DATA ARCHITECTURE AND MODELS

original reports abstract

Developing a Data and Analytics Platform to Enable a Breast Cancer Learning Health System at a Regional Cancer Center

Jeremy Petch, PhD, MA[1,2,3,4]; Joel Kempainnen[1]; Christopher Pettengell, MD[5]; Steven Aviv, BSc[5]; Bill Butler, BSc(hons)[6]; Greg Pond, PhD, MSc[7]; Ashirbani Saha, PhD, MEng[1,7,8]; Jessica Bogach, MD, MSc[9]; Alexandria Allard-Coutu, MD[9]; Peter Sztur, BEng[1]; Jonathan Ranisau, MASc[1]; and Mark Levine, MD, MSc[6,7]

PURPOSE This study documents the creation of automated, longitudinal, and prospective data and analytics platform for breast cancer at a regional cancer center. This platform combines principles of data warehousing with natural language processing (NLP) to provide the integrated, timely, meaningful, high-quality, and actionable data required to establish a learning health system.

METHODS Data from six hospital information systems and one external data source were integrated on a nightly basis by automated extract/transform/load jobs. Free-text clinical documentation was processed using a commercial NLP engine.

RESULTS The platform contains 141 data elements of 7,019 patients with newly diagnosed breast cancer who received care at our regional cancer center from January 1, 2014, to June 3, 2022. Daily updating of the database takes an average of 56 minutes. Evaluation of the tuning of NLP jobs found overall high performance, with an F1 of 1.0 for 19 variables, with a further 16 variables with an F1 of > 0.95.

CONCLUSION This study describes how data warehousing combined with NLP can be used to create a prospective data and analytics platform to enable a learning health system. Although upfront time investment required to create the platform was considerable, now that it has been developed, daily data processing is completed automatically in less than an hour.

JCO Clin Cancer Inform 7:e2200182. © 2023 by American Society of Clinical Oncology

BACKGROUND

In 2007, the Institutes of Medicine coined the term Learning Health System to describe a framework for organizing health care institutions around a virtuous cycle in which clinical care and research continuously inform one another, enabled by a modern informatics infrastructure that operationalizes routinely collected health data for research and quality improvement.[1] Although most health care systems now use electronic health records (EHR), many lack the necessary data and analytics infrastructure to use EHR data in real time for continuous improvement, innovation, and knowledge generation, especially at the regional and national levels.[2-6]

For routinely collected data to be effective in enabling a learning health system, they generally must be integrated, timely, meaningful, high-quality, and actionable.[5] These conditions can be difficult to meet. For example, in the context of regional cancer centers, a single patient's record may be split among all the different health care organizations where they have received care. Regional data repositories designed to address this challenge often experience delays in receiving and cleaning data and may be missing important clinical,[3] social determinant,[4,7] or patient-reported outcome data.[5] Indeed, important information about a patient's condition is often recorded only in the form of free-text clinical documentation (such as consult notes and radiology reports), which is historically needed to be abstracted via manual review.[3] Furthermore, clinical documentation standards often vary, which can introduce data quality issues.[8-10] Privacy requirements or inadequate data governance frameworks can also impede access to these data, limiting their actionability.[11,12]

Despite the challenges, a number of efforts have been undertaken to create the necessary data and analytics infrastructure to enable a learning health system in oncology.[13-18] Many of these have used a data warehousing or data lake approach, in which data are automatically transferred on a regular schedule from disparate sources to a centralized data repository that is designed for analytics.[19-26] When properly designed, data warehouses and data lakes can significantly improve the timeliness and actionability of data although they require significant investment in time and resources to create.

ASSOCIATED CONTENT

Data Supplement

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on February 10, 2023 and published at ascopubs.org/journal/cci on March 31, 2023: DOI https://doi.org/10.1200/CCI.22.00182

CONTEXT

Key Objective

How can oncology centers leverage their clinical information systems to unlock the integrated, timely, meaningful, high-quality, and actionable data required to establish a learning health system?

Knowledge Generated

A combination of data warehousing and natural language processing was used to create a patient-level, prospective, longitudinal data and analytics platform at a regional cancer center. This data and analytics platform includes comprehensive clinical data about 7,000+ patients with newly diagnosed breast cancer seen over the past eight years and can be used for both research and quality improvement.

