

Design-phase prediction of potential cancer clinical trial accrual success using a research data mart

Jack W London,^{1,2} Luanne Balestrucci,³ Devjani Chatterjee,¹ Tingting Zhan⁴

¹Kimmel Cancer Center, Thomas Jefferson University, Philadelphia, Pennsylvania, USA

²Department of Cancer Biology, Thomas Jefferson University, Philadelphia, Pennsylvania, USA

³Jefferson Graduate School of Biomedical Sciences, Thomas Jefferson University, Philadelphia, Pennsylvania, USA

⁴Department of Pharmacology & Experimental Therapeutics, Thomas Jefferson University, Philadelphia, Pennsylvania, USA

Correspondence to

Dr Jack London, Kimmel Cancer Center, Thomas Jefferson University, 233 S. 10th Street, Room 808 BLSB, Philadelphia, PA 19107, USA; Jack.london@jefferson.edu

Received 27 March 2013

Revised 22 May 2013

Accepted 28 June 2013

Published Online First

14 July 2013

ABSTRACT

Background Many cancer interventional clinical trials are not completed because the required number of eligible patients are not enrolled.

Objective To assess the value of using a research data mart (RDM) during the design of cancer clinical trials as a predictor of potential patient accrual, so that less trials fail to meet enrollment requirements.

Materials and methods The eligibility criteria for 90 interventional cancer trials were translated into i2b2 RDM queries and cohort sizes obtained for the 2 years prior to the trial initiation. These RDM cohort numbers were compared to the trial accrual requirements, generating predictions of accrual success. These predictions were then compared to the actual accrual performance to evaluate the ability of this methodology to predict the trials' likelihood of enrolling sufficient patients.

Results Our methodology predicted successful accrual (specificity) with 0.969 (=31/32 trials) accuracy (95% CI 0.908 to 1) and predicted failed accrual (sensitivity) with 0.397 (=23/58 trials) accuracy (95% CI 0.271 to 0.522). The positive predictive value, or precision rate, is 0.958 (=23/24) (95% CI 0.878 to 1).

Discussion A prediction of 'failed accrual' by this methodology is very reliable, whereas a prediction of accrual success is less so, as causes of accrual failure other than an insufficient eligible patient pool are not considered.

Conclusions The application of this methodology to cancer clinical design would significantly improve cancer clinical research by reducing the costly efforts expended initiating trials that predictably will fail to meet accrual requirements.

BACKGROUND AND SIGNIFICANCE

Translational research is the process of transforming basic science laboratory discoveries into advancements in the clinical treatment of diseases. Interventional clinical trials research is the step in this process which evaluates the safety and efficacy of proposed new treatments. Clinical trials are scientifically organized studies with controlled variables that allow meaningful results. The proper characterization of the individuals who will be accepted as participants in the study is crucial to assuring that the trial outcomes can be clearly understood. The 'eligibility rules' for participants in these trials are therefore often extensive and sometimes complex. Statistical analysis of the study design is employed to calculate the number of eligible participants needed for the trial results to be considered valid. This required number of participants is the trial's target accrual, or minimum accrual goal.

As important as interventional clinical trials are in translational research, these studies may never accrue the statistically required number of participants to complete the study's research plan. An Institute of Medicine (IOM) report on cancer cooperative group trials found that 40% were never completed because of failure to achieve minimum accrual goals.¹ The IOM report states, 'The ultimate inefficiency is a clinical trial that is never completed because of insufficient patient accrual, and this happens far too often.' These non-accruing trials are often kept open for many months before closure, consuming personnel resources in their setup and operation at a significant cost to institutions, without providing any return in definitive research findings. Furthermore, while many of these trials register zero patients, others accrue some patients, resulting in thousands of patients nationwide who are recruited to unproductive research studies.² A number of studies have investigated barriers to clinical trial accrual, and reported various physician-related and patient-related obstacles.³⁻⁹ Physician barriers cited include inadequate reimbursement, lack of support resources, the irrelevance of available studies to the practice population, and treatment preferences. Patient barriers cited include concerns and uncertainty about treatments, treatment preferences, unavailability of an appropriate trial, lack of awareness of trials, and transportation and other logistical constraints. These cited studies all have focused on accrual issues occurring *after* trial activation. Recently, however, Schroen *et al*¹⁰ have evaluated cooperative group trial processes occurring *prior* to trial activation that impact subsequent accrual for that trial. They concluded that trials designed to more accurately reflect the interests of participating physicians and potential patients, and that develop more accurate strategies for accrual prediction, would have increased success in trial accrual. A survey of trial lead statisticians and study chairs concluded that cooperative group trial accrual predictions were usually based upon the group's previous experience with a certain disease type, stage, or treatment, or estimates by site principal investigators.¹¹ This paper concluded that 'better accrual prediction methods are critically needed.'

