JAMIA Open, 1(2), 2018, 166–171
doi: 10.1093/jamiaopen/ooy023
Advance Access Publication Date: 10 July 2018
Case Report

Case Report

A Curated Cancer Clinical Outcomes Database (C3OD) for accelerating patient recruitment in cancer clinical trials

Dinesh Pal Mudaranthakam,[1,2,†] Jeffrey Thompson,[1,2,†] Jinxiang Hu,[1,2] Dong Pei,[1] Shanthan Reddy Chintala,[1] Michele Park,[2] Brooke L. Fridley,[3] Byron Gajewski,[1,2] Devin C. Koestler,[1,2,†] and Matthew S. Mayo[1,2,†]

[1]Department of Biostatistics, University of Kansas Medical Center, Kansas City, Kansas, USA, [2]University of Kansas Cancer Center, Kansas City, Kansas, USA and [3]Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, Florida, USA

[†]These authors contributed equally to this article.
Corresponding Author: Dinesh Pal Mudaranthakam, MS, Department of Biostatistics, University of Kansas Medical Center, 3901 Rainbow Blvd., Kansas City, KS 66160 (dmudaranthakam@kumc.edu).

Received 17 January 2018; Revised 29 April 2018; Editorial Decision 29 May 2018; Accepted 29 May 2018

ABSTRACT

Data used to determine patient eligibility for cancer clinical trials often come from disparate sources that are typically maintained by different groups within an institution, use differing technologies, and are stored in different formats. Collecting data and resolving inconsistencies across sources increase the time it takes to screen eligible patients, potentially delaying study completion. To address these challenges, the Biostatistics and Informatics Shared Resource at The University of Kansas Cancer Center developed the Curated Cancer Clinical Outcomes Database (C3OD). C3OD merges data from the electronic medical record, tumor registry, bio-specimen and data registry, and allows querying through a single unified platform. By centralizing access and maintaining appropriate controls, C3OD allows researchers to more rapidly obtain detailed information about each patient in order to accelerate eligibility screening. This case report describes the design of this informatics platform as well as initial assessments of its reliability and usability.

Key words: clinical research informatics, automated eligibility screening, patient recruitment

INTRODUCTION

Among the many issues faced by investigators in the design and execution of clinical trials, patient recruitment is recognized as one of the foremost challenges.[1–3] The recruitment process, particularly patient screening, is especially challenging and time consuming for oncology studies.[4] Determining eligibility often involves leveraging data from multiple different sources; for example, lab metrics criteria, such as prostate specific antigen and/or absolute neutrophil count, treatment history, and previous medications. Although some of this information is contained within a patient's electronic medical record (EMR), it might not be the most up to date information and may contain discrepancies that slow or hinder patient recruitment. In many cases, EMR data alone is insufficient for screening patients for cancer clinical trials.[5,6] Inefficient patient recruitment lengthens the study duration, increases cost,[7,8] and can lead to early termination of a trial.[9] For these reasons, innovative informatics solutions that accelerate patient screening and recruitment are critical.

Eligibility requirements for a clinical trial are specified in terms of inclusion and exclusion (IE) criteria, which represent detailed descriptions of characteristics that patients must or must not meet in order to participate in the trial. Despite the now widely recognized advantages of using EMR information to expedite trial recruitment, eligibility screening is still conducted manually in most instances.[10–12] Manual screening is a tedious and time-consuming process that results in considerable financial burden for an institution.[13] One obstacle standing in the way of automated screening for cancer clinical trials is that inclusion/exclusion criteria are often contained in disparate data sources. These data sources are typically maintained by different groups within an institution, use different technologies, and are stored in different formats. For example, at our institution, The University of Kansas Cancer Center (KUCC), the pathology and pharmacy department store information using separate software applications outside of the primary hospital EMR system. In order to identify a list of eligible participants, one often needs to jointly review multiple modules within the EMR, tumor registry, bio-specimen information, and patient surveys. Inconsistencies across data sources increase the time it takes to screen eligible patients, potentially delaying study completion. Platforms that enable querying across multiple disparate data sources have the potential to greatly expedite patient identification and improve the efficiency of clinical trials.