Relevance

This study describes an approach through which cancer centers can improve their ability to use routinely collected clinical data to enable a learning health system in which clinical care and research continuously inform one another.

Although data integration through warehouses and lakes can make data more timely and actionable, they are not typically equipped to facilitate the analysis of free-text documentation where so much clinically meaningful information resides. This has spurred recent interest in natural language processing (NLP), a branch of artificial intelligence, for the extraction of structured data from clinical text. Although NLP rarely achieves perfect accuracy, research suggests that it can achieve similar performance to manual chart abstraction.[27-36] There is increasing interest in blending NLP with traditional data warehousing to create a comprehensive data set for research and quality improvement although to date, relatively few institutions have developed these systems,[37-39] and to the best of our knowledge, none currently exists for breast cancer.

The Juravinski Cancer Center (JCC) is a regional cancer center serving a catchment area of approximately 2.5 million people and receiving referrals from 10 community hospitals. Data integration challenges at this center were similar to those faced by other regional cancer centers. Patient records were distributed across six clinical information systems, and a significant amount of information was stored exclusively as free text in clinical notes and radiology reports. Medical records of referred patients were often incomplete or transferred as text files with no structure. Although a regional viewer was available for clinicians to see records in other hospitals' EHRs, these records were view-only and not available in a repository for research or quality improvement purposes.[40] As a result of these factors, making use of routinely collected health data was a difficult and time-consuming task.

This article documents the establishment of data and analytics platform for a breast cancer learning health system at the JCC in Ontario, Canada. This platform automatically extracts patient data—including social determinants and patient reported outcomes—from disparate clinical information systems on a nightly basis, uses NLP to extract structured data from free-text documentation, and integrates them into a single up-to-date, longitudinal, prospective data model.

METHODS

Stakeholder Engagement

One of the primary barriers to the kinds of quality improvement initiatives enabled through learning health systems can be stakeholder resistance.[41,42] To help address this barrier from the outset, we set about engaging stakeholders around the hospital to contribute to shaping the vision and mission of the data and analytics platform. These stakeholders included clinicians, managers, quality improvement teams, data analysts, researchers, information technologists, privacy officers, and executive sponsors. This engagement extended to helping determine which data elements would be included in the platform since we reasoned that buy-in would be highest if all stakeholder groups could make effective use of the platform to meet their own goals. Stakeholders were invited to provide input into a briefing note that outlined the vision and mission for the platform, along with proposed methods, privacy/security protections, and data elements. Key stakeholders were asked to add their names to the briefing note once their input had been included, and the note was then reviewed and approved by the hospital's data and analytics governance committee.

Data Sources

Once a list of data elements had been identified by our stakeholders, we set about identifying where these data resided in our hospital's informatics environment. With help from our information technology and decision support departments, we identified that data of interest originated in six distinct clinical information systems.

MEDITECH. The hospital's primary EHR and stored clinical documentation and data on patient demographics and encounters.

Hamilton regional laboratory medicine program. Regional laboratory information system that stored pathology reports.

PowerScribe. Regional radiology reporting platform that stored medical imaging reports.

MOSAIQ. Oncology clinical information system that stored data on radiation planning.

Oncology patient information system. Province-wide clinical information system that stored data on systemic therapy.[43]

Your symptoms matter. Province-wide electronic information system that stored data on patient-reported outcome measures via the Edmonton System Assessment System.[44]

We then mapped the data flows between these distinct systems and identified that copies of the data of interest from PowerScribe and the Hamilton Regional Laboratory Medicine Program were stored in MEDITECH and copies of data from the Oncology Patient Information System and Your Symptoms Matter were stored in a MOSAIQ data mart. We examined the copied data to verify that they were complete and useable for our purposes.

In addition to these six systems, we also identified that information on social determinants of health could be acquired from the Ontario Marginalization Index,[45] a Canadian deprivation-based index similar to the Multidimensional Deprivation Index developed by the US Census Bureau.[46]

Architecture and Data Flows

Architecture was developed in consultation with our hospital's information technology department. The hospital was in the process of developing a data warehouse for operational and financial analytics, so we elected to use the same design patterns for both systems to minimize operational overhead and provide the option of merging the resources in the future. We thus adopted Microsoft SQL server for primary data storage, Microsoft Server Integration Services (SSIS) for developing and managing extract/transform/load jobs, T-SQL for stored procedures, and estrogen receptor/Studio for data modeling. The NLP software DARWEN was run on Docker. A deidentified copy of the data was stored in a separate research informatics environment using PostgreSQL. The flow of data is illustrated in Figure 1, and a detailed description of data handling is provided in the Data Supplement.

