Therapeutic Innovation & Regulatory Science (2021) 55:1006–1012
https://doi.org/10.1007/s43441-021-00301-z

ORIGINAL RESEARCH

Characterizing Pain Points in Clinical Data Management and Assessing the Impact of Mid-Study Updates

Beth Harper, MBA[1] · Zachary Smith, MA[2] · Jane Snowdon, PhD, FAMIA[3] · Robert DiCicco, PharmD[3] · Rezzan Hekmat, MSH[3] · Van Willis, PhD[3] · Dilhan Weeraratne, PhD[3] · Ken Getz, MBA[2]

Received: 8 February 2021 / Accepted: 20 April 2021 / Published online: 7 May 2021

Abstract

Background The causes, degree and disruptive nature of mid-study database updates and other pain points were evaluated to understand if and how the clinical data management function is managing rapid growth in data volume and diversity.

Methods Tufts Center for the Study of Drug Development (Tufts CSDD)—in collaboration with IBM Watson Health—conducted an online global survey between September and October 2020.

Results One hundred ninety four verified responses were analyzed. Planned and unplanned mid-study updates were the top challenges mentioned and their management was time intensive. Respondents reported an average of 4.1 planned and 3.7 unplanned mid-study updates per clinical trial.

Conclusion Mid-study database updates are disruptive and present a major opportunity to accelerate cycle times and improve efficiency, particularly as protocol designs become more flexible and the diversity of data, most notably unstructured data, increases.

Keywords Clinical data management · Clinical data science · Mid-study updates

Introduction

The clinical data management function faces unprecedented challenges with the rapid growth in data volume coming from a growing number of diverse data sources including mobile and wearable devices and real-world data. A growing percentage of data is coming directly from patients themselves and is unstructured in nature [1]. These trends, combined with more complex studies and non-traditional study designs, require continuous improvement in all aspects of data management. Adaptive and flexible designs, supported by hybrid, decentralized clinical trial models informed by risk-based assessments and augmented analytics are but a few of the many ways that drug development is evolving.

A recent white paper by the Society for Clinical Data Management (SCDM) challenged the drug development enterprise to nurture “true technology innovation” [2]. Recommendations included the development of fit-for-purpose data standards and the expansion of intelligent clinical data management systems that allow clinical data scientists to not only collect and integrate data, but also interact with them. The paper further recognized that while electronic data capture (EDC) remains the primary data collection platform in the industry, more flexibility is needed within these systems, in particular with regard to enabling fast mid-study changes (i.e., “being able to build, test, validate and push live adaptations out within days”), an emerging capability deemed essential for future drug development activity. The COVID-19 pandemic has brought new meaning to the concepts of unplanned and unforeseen mid-study changes along with a greater sense of urgency to address many of the recommendations in the SCDM white paper.

Several recent studies have measured overall data management cycle times and outlined specific initiatives that organizations are taking to improve their capabilities and drive efficiency [3, 4]. Key practices include establishing formal data and data governance strategies, implementing

✉ Zachary Smith
Zachary.Smith605922@tufts.edu

[1] Clinical Performance Partners, Aurora, IL, USA

[2] Tufts CSDD, 75 Kneeland Street- Floor 11, Boston, MA, USA

[3] IBM Watson Health, Cambridge, MA, USA

a central data platform or hub/lake, developing data science capabilities, and investing in more sophisticated data technology infrastructure. Other studies in the literature have recognized new roles and competencies that are needed to support digital transformation throughout the drug development process [5, 6].

One area that has not been quantified and explored is the evolving challenges associated with this transformation and the preparedness of organizations to manage them. In late 2020, the Tufts Center for the Study of Drug Development (Tufts CSDD), in collaboration with and funded by IBM Watson Health, conducted a study to assess clinical data management pain points in detail, ascertain the impact that these challenges have on key performance metrics, and gain insights into how the industry is addressing and responding to them.