Although recent years have seen significant innovation aimed at improving the efficiency of patient screening,[10,14–19] most previous systems have not been specifically tailored to cancer patients, and for those that have, such focused systems have seen some of the greatest success in adoption.[14] To automate and improve patient screening for clinical trials conducted at KUCC, we developed the Curated Cancer Clinical Outcomes Database (C3OD). Electronically discovered information, such as patient history, physician encounters, demographics, medication logs, diagnosis from the tumor registry, and bio-specimen data are curated directly from the source system into C3OD repository. The repository currently holds data from two disease working groups, breast and prostate cancer. In this case report, we describe the approach taken to build the C3OD database, and examine its reliability and usability.

MATERIALS AND METHODS

Data sources

The C3OD database is populated with data that are extracted from the tumor registry with 'NAACCR 16C' and EMR, as represented in Figure 1A. The tumor registry is typically 6–8 months behind as the abstraction and diagnosis confirmation is a time consuming and laborious process. Raw data from these sources are curated, which makes it easy for researchers to execute the query using the user interface. The tumor registry contains information on tumor anatomic site, histology, and other disease characteristics, whereas the EMR system contains information on patient demographics, family history, diagnosis, and comorbidities. Based on the curation level and source, data were classified into five different class variables (Figure 1B) with most of the core variables being consistent across different diseases. As a first step, C3OD was populated with Class I variables.

C3OD workflow

C3OD is composed of two individual databases, identified and de-Identified databases. These two databases are populated in three stages. This arrangement facilitates data pre-processing and unification, and serves to insulate protected health information (PHI) from unauthorized users.

The extraction, transformation, and load process is a semi-automated process carried out by the database administrator, and once populated, it undergoes validation by the application administrator. C3OD is not intended to be real-time; it will always lag from the EMR. As the process to clean and reconcile the different data sources is lengthy (∼2 weeks), C3OD is updated once every month.

Raw data are extracted from tumor registry and the EMR. This data arrives as a flat file on the C3OD server, which is then curated using a Statistical Analysis Software (SAS) script to align the variable names and map the data elements to allow for more user-friendly filtering and querying over the data. For example, gender is often coded as one for female, two for male, and three for unknown. However, this coding needs to be processed into a user-friendly format to allow the end user to query using the actual labels (ie, male, female, unknown) versus the coding they represent (ie, 1, 2, and 3). Similarly, the tumor registry fields are reformatted to be easily searched by grouping the type of variable. For example, all the demographic variables are grouped together, as are all the treatment variables. Variables used from tumor registry to populate C3OD can be found in Supplementary File S1. The tumor registry started abstracting data on cancer patients in 2004, and similarly, the University of Kansas Health System implemented its EMR system in the early 2000's. Because of the high volume of patient information housed in the EMR system, C3OD is populated only with patients who have specific diagnoses. The initial data feed includes only breast cancer and prostate cancer; however, this will be sequentially expanded to include other cancer types. In addition, all patients from the tumor registry are populated into the C3OD system. The medical record number (MRN) is the key to link the patient information from different systems, including the manually abstracted data. When there are discrepancies in overlapping fields like race, date of birth, and sex, the number of times a certain value appears in the system, along with where the value originated from, are considered before reaching a decision on what value to retain (some systems are weighted more than others depending on the field). Once the data are cleaned, the patient MRN is substituted with a random value. This anonymized version of the database (de-identified database) is the instance that is queried by the end user. Every time the data within C3OD is changed, a new version number is assigned, and all versions are archived to enable reproducible research.