To maximize the timeliness of data while minimizing the chance of performance degradation on clinical systems, the extract/transform/load jobs were programmed to automatically run every night at 2:30 am. Our data engineer used Microsoft SSIS' change data capture features to extract only new or modified data each night.

Since data in the data and analytics platform could be used for clinical and quality improvement purposes, it included personal health information. In consultation with our privacy office, we created a deidentified version of this database to be used for research purposes. The intent of this approach was to improve efficiency and reliability by creating a rigorous deidentification procedure up-front, rather than requiring an analyst to deidentify data on a project-by-project basis in the future. To deidentify the database, we removed direct identifiers (name, health card number, etc) and modified quasi-identifiers (eg, data elements that contained specific dates were modified to use days from diagnosis).

System performance for data extraction, transformation, and loading was monitored and evaluated using system logs recorded by Microsoft SSIS.

NLP

We used NLP for data extraction in two scenarios: first, when no structured data were available for a particular data element (eg, comorbidities) for any of our patients; second, when structured data were available for some patients, but not all. The second scenario occurred because some patients received all their care at our regional cancer center, whereas others were referred only after being diagnosed or having received some treatment at a referring hospital. For patients who received all their care at our cancer center, we had structured data on estrogen receptor status, progesterone receptor status, and human epidermal growth factor receptor 2 (HER2) status from synoptic reports. However, for patients referred after their diagnosis, we used NLP to extract these data from free-text clinical documentation. In these cases, we elected to use structured data from synoptic reports when they were available and to fill in the gaps with NLP when they were not.

We used DARWEN, a commercially available medical NLP engine, to extract structured data from unstructured clinical documentation. DARWEN uses a proprietary combination of linguistic rules–based algorithms and deep learning models to perform data extraction. Its operations and performance have been described previously.[27,28]

For data elements where structured data were not available for any patients, ground truth for model development and evaluation was established through manual chart review. Chart abstraction rules were drafted by a clinical expert and refined in collaboration with the JCC's Breast Cancer Disease Site Group, which included medical, radiation, and surgical oncologists. Manual extraction of 200 randomly selected charts was performed by two trained chart reviewers. 100 abstracted charts were used for model training, with 50 reserved for validation and 50 held back for final testing. In addition to this approach, we were able to conduct a further test for estrogen receptor status, progesterone receptor status, and HER2 status using structured data from JCC pathologists' synoptic reports as ground truth. This approach allowed for a very large test set for performance estimation in cases where NLP was used to fill in the gaps for patients whose pathology workup was performed at a referring hospital.

To ensure that the data produced by NLP were well tuned to our local data and thus of sufficiently high-quality to be used in research studies, we evaluated its performance by comparing it against manual chart abstraction for the held-out test set (n = 50). F1 score, the harmonic mean of sensitivity and positive predictive value, was used as the primary evaluation metric. Secondary metrics included