Methods

Study Design and Participants

Tufts CSDD and IBM Watson Health developed an online survey on Qualtrics to collect perceptions and experiences of clinical data management functions supported by clinical data management solutions. The survey was distributed world-wide to email addresses acquired through an email list service. The survey included two filter questions in the background section at the start of the survey in order to ensure that respondents had the relevant experience required. Respondents who indicated that their organization did not conduct any Phase I-III clinical studies; or who indicated that they did not use EDC solutions nor were they responsible for EDC solutions at their organization were directed to the end of the survey without responding to the questions used in this analysis. The survey took approximately 20–30 min to complete and was organized as follows:

1. Seven questions on respondent background (position title, years in position, etc.)
2. Nine questions on data collection and management experience within their organization
3. Four questions on the features and functionality of EDC systems used at their organization (including a question about EDC systems' performance and effectiveness during the COVID-19 pandemic)
4. Three questions on overall and specific data management cycle times
5. Two questions on data sources and integration
6. Two questions on satisfaction with EDC solution service providers

At the conclusion of the survey, respondents were provided a link to an additional survey where they were given the option to provide contact information. The use of a second, separate survey allowed complete de-identification of responses and prevented the linking of contact information provided to any particular survey response. Survey responses were collected between September 10, 2020 through October 30, 2020. Respondents were invited to participate in the survey via email. The only compensation for responding was the offer of a summary of survey results if the respondent requested it and provided their contact information after completing the survey. On October 30, 2020 the survey was closed and any unfinished survey responses were closed and recorded. Survey responses were anonymized and no metadata was collected. The Tufts CSDD team removed values suspected of being input errors (i.e., 0-day cycle times, 90 planned mid-study updates per trial).

Key Definitions

Given our focus on mid-study updates, the following definitions were provided to survey respondents:

Database planned mid-study updates—*Planned* study-specific database amendments and post-production changes as outlined in the study protocol or standard operating procedure (SOP).

Database unplanned mid-study updates—*Unplanned* study-specific database amendments and post-production changes not outlined in the original study protocol or SOP.

Sub-groups and Data Analysis

Sub-groups for analysis were based on responses to specific questions. "How many Phase I-III clinical trials… does your organization initiate each year?" was used as a proxy for company size. Respondents reporting 50 or more trials per year were considered "Larger," and respondents reporting fewer than 50 trials per year were considered "Smaller." Respondent experience with their primary EDC was split at the median. Those reporting having used their primary EDC for more than 6 years were considered "Expert," and respondents with 6 or fewer years' experience with their primary EDC were considered "Beginner." Another sub-group was formed based on whether respondents reported were satisfied with their primary EDC's handling of unplanned mid-study updates ("Satisfied" vs. "Unsatisfied"). Survey responses used to create Table 3 were originally on a 4-point Likert scale, but these responses were later combined into binary responses (i.e. "Extremely Difficult" and "Somewhat Difficult" were combined into "Difficult;" "Very Easy" and "Easy" were combined into "Easy"). Responses regarding the types of trials in which specific functions were used were also combined into three categories—"Function is Used," "Function is Available But Not Used," and "Function Is Not Available." "Number of Pain Points Experienced" was calculated by counting the number of pain points reported by each respondent.

Frequencies, means, medians, coefficients of variation, and ranges were calculated for responses. ANOVA and Chi-Square were used to test for significant differences between subgroups. For comparison purposes, cycle time data was drawn from a previous Tufts CSDD study conducted in 2017 [2]. No significance tests were conducted between the 2020 and 2017 studies. Data was stored on a secure drive as a .csv and analyzed using SAS 9.4.

Results

Verified survey responses were collected from 194 respondents across a broad range of organization types and roles. Respondents from Sponsor or consulting organizations represented the largest group (59.0%), followed by Academic Health Systems, Hospitals and Non-Profit organizations (16.7%), Contract Research Organizations (13.5%), Research Sites (4.5%), Medical Device Manufacturers (1.9%) and Other organizations (4.5%). The top four respondent roles represented individuals working in Clinical Operations (26.5%), Data Management (18.2%), Executive Leadership (18.2%) and Clinical Development / Clinical Scientists (17.4%). Individuals working as Investigative Site Staff, in Study Monitoring, Vendor Management, Clinical IT or as Biostatisticians / Data Scientists were each represented by less than four percent of respondents. The mean number of years of experience was 10.8 with a coefficient of variation (CoV) of 0.75.

Three-fourths (76.9%) of respondents work for Smaller organizations. The remainder (23.1%) work for Larger organizations. There was a wide variance in terms of the percentage of data management work that was outsourced by these organizations with a range of zero to 100%. The mean was 70.7% with a CoV of 0.48. Responses reflected the experiences of organizations conducting trials across all continents and regions.

