Characterizing Basic and Complex Usage of i2b2 at an Academic Medical Center

Evan T. Sholle, M.S.¹, Marika Cusick, M.S.¹, Marcos A. Davila, M.S.¹, Joseph Kabariti, B.S.¹, Steven Flores, B.S.¹, Thomas R. Campion, Jr. Ph.D.¹,²,³,⁴

¹Information Technologies and Services Department, Weill Cornell Medicine, New York, NY; ²Department of Healthcare Policy and Research, Weill Cornell Medicine, New York, NY; ³Clinical and Translational Science Center, Weill Cornell Medicine, New York, NY; ⁴Department of Pediatrics, Weill Cornell Medicine, New York, NY;

Abstract

Developed to enable basic queries for cohort discovery, i2b2 has evolved to support complex queries. Little is known whether query sophistication – and the informatics resources required to support it – addresses researcher needs. In three years at our institution, 609 researchers ran 6,662 queries and requested re-identification of 80 patient cohorts to support specific studies. After characterizing all queries as “basic” or “complex” with respect to use of sophisticated query features, we found that the majority of all queries, and the majority of queries resulting in a request for cohort re-identification, did not use complex i2b2 features. Data domains that required extensive effort to implement saw relatively little use compared to common domains (e.g., diagnoses). These findings suggest that efforts to ensure the performance of basic queries using common data domains may better serve the needs of the research community than efforts to integrate novel domains or introduce complex new features.

Introduction

Secondary use of electronic health record (EHR) data is critical to clinical and translational research (1,2). To provide investigators with self-service access to electronic patient data for cohort identification, institutions have deployed tools developed locally (3–7) and by consortia (8,9). Of consortium-based efforts, Informatics for Integrating Biology and the Bedside (i2b2) is perhaps most widely-known, with more than 110 institutions having adopted the tool since 2004 (10,11). From adoption in individual sites, i2b2 has spawned multi-institutional cohort discovery efforts at the regional (12,13) and federal (14) levels using the Shared Health Information Resource Network (SHRINE) (15) as well as clinical trial recruitment activities at the global level by biopharmaceutical companies (16). Furthermore, informatics researchers have described sophisticated efforts to streamline i2b2 activities including but not limited to data acquisition (17–20), representation (21,22), and visualization, (23–25) as well as natural language processing and query performance improvement (26,27). Most critically, clinical researchers have cited i2b2 as facilitating cohort identification in studies of major depressive disorder (28), migraine (29), eosinophilic esophagitis (30,31), hearing loss (32), and hospital-acquired acute kidney infection, (33) among other disease areas.

Although existing investigations have demonstrated adoption, extension, and scientific value of i2b2, little is known about how clinical and translational researchers perform i2b2 queries. Existing literature is limited to studies of i2b2 support for specific use cases selected by informatics researchers (34,35) and enterprise-wide usage by non-informatics personnel in two sites (36,37). Notably, the studies of enterprise-wide i2b2 usage used less comprehensive measures and examined shorter time periods compared to evaluations of early self-service clinical data query tools (38,39). Additionally, two early i2b2 studies revealed deficits that resulted in development of new features enabling complex queries, such as extraction of data in CDISC ODM format and the ability to specify temporal sequences of events, addressing use cases devised by informatics professionals (34,35).

However, to the best of our knowledge, no studies have characterized the sophistication of everyday i2b2 queries nor examined the relationship between query sophistication and resultant scientific activity of users. In its default configuration, i2b2 offers basic query functionality, where users assemble searches by dragging clinical concepts into a maximum of three groups. Additionally, i2b2 provides complex query functionality, including the ability to define parameters where observations occurred during the same financial encounter or exhibited other sophisticated temporal relationships.

Understanding how clinical and translational researchers with and without informatics expertise use i2b2 to perform basic and complex queries can inform how informatics professionals deliver service, measure effects, and describe value of i2b2 and other approaches for secondary use of EHR data. Usage patterns, intended downstream application,
and individual expertise may differ substantially across i2b2 users, whose familiarity with the intricacies of health
data and degree of comfort with interactive web applications may vary. We hypothesized that users would perform
more basic queries than complex queries, and that basic queries would yield more requests for cohort reidentification
than complex queries.