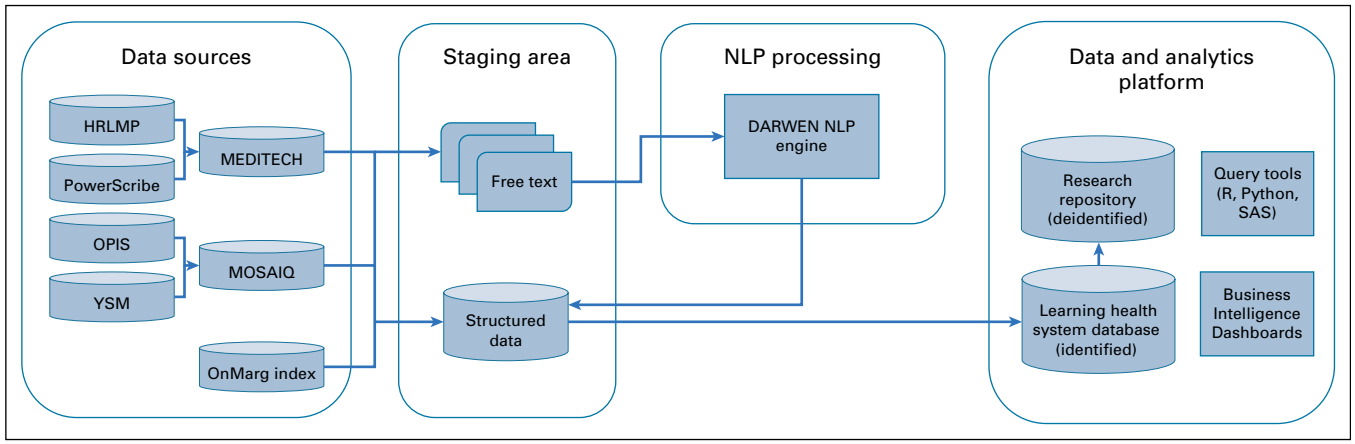

FIG 1. Data flow diagram. Nightly extract/transform/load jobs developed in Microsoft SSIS transfer free-text clinical documentation and structured data from source systems to a staging area. Free-text documentation is processed by an NLP engine, with structured data output loaded into a SQL database in the staging area. A second stage of extract/transform/load operations transfers data to the learning health system data and analytics platform. NLP, natural language processing; SSIS, Microsoft Server Integration Services.

sensitivity (recall), specificity, positive predictive value (precision), negative predictive value, and overall accuracy. We also conducted a detailed manual error analysis for the variables with the lowest performance.

Ongoing Quality Assurance

Performance decay of artificial intelligence systems because of data drift or concept drift is a growing concern in health care.[47,48] To control for potential performance decay of our NLP models, we launched an ongoing quality assurance program, in which a random selection of charts is reviewed semiannually by two oncologists. Results of the chart review are compared against NLP output, and any deviation from baseline NLP performance is flagged for follow-up by our technical team. In addition, system logs are reviewed regularly to identify any failures with extract/transform/load operations. To ensure that this process can be carried out efficiently, we have created a browser-based chart abstraction tool that allows chart abstractors to simultaneously review clinical documentation while filling out a standardized chart abstraction form in a single browser window (the screenshot is included in the Data Supplement).

RESULTS

This work culminated in an automated, longitudinal, prospective data and analytics platform that provides access to integrated, timely, meaningful, high-quality, and actionable data for research, quality improvements, and other learning health system activities. The platform contains 141 data elements of 7,019 patients with newly diagnosed breast cancer who received care at the JCC from January 1, 2014, to June 3, 2022.

Data elements in the platform are organized into tables on the basis of subject areas, which include both data elements originating from structured databases and those extracted with NLP (Fig 2). A detailed data dictionary is included in the Data Supplement. All data elements can be joined at the patient level using PatientID as a key. Data in the platform are longitudinal, which allows for the visualization and analysis of patients' entire care journey as a timeline (Fig 3). Data in the platform can be accessed through dashboards built in the organization's enterprise business intelligence tool (Tableau) or using analysis tools such as R, Python, and SAS.

System Performance

Daily updation of the database takes place every day at 2:30 am to minimize impact on source systems and our hospital network. Over the month of May 2022, the average runtime of daily update jobs was 56 minutes. Extract/transform/load jobs were consistent at approximately 40 minutes, with the primary source of variability coming from daily NLP processing. This variability was driven by clinic schedules, with daily patient volumes ranging from 80 to 300. The initial NLP run used to populate the database with records from over 7,000 patients took 24 hours. The extract/transform/load and NLP jobs were run on a server with a four core Intel Gold 6248 @ 2.50 GHz CPU and 16 GB of RAM.