Respondents report that a wide variety of EDC systems are being used. Many organizations (73.9%) are using multiple EDC solutions; 26.1% report using only a single EDC solution. Respondents who use more than one EDC solution were asked to consider their primary solution (the solution they use most frequently) when responding to several questions, however, this analysis focused mainly on understanding the use, limitations, and impact of their collective EDC solution experience.

Cycle Time Experience and Pain Points

Study close-out cycle times—historically one of the most common data management performance metrics gathered—has not changed during the past three years. The duration from last patient last visit (LPLV) to database lock (DBL) remains at approximately 37 days (ranging from 2 to 120 days, with a median of 30 days). Larger companies reported an increase from 33.7 to 36.0 days on average; Smaller companies reported a decrease in cycle time from 42.7 to 37.5 days (Refer to Fig. 1).

One-third of companies responding reported that the study database is released before the occurrence of the first patient's first visit (FPFV). This preparation has not changed since the 2017 survey.

Responding companies report experiencing on average 3.7 unplanned and 4.1 planned updates per study (Refer to Table 1). The coefficient of variation around these means is very high (1.82–1.84) indicating wide differences in the average number of mid-study updates across organizations.

A number of causes of mid-study updates were mentioned. Protocol amendments and intentional or pre-planned updates were among top causes noted by 84.7% and 62.9% of respondents, respectively. A majority (70%) mentioned sponsor requests as a primary cause of updates. Approximately one-third of respondents noted feedback from users (37.7%), misinterpretation of or new knowledge about the protocol (36.6%), enablement of new features (36.6%) and requests by sites (32.9%) as primary causes of mid-study updates.

A more granular look at the data management lifecycle is presented in Fig. 2. The overall cycle time from initial protocol approval to Database Go Live is about ten weeks (69.4 days) on average. The time to manage mid-study updates is a little over 4 weeks (28.5 days for planned, 29.9 days for unplanned). Smaller companies report shorter durations for early stage data management lifecycle milestones; Larger companies report shorter relative durations for late-stage

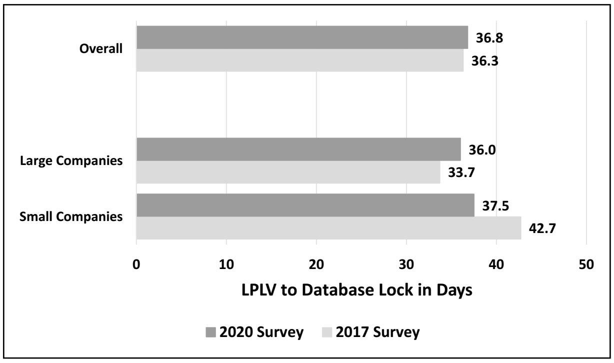

Fig. 1 Comparison of last patient last visit (LPLV) to database lock cycle times across current and 2017 survey

Table 1 Frequency of mid-study updates

Update type	*n*	Mean updates per Study (CoV)	Median	Range
Planned mid-study updates	53	4.1 (1.84)	2	0–45
Unplanned mid-study updates	56	3.7 (1.82)	2	0–45

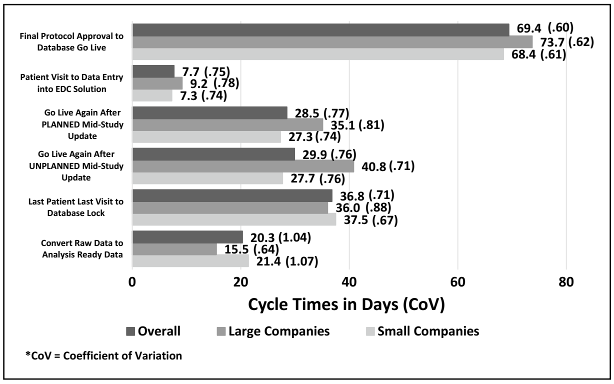

Fig. 2 Clinical data management cycle time for various segments of the lifecycle

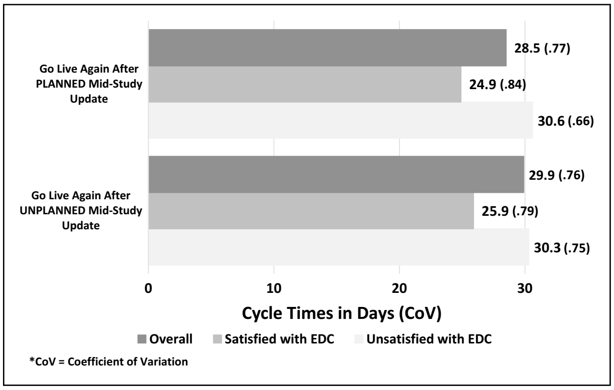

Fig. 3 Cycle time advantages reported by those more or less satisfied with their electronic data capture (EDC) solution's ability to manage mid-study updates

milestones (e.g., converting raw data to analysis data sets and LPLV to database lock).