The growing deployment of clinical data warehouses and associated research data marts (RDMs) at academic medical centers provides the resources for defining cohorts of the institution's patient population having certain characteristics.¹² These data repositories have been mined to identify patients potentially eligible for already opened clinical trials.^{13 14} However, this project seeks to use

To cite: London JW, Balestrucci L, Chatterjee D, *et al.* *J Am Med Inform Assoc* 2013;**20**:e260–e266.

these resources to assess whether a clinical trial *being developed* can reasonably be expected to reach its target accrual goal, given the proposed eligibility criteria.

OBJECTIVE

To significantly increase interventional clinical trial research productivity, and reduce wasted expenditure of translational research resources and participant engagement, it is necessary to minimize the likelihood that interventional clinical trials will fail to complete because of an insufficient pool of eligible participants at the performance site. This paper describes the evaluation of methodology designed to provide clinical researchers with estimates of potential trial accrual by applying the proposed trial eligibility criteria to recent institutional patient populations defined by RDMs that are derived from de-identified clinical warehouse data. By indicating *before* a trial is opened whether successful accrual is likely, based on recent patient contact, this methodology would address the critical clinical trial research problem of wasted effort and resources when trials fail to reach completion because of insufficient participant accrual.

MATERIAL AND METHODS

This study was carried out at the Kimmel Cancer Center (KCC) at Thomas Jefferson University in Philadelphia. This National Cancer Institute-designated cancer center has approximately 150 cancer trials open for accrual at any given time. KCC clinical investigators have authored 176 investigator-initiated trials in the past 6 years, and 136 cancer cooperative group trials were opened at the center. KCC has access to patients from the Jefferson Health System, the largest health system in the Philadelphia metropolitan area, headed by the KCC clinical institution, Thomas Jefferson University Hospital (TJUH).

Jefferson i2b2 research data mart

The overall objective of this study was to evaluate whether accrual for proposed cancer clinical trials could be predicted by performing cohort queries that are based on the trial's eligibility criteria on recent patient data in Jefferson's RDM, created from de-identified integrated hospital clinical, tumor registry, and specimen data. The RDM has an i2b2 framework and provides investigators with a query tool for cohort identification and hypothesis generation. The 'informatics for integrating biology and the bedside,' or i2b2, framework was developed at the National Institutes of Health-funded National Center for Biomedical Computing Informatics based at Partners HealthCare System in Boston.¹⁵⁻¹⁷ In the Jefferson i2b2 RDM deployment, hospital, specimen, and cancer registry data are integrated in this de-identified RDM, with data refreshes occurring monthly.¹⁸⁻²⁰ Currently, the database contains over 28 million observations on about 350 000 patients, with more than 800 000 specimens. Specimen and cancer registry data go back to 1988, while other clinical data are from 2008 to the present. The clinical data are extracted from the TJUH clinical data warehouse and include observations on patient demographics, diagnoses, procedures, laboratory test results, medications, hospitalizations, and vital signs. The specimen data are extracted from caTissue and include anatomic origin, class (eg, fluid, solid tissue), type (eg, frozen, paraffin-embedded), and when appropriate, pathological status (eg, malignant, normal adjacent tissue). The tumor registry data are extracted from a commercial system and include primary cancer diagnosis, tumor histology, stage, grade, and site specific factors (eg, breast cancer patient ER, PR, HER2 receptor status, prostate patient Gleason score), and the patient's cancer recurrence, survival, and treatment data.

i2b2 uses a readily extensible open-source data model to integrate disparate data sources. The flexibility of the i2b2 framework stems largely from its use of metadata to describe and catalog the concepts contained within the database. The data model itself is described as a 'star schema' with central fact tables containing measures and quantities linked to surrounding dimension tables containing descriptors or concepts.²¹ Querying the i2b2 database will yield cohorts of individuals having 'observation-facts' corresponding to the concepts contained in the deployed data model. The diagram in figure 1 describes the hierarchical organization of data within the ontological groups. For clarity, this diagram does not show all the main data sets in Jefferson's i2b2 deployment, and only a few hierarchies are expanded to show higher level subcategories. Figure 2 is a screen shot of the i2b2 query web client interface with Jefferson's main data set concepts displayed (and the tumor registry data set expanded to show its main subclasses). Queries are formed by expanding the categories to the level of granularity needed (clicking the '+' icon), and 'dragging' the desired concept into one of the 'group' columns on the right. Queries yield sets of patients having 'observation-facts' corresponding to the concepts dropped in all the columns. Multiple concepts in a single group column are logically 'OR'ed, while concepts in different columns are logically 'AND'ed. Thus if the tumor registry ontology is expanded and the primary cancer diagnosis concept (resolved to 'breast cancer') is dragged into 'Group 1' and the biospecimen concept for 'frozen breast specimen' is dropped into 'Group 2,' the query would select all patients who have been diagnosed with primary breast cancer and have banked frozen breast specimens. Individuals having multiple occurrences of an observation-fact, such as four frozen breast specimens, are only included once in the resulting cohort.