Methods

Setting

The Weill Cornell Medicine (WCM) Physician Organization constitutes a multi-specialty group practice with over
900 physicians serving more than 2 million patients at more than 20 clinics across the New York City area. All WCM
physicians have admitting privileges to NewYork-Presbyterian Hospital (NYP), a long-time teaching affiliate. In
addition to clinical care, WCM serves as the medical education and biomedical research locus of Cornell University,
for which WCM houses a Clinical and Translational Science Award (CTSA) hub and several core facilities.

For documenting clinical care, WCM physicians use EpicCare Ambulatory in outpatient clinics and Allscripts Sunrise
Clinical Manager in the inpatient setting. Separate information technology teams from WCM and NYP oversee the
outpatient and inpatient clinical systems. As described elsewhere, the WCM Research Informatics group enables
secondary use of data from institutional EHR systems as well as support of research-specific applications (40). To
support i2b2, we transform data acquired from disparate source systems on a monthly basis and use it to populate two
i2b2 instances: a master, identified instance, termed “Red,” used exclusively by members of the Research Informatics
group for testing and cohort re-identification; and a publicly accessible instance, termed “Green,” that has been
subjected to a de-identification process in accordance with the Safe Harbor definition of the Health Insurance
Portability and Accountability Act (HIPAA).

The cohort discovery workflow starts with the construction of an initial query on the de-identified “Green” instance
of i2b2. Researchers are encouraged to contact informatics staff for support in development of individual queries, but
the tool is offered on a self-service basis. Before obtaining an IRB protocol, users are limited to obtaining demographic
breakdowns of patient cohorts and obfuscated patient counts, in accordance with the de-identified nature of the tool.
No plugins are available in the publicly accessible “Green” instance. However, after a user has defined the cohort of
interest, they may submit a formal request to informatics staff for cohort reidentification, at which point a member of
the Research Informatics team reviews the request and corresponding IRB protocol, then, assuming approval,
replicates the query on the identified “Red” i2b2 instance, using the “ExportXLS” plugin to extract a list of medical
record numbers to be provided to the requester.

Data collection

The de-identified “Green” instance of i2b2 has been available to researchers at our institution since October 2015.
When the tool was first made available to the research community, available query features were limited to patient
demographic information, diagnoses, medications, and procedures from our local outpatient EHR system, and drew
from 500 million rows of data for close to 2 million patients. Since then, we have expanded the catchment of this
instance considerably to over 2.5 billion rows of data for approximately 3 million patients. Today users can run queries
on a number of additional clinical concepts, including clinical trials enrollment, data from the institutional tumor
registry, lab results, genomic data from next-generation sequencing (NGS) panels, family history, allergies, eye exams,
and data elements derived from natural language processing tools. Additionally, users can now query data from
inpatient systems, including both the inpatient EHR and perioperative ancillary systems. Data are mapped to
standardized reference terminologies and represented in ontologies consistent with the hierarchical representations
specified by ICD, LOINC, and other pertinent standards. Using i2b2 modifiers, users can distinguish between
numerous outpatient and inpatient systems (e.g., inpatient billing diagnosis versus outpatient problem list).

We obtained i2b2 user metadata and query logs from October 2015 through December 2018. To characterize the usage
of i2b2 at our institution, we conducted a series of SQL queries against these tables, importing the resulting data into
Python for subsequent analysis.

Evaluation

The first series of queries sought to characterize the types of queries users ran, stratifying them into basic and complex
queries. Basic queries were defined as those using three or fewer groups and not stipulating any temporal relationship
between groups. Complex queries were defined as those either stipulating a temporal relationship (e.g. “occurs at same
financial encounter”) between groups or using more than three groups. For each query executed by a member of the
research community (i.e., excluding queries run by members of the Research Informatics team), we first stratified the query as basic or complex according to this definition. We then determined the number of clinical concept domains (e.g., diagnoses, medications, procedures) used by each query. While domains did not bear directly on the characterization of queries as basic or complex, as a basic query could include multiple domains in a single group, it nonetheless allowed us to characterize the extent to which users queried the extent of all data domains made available for building queries.

By combining i2b2 metadata with institutional metadata, we categorized users into three primary roles: staff, faculty, and trainees. Faculty constituted any individual with a formal faculty appointment, ranging from chaired professors to adjunct and voluntary faculty members. Trainees constituted medical students, residents, and fellows, while staff constituted the remainder of our users, including clinical research coordinators, research associates without faculty appointments, and revenue cycle personnel.

To measure scientific activity of i2b2 users, we reviewed internal records of IRB-approved requests to re-identify individual patient cohorts determined via i2b2 query. For each request, we characterized the i2b2 query that produced it as either “basic” or “complex” according to the previously provided definitions.