The interactive nature of the user interface is what makes C3OD unique compare to the other query tools. The web application is deployed on a Tomcat web server, which is housed under a web application server. The web application server connects to the database that is housed under the de-identified database server to execute queries. Users can query using the web application server to determine if an adequate patient population exists for their study, but they cannot access protected data until the system confirms they have appropriate Institutional Review Board (IRB) and Data Access Committee (DAC) approval. An overview of this process is shown in Figure 1A.

Investigator initiated trials

Investigator Initiated Trials (IIT) are conducted under the direction of a clinical investigator who is often a physician at an academic institution. To examine the reliability of C3OD, seven recently concluded IITs conducted at KUCC were randomly selected out of 50 ongoing and recently completed trials. These seven studies had different primary cancer sites as their research focus, and included breast ($n = 3$), non-Hodgkin lymphoma ($n = 3$), and multiple myeloma ($n = 1$). The inclusion/exclusion criteria for each of the seven selected studies is given in Supplementary File S2.

C3OD reliability assessment

For each of the seven selected studies, a list of patient MRNs were retrieved from the C3OD system using the inclusion/exclusion criteria. To assess its reliability, the list generated from C3OD was

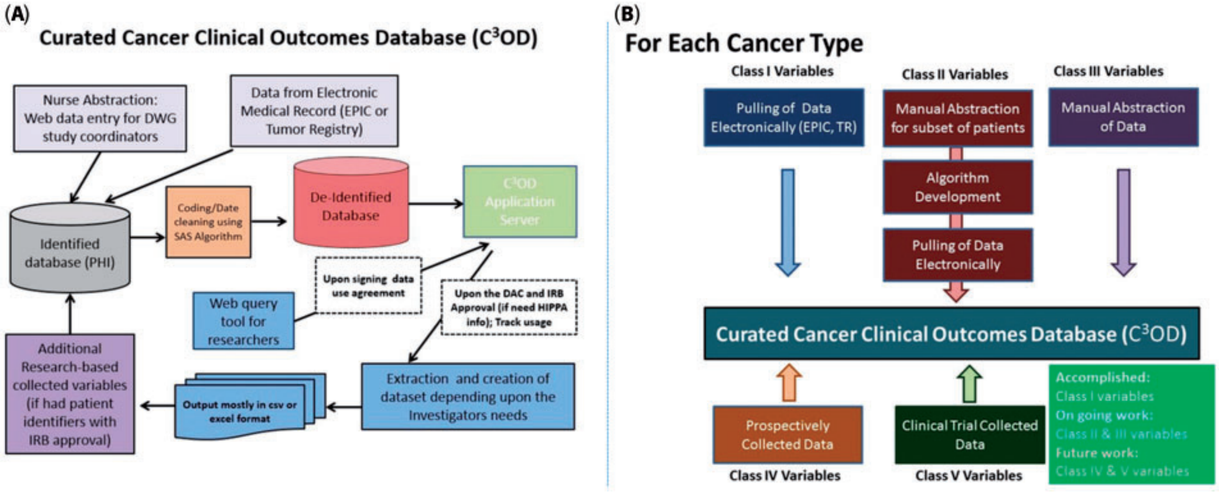
(A)
Curated Cancer Clinical Outcomes Database (C^{3}OD)
Nurse Abstraction: Web data entry for DWG study coordinators
Data from Electronic Medical Record (EPIC or Tumor Registry)
Identified database (PHI)
Coding/Date cleaning using SAS Algorithm
De-Identified Database
C^{3}OD Application Server
Upon signing data use agreement
Upon the DAC and IRB Approval (if need HIPPA info); Track usage
Web query tool for researchers
Additional Research-based collected variables (if had patient identifiers with IRB approval)
Output mostly in csv or excel format
Extraction and creation of dataset depending upon the Investigators needs
(B)
For Each Cancer Type
Class I Variables
Class II Variables
Class III Variables
Pulling of Data Electronically (EPIC, TR)
Manual Abstraction for subset of patients
Algorithm Development
Pulling of Data Electronically
Manual Abstraction of Data
Curated Cancer Clinical Outcomes Database (C^{3}OD)
Prospectively Collected Data
Clinical Trial Collected Data
Class IV Variables
Class V Variables
Accomplished:
Class I variables
On going work:
Class II & III variables
Future work:
Class IV & V variables