Higher variation around mean durations is generally observed in the later-stage clinical data management cycle times—most notably study close-out.

Planned and unplanned mid-study updates are the top two pain-points experienced by respondents. More than half (56.5%) mention planned and 48.4% mention unplanned, mid-study database updates. Nearly half (43.6)% of respondents mention flexibility and customization among their top three pain points. Other pain points were reported by less than one-third of respondents: Database Go Live delays and lack of integrated patient engagement/electronic clinical outcome assessment (eCOA) applications (32.3%); solution cost (30.7%); data incompatibility between platforms (25.8%); customer support problems (22.6%); and solution complexity/slow user learning curve (19.4%).

EDC Solution Satisfaction, Functionality and Utilization

The majority of respondents reported high satisfaction with their primary EDC solution's ability to perform a wide range of tasks with lower relative levels of satisfaction associated with the management of mid-study updates.

Nearly all respondents—98.3% and 92.9%, respectively—were satisfied with the data collection and database design functionality of their EDC solution. High satisfaction was also reported for database closeout (89.5%) and data processing and cleaning capabilities (83.3%). Most respondents (88.1%) were satisfied with the overall process (from database design to database closeout) but less so with the associated cycle time (only 75.4% being satisfied with the time from protocol approval, to Go Live, to database closeout). Less than eighty percent (79.7%) of respondents were satisfied with the ability of their EDC solution to handle planned mid-study updates and even fewer (67.2%) with the ability to handle unplanned mid-study updates.

Organizations that were more satisfied with the ability of their EDC solution to handle mid-study updates reported a roughly five-day speed advantage per update for both planned and unplanned mid-study updates (Refer to Fig. 3).

The majority of respondents were also satisfied with the ability of their EDC solution to handle their needs during the pandemic. Eighty-six percent found that their primary EDC solution met their needs for non-COVID-19 clinical trials with only slightly less (77.5%) indicating that their solution met their needs during the COVID-19 pandemic or could meet their needs in the event of future similar outbreaks. A majority (83.7%) also found that their EDC solution empowered them to manage data in remote work environments, a particularly important feature necessitated by the pandemic.

Respondents indicated that most EDC solutions offer a wide range of functionality (see Fig. 4). Data Integration capabilities were reported to be available in the majority (83.0%) of solutions, along with typical data management capabilities, including electronic data capture (98.2%), query management (96.5%) and study-level reporting (94.7%). While not directly related to data integration, numerous

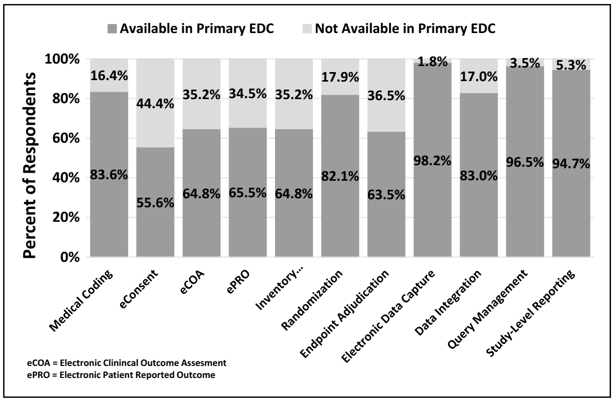

Fig. 4 Availability of functions within electronic data capture (EDC) solutions

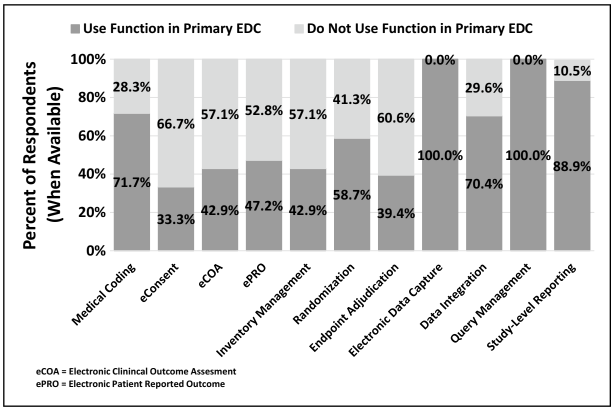

Fig. 5 Utilization of functions within the electronic data capture (EDC) solutions