**Results**

During the study period, 609 users—164 (27%) faculty, 219 (36%) trainees, and 226 (36.6%) staff—performed a total of 6662 queries in i2b2. Of these, 4769 (71.6%) constituted basic usage in that they made use of three or fewer groups and did not specify a temporal relationship between the clinical concepts used to build the query. Of the 80 queries that resulted in an IRB-approved request for cohort re-identification, 68 (85%) were basic and 12 (15%) were complex (made use of more than three groups or changed the default temporal constraint). 30 (37.5%) were executed by faculty, 28 (35%) were executed by trainees, and 22 (27.5%) were executed by staff.

Of the 6662 queries users ran during the study period, more than 90% used three or fewer clinical concept domains (e.g., diagnoses, procedures), regardless of the number of groups they used. A plurality of queries made use of only one domain, as detailed in Table 2.

**Table 2. Stratification of number of domains used to build patient cohorts**

<table>
<thead>
<tr>
<th>Number of domains in query</th>
<th>Count of queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2913 (44%)</td>
</tr>
<tr>
<td>2</td>
<td>2603 (39%)</td>
</tr>
<tr>
<td>3</td>
<td>711 (11%)</td>
</tr>
<tr>
<td>4</td>
<td>316 (5%)</td>
</tr>
<tr>
<td>5</td>
<td>105 (2%)</td>
</tr>
<tr>
<td>6</td>
<td>12 (&lt;1%)</td>
</tr>
<tr>
<td>7</td>
<td>2 (&lt;1%)</td>
</tr>
</tbody>
</table>

As shown in Table 4, the most common domain used in queries was diagnoses (using ICD-9 and ICD-10 codes), with medications and demographics making up a distant second and third. Genomics, family history, and data points abstracted from the tumor registry, including ICD-O codes, represented the domains least frequently utilized in building queries.

**Table 3. Stratification of domains used to build patient cohorts by number and relative frequency of domain.** Since queries may make use of more than one domain, the total sum of percentages adds up to more than 100%

<table>
<thead>
<tr>
<th>Domain</th>
<th>Count of queries using domain</th>
<th>Percentage of total queries using domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnoses</td>
<td>5096</td>
<td>76.49</td>
</tr>
<tr>
<td>Medications</td>
<td>1621</td>
<td>24.33</td>
</tr>
<tr>
<td>Demographics</td>
<td>1590</td>
<td>23.87</td>
</tr>
<tr>
<td>Encounters</td>
<td>999</td>
<td>15.00</td>
</tr>
<tr>
<td>Procedures</td>
<td>984</td>
<td>14.77</td>
</tr>
<tr>
<td>Labs</td>
<td>906</td>
<td>13.60</td>
</tr>
<tr>
<td>Previous Query</td>
<td>360</td>
<td>5.40</td>
</tr>
<tr>
<td>NLP</td>
<td>229</td>
<td>3.44</td>
</tr>
</tbody>
</table>
Of diagnoses, the most commonly-queried clinical concept domain, neoplastic disease was by far the most common criteria used, occurring in almost 30% of all queries run during the analysis period (Table 5). External causes of morbidity/mortality and endocrine diseases occurred second- and third-most frequently.