Figure 1. (A) C3OD architecture. The C3OD database consists of two separate databases. The identified database that contains raw data from nurse abstraction, data from electronic medical record, and research-based collected variables. The de-Identified database contains data that has been cleaned by algorithm. The application server houses the webserver where the GUI is hosted. Using this GUI, investigators can query for potential participants. (B) Variable types from different data sources that feed into C3OD.

then compared with the MRN of patients that were actually enrolled in each study. The informatician who executed the queries in C3OD was blinded to the list of patients who were actually enrolled in the seven considered studies. For each study, the steps below were followed. The informatics specialist:

1. Selected a clinical trial and identified its inclusion/exclusion criteria.
2. Worked with a clinician to translate the inclusion/exclusion criteria into database friendly criteria. For example, the inclusion criteria of "Performance status of two or better" can be translated into "Quality of survival = 0, 1, 2, 8 or 9"
3. Submitted queries to the C3OD database based on inclusion/exclusion criteria.
4. Saved the patient count after each query. This is used to show how inclusion/exclusion criteria influence cohort size.

The final cumulative patient-count satisfying the inclusion/exclusion criteria represents the potential cohort size for the study.

C30D usability assessment

To assess the usability of the C3OD, we surveyed 10 clinical research coordinators about their experiences using the C3OD. The purpose of the survey was to determine user satisfaction, ease of use, effectiveness, and efficiency of C3OD. The survey (Supplementary File S3) was adapted from the Computer System Usability Questionnaire.[20] Coordinators were asked to complete the survey online and were given a month to submit their responses. There were 21 questions in the survey, 19 of them were likert-scale-type questions with responses ranging from strongly agree (1) to strongly disagree (7). Not applicable (NA) was included in each question as an option (8). Questions 20 and 21 were open-ended questions asking the users to list the most negative and positive aspects of C3OD.

RESULTS

Of the three breast cancer studies that were selected for C3OD system validation, execution of the inclusion/exclusion criteria for study NCT00491816 were recorded in mp4 format to illustrate user-interaction with C3OD (Supplementary File S4). NCT00491816 is a *Phase II Trial of neoadjuvant erlotinib plus chemotherapy for treatment of ER negative, PgR negative and HER-2 negative primary breast cancer*. Users would begin by logging into C3OD via the web user interface. Next, users would sequentially select the corresponding variables and their appropriate values based on the inclusion/exclusion criteria for the study of interest. Execution of each inclusion/exclusion criterion for NCT00491816 (Supplementary File S2) and the resulting cohort size is depicted in Figure 2. At each step when an inclusion/exclusion criterion is queried, the potential cohort size remains the same or gets smaller, narrowing the selection only to the patients that are qualified.

In order assess the reliability of C3OD, a blinded informatician used the inclusion/exclusion criteria from the seven selected clinical trials to query in C3OD, resulting in a list of eligible patients' MRN. Table 1 contains the results of our assessment. Across all seven studies considered here, C3OD successfully retrieved the MRN of patients who were actually recruited and enrolled in each study, with most retrieval rates over 80%. Upon further examination, we discovered that for cases in which C3OD did not retrieve the MRN of a recruited patient (studies NCT01611090 and NCT00491816), the PI had granted an eligibility waiver for those enrolled that did not meet all inclusion/exclusion criteria.

Four of the 10 clinical research coordinators responded to the C3OD usability survey. Responses were mixed among those that responded, with half of coordinators reporting positive experiences and half expressing neutral or unfavorable experiences using C3OD (Figure 3). Two of the four coordinators agreed on all questions that the C3OD was satisfactory, organized, easy to use, efficient and effective, gives clear error message, and interface was pleasant. One coordinator was neutral or disagreed on most of the questions mentioned above, and one coordinator disagreed on most questions but thought the C3OD information was organized and liked the interface. Coordinators pointed out in the open-ended question "C3OD was useful and easy to export to excel forms".

