Journal of Clinical and Translational Science

www.cambridge.org/cts

Research Methods and Technology
Brief Report

Cite this article: Zimolzak AJ, Davila JA, Punugoti V, Balasubramanyam A, Klotman PE, Petersen LA, Rochat RH, Liao G, Laubscher RR, Leiber L, and Amos CI. Lessons learned from an enterprise-wide clinical datathon. *Journal of Clinical and Translational Science* **6**: e125, 1–5. doi: 10.1017/cts.2022.450

Received: 17 February 2022
Revised: 12 July 2022
Accepted: 18 August 2022

Keywords:
Diffusion of innovation; technology assessment; biomedical; cooperative behavior; interdisciplinary communication; information systems

Address for correspondence:
C. I. Amos, PhD, Dan L. Duncan Institute for Clinical and Translational Research, One Baylor Plaza, BCM 451, Houston, TX 77030, USA.
Email: chrisa@bcm.edu

ACTS
ASSOCIATION FOR CLINICAL AND TRANSLATIONAL SCIENCE

Clinical Research FORUM
Analysis. Advocacy. Action.

Lessons learned from an enterprise-wide clinical datathon

Andrew J. Zimolzak[1,2], Jessica A. Davila[1,2], Vamshi Punugoti[1], Ashok Balasubramanyam[1], Paul E. Klotman[1], Laura A. Petersen[1,2], Ryan H. Rochat[3,4], Gloria Liao[1], Rory R. Laubscher[1], Lee Leiber[1] and Christopher I. Amos[1]

[1]Department of Medicine, Baylor College of Medicine, Houston, TX, USA; [2]Center for Innovations in Quality, Effectiveness and Safety, Michael E. DeBakey Veterans Affairs Medical Center and Baylor College of Medicine, Houston, TX, USA; [3]Texas Children's Hospital, Houston, TX, USA and [4]Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA

Abstract

In 2020, Baylor College of Medicine held a datathon to inform potential users of a new data warehouse, allow users to address clinical questions, identify warehouse capabilities and limitations, foster collaborations, and engage trainees. Senior faculty selected proposals based on feasibility and impact. Selectees worked with Information Technology for 2 months and presented findings. A survey of participants showed diverse levels of experience, high perceived value of the datathon, high rates of collaboration, and significant increases in knowledge. A datathon can promote familiarity with a new data warehouse, guide data warehouse improvement, and promote collaboration.

Introduction

There is a growing amount of electronic health data available that can be used to study healthcare delivery and conduct quality improvement. To facilitate innovative use of these data sources, "datathon" events have increased in popularity to address clinical questions and facilitate multidisciplinary collaborations [1]. The term "hackathon" was first used around 1999 and referred to several days of intense software development and innovation. This term later broadened to include innovation in the life sciences, and the first PubMed citation using the term "hackathon" is in 2011 [2], the same year as the founding of the MIT Hacking Medicine group [3]. Subsequent terminology included "data marathon" [1] and finally "datathon" in 2015 [4].

In 2020, Baylor College of Medicine (BCM) held a virtual datathon event focused on addressing clinical and quality improvement questions using the BCM Enterprise Data Warehouse. As a new data resource at BCM, our goals were to (1) inform potential users of available data resources, (2) leverage local data to address clinical questions, (3) test local data and identify limitations of data use, and (4) facilitate cross-institutional collaborations among faculty, postdoctoral trainees, and students.

Methods

BCM is located in the Texas Medical Center of Houston. Its clinical faculty are on staff at affiliate institutions that include Harris Health, Baylor St. Luke's Medical Center, Texas Children's Hospital, and the Michael E. DeBakey VA Medical Center. In addition, BCM supports an extensive faculty group practice called Baylor Medicine. Unique electronic health records are used and maintained at each affiliate, with varying policies for data governance and data access. The BCM enterprise data warehouse was established to integrate these data sources and serve as a central data repository, containing data from multiple source systems organized into data marts and data lakes. The warehouse maintains clinical data on 4.3 million unique individuals, 151 million encounters, 122 million lab tests, and over 800 thousand radiology images.