other functions were noted to be available within the EDC solutions. Randomization and medical coding functions were reported to be widely (> 80%) available. Endpoint adjudication, eCOA, electronic patient reported outcomes (ePRO), Inventory Management and eConsent were available in over half of solutions.

Although most systems offered a wide variety of functionality, utilization of these functions varied widely. Figure 5 shows the range of utilization from 33.3% for eConsent to full utilization (100%) of EDC and query management functionality. Approximately, two-thirds (70.4%) of respondents reported using the data integration functionality within their EDC solution.

Data Integration and Analysis

EDC was noted as being the most commonly used tool for managing electronic case report forms (eCRF), local lab data, quality of life (QoL) data and medical images (see Table 2). SAS was the predominant tool used for most other data sources. Investment in or development of integrated data platforms or hubs is low with < 20% reporting their use to integrate, organize, review or analyze data from multiple disparate sources. Consistent with findings from a recent Tufts CSDD study, Excel, the least sophisticated tool from a data analytics standpoint, is still widely used (3). For many respondents in the 2020 survey, mobile health, genomic and proteomics data were not being collected in their primary study database.

A variety of factors were identified as barriers to the transfer of data from third parties into the EDC. Top among

Table 2 Tools and systems being used to integrate, organize, review, or analyze data

		Percent of respondents using each tool					
Source of data	*n*	Data lake or data hub (%)	EDC (%)	Excel (%)	SAS (%)	Other tool (%)	Not applicable (%)
eCRF data	59	18.6	**79.7**	30.5	47.5	15.3	3.4
Local laboratory data	57	5.3	**54.4**	29.8	38.6	15.8	14.0
Central laboratory data	58	13.8	34.5	29.3	**48.3**	17.2	3.4
Biomarker data	58	12.1	24.1	27.6	**46.6**	12.1	15.5
Pharmacokinetic data	58	12.1	19.0	31.0	**44.8**	15.5	15.5
Pharmacodynamic data	57	12.3	19.3	24.6	**45.6**	10.5	22.8
Mobile health data	52	7.7	23.1	15.4	**28.8**	11.5	*46.2*
Genomic/proteomic data	54	9.3	13.0	14.8	**25.9**	7.4	*46.3*
Quality of life data	57	10.5	**50.9**	12.3	38.6	17.5	14.0
Medical images	56	10.7	**26.8**	3.6	19.6	19.6	*33.9*
ePRO data	56	14.3	30.4	14.3	**35.7**	14.3	26.8
Other data	7	0.0	0.0	0.0	14.3	14.3	*85.7*

Bold most commonly used tool for source of data

Italic most common response was N/A

EDC electronic data capture, *eCRF* electronic case report form, *ePRO* electronic patient reported outcome them, as noted by two-thirds of the respondents, were limitations of the EDC solution (66.2%), the cost or ease of integration (60.6%) and the technical demands this effort placed on internal (66.2%) or external (56.3%) support staff. EDC solution transaction and EDC solution performance degradation were also noted by more than half the respondents (52.9% and 50.7%, respectively) as top factors impacting efficient data transfer.

Despite the barriers to accessing data from third parties, most respondents found data from a wide variety of sources to be easy to integrate (see Table 3). Data integration was most commonly reported as easy for eCRF, Central Lab, ePRO, Pharmacodynamic, QoL, Biomarker, Pharmacokinetic and Local Lab Data. Only Medical Images, Mobile Health Data and Genomic / Proteomic Data were commonly noted as difficult to integrate.