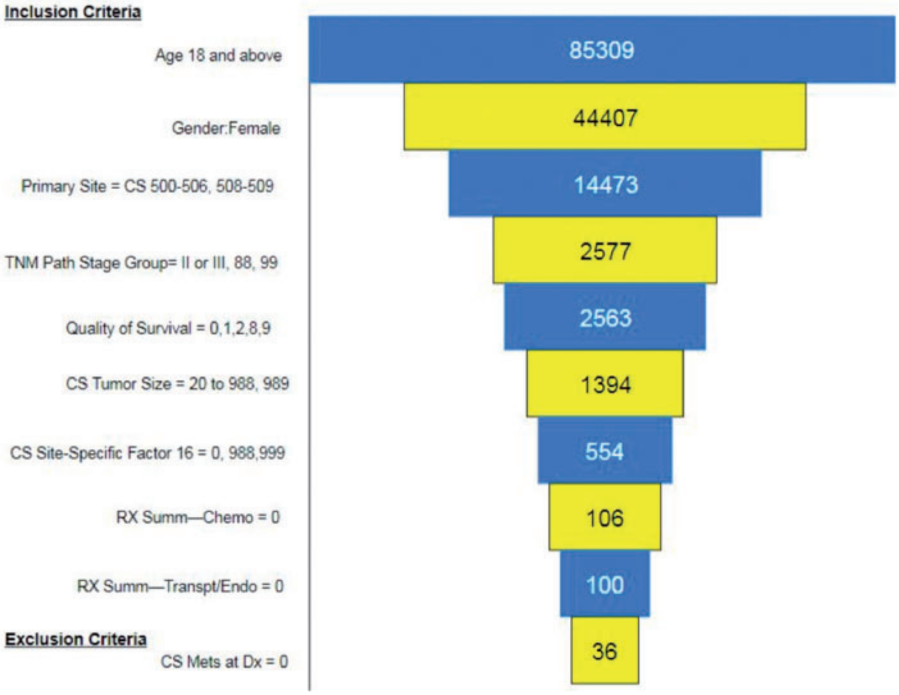

Figure 2. C3OD-derived cohort sizes based on sequential queries of the inclusion/exclusion criteria for study NCT00491816.

Table 1. Patient-recruitment reliability assessment of C3OD

Trial	Actual number of patients enrolled	Number based on C3OD screening	Percentage of patients identified by C3OD among those actually enrolled in the selected trials (%)
NCT00433511	15	19	100
NCT02595320	36	42	100
NCT00491816	32	39	81
NCT02136134	3	3	100
NCT01974440	3	5	100
NCT01779791	4	14	100
NCT01611090	5	8	80

DISCUSSION

C3OD is an innovative new tool enabling rapid identification of potential participants for clinical trials, protecting patient information, and connecting investigators with the best available data for their studies. By creating a single interface for accessing many disparate data sources, we have taken an important step towards simplifying the process of initiating and completing clinical trials. C3OD provides multiple levels of access, allowing investigators to determine study feasibility before submitting a protocol. In situations when the cohort size is not sufficient, querying C3OD for additional populations could assist researchers in adjusting their inclusion/exclusion criteria, thereby achieving the desired sample size. Once a study is funded and the study team has IRB approval to move forward with patient enrollment, the study coordinator will request PHI information from eligible patients (patients opt IN for participating in future clinical trials). Using this list, the coordinator would then contact the patients and recruit them for final screening and study enrollment.

Based on our experience, most of the eligibility criteria for cancer clinical trials are contained in electronic sources. However, some information is stored under different systems, including: pathology software, lab software, etc., which are silo systems and not part of the EMR. In addition, at times, the data is in free-text format, which is not readily amenable for analytics. A major challenge faced by recruiters at KUCC involves the identification of patients with a certain tumor subtype; for example, breast cancer hormone receptor status. This information is typically contained in pathology reports in free-text format. In order to curate the data, KUCC has decided to have nurse abstractors manually retrieve this information from patient records in order to streamline and structure the information in a more user-friendly format (Class III variables in Figure 1B). In addition to augmenting C3OD to incorporate Class III variables, we are also in the process of developing a natural language processing algorithm with Python and R that parses free-text in pathology reports (Class II variables, Figure 1B). Including Class II and III variables extends the coverage of inclusion/exclusion criteria, and in doing so, decreases the need for manual validation.

Naturally, a number of challenges were encountered in the development of C3OD. Chief among them was the need to convince clinicians and other stakeholders of the benefit to sharing EMR data with researchers. Therefore, in collaboration with these stakeholders, we developed a data share agreement that clearly delineates how data will be used and the responsibilities for all those involved. Another challenge we faced was the development of a sophisticated security model where every potential user is vetted through the appropriate channels before they can begin using the query tool. We took advantage of the need to provide secure access, to better uphold our responsibilities with the data, by automating the process by which we check that potential users of the system have the required responsible conduct of research and other certifications (eg CITI), and that they have appropriate appointments (eg KUCC). Finally, in order to curate a list of eligibility screening variables, our