NLP Performance

NLP performance is described in Table 1, and the distribution of labels for variables extracted by NLP is reported in the Data Supplement. An F1 of 1.0 was achieved for 19 variables, with a further 16 variables with an F1 of > 0.95. These results are consistent with previous validation studies of the DARWEN NLP engine.[27,28]

The lowest F1 was for detecting venous thromboembolism (0.57), which in this case was related to a lower positive predictive value (precision), a result not entirely unexpected given the rarity of this complication (there were only two cases in our test data set). Our manual error analysis of these

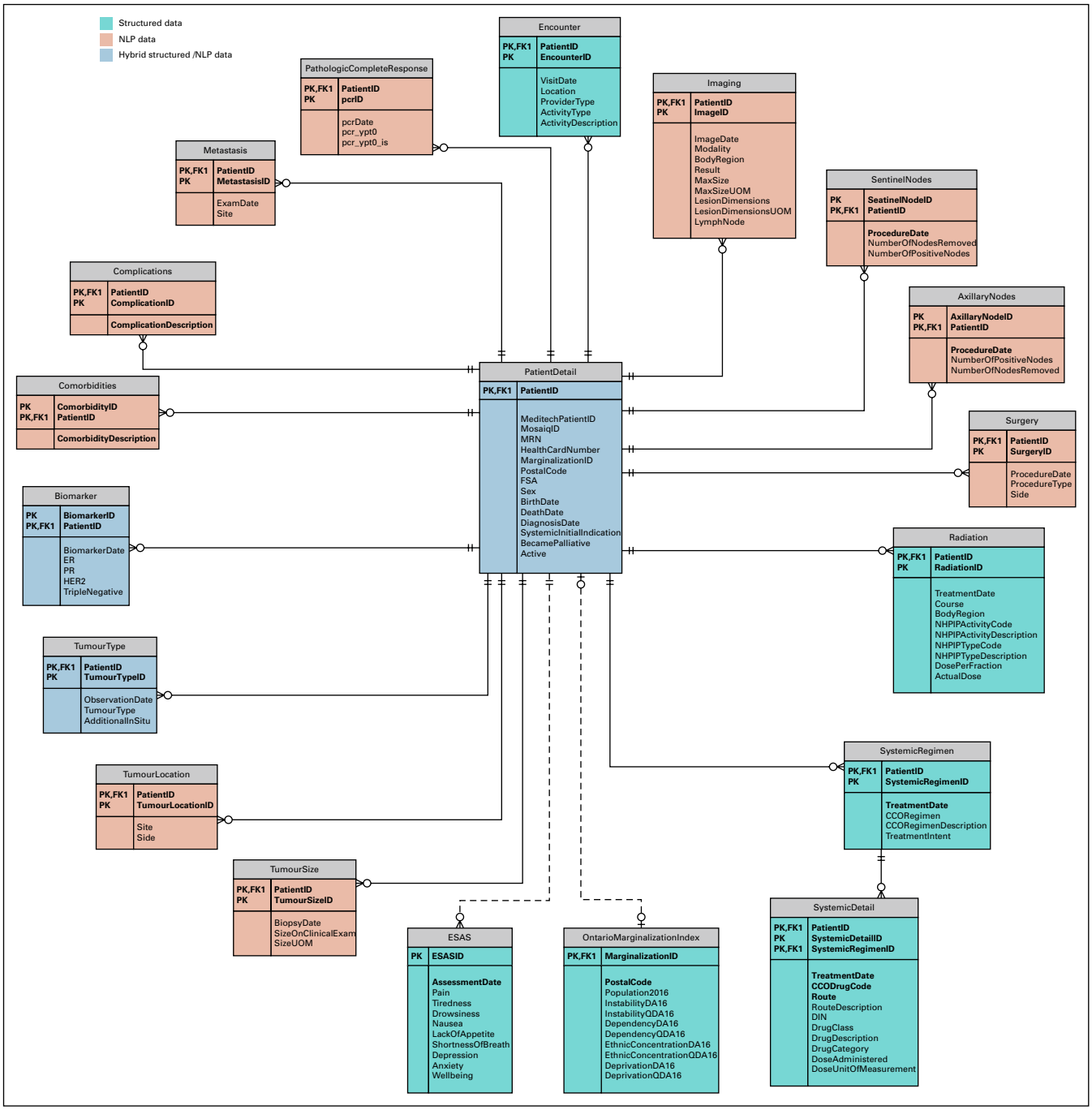

FIG 2. Data and analytics platform conceptual schema. This figure illustrates the data elements available for each subject area in the platform. All data elements can be joined at the patient level. Tables where data elements originated from structured data are given in teal. Tables with data elements entirely extracted using NLP from free-text clinical documentation are given in red. Tables that used a hybrid of NLP and structured data extraction are given in blue. (This figure is provided as a vector graphic so that it is legible when zoom in on.) NLP, natural language processing.

false-positive cases found that the NLP had missed negating clauses (ie, no evidence of venous thromboembolism).