The BCM datathon had institutional support from multiple leaders, including the President of BCM, Office of Research leadership, and the Chief Information Officer who leads Information Technology (IT). Clinical champions were identified from affiliate institutions. A datathon planning committee was established to determine event policies, facilitate cross-institutional collaborations, and select potential projects to participate in the event. Because this would be a virtual event due to the COVID-19 pandemic, the committee was responsible for planning the logistics of the virtual event. Unlike traditional in-person 2-day datathon events, teams had more time to work on projects before their presentations. To identify appropriate projects for the datathon, the planning committee reviewed preliminary data from the BCM data warehouse. They conducted four test-case projects to ensure data delivery processes and usability. Test cases were conducted in healthcare quality and population health domains, including hypertension control, mammography rates, anesthesia, and an inpatient warning score.

A standardized application process for datathon participation was developed. Only BCM faculty, trainees, and staff were eligible to participate. A multidisciplinary panel of senior faculty from across the College reviewed proposed project applications based on seven items/dimensions in a standardized rubric (presented in supplementary table 1). The committee selected proposals based on feasibility and relevance to quality improvement or population health. Since datathon projects were considered healthcare operations for this event and no identifiable information was provided, institutional review board approval was not required for the datathon. Institutional review was required if teams wished to analyze data after the datathon or disseminate findings.

Data extraction began immediately after project selection. Teams submitted data requests to BCM IT, and project teams were provided a schedule for data availability. Project teams had iterative consultations with IT to refine parameters for data extraction. Microsoft Teams was used for real-time collaboration. Our datathon was innovative because it involved sustained effort from BCM IT rather than a few days of intense work. A final dataset was provided to teams for analysis in early October 2020; final presentations occurred on October 27. A six-person judging committee, including faculty from the affiliated hospitals, gave four awards (most clinically innovative, most innovative use of data, excellence in collaboration, greatest potential for impact on patient care). The judging committee and coauthors determined opportunities and challenges by reviewing participants' work and comments during the datathon.

We developed a 28-question survey to study datathon participants' experience. Questions covered participants' background experience, datathon team characteristics and collaborations, knowledge before and after the datathon, the perceived value of the datathon, and plans for future work, such as dissemination of findings. Questions were multiple choice or Likert-type ordinal rating scales with an optional free-response question. The survey was administered using REDCap software online [5]. The data underlying this article will be shared on reasonable request to the corresponding author.

Table 1. *Survey respondent characteristics*

Characteristic	Sub-category	Measure
Academic rank	Student	3 (10.7)
	Fellow	2 (7.1)
	Staff	7 (25.0)
	Assistant professor	5 (17.9)
	Associate professor	5 (17.9)
	Full professor	6 (21.4)
Prior datathon participation	–	14 (51.9)
Role	Team lead	13 (46.4)
	Clinician	9 (32.1)
	Chart reviewer	6 (21.4)
	Statistics	6 (21.4)
	Data manager	4 (14.3)
	Data warehouse	3 (10.7)
	Learner	3 (10.7)
	Data scientist	2 (7.1)
	Other	1 (3.6)
Years at BCM, median (Q1–Q3)	–	5 (3–8)
Team size, median (Q1–Q3)	–	5 (4–7)
Person-hours effort, median (Q1–Q3)	–	20 (10–45)
Percent spent with IT, median (Q1–Q3)	–	10 (2.5–45)
Prior electronic health record data experience	1 (very little)	5 (18.5)
	2	4 (14.8)
	3 (moderate)	8 (29.6)
	4	6 (22.2)
	5 (a great deal)	4 (14.8)

Values are presented as count (percent) unless otherwise indicated. Roles add to more than 100% because respondents could report more than one role. Abbreviations: IT = information technology; BCM = Baylor College of Medicine; Q = quartile.