Retrospective study design

To determine the ability of the RDM to predict accrual for prospective trials, we retrospectively used the RDM to obtain patient populations for 2 years prior to recent trials and compared these cohort sizes to the actual accrual observed after the trial was opened. We initially considered all 110 interventional cancer trials opened at KCC in the years 2008, 2009, and 2010, since these have been open for at least 2 years and their accrual performance could be evaluated. We constructed RDM cohort size queries corresponding to the trial eligibility criteria for the 2 years prior to each trial's opening (ie, we considered TJUH patient populations from 2007 and 2008 for trials opened in 2009). Our study set was reduced to 90 trials because important criteria for 20 trials could not be mapped to the current scope of data in our RDM (this inability to map some eligibility criteria to RDM concepts will be discussed in detail in the 'Discussion' section). We computed an annual cohort size by averaging the 2-year totals. We then compared our RDM annual cohort size for the 2 years preceding a trial's opening to the annual target goal for that trial. Since we assumed that 50% of eligible participants would enroll in a study (this assumption will be explored in the 'Discussion' section), the RDM cohort would have to be at least twice the accrual goal for a prediction of successful trial accrual. We defined a trial's actual accrual performance as successful if it accrued at least 80% of its target enrollment, since requiring achievement of 100% of the accrual goal is too rigorous, particularly for trials that are still open and given that participation rates vary year-to-year. We then assessed the predictive ability of our RDM for trial accrual, reporting the sensitivity and specificity, since we have a 'gold standard' (ie, a historical outcome) to compare with our predictions.

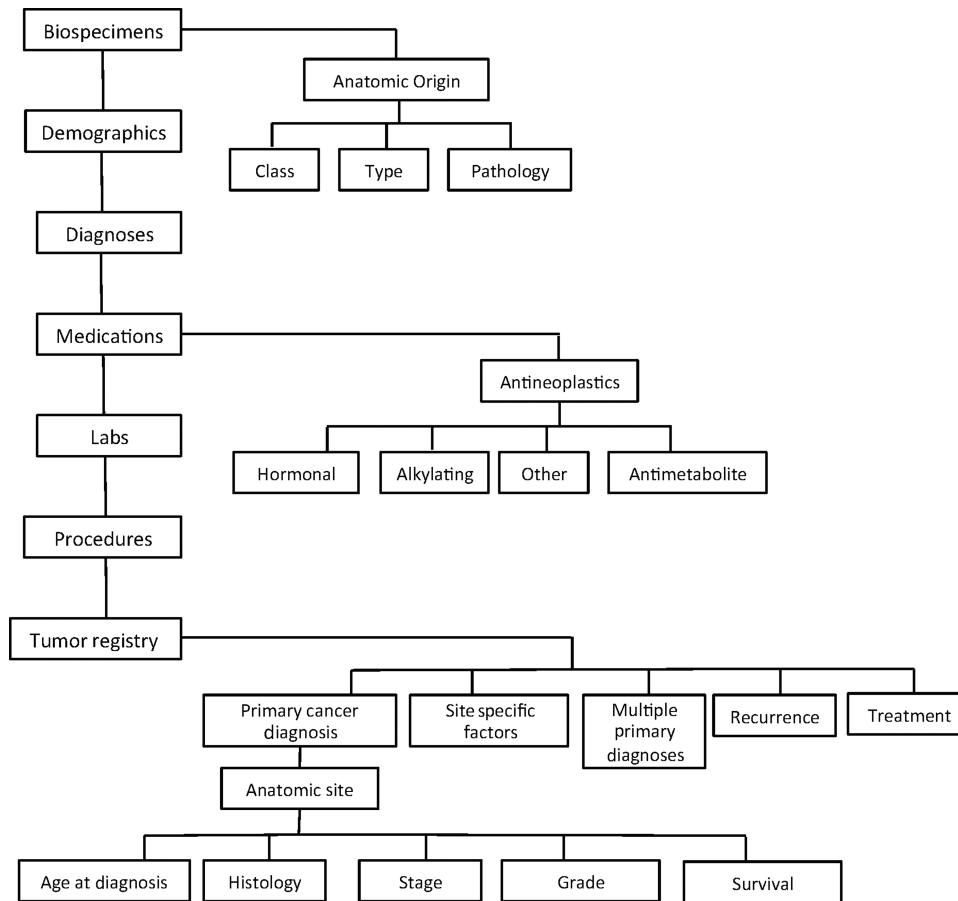


Figure 1 The hierarchical organization of data within the i2b2 ontological groups. For clarity, this diagram does not show all the main data sets in Jefferson's i2b2 deployment, and only a few hierarchies are expanded to show higher level subcategories.