**Table 4.** Breakdown of diagnoses used to build patient cohorts by ICD-9/ICD-10 grouping

<table>
<thead>
<tr>
<th>ICD9</th>
<th>ICD10</th>
<th>Description (ICD-9 and ICD-10)</th>
<th>Count of queries</th>
<th>Percentage of all queries using diagnostic group</th>
</tr>
</thead>
<tbody>
<tr>
<td>140–239</td>
<td>C00 – D49</td>
<td>Neoplasms</td>
<td>1417</td>
<td>29.97%</td>
</tr>
<tr>
<td>E and V</td>
<td>V00 – V99 / Z00 – Z99</td>
<td>External causes of injury and supplemental classification</td>
<td>707</td>
<td>14.95%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>External causes of morbidity</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Factors influencing health status and contact with health services</td>
<td></td>
<td></td>
</tr>
<tr>
<td>240–279</td>
<td>E00 – E89</td>
<td>Endocrine, nutritional, and metabolic diseases</td>
<td>617</td>
<td>13.05%</td>
</tr>
<tr>
<td>390–459</td>
<td>I00 – I99</td>
<td>Diseases of the circulatory system</td>
<td>552</td>
<td>11.68%</td>
</tr>
<tr>
<td>520–579</td>
<td>K00 – K95</td>
<td>Diseases of the digestive system</td>
<td>491</td>
<td>10.39%</td>
</tr>
<tr>
<td>280–289</td>
<td>D50 – D89</td>
<td>Diseases of the blood and blood-forming organs and certain disorders involving the immune system</td>
<td>373</td>
<td>7.89%</td>
</tr>
<tr>
<td>290–319</td>
<td>F01 – F99</td>
<td>Mental disorders</td>
<td>346</td>
<td>7.32%</td>
</tr>
<tr>
<td>460–519</td>
<td>J00 – J99</td>
<td>Diseases of the respiratory system</td>
<td>304</td>
<td>6.43%</td>
</tr>
<tr>
<td>320–359</td>
<td>G00 – G99</td>
<td>Diseases of the nervous system</td>
<td>250</td>
<td>5.21%</td>
</tr>
<tr>
<td>800–999</td>
<td>S00 – T88</td>
<td>Injury and poisoning</td>
<td>244</td>
<td>5.16%</td>
</tr>
<tr>
<td>780–799</td>
<td>R00 – R99</td>
<td>Symptoms, signs, and ill-defined conditions</td>
<td>159</td>
<td>3.36%</td>
</tr>
<tr>
<td>001–139</td>
<td>A00 – B99</td>
<td>Infectious and parasitic diseases</td>
<td>156</td>
<td>3.30%</td>
</tr>
<tr>
<td>580–629</td>
<td>N00 – N99</td>
<td>Diseases of the genitourinary system</td>
<td>109</td>
<td>2.31%</td>
</tr>
<tr>
<td>740–759</td>
<td>Q00 – Q99</td>
<td>Congenital anomalies</td>
<td>84</td>
<td>1.78%</td>
</tr>
<tr>
<td>710–739</td>
<td>M00 – M99</td>
<td>Diseases of the musculoskeletal system and connective tissue</td>
<td>82</td>
<td>1.74%</td>
</tr>
<tr>
<td>680–709</td>
<td>L00 – L99</td>
<td>Diseases of the skin and subcutaneous tissue</td>
<td>80</td>
<td>1.69%</td>
</tr>
<tr>
<td>630–679</td>
<td>O00 – O9A</td>
<td>Pregnancy, childbirth, and the puerperium</td>
<td>41</td>
<td>0.87%</td>
</tr>
<tr>
<td>760–779</td>
<td>P00 – P96</td>
<td>Certain conditions originating in the perinatal period</td>
<td>27</td>
<td>0.572%</td>
</tr>
<tr>
<td>360–389</td>
<td>H00 – H59</td>
<td>Diseases of the sense systems</td>
<td>25</td>
<td>0.53%</td>
</tr>
<tr>
<td></td>
<td>H60 – H94</td>
<td>Diseases of the eye and adnexa</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Diseases of the ear and the mastoid process</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Discussion

Upon conducting an analysis of i2b2 usage, we found that users were most likely to use the tool to identify patient cohorts based on relatively simple, predominantly diagnostic criteria. Utilization of the tool was relatively evenly distributed across staff, faculty, and trainees. Among queries that resulted in the re-identification of a patient cohort, most did not make use of complex i2b2 features. Features that required extensive development effort to integrate into the i2b2 data model did not see widespread adoption.

i2b2 has seen adoption for use cases beyond its original intended purpose. Both locally and through collaborations organized under the auspices of the i2b2 consortium, informatics professionals have developed, implemented, and evaluated plugins that extend the capabilities of the platform, including the execution of basic statistical analysis of relative fact prevalence between cohorts (41), the implementation of infobuttons (42), and others. However, the primary utility of i2b2 is its ability to provide a self-service platform whereby investigators can easily build queries by dragging clinical concepts from a hierarchical ontology into a series of groups that define Boolean logic for a generated SQL query against the underlying i2b2 data model. Queries can range from simple (“How many patients have ever been assigned an ICD-9 code of 250.00?”) to complex (“Of patients who had lithotripsy in 2016, how many had a hemoglobin value over 10.5 g/dL within 10 days of their first ever exposure to hydroxyurea?”)