Results

A total of 33 project teams submitted proposals, and 13 were selected (3 outpatient, 8 inpatient, 2 combined inpatient/outpatient, 12 adult, and 1 pediatric). The following topic areas were covered: early warning for acute kidney injury (AKI) after surgery, seasonality of AKI and respiratory viruses, designing surgical instruments using radiographs, characteristics of COVID patients with complications, referrals and quality in chronic kidney disease using geoanalytics, blood pressure variability and intracerebral hemorrhage outcomes, predictors of COVID outcomes, fluid balance effect on outcomes in patients with subarachnoid hemorrhage, complications of cancer treatment, smoking and cancer screening, inappropriate recording of antimicrobial allergies, and disparities in fragility fracture care between two hospitals. Details about the departmental membership of these teams are presented in supplementary table 2.

Surveys were sent to all 67 participants, of whom 28 initiated the survey, and 25 completed all questions. Baseline characteristics of respondents are shown in Table 1. In brief, academic ranks included three students, two fellows, five assistant professors, five associate professors, six full professors, and seven other job titles. Fourteen (52%) had prior experience participating in a datathon or hackathon. Median years at BCM was 5, median team size was five individuals, median effort was 20 person-hours, and median percentage spent with IT was 10%. Respondents reported electronic health record data experience that was relatively evenly distributed across levels.

Figure 1 shows collaboration results. Seven survey respondents (27%) had never worked with their team before. Twenty (87%) reported collaborating outside their department, and twenty-one (91%) reported collaborating with new people. Only 1 respondent reported that their team fully completed their project, whereas 11 reported partial completion. Six reported that they fully answered their research question, and eleven reported partial completion.

https://doi.org/10.1017/cts.2022.450 Published online by Cambridge University Press