Example of methodology

To demonstrate the methodology we will consider a national trial (opened at many US cancer centers) with straightforward

eligibility criteria: 'A randomized phase III comparison of standard-dose versus high-dose conformal radiotherapy with concurrent and consolidation carboplatin/paclitaxel +/-

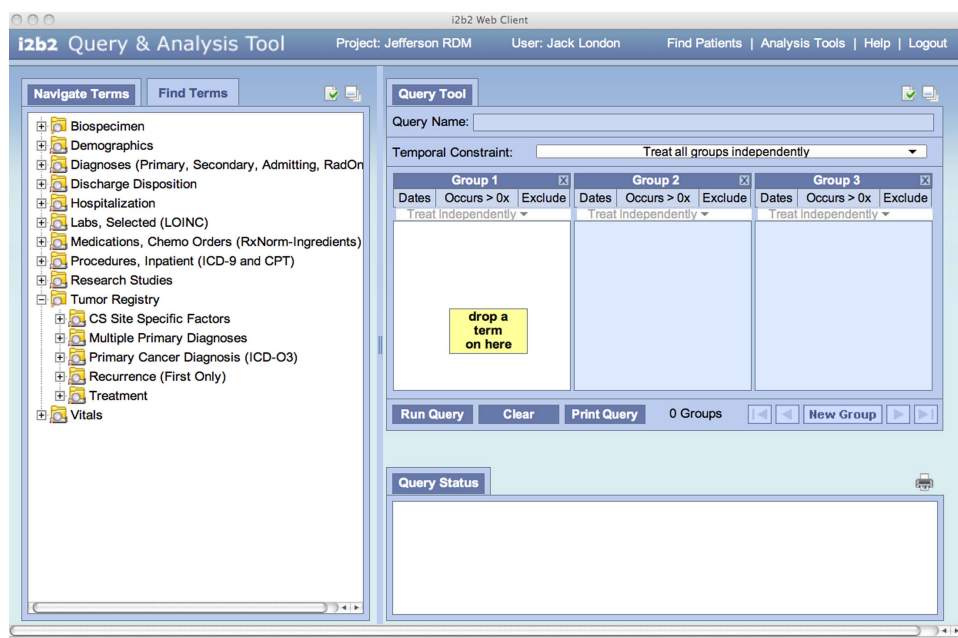


Figure 2 The i2b2 'drag and drop' query tool display, with our main ontology headings displayed on the left. The 'Tumor Registry' ontology is expanded to show its main sub-headings.

cetuximab in patients with stage IIIA/IIIB non-small cell lung cancer.' This trial is studying high-dose radiation therapy given together with cetuximab and chemotherapy to see how well it works compared with standard-dose radiation therapy and chemotherapy in treating patients with newly diagnosed stage III non-small cell lung cancer. The primary trial eligibility rules are:

1. Subject must have cytologically or histologically confirmed non-small cell lung cancer.
2. Subject must have stage IIIA or stage IIIB disease.
3. Subject must *not* have N3 disease (for non-small cell lung cancer this means no involvement of certain lymph nodes on the other side of the chest from the tumor.)

The trial design also stipulates a minimum annual target accrual of four patients, with a total target accrual of 20 patients for KCC.

Translate eligibility criteria to query and obtain RDM cohort size

To start, we select patients satisfying the first criterion, those having a primary diagnosis of non-small cell lung cancer, by navigating the i2b2 ontological trees

Tumor Registry → *Primary Cancer Diagnosis* → *Respiratory System and Intrathoracic Organs* → *Bronchus and Lung* → *Histology* → *Non-small cell carcinoma*

After dragging the non-small cell carcinoma histology into the first query clause column, we similarly navigate the lung cancer 'stage' observations and drag American Joint Committee on

Cancer (AJCC) stage to the second column, selecting values corresponding to stages 'IIIA' and 'IIIB.' Finally, we expand both the clinical and pathological TNM classifications, and drag TNM 'N3' classifications into the third column, setting the 'Exclude' button, since we want to exclude patients with N3 disease. The translation of the eligibility criteria into a query definition is now complete.

Figure 3 shows the resulting i2b2 query tool display, with clinical trial eligibility criteria translated into query parameters. The constraints in groups 1 through 3 define the desired cohort as patients having non-small cell carcinoma of the lung, stages 3A or 3B, with both clinical and pathological TNM N3 classifications excluded. Before the query is evaluated, constraints can be placed on the date range for the patient observations, and also that all criteria apply to the same tumor for a patient. For this study we set the date range to the 2 years prior to the trial initiation and obtained the total number of patients meeting the eligibility requirements. Dividing this total by 2 provided the average number of eligible patients seen in each of the 2 years before the trial opened. In this particular example, the total cohort was 22 patients, or an average of 11 patients per year.

Compare predicted accrual to actual accrual

The designation of the actual accrual performance for this trial as either 'success' or 'failed' is based on the annual target accrual defined in the trial design and the number of years the trial has been open. For this trial the annual target is four

The screenshot displays the i2b2 Query & Analysis Tool interface. On the left, a tree view shows the navigation path: CS Site Specific Factors → Multiple Primary Diagnoses → Primary Cancer Diagnosis (ICD-O3) → BLOOD, BONE MARROW, HEMATOPOIETIC → BONES, JOINTS AND ARTICULAR CARTILAGE → BRONCHUS AND LUNG → Histology → Stage, Grade, Behavior → AJCC Best Stage - 7428 → AJCC Clinical Stage - 7428 → AJCC Pathological Stage - 7428 → Behavior (benign, malignant, in situ) → Grade (differentiation) - 7428 → TNM Clinical → Clinical M (metastasis) - 7428 → Clinical N (nodes) - 7428 → Clinical T (tumor) - 7428 → TNM Pathological → Pathological M (metastasis) - 7428 → Pathological N (nodes) - 7428 → Pathological T (tumor) - 7428.

The central Query Tool area shows three groups of criteria:

- Group 1: Non-small cell carcinoma -
- Group 2: Stage = ("Stage 3A", "Stage 3B")
- Group 3: Clinical N (nodes) = ("N3")
Pathological N (nodes) = ("N3")

The criteria are connected by AND operators. The interface also shows a 'Run Query' button and a 'Query Status' section at the bottom indicating the query is finished with a patient count of 22.

Figure 3 i2b2 query tool web client with clinical trial eligibility criteria translated into query parameters. The constraints in groups 1 through 3 define the desired cohort as patients having non-small carcinoma of the lung, stages 3A or 3B, with both clinical and pathological TNM N3 classifications excluded.

patients/year, which would be 16 patients on study after the 4 years this trial has been open. (As stated above, we chose trials that have been open for at least 2 years so that their accrual performance can be evaluated.) Multiplying this target accrual by the 80% threshold factor for defining successful accrual stipulates that at least 13 patients should have been enrolled when we performed our analysis. Since in reality 18 patients had been enrolled, this trial was designated as an accrual ‘success.’

To obtain a prediction based on our RDM cohort query, we see that 22 eligible patients were found in the RDM for the 2 years prior to the initiation of this trial, or an average of 11 patients per year. The trial’s annual requirement is four patients per year. Since we arbitrarily assumed a participation rate of 50%, we would need a patient pool of eight patients per year to have four patients enroll. Since the recent average yearly patient population eligible for this trial (11 patients) exceeds the pool needed to meet the annual requirement (eight patients), this trial was predicted to be an accrual ‘success.’ For this trial, the methodology correctly predicted that the trial would successfully accrue patients. For other trials, the methodology correctly predicted accrual failure. As will be discussed below, incorrect predictions of success and failure also resulted.

RESULTS

To assess the predictive precision of our proposed project, a contingency table (table 1) was produced for the 90 trials analyzed. A trial was denoted as *potentially successful* in meeting its annual target accrual (‘Predicted success’ row in table 1) if the retrospective i2b2 cohort analysis indicated sufficient patients for the trial. A trial was denoted as *actually successful* in meeting its annual target accrual if the trial satisfactorily approached the protocol’s stated target annual accrual (‘Actual success’ column in table 1).

Our methodology has 0.969 (=31/32 trials) accuracy (95% CI 0.908 to 1) for predicting successful accrual (ie, specificity) and 0.397 (=23/58 trials) accuracy (95% CI 0.271 to 0.522) for predicting failed accrual (ie, sensitivity). The positive predictive value, or precision rate, is 0.958 (=23/24 trials) (95% CI 0.878 to 1).

DISCUSSION

Mapping trial eligibility criteria to RDM concepts

This study focused on interventional cancer clinical trials. Since ‘cancer’ can be many diseases, the eligibility criteria for these trials describe the disease focus of the trial, which can include the primary disease site (eg, lung), the histology of the tumor (eg, adenocarcinoma), and the extent of the disease—its stage (eg, T1 N2 M0, stage III), grade (eg, grade III), and behavior (eg, in situ). Additional diagnostic characteristics of eligible trial participants can include tumor cell oncogene (eg, KRAS) and cell receptor (eg, HER2) expression. Cancer trial eligibility criteria may also constrain participation based on a patient’s treatment options, such as a trial only being open to

individuals with ‘un-resectable’ tumors—those that cannot be completely removed by surgery. Thus the eligibility criteria for a trial in this study included the statement, ‘Stage IIIA or IIIB un-resectable non-small cell lung cancer—excluding patients with N3 disease,’ defining the primary disease, extent, and treatment options that a patient must have for trial participation.

‘Tumor registry’ concepts were therefore the i2b2 RDM ontological group used most for this study, with primary disease site, histology, and stage being the most commonly utilized factors. The ICD-O-3 coded tumor registry primary cancer diagnosis concepts (primary disease site, histology, and behavior) are often more precise and correspond more closely to the eligibility criteria generated by cancer clinical researchers than hospital ICD-9 diagnostic codes found in the ‘diagnosis’ ontological group containing concepts for all diseases, not just cancer. Cancer clinical trial eligibility criteria also commonly reference the AJCC disease stage classification system which are included in the i2b2 tumor registry ontology. However, other RDM i2b2 groups were accessed in mapping eligibility rules: ‘medications’ for chemotherapy drugs administered, ‘procedures’ for more precise description of surgical procedures performed, ‘labs’ for clinical laboratory test values (eg, PSA), and ‘demographics’ for race and ethnicity. Of course, applying this study’s methodology to non-cancer trials would shift the RDM focus to other groups: a clinical trial focused on diabetic patients might likely utilize the RDM ICD-9 coded ‘diagnosis’ concepts to identify patients with uncontrolled adult-onset type II diabetes mellitus and the RDM LOINC-coded ‘lab’ test concepts for hemoglobin A1c measurements.

As indicated in the ‘Retrospective study design’ section above, 20 interventional cancer trials were not included in our study because we could not map important eligibility criteria to concepts then existing in our i2b2 RDM. The most frequently encountered problem was eligibility criteria specifying observations of recurrence of cancer in anatomic sites distant from the site of the primary tumor. For example, several trials were focused on evaluating treatment of cancer that had metastasized to the liver. While the tumor registry ontology and data we had deployed could identify patients with recurrence to the liver *only*, identification of recurrence to the liver as well as to *other* sites was not possible, and therefore we could not identify *all* patients with liver metastasis. We likewise lacked the ability to identify the patients’ initial metastatic disease status, and patients with certain gene mutations (eg, FLT3 mutations in acute myeloid leukemia patients). We are addressing these deficiencies by adding metastatic disease and cancer disease-specific concepts and data to our RDM. It was also not possible to apply a constraint based on a temporal relationship (eg, neo-adjuvant chemotherapy) with our current RDM. (We plan to address these temporal issues by creating new data elements directly expressing temporal relationships, computed from the dates of the treatment data.) Eight of the 20 interventional trials not included were dropped because they were only opened at our community hospital network affiliates whose patient population data were not included in our RDM.

Predictive value and sensitivity of this methodology

Our results show that the methodology, while having an excellent positive predictive value (95.8%, 23 of the 24 trials predicted to fail, actually failed), is not good at predicting failed accrual (39.7%, only predicted 23 of the 58 failed trials). In other words, if the methodology predicts ‘failed accrual,’ then we should trust this prediction and should not proceed to open the trial with its current eligibility criteria; however, a prediction

Table 1 Contingency table comparing i2b2 accrual predictions with actual accrual success, assuming only 50% of potential participants identified by i2b2 are enrolled

Accrual requirements met?	Actual success	Actual failure
Predicted success	31	35
Predicted failure	1	23

of accrual success using this method is no guarantee that target goals will be met.

A benefit of analyzing potential trial accrual *during the protocol design phase* is that it offers an opportunity to ‘tweak’ eligibility rules when insufficient patient cohorts are found. A change in participation criteria that does not impact significantly on the scientific objectives of the trial may provide a sufficiently large potential patient pool. Once a trial has been opened, modifications to participation criteria require institutional review board (IRB)-reviewed protocol amendment. Therefore, changing the protocol design is much more efficiently done prior to initial IRB review and trial initiation. Cancer clinical trials usually undergo a two-stage review process. First, the trial is reviewed for scientific merit. Those trials receiving scientific review approval are then reviewed by the IRB for conformance with ethical and legal requirements for human subject research. It is our plan that the methodology presented here for estimating a trial’s accrual potential be performed as part of the initial scientific review process. If performed at this point in the pre-initiation phase of a clinical trial, those having little likelihood of achieving their participation requirements will either be rejected or have their design modified to improve their prospect of successful completion. Not opening the 23 trials that were correctly predicted to fail to accrue sufficient participants over the 3 years studied would have prevented the waste of about \$200 000 in trial startup costs alone, and the participation of 57 patients in studies which did not contribute to advancing science or clinical care.

Analysis of the single instance of a successfully accruing trial that our methodology incorrectly predicted would fail showed that, for that trial, our assumption of only 50% participation by eligible patients was too low: the trial principal investigator, a department chair, was very supportive of clinical research and had significant influence over faculty behavior. Furthermore, the study did not compete with any other trial, and did not pose any controversial decisions for potential participants. If we had assumed 75% participation, our methodology would have correctly predicted success. The cancer center scientific review committee members, aware of occasional extenuating circumstances affecting patient participation rates, might know to apply a higher participation rate assumption.

We further explored the effect our participation rate assumption had on our results by calculating the sensitivity and specificity corresponding to various assumed participation rates. For various participation rates, table 2 shows the specificity (ie, the probability of the methodology predicting successful accrual for a trial that actually accrues enough participants) as well as the sensitivity (ie, the probability of the methodology predicting failed accrual for a trial which actually fails to accrue sufficient patients). The corresponding receiver operating characteristic (ROC) curve, which is shown in figure 4, confirms that our methodology is very specific (it predicted successful accrual for 31 of 32 trials that actually successfully accrued enough patients), but not very sensitive (it predicted failed accrual for only 23 of 58 trials that actually failed to accrue enough patients), and that our assumed 50% enrollment rate (‘open circle’ data point on the ROC curve) offers a reasonable balance between specificity and sensitivity.

The lack of sensitivity of our methodology—its inability to reliably predict accrual failure—stems from the focus of our process on having a patient population satisfying the eligibility criteria, while not taking into account a number of other factors that are known to affect trial participation. As summarized by Schroen *et al*,¹⁰ physician-related factors include inadequate

Table 2 Specificity and sensitivity of our methodology for various assumed participation rates of eligible patients (assumed 50% rate bolded)

% Of eligible patients who enroll in a trial	Specificity	Sensitivity
10	0.500	0.776
20	0.750	0.638
25	0.813	0.534
30	0.844	0.534
40	0.938	0.448
50	0.969	0.397
60	0.969	0.379
70	1.000	0.328
75	1.000	0.310
80	1.000	0.310
90	1.000	0.276

reimbursement and support resources, irrelevance of available studies to the practice population, and treatment preferences, while patient-related factors include treatment preferences and logistical considerations. At an academic medical center such as Jefferson, only the last factor, treatment preference, would be expected to account for physician reluctance to participate in clinical research. Physicians will not put their patients on a trial when, in their professional judgment, treatment under a competing trial or standard of care would be better for that individual. However, the patient-related factor of treatment preference in trial non-participation is prevalent both at academic centers and in the community: patients may be concerned about the uncertainty of treatment in a randomized trial, or reluctant to undergo trial toxicities given the anticipated benefit to them. Logistical factors—such as transportation to a central treatment facility and/or additional tests—can also inhibit patient participation in trials. Our methodology could conceivably be extended to include some of these logistical factors by, for

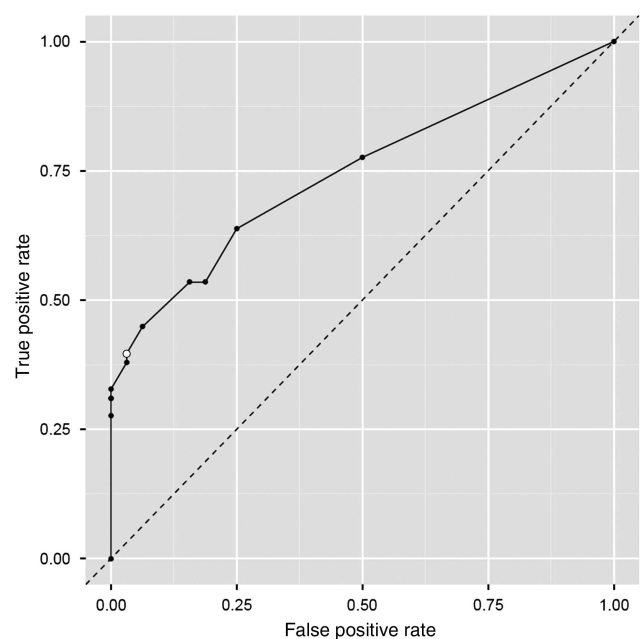


Figure 4 Receiver operating characteristic (ROC) curve based on various assumed participation rates. The ‘open circle’ data point corresponds to an assumed participation rate of 50%.

instance, including demographic constraints (HIPAA permissible zip code catchment areas) in analyzing potential patient populations. Going further, the inclusion of socio-demographic data in the RDM could be used when such measures are found to be indicative of trial participation. In the final analysis, however, it does not matter whether it is the physician or the patient who finds the trial unattractive—in either case a very low or zero actual trial participation rate will translate into an unsuccessfully accruing trial no matter how large the potential eligible patient population.

CONCLUSIONS

This study has shown that a RDM having integrated data from an institution's clinical data warehouse and other sources can be used to analyze potential accrual for proposed clinical trials by finding cohort sizes of recent patients satisfying the major eligibility rules for the trial being designed. A predicted failure to reach a desired participation target is usually indicative of future actual failure, although extenuating circumstances, such as a new clinical recruit or strong physician support for the trial, may mitigate this negative forecast. Pre-initiation protocol review committees hopefully would be aware of these other considerations. If this methodology predicts trial accrual success, it is highly unlikely that the trial will fail because of a lack of eligible patients. However, accrual failure may still occur because of various other causes such as competing trials or patient reluctance.

We plan to refine this methodology by including more data in our RDM corresponding to trial eligibility requirements, as well as including cohort analyses of existing competing trials to more accurately anticipate the combined needs of proposed and existing studies. We are also investigating automating the mapping of trial eligibility rules to RDM queries. This work seeks to provide clinical researchers developing clinical trials with rigorous guidance for eligibility rule design by use of a software application that draws on the RDM i2b2 concepts. This will move the analysis of potential accrual to even earlier in the design process than the scientific review, with further reduction of wasted research staff effort.