Significant development effort is required to implement i2b2 features and integrate new clinical data elements into the platform’s catchment, often requiring deep engagement with source systems in order to accurately and completely map the data into the i2b2 model (21). As is likely the case for many health care providers, effort from skilled programmers with informatics expertise is a limited resource at our site - in considering effort allocation across platforms, the need to consider return on investment is paramount. In considering the degree of effort required to implement some of the least-frequently used ontology items (e.g. tumor registry data and genomic data, both of which required extensive development effort and the support of a third-party vendor to implement) it became immediately apparent that a user-driven approach might seek to extend the usability of the tool in areas that directly impact day-to-day use of the platform rather than focusing on the intake of new data domains without explicit stakeholder impetus. While it is possible that user training on complex i2b2 features might increase utilization, it is also possible that current utilization patterns accurately reflect the research priority of the user base.

Of particular note, queries targeting diagnoses of neoplasms were considerably more common than other conditions. It is possible that a small handful of users focusing on these conditions may be disproportionately impacting the distribution of diagnosis queries – other sites considering usage patterns within informatics platforms such as i2b2 may wish to consider the extent to which aggregate statistics of user behavior are impacted by “power users” of the platform, and how this may bear on any higher-order decisions made on the basis of such an analysis.

In comparison with previous work, which has focused on evaluating the implementation and dissemination of i2b2 from a systems standpoint, our work sought to quantify its utilization at a single site. Bell et al.’s 2017 work, the only quantitative analysis of i2b2 user data of which we are aware, surveyed a random sample of i2b2 users in an attempt to garner their perceptions of the plausibility of i2b2-derived counts, as well as their impressions of the overall usability of the tool. In keeping with the recommendations issued by Murphy et al (10) our findings here suggest the prioritization of simple data elements by individual institutions as the i2b2 consortium seeks to promote broader dissemination of the tool to support the global biomedical research enterprise.

Intrinsic limitations of the methodology of this analysis include, first and foremost, the lack of a direct vantage point into users’ goals and directives in using the system. While direct observational research, such as a survey or focus groups of users, was beyond the scope of the current study, further research may seek to supplement the results of this analysis with data derived directly from user report. Additionally, the raw query counts used to conduct this analysis may serve as, at best, a poor proxy for actual intended use of the system, as users struggling to understand the capabilities and limitations of the tool may run dozens or hundreds of queries while seeking to hone in on a target concept, whereas other, more skilled users (including those with prior i2b2 experience from another institution) may need only one attempt to derive the cohort they sought to identify. Further research may seek to address these limitations by decomposing user behavior into discrete episodes, treating consecutive queries within the same login session as single units. Additionally, we may seek to evaluate i2b2 query behavior in the context of bibliometrics, seeking to determine whether complex queries are more likely to result in scientific output. Finally, although we conducted this study at a single site, the methods employed to analyze and characterize user behavior can likely generalize to other sites, permitting a multi-site comparison. Regardless of these limitations, our analysis of cohorts for which users sought reidentification remains pertinent – in keeping with larger, query-based trends, users mostly sought to re-identify relatively simple queries.
Complex i2b2 use cases requiring extensive development effort to implement and validate may not address user needs. i2b2 was conceptualized, first and foremost, as a way to enable clinical and translational researchers to make use of large medical data sets to further their research. Since implementation and maintenance effort on behalf of informatics teams is a fungible and finite resource, it stands to reason that targeting this effort on plugins, concepts, and integrations that do not align with the priorities of the research enterprise may divert resources from other domains, such as user experience/usability and data quality, limiting the extent to which i2b2 can be of maximum utility to the research enterprise. Informatics groups may also seek to address these findings in other ways: for example, by disabling the ability to run complex queries entirely or by flagging repeated complex queries for attention by a qualified data analyst.

In evaluating usage and downstream scientific impact of our i2b2 instance, we found that the majority of users were primarily leveraging the simplest capacities of the platform, running straightforward queries primarily built on diagnostic facts. Of the queries that ultimately led to further user engagement (including requests for re-identification and, ultimately, scientific output), the majority were relatively simple queries, suggesting that informatics groups at academic medical centers may wish to tailor their efforts accordingly.

Acknowledgments

This study received support from NewYork-Presbyterian Hospital (NYPH) and Weill Cornell Medical College (WCMC), including the Clinical and Translational Science Center (CTSC) (UL1 TR000457) and Joint Clinical Trials Office (JCTO).

References


http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4525245/

38. Murphy SN, Morgan MM, Barnett GO, Chueh HC. Optimizing healthcare research data warehouse design through past COSTAR query analysis. Proc AMIA Symp [Internet]. 1999;892–6. Available from:


