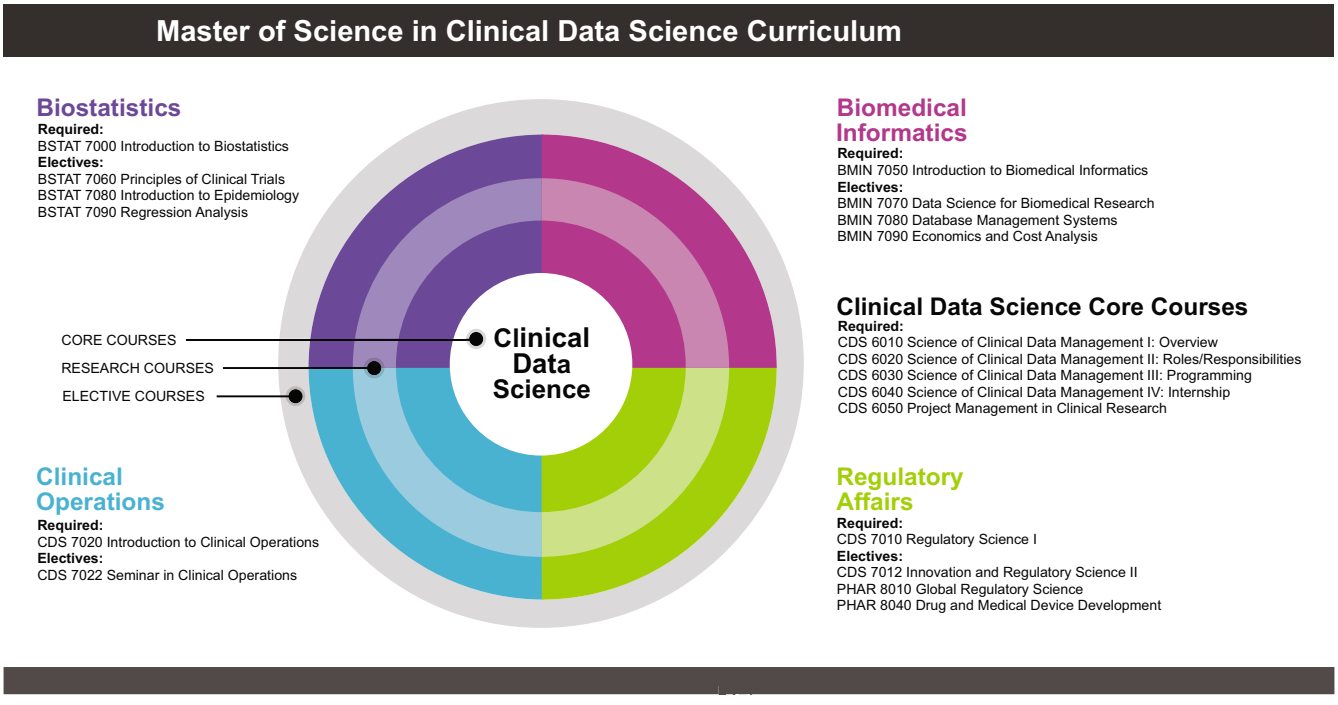

FIGURE 4 Mapping of clinical data science courses onto the theoretical framework.

6040). The fact that the content is taught in contiguous classes at successively deeper levels of sophistication derives from Bruner's concept of a spiral curriculum, a strategy that the American Association for the Advancement of Science has championed for years regarding scientific literacy in K-12 programs nationwide.[27–29] One would also expect to see this continuity between each research course and its subsequent elective, among others. The same interrelatedness may also be applied to courses taken alongside one another, referred to as horizontal articulation, when familiarizing students with database development in CDS 6020, its applications to artificial intelligence and machine learning in BMIN 7050, and its impact on human subjects protections in CDS 7010, content covered in the required research courses.

Multilayered instruction

Planned connections among courses, whether vertical or horizontal, constitute only part of the curriculum's connective tissue. Because students enter a given class with very different backgrounds and knowledge sets, instructors must be cognizant of the need to reach students at different levels using multiple layers of instruction when teaching technically complex material. In their synthesis of 228 meta-analyses of learning in the classroom, Hattie and Donoghue have organized the transmittal of information into three levels: surface learning (emphasis on *skill* development), deep learning (emphasis on formal, *discipline-specific* learning), and transfer learning (emphasis on *complex problem solving*).[30] Consistent with employer-based, continuing education seminars and modules, most education and training models thus far have been focused on routinized skill development more so than deep or transfer learning, the types of learning needed for critical thinking and complex problem solving that rely on relationships among related knowledge bases. By bringing together principles and content from their various classes, the students develop the ability to analyze, synthesize, and apply principles and content learned in the classroom to new and unfamiliar problems encountered in the field.