The other lower F1 scores were primarily related to detecting comorbidities, specifically atrial fibrillation (0.80), chronic obstructive pulmonary disorder (0.80), and stroke (0.86). For all three of these conditions, this was driven by lower sensitivity, indicating that the NLP missed some cases. On manual error analysis, we noted that all the missed cases occurred when patients had four or five comorbidities. In these cases, the NLP successfully detected three or four of the comorbidities, but missed a fourth or fifth.

In addition, we completed the first cycle of semiannual quality assurance before publication. This activity identified an anomaly in our radiation planning data. Before launch,

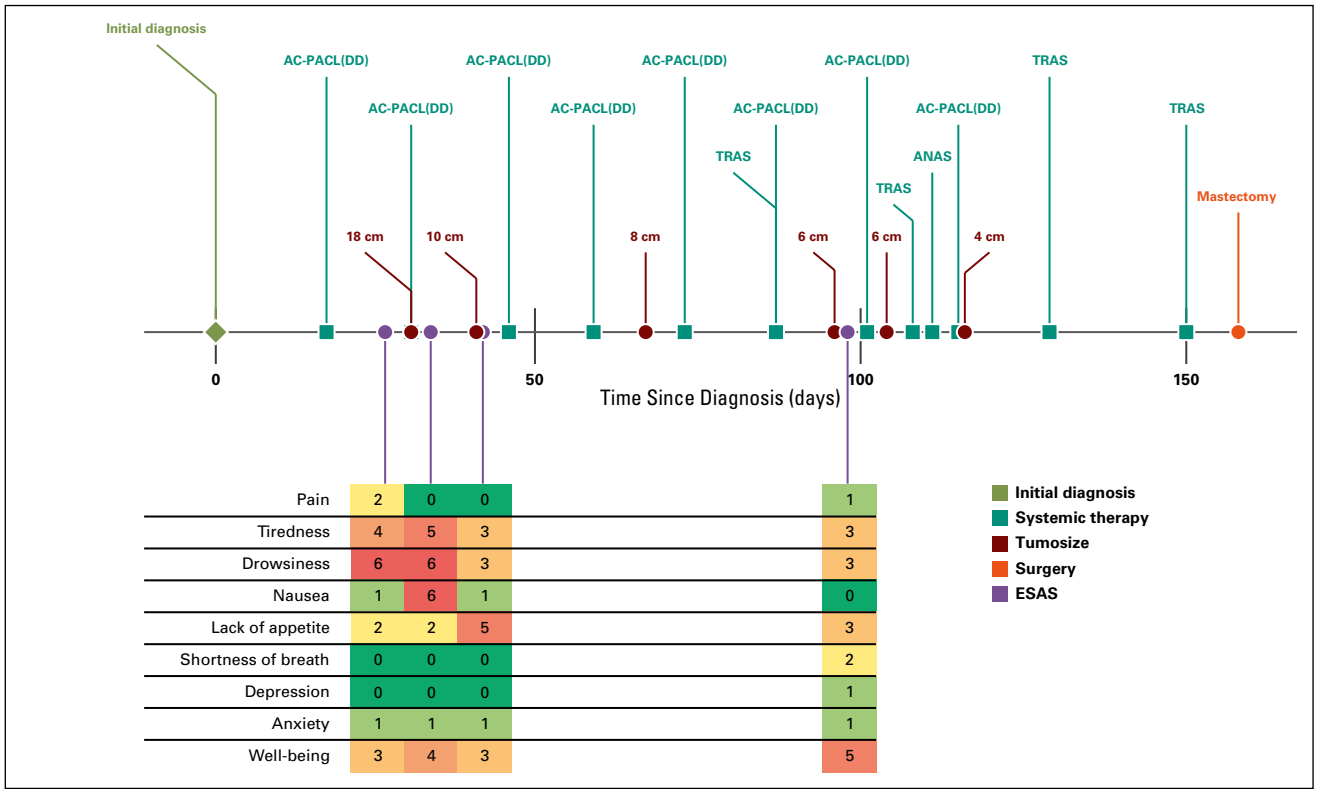

FIG 3. Longitudinal view of a patient journey. This figure illustrates a portion of a single patient's deidentified data longitudinally. The x-axis displays days since diagnosis. As illustrated here, this patient received dose dense doxorubicin with paclitaxel (Cancer Care Ontario Code AC-PACL [DD]) along with TRAS and ANAS. Their tumor was observed to have shrunk from 18 cubic centimeters to 4 over the 12 weeks of systemic therapy. Symptomology was recorded using the ESAS, and most symptoms—with the exception of well-being—were worse earlier in the course of treatment. For readability, the data illustrated in this figure only include a fraction of what is available for every patient. ANAS, anastrozole; ESAS, Edmonton System Assessment System; TRAS, trastuzumab.

we conducted extensive tests on our extract/transform/load jobs to ensure that they were stable and performed source to target verification to ensure that data in the platform were an accurate representation of data in the source systems. However, during routine quality monitoring, we noted that some older radiation data were changing in ways that did not make clinical sense. On investigation, we discovered that although our platform's extract/transform/load jobs were operating correctly, data in the source system table we were pulling from were unstable because of an error in their code. We were able to resolve the issue by working with our IT team to identify alternate tables within the source system that were unaffected by the error.

DISCUSSION