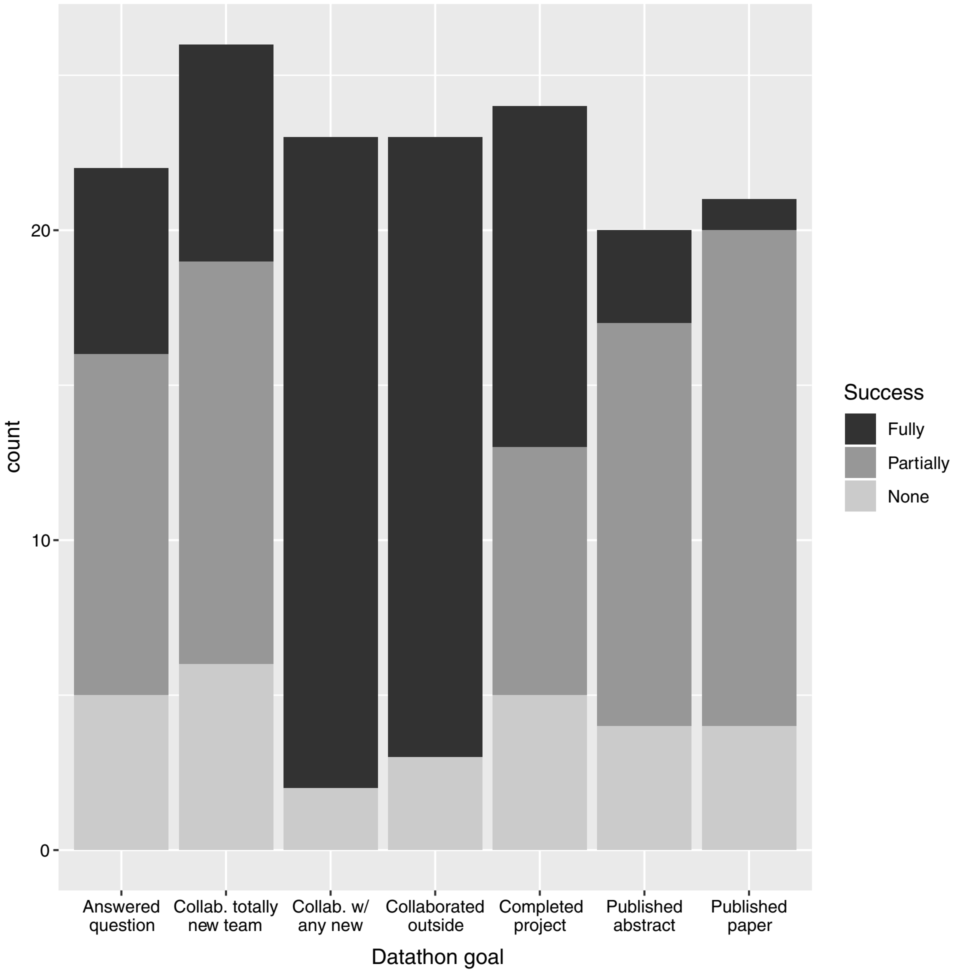

Fig. 1. Datathon goals achieved by survey respondents. Datathon participants responded to a survey that collected self-assessed success (fully, partially, or unsuccessful) across seven specific goals of the datathon. Goals included answering the original research questions and dissemination of findings. Collaboration was measured in three questions: worked with datathon team before (none, some, all), collaborated with new people (yes/no), and collaborated outside home department (yes/no). Total counts vary by question if respondents did not answer all survey questions.

Regarding dissemination of findings, 3 reported writing an abstract, 13 reported a planned abstract, 16 reported a planned paper, and 1 reported writing a paper (Fig. 1).

On a scale of 1–5, modal difficulty in obtaining data was 3, modal difficulty in working with data was 3, modal response to considering this a valuable experience was 5 (56%), and most (65%) reported a score of 5 for intent to participate in future datathons. Many (46%) reported strong intentions (5) to conduct future studies using BCM Enterprise Data. In all three dimensions of self-reported knowledge (how to use the data warehouse, data availability, and data warehouse limitations), paired pre/post-analyses showed significant increases (Fig. 2).

Discussion

This datathon met our goals by familiarizing investigators with available data resources and limitations, familiarizing IT with investigators' requests, and identifying opportunities for improvement in the secondary use of clinical data. The lessons learned will allow BCM to address relevant and timely research questions in the future using locally integrated data sources.

During the datathon, participants across the College successfully analyzed real-world data and reported their experiences. We achieved our goal of encouraging participants to work with new collaborators. Our datathon used a novel time frame of several weeks, allowing asynchronous participation by team members whose clinical schedules did not align. The IT collaboration platform allowed us to pivot rapidly to remote work when the pandemic began. One notable success of this event was the cross-institution collaboration to address important clinical questions using innovative methods. Clinical questions covered the spectrum from inpatient critical care to outpatient population health, from specialized to general, and data types encompassing traditional structured to critical care flowsheets to imaging. We successfully recruited a broad range of faculty with varying academic ranks and experience working with electronic health data (Table 1). Staff with diverse experience across the organization engaged in many projects, leading to timely and innovative project ideas.

There were several opportunities to increase the efficiency of working with IT staff and accessing project data. The increased use of a common platform (e.g., Microsoft Teams, OneDrive) for communication and collaborative editing of files would support

https://doi.org/10.1017/cts.2022.450 Published online by Cambridge University Press

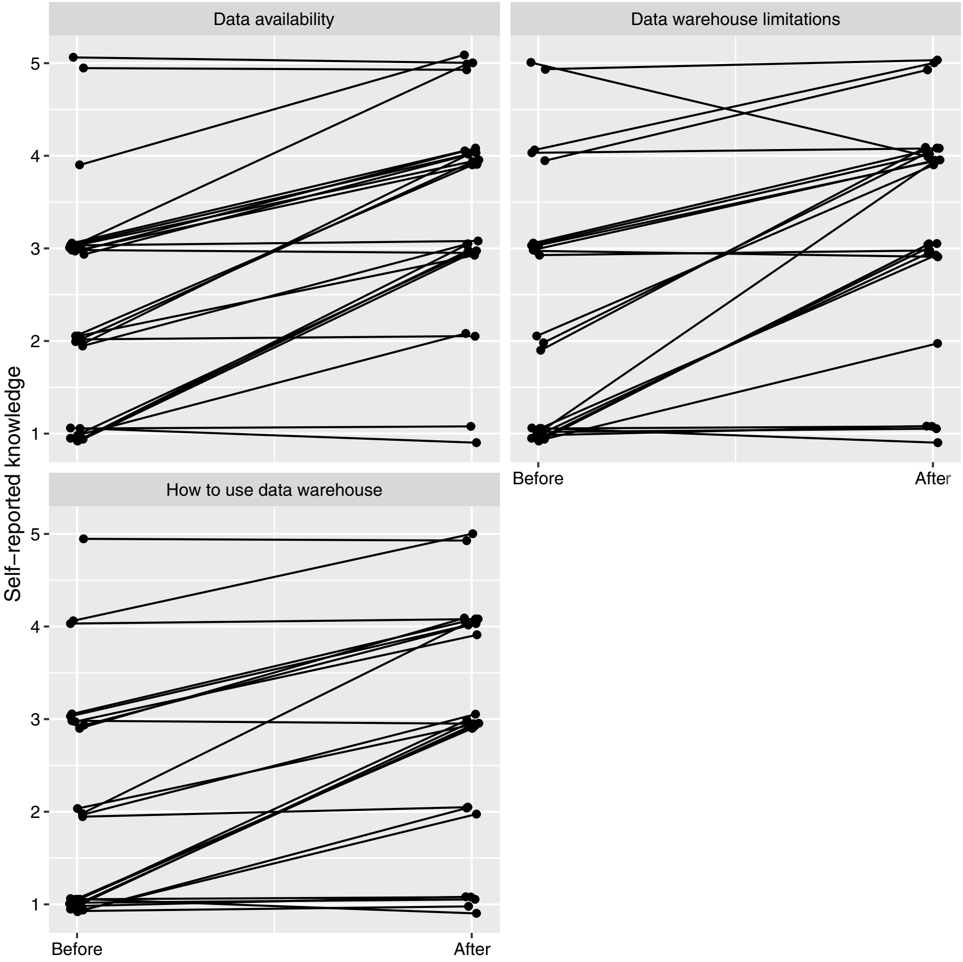

Fig. 2. Effect of datathon on respondents' self-reported knowledge. Three dimensions of data warehouse knowledge were surveyed with higher values indicating more knowledge. One line segment represents one respondent and connects the pre- and post-datathon responses. *P* values were 0.003 for knowledge about data availability, 0.010 for understanding data warehouse limitations, and 0.015 for knowledge about how to use the warehouse (chi-squared test for trend in proportions).

more real-time interactions between IT and clinicians/researchers. Second, the questions posed by teams were highly dependent on data availability, even among the 13 selected projects (e.g., accurate and complete ascertainment of narrow inclusion criteria was a challenging to some projects such as number 4 and 7 in supplementary table 2). This presents an opportunity for self-service tools like Epic SlicerDicer and i2b2 to provide insight into data availability [6]. Given the strong interest in their adoption, these software tools were implemented with training shortly after the datathon ended. Third, most projects required extensive data cleaning, a common phenomenon when using clinical data for secondary analyses [7]. Increased integration of data sources could reduce time and effort spent on data cleaning, but projects still need to budget time for cleaning. Integration and patient linkage across data sources should also be automated as much as possible. We found that large, high-resolution datasets (e.g., critical care vital signs) were challenging to curate; however, these granular data were valuable when analyzed appropriately. Lastly, an executive champion was critical. In our case, the President of the College championed the event for faculty and learners and dedicated IT resources to the project.

We identified several areas of opportunity for future datathons. First, we observed communication barriers between IT and clinical staff. Clinical information in the electronic health record used by the provider could not be available in the same format in the data warehouse. Education and detailed resources for users about data availability could support the alignment of expectations. Differences in the understanding of technical language between IT staff and clinicians also caused communication gaps. Interactive tools [8] or wider inclusion of biomedical informaticians (staff with cross-training in clinical and IT domains [9]) could improve efficiency. While these barriers were greater for less experienced users, team members with expertise in clinical and IT field-guided projects effectively. Second, teams did not have access to self-service tools such as i2b2 and Epic SlicerDicer to retrieve and analyze data. Agreeing upon a standard set of tools at the beginning of a project coupled with targeted instruction can improve understanding (e.g., indicating at an earlier stage that inclusion criteria may result in too small of a cohort) and support efficient transitions between project tasks. As a result of the survey, we have developed a series of instruction sessions for the next iteration of the datathon. A governance software solution was purchased and installed to improve the integration and tracking of data for large-scale projects like the datathon. The software will support data integration across the affiliated sites that provide data to BCM, management of requests for data, and central tracking of data mart delivery. Lastly, projects requiring imaging data presented a unique challenge. The software demands of image

https://doi.org/10.1017/cts.2022.450 Published online by Cambridge University Press

retrieval and analysis suggest a need for a dedicated platform for imaging analysis to support future work.

Our study has several limitations. Some teams had more time for data analysis than others due to variations in the time required to extract specific datasets from the BCM Enterprise Data Warehouse. The survey response rate was slightly less than 50%. We were also unable to link individual survey responses to specific teams. Lastly, participants completed all survey questions only after the datathon concluded, relying on possibly imperfect recall of their knowledge level before the event (Fig. 2). Future datathons will administer a survey in two parts (before and after the event).

Conclusions

Our findings support a new and innovative use of a datathon: to kick off a new data warehouse and familiarize users with its capabilities. The participants in the datathon were satisfied with their experience participating in the datathon, the new collaborations formed, and their increased knowledge about data resources at BCM because of the datathon. Additionally, respondents' comments helped identify gaps in data knowledge and delivery that we will address through training and improved data governance strategies. We are in the process of planning another datathon and will continue to learn from challenges and leverage opportunities to help BCM increase the use of valuable clinical data to improve quality of care.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/cts.2022.450

Acknowledgements. Dr. Zimolzak is supported by the Dan L. Duncan Institute for Clinical and Translational Research, the Houston VA Health Services Research and Development (HSR&D) Center for Innovations in Quality, Effectiveness, and Safety (CIN13-413), and the Gordon and Betty Moore Foundation (GMBF 5498 and GBMF 8838). Datathon activities were supported in part by the Baylor College of Medicine President's Circle.

Disclosures. The authors have no conflicts of interest to report.

References

1. **Badawi O, Brennan T, Celi L, *et al.*** Making big data useful for health care: a summary of the inaugural MIT critical data conference. *JMIR Medical Informatics* 2014; **2**(2): e22.
2. **Novère NL, Hucka M, Anwar N, *et al.*** Meeting report from the first meetings of the Computational Modeling in Biology Network (COMBINE). *Standards in Genomic Sciences* 2011; **5**(2): 230–242.
3. **DePasse JW, Carroll R, Ippolito A, *et al.*** Less noise, more hacking: how to deploy principles from MIT's hacking medicine to accelerate health care. *International Journal of Technology Assessment in Health Care* 2014; **30**(3): 260–264.
4. **Barash CI, Elliston K, Potenzone R.** tranSMART Foundation Datathon 1.0: the cross neurodegenerative diseases challenge. *Applied & Translational Genomics* 2015; **6**: 42–44.
5. **Harris PA, Taylor R, Thielke R, *et al.*** Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics* 2009; **42**(2): 377–381.
6. **Murphy SN, Weber G, Mendis M, *et al.*** Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association* 2010; **17**(2): 124–130.
7. **MIT Critical Data**. *Secondary Analysis of Electronic Health Records*. Cham, Switzerland: Springer, 2016, pp. 115–120, 143–144.
8. **Fillmore N, Do N, Brophy M, *et al.*** Interactive machine learning for laboratory data integration. *Studies in Health Technology and Informatics* 2019; **264**: 133–137.
9. **Friedman CP.** What informatics is and isn't. *Journal of the American Medical Informatics Association* 2012; **20**(2): 224–226.

https://doi.org/10.1017/cts.2022.450 Published online by Cambridge University Press