Table 3 Ease of data integration across multiple data sources

		Percent of respondents	
Source of data	*n*	Easy (%)	Difficult (%)
eCRF data	56	89.3	10.7
Central laboratory data	54	68.5	31.5
ePRO data	37	62.2	37.8
Pharmacodynamic data	42	61.9	38.1
Quality of life data	46	60.9	39.1
Biomarker data	47	59.6	40.4
Pharmacokinetic data	45	55.6	44.4
Local laboratory data	48	54.2	45.8
Medical images	34	50.0	50.0
Mobile health data	26	38.5	61.5
Genomic/proteomic data	27	22.2	77.8
Other data	1	100	0.0

eCRF electronic case report form, *ePRO* electronic patient reported outcome

Discussion

A number of data management challenges and pain points were identified in this study. Top pain points are associated with planned and unplanned mid-study updates. Mid-study updates appear to be common since an average of 3–4 were reported per study in our survey. Other major pain points include Database Go Live delays, lack of integrated outcome assessments and technology solution costs. These challenges were comparable regardless of company size or user years of experience.

Study close-out cycle times have not improved and remain the data management cycle with the highest observed variance, suggesting inconsistent and unpredictable experience. The more granular cycle times present new benchmarks and elucidate the time-intensive requirements of managing data volume and diversity at this time.

Satisfaction with EDC solutions and their capabilities remains high, and EDC solutions offer a variety of features and functionality. Although widely available, utilization of this functionality is variable and suggests—consistent with the SCDM white paper—that EDC systems have not been designed to be the central study data repository receiving and loading all external data [4].

This study found that data integration capabilities are available in the majority of EDC solutions. While some challenges were noted integrating data from third party sources, this didn't appear to be a significant barrier and, for the most part, respondents reported data integration within their EDC solution to be easy. Consistent with other studies, data integration tools vary widely with a number of unsophisticated ones (e.g., Excel) still in use. Future research is needed to demonstrate whether a larger number of operations performed on a single platform will result in time savings and improved efficiency.

Mid-study updates are the most common pain point and the area where organizations are the least satisfied with the capabilities of their EDC solutions. While those who were more satisfied in this area achieved a 5-day cycle time advantage in the go live time after database release per update, it's unknown to what extent this is due to the functionality of the EDC solution or other internal processes. Future research will look to gather insights into best practices associated with managing planned and unplanned mid-study update processes. Additional research into the pain points associated with other areas such as the removal of data silos and end-to-end harmonization may also yield beneficial results.

Conclusion

The growing volume and diversity of clinical trial data is inevitable as drug developers look to gather data on more stratified patient populations, rely on more sources for clinical research and patient health data, and conduct more operationally complex protocols. At the same time, drug development timelines and efficiency continue to worsen. In response, the clinical data management function—in collaboration with clinical data solutions providers —is expected to evolve substantially to address these challenges and leverage the value of rich data. The pandemic has fast-tracked the adoption of decentralized trials, which translates to new patient engagement approaches and technology.

This study is the first-of-its kind to characterize pain points associated with clinical data volume and diversity and to capture the specific impact of mid-study updates. We can expect the occurrence of mid-study updates, and demand for practices and solutions to more effectively manage them, to increase in the future.

Acknowledgements

We would like to acknowledge the support of Gretchen Jackson, MD, PhD, Nawshin Kutub, PhD, Brett South, PhD, and Courtney VanHouten, PhD in the design and execution of this study.

Funding

IBM Watson Health.

Declarations

Conflict of interest

No potential conflicts were declared.

Reference

1. Lamberti MJ, Kubick W, Awatin J, McCormick, et al. The use of real world evidence and data in clinical research and post approval safety studies. Ther Innov Regul Sci 2018
2. Wilkinson M, Young R, Harper B, Machion G, Getz K. Baseline assessment of the evolving 2019 eClinical Landscape. Ther Innov Regul Sci. 2018;53(1):869–76.
3. Harper B, Wilkinson M, Indupuri R, Rocchio R, Getz K. Evolving clinical data strategies and tactics in response to digital transformation. Ther Innov Regul Sci. https://link.springer.com/article/10.1007%2Fs43441-020-00213-4. Accessed 23 Dec 2020.
4. Society for Clinical Data Management White Paper. The evolution of clinical data management to clinical data science (Part 2: the technology enablers). https://scdm.org/white-papers/. Accessed 20 Dec 2020.
5. Society for Clinical Data Management White Paper. The evolution of clinical data management to clinical data science (Part 3: the evolution of the CDM role). https://scdm.org/white-papers/. Accessed 20 Dec 2020.
6. Tufts Center for the Study of Drug Development. Drug development workforce in the age of digital transformation. https://csdd.tufts.edu/white-papers. Accessed 29 Dec 2020.