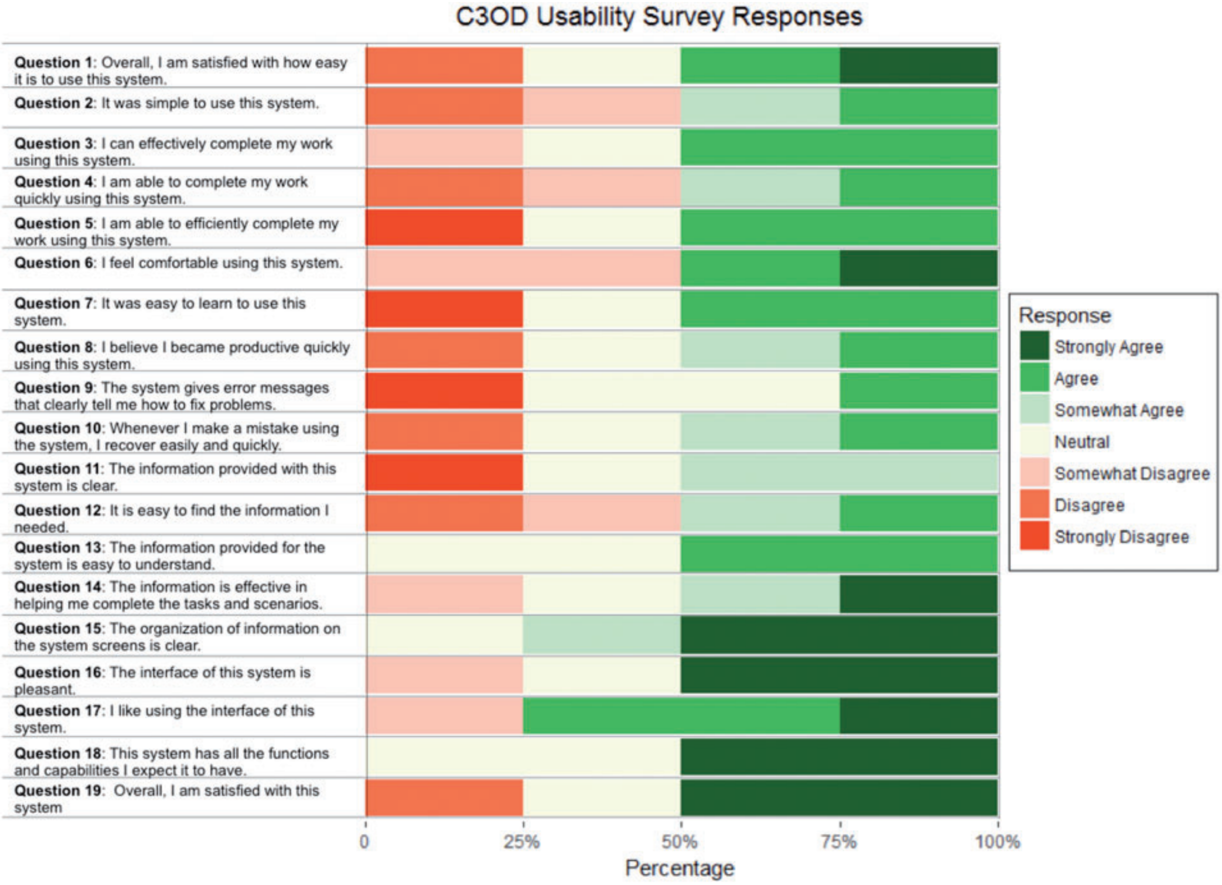

Figure 3. Results of the C3OD usability survey. Responses are based on the $n = 4$ research co-ordinators that responded to the C3OD usability survey.

informatics team has been collaborating with cancer disease physicians to identify the most relevant variables for different cancer types.

As with any platform, C3OD has certain strengths and limitations. At the present moment, the C3OD repository is populated with data from only two disease working groups (eg, breast and prostate cancer). However, work is already underway to expand the repository to include other disease working groups. Another potential limitation of the current implementation of C3OD is its update lag time, as previously described. While the update lag time is a limiting factor, this is counterbalanced by the fact that monthly feeds (as opposed to more frequent feeds) avoids overburdening the primary hospital system. Finally, an outstanding area for improvement involves augmenting the existing query interface to make it more user-friendly. Despite these limitations, C3OD has a number of strengths, including: expedited patient screening, accuracy in identifying eligible patients for outcomes-based research, and a system that automatically checks to ensure that users have the required trainings and authorizations.

In summary, we feel that this case report serves two primary purposes. First, by demonstrating that C3OD can be used to reliably screen patients for cancer clinical trials at a fraction of time it would it take using the legacy method, we provide further evidence that it is worth the effort for institutions to implement similar systems. In addition, we described some of the key challenges that we faced in the development of C3OD along with its limitations, which may be of value for institutions that are currently developing (or planning to develop) systems that combine and allow querying over multiple disparate data sources. While several challenges remain in refining and improving C3OD, its high reliability, accuracy, and security position it as a viable tool for accelerating patient recruitment for cancer clinical trials.

FUNDING

This study was supported by the National Cancer Institute (NCI) Cancer Center Support Grant P30 CA168524 and used the Biostatistics and Informatics Shared Resource (BISR).

CONTRIBUTIONS

D.P.M. is the architect and developer of the curated clinical outcomes (C3OD) database. D.P.M and J.T. oversaw all aspects of drafting, revision, and final approval of the manuscript. J.H. contributed to the C3OD usability survey and participated in the drafting and revising of the manuscript. D.P. participated in the drafting and revising of the manuscript. S.R.C assisted in the development and reliability testing of C30D, participated in the drafting, and revising the manuscript. M.P. co-ordinated the usability testing of C30D and participated in drafting and revising of the manuscript. B.L.F. made significant intellectual contributions in regards to the development and design of C30D and participated in the drafting and revising of the manuscript. B.G. participated in the drafting and revising of the manuscript. D.C.K helped organize the manuscript preparation efforts and oversaw all aspects in the drafting and revising of the manuscript. M.S.M. provided advice on design, development, and operational issues related to C3OD implementation and participated in the drafting and revising of the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association Open* online.

Conflict of interest statement. None declared.