Application of this methodology during the protocol design phase, prior to trial initiation, would significantly contribute to reducing the wasted effort and resources currently expended on clinical trials that do not reach completion because of lack of patient participation. A very significant amount of expenditures for staff time alone will be saved by this design phase analysis. Furthermore, this would reduce the fruitless research contributions of patients who enroll in uncompleted trials.

Contributors JWL conceived of the study and is responsible for the study design. LB and DC are responsible for the data acquisition. TZ, DC, and JWL are responsible for analyzing and interpreting study data and results. All authors contributed to the manuscript preparation.

Funding This work was partially supported by National Cancer Institute grant P30CA056036 and by Kimmel Cancer Center, Thomas Jefferson University institutional funds.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The i2b2 query tool is open source software, available at i2b2.org. JWL may be contacted regarding details of the i2b2 ontologies deployed at Thomas Jefferson University.

REFERENCES

- Nass SJ, Moses HL, Mendelsohn J, eds. Committee on Cancer Clinical Trials and the NCI Cooperative Group Program Board on Health Care Services. *A National Cancer Clinical Trials System for the 21st Century: Reinvigorating the NCI Cooperative Group Program*. Washington, DC: National Academies Press, 2010.
- Cheng S, Dietrich M, Finnigan S, et al. A sense of urgency: evaluating the link between clinical trial development time and the accrual performance of CTEP-sponsored studies. *2009 ASCO Annual Meeting Proceedings; J Clin Oncol*, 2009;27:CRA6509.
- Lara PN Jr, Higdon R, Lim N, et al. Prospective evaluation of cancer clinical trial accrual patterns: identifying potential barriers to enrollment. *J Clin Oncol* 2001;19:1728–33.
- Prescott RJ, Counsell CE, Gillespie WJ, et al. Factors that limit the quality, number and progress of randomised controlled trials. *Health Technol Assess* 1999;3:1–143.
- Grunfeld E, Zitzelsberger L, Cristine M, et al. Barriers and facilitators to enrollment in cancer clinical trials: qualitative study of the perspectives of clinical research associates. *Cancer* 2002;95:1577–83.
- Ross S, Grant A, Counsell C, et al. Barriers to participation in randomised controlled trials: a systematic review. *J Clin Epidemiol* 1999;52:1143–56.
- Ellis PM. Attitudes towards and participation in randomized clinical trials in oncology: a review of the literature. *Ann Oncol* 2000;11:939–45.
- Mapstone J, Elbourne D, Roberts IG. Strategies to improve recruitment to research studies. *Cochrane Database Syst Rev* 2007;(2):MR000013.
- Mills EJ, Seely D, Rachlis B, et al. Barriers to participation in clinical trials of cancer: a meta-analysis and systematic review of patient-reported factors. *Lancet Oncol* 2006;7:141–8.
- Schroen AT, Petroni GR, Wang H, et al. Preliminary evaluation of factors associated with premature trial closure and feasibility of accrual benchmarks in phase III oncology trials. *Clin Trials* 2010;7:312–21.
- Schroen AT, Petroni GR, Wang H, et al. Challenges to accrual predictions to phase III cancer clinical trials: a survey of study chairs and lead statisticians of 248 NCI-sponsored trials. *Clin Trials* 2011;8:591–600.
- Mackenzie SL, Wyatt MC, Schuff R, et al. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. *J Am Med Inform Assoc* 2012;19(e1):e119–24.
- Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. *J Am Med Inform Assoc* 2012;19:181–5.
- Ferranti JM, Gilbert W, McCall J, et al. The design and implementation of an open-source, data-driven cohort recruitment system: the Duke Integrated Subject Cohort and Enrollment Research Network (DISCERN). *J Am Med Inform Assoc* 2012;19(e1):e68–75.
- Murphy SN, Mendis M, Hackett K, et al. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. *AMIA Annu Symp Proc* 2007;11:548–52.
- Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *JAMIA* 2010;17:124–30.
- Abend A, Housman D, Johnson B. Integrating clinical data into the i2b2 repository. *Summit on Transl Bioinforma*; 2009;2009:1–5.
- London JW, Housman D, Sagenich C, et al. Integration of Research Biospecimen Annotation with Clinical Data in an i2b2 Research Data Mart. *Am Med Inform Assoc Annu Symp* 2010;2010:1151.
- London JW, Chatterjee D. Implications of observation-fact modifiers to i2b2 ontologies. *IEEE International Conference on Bioinformatics & Biomedicine*; 2011;2011:929–30.
- London JW, Chatterjee D. Using clinical data mining to improve completion of clinical trials. *American Medical Informatics Association 2012 Summit on Clinical Research Informatics*; 2012;2012:122.
- <http://www.dwhworld.com/dwh-schemas> (accessed 16 Mar 2013).