Having a curriculum on paper as a blueprint for success is only the first step. For the implementation to be successful, the curriculum must be housed within a strong academic unit, with strong and visionary faculty to make the curriculum come alive. In the initial stages, program administrators may wish to start with five to seven full-time faculty members who can teach courses, advise students, and advocate for the program (e.g., budgets, floor space, recruiting, resources, general public relations). Part-time/adjunct faculty members from the community will also be an important asset to the program and serve as valuable resources to the students and faculty alike. Having one faculty member dedicated to internship placements and supervisions and a second faculty member devoted to advocating for international students is a must. But, even at this staffing level, resources may be constrained and will need to grow as student recruitment and retention grows. Having a supportive institutional administrator to advocate within the administrative hierarchy is an absolute necessity. But, in the end, there is no substitute for having faculty who can serve as good teachers, writers, mentors, and role models for the students and developing professionals. For the program to be successful teaching students for academe, industry, and government, having liaisons at each of the aforementioned settings is essential.

When it comes to programmatic enhancements that propel the program to the highest levels possible, the program must have assets that fledgling programs do not. Two characteristics define this type of program: first, is a multifactored delivery mechanism for the curriculum, one that includes a traditional brick-and-mortar presence, the ability to deliver courses virtually to remain competitive with today's best e-learning programs, and a strong asynchronous component to reach students who simply cannot make it to campus as is the case with many national or international students. The challenge of course will be to find a way to have equal rigor across all three platforms. A second characteristic of a market-leading program is the presence of a transition process for a program's new graduates. That is, a program that informally supports its students following graduation as they acclimate to the new world of scientific practice. Such a program can also offer potential employers a resource for help, guidance, and support, as they begin to open their doors to the newly developing professionals.

CONCLUSIONS

The purpose of this article was to propose and provide a blueprint for a graduate-level curriculum in clinical data science, devoted to the measurement, acquisition, care, treatment, and inferencing of clinical research data. This new “clinical” data science curriculum will prepare students for work in academic, industry, and government research settings and may be reproduced by any institution interested in and capable of offering such a program.

Among the notable strengths of the model are a formal sanctioning of the profession among academic medical centers at the national and international levels; a program that is built upon principles of sound educational theory

and practice; content that is mapped to core competencies of professional societies; and coursework that serves as the knowledge base for the new discipline that is based upon the established disciplines of biostatistics, biomedical informatics, clinical operations, and regulatory affairs. Formal coursework in project management and clinical research ethics as well as an internship at a leading research organization will elevate the preparation beyond what modularized programs can offer. The coursework will not only increase the size of each student's professional toolbox, but the depth of it as well, reframing the training from that of a technician to one of a scientist—to someone who asks 'why' instead of 'how'—or 'how' can we do it better? In essence, training the next generation of professionals to better respond to the problems of today for the benefit of tomorrow can be a challenging but highly rewarding task. The curriculum presented here is not only an investment in people, their knowledge, and the profession of clinical data science, but an investment in the entire scientific fabric of clinical research. In short, it is knowledge for the world.

ACKNOWLEDGMENTS

I would like to thank the dozens of faculty, staff, students, and administrators with whom I met over the past 2 years to brainstorm ideas regarding this new curriculum. I would also like to offer a special note of thanks to the members of my advisory committee, who were willing to meet with me many, many times to help steer me in the right direction with respect to 'best practice.' Those advisory board members are: Ms. Maria Craze, Executive Director, Global Data Operations, Merck & Co.; Dr. Peter Hlebowitsh, Dean and Professor, College of Education, The University of Alabama; Ms. Kit Howard, Senior Director, Standards and Education, Clinical Data Interchange Standards Consortium (CDISC); Mr. Reza Rostami, Director of Quality Management, Duke Clinical Research Institute, Duke University; Mr. Peter Stokman, Business Lead, Data Review and Operational Insights, Bayer Global; Dr. Meredith Nahm Zozus, Division Chief and Director, Clinical Research Informatics, University of Texas Health Science Center at San Antonio; and two unnamed senior government administrators for whom federal policy precludes naming. In addition, I would like to thank Dr. Maurizio Macaluso, Cincinnati Children's Hospital, for his insightful edits and suggestions that improved the readability of the article, and Ms. Linda Ittenbach for her artistic and graphic expertise that proved so valuable in the presentation of this content.

FUNDING INFORMATION

Development of this curriculum was made possible by a sabbatical award to the primary author by the Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital.

CONFLICT OF INTEREST STATEMENT

The author has declared no competing interests for this work.

ORCID

Richard F. Ittenbach https://orcid.org/0000-0002-7914-4298

REFERENCES

1. 21 U.S. Code 301. Federal Food, Drug, and Cosmetic Act. Washington, DC: U.S. Government Printing Office. 1938 Ch. 675, Sec.1, 52 Stat. 1040.
2. Greene JA, Podolsky SH. Reform, regulation, and pharmaceuticals—the Kefauver–Harris Amendments at 50. *N Engl J Med*. 2012;367(16):1481-1483.
3. Meadows M. Promoting safe and effective drugs for 100 years. *FDA Consum*. 2006;40(1):14-20.
4. U.S. Department of Labor. *Clinical Data Manager: 15-2051.02*. 2023. https://www.onetonline.org/link/summary/15-2051.02. Accessed March 8, 2023.
5. Zerhouni EA. Translational and clinical science – time for a new vision. *N Engl J Med*. 2005;353(15):1621-1623. doi:10.1056/NEJMsb053723
6. Austin CP. Opportunities and challenges in translational science. *Clin Transl Sci*. 2021;14(5):1629-1647.
7. ClinicalTrials.gov. National Insitutes of Health, National Library of Medicine. 2023. Updated March 9, 2023. https://clinicaltrials.gov/ct2/resources/trends#LocationsOfRegisteredStudies. Accessed March 11, 2023.
8. Ittenbach RF. Practice of clinical data management worldwide: introduction to the special issue. *J Soc Clin Data Manag*. 2021;1(3):1-2. doi:10.47912/jscdm.146
9. Boichuk V, Glushakov S. The untapped potential of clinical data management in Ukraine: a novel training program case study. *J Soc Clin Data Manag*. 2021;1(3):1-4. doi:10.47912/jscdm.43
10. Houston L, Probst Y. Clinical data management: a review of current practices in Australia. *J Soc Clin Data Manag*. 2021;1(3):1-5. doi:10.47912/jscdm.62
11. Yamaguchi T, Miyaji T, Suganami H, Hayashi Y, Committee SJS. Clinical data management in Japan: past, present, and future. *J Soc Clin Data Manag*. 2021;1(3):1-6. doi:10.47912/jscdm.45
12. Banach MA, Fendt KH, Proeve J, Plummer D, Qureshi S, Limaye N. Clinical data management in the United States: where we have been and where we are going. *J Soc Clin Data Manag*. 2021;1(1):14,1-6. doi:10.47912/jscdm.61
13. Tyler RW. *Basic Principles of Curriculum and Instruction*. University of Chicago Press; 2013.
14. Valenta AL, Berner ES, Boren SA, et al. AMIA Board White Paper: AMIA 2017 core competencies for applied health informatics education at the master's degree level. *J Am Med Inform Assoc*. 2018;25(12):1657-1668.
15. Joint Task Force for Clinical Trial Competency. Domains and Leveled Core Competencies. Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard. 2022. https://mrctcenter.org/clinical-trial-competency/framework/domains/. Accessed March 6, 2022.

16. Sonstein SA, Jones CT. Joint Task Force for Clinical Trial Competency and clinical research professional workforce development. *Front Pharmacol*. 2018;9:1148. doi:10.3389/fphar.2018.01148
17. Zozus MN, Lazarov A, Smith LR, et al. Analysis of professional competencies for the clinical research data management profession: implications for training and professional certification. *J Am Med Inform Assoc*. 2017;24(4):737-745.
18. Ittenbach RF. Managing Scientific Data: Changing the Scientific Perspective. Presented at: Society for Clinical Data Management Annual Conference; September 30, 2019; Baltimore, MD. 2019.
19. Foulkes P. *The Intelligent Use of Big Data on an Industrial Scale*. insideBIGDATA; 2017. https://insidebigdata.com/white-paper/guide-big-data-industrial-scale/. Accessed March 3, 2023.
20. Zozus MN. Clinical data management: what are we missing? Presented at: 26th Drug Information Association Japan Annual Workshop for Clinical Data Management; February 27–28, 2023; Tokyo, Japan.
21. Patient Centered Outcomes Research Institute (PCORI). *2020 Annual Report*. PCORI Institute. https://www.pcori.org/sites/default/files/PCORI-Annual-Report-2020.pdf. Accessed March 3, 2023.
22. Patient Centered Outcomes Research Institute (PCORI). *2021 Annual Report*. PCORI Institute. https://www.pcori.org/sites/default/files/PCORI-Annual-Report-2021.pdf. Accessed March 3, 2023.
23. Wheeland DG. Final NIH genomic data sharing policy. *Fed Regist*. 2014;79:51345-51354.
24. Aitken M, Kleinrock M, Connelly N, Pritchett J, Kern J. *Global Trends in R & D: Overview Through 2021*. Vol 2022. IQVIA Institute for Human Data Science;2022:1-70.
25. Lu Z. Clinical data management: current status, challenges, and future directions from industry perspectives. *Open Access J Clin Trials*. 2010;2:93-105.
26. Hlebowitsh PS. *Designing the School Curriculum*. Allyn & Bacon; 2005.
27. Harden RM. What is a spiral curriculum? *Med Teach*. 1999;21(2):141-143.
28. Bruner JS. *The Process of Education (Revised)*. Harvard University Press; 2009.
29. American Association for the Advancement of Science (AAAS). Project 2061. American Association for the Advancement of Science. https://www.aaas.org/programs/project-2061. Accessed February 27, 2023.
30. Hattie JA, Donoghue GM. Learning strategies: a synthesis and conceptual model. *Nature*. 2016;1(1):1-13.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ittenbach RF. From clinical data management to clinical data science: Time for a new educational model. *Clin Transl Sci*. 2023;00:1-12. doi:10.1111/cts.13545