ACKNOWLEDGEMENTS

We would like to extend our gratitude to Drs. Roy Jensen and Andrew Godwin for the indispensable guidance and advice they provided on the design and development of C3OD. Additionally, we thank Velos Inc. for contributing the software platform underlying C3OD.

REFERENCES

1. Ferland D, Fortin PR. Recruitment strategies in superiority trials in SLE: lessons from the study of methotrexate in lupus erythematosus (SMILE). *Lupus* 1999; 8 (8): 606–11.
2. Lovato LC, Hill K, Hertert S, *et al*. Recruitment for controlled clinical trials: literature summary and annotated bibliography. *Control Clin Trials* 1997; 18 (4): 328–52.
3. Collins JF, Williford WO, Weiss DG, *et al*. Planning patient recruitment: fantasy and reality. *Stat Med* 1984; 3: 435–43.
4. Massett HA, Mishkin G, Rubinstein L, *et al*. Challenges facing early phase trials sponsored by the National Cancer Institute: an analysis of corrective action plans to improve accrual. *Clin Cancer Res* 2016; 22 (22): 5408–16.
5. Madden JM, Lakoma MD, Rusinak D, *et al*. Missing clinical and behavioral health data in a large electronic health record (EHR) system. *J Am Med Inform Assoc* 2016; 23 (6): 1143–9.
6. Kopcke F, Trinczek B, Majeed RW, *et al*. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. *BMC Med Inform Decis Mak* 2013; 13: 37.
7. Gul RB, Ali PA. Clinical trials: the challenge of recruitment and retention of participants. *J Clin Nurs* 2010; 19 (1–2): 227–33.
8. Hunninghake DB, Darby CA, Probstfield JL. Recruitment experience in clinical trials: literature summary and annotated bibliography. *Control Clin Trials* 1987; 8 (4 Suppl): 6S–30S.
9. Prescott RJ, Counsell CE, Gillespie WJ, *et al*. Factors that limit the quality, number and progress of randomised controlled trials. *Health Technol Assess* 1999; 3 (20): 1–143.
10. Thadani SR, Weng C, Bigger JT, *et al*. Electronic screening improves efficiency in clinical trial recruitment. *J Am Med Inform Assoc* 2009; 16: 869–73.
11. Embi PJ, Payne PR. Clinical research informatics: challenges, opportunities and definition for an emerging domain. *J Am Med Inform Assoc* 2009; 16: 316–27.
12. Embi PJ, Jain A, Harris CM. Physicians' perceptions of an electronic health record-based clinical trial alert approach to subject recruitment: a survey. *BMC Med Inform Decis Mak* 2008; 8: 13.
13. Penberthy LT, Dahman BA, Petkov VI, *et al*. Effort required in eligibility screening for clinical trials. *J Oncol Pract* 2012; 8 (6): 365–70.
14. Eubank MH, Hyman DM, Kanakamedala AD, *et al*. Automated eligibility screening and monitoring for genotype-driven precision oncology trials. *J Am Med Inform Assoc* 2016; 23 (4): 777–81.
15. Shivade C, Hebert C, Lopetegui M, *et al*. Textual inference for eligibility criteria resolution in clinical trials. *J Biomed Inform* 2015; 58 Suppl: S211–8.
16. Ni Y, Wright J, Perentesis J, *et al*. Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients. *BMC Med Inform Decis Mak* 2015; 15: 28. :
17. Ni Y, Kennebeck S, Dexheimer JW, *et al*. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. *J Am Med Inform Assoc* 2015; 22 (1): 166–78.
18. Weng C, Batres C, Borda T, *et al*. A real-time screening alert improves patient recruitment efficiency. *AMIA Annu Symp Proc* 2011; 2011: 1489–98.
19. Embi PJ, Jain A, Clark J, *et al*. Effect of a clinical trial alert system on physician participation in trial recruitment. *Arch Intern Med* 2005; 165 (19): 2272–7.
20. Lewis JR. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *Int J Human Comput Interact* 1995; 7 (1): 57–78.