This case study illustrates how a data and analytics platform can be created at a regional cancer center to enable the kind of research and quality improvement activities that exemplify a learning health system. Use of a data warehousing approach provided data that are integrated, timely, and actionable, whereas incorporation of automated NLP allowed for the extraction of high-quality, clinically meaningful data that would have otherwise been accessible only through time-consuming manual chart review. Although our previous pilot study established that this was feasible in breast cancer on a small sample and with a restricted set of data elements, the current study documents how such a system can be delivered at scale.[3]

This study highlights the importance of ongoing quality assurance of artificial intelligence deployments in health care. Although extensive testing before launch can catch most defects, like concept drift and data drift, the instability in the radiation planning data that our quality assurance activities uncovered could only be detected through ongoing monitoring. Although ongoing quality assurance requires resources, without this kind of monitoring, our platform would have gone on faithfully reproducing erroneous data from a defective source system table.

This project was performed at a single center with its own unique challenges with respect to clinical informatics and regional data sharing. Thus, both extract/transform/load operations and NLP models would require adaptation and tuning to be deployed at another center. Similarly, like other regional cancer centers, the JCC typically cares for patients with more advanced disease than referring hospitals, so the

TABLE 1. Evaluation of Natural Language Processing Compared With Manual Chart Review

Variable	Possible Values	F1 (95% CI)	Sensitivity/ Recall	Specificity	Positive Predictive Value/ Precision	Negative Predicative Value	Overall Accuracy
Imaging							
CT abdomen/pelvis date	dd/mm/yyyy	1.00	1.00	1.00	1.00	1.00	1.00
CT abdomen/pelvis result	Metastasis, no metastasis, unknown	0.96 (0.88 to 0.96)	0.96	0.99	0.96	0.99	0.98
CT abdomen/pelvis lymph node	Node absent, node present, unknown	0.96 (0.9 to 0.98)	0.96	0.99	0.96	0.99	0.98
U/S abdomen/pelvis date	dd/mm/yyyy	1.00	1.00	1.00	1.00	1.00	1.00
U/S abdomen/pelvis result	Metastasis, no metastasis, unknown	0.98 (0.94 to 1.0)	0.98	0.99	0.98	0.99	0.99
Bone scan date	dd/mm/yyyy	1.00	1.00	1.00	1.00	1.00	1.00
Bone scan result	Metastasis, no metastasis, unknown	0.98 (0.94 to 0.98)	0.98	0.99	0.98	0.99	0.99
CT chest date	dd/mm/yyyy	1.00	1.00	1.00	1.00	1.00	1.00
CT chest result	Metastasis, no metastasis, unknown	1.00	1.00	1.00	1.00	1.00	1.00
CT chest lymph node	Node absent, node present, unknown	0.98 (0.94 to 0.98)	0.98	0.99	0.98	0.99	0.99
Chest X-ray date	dd/mm/yyyy	1.00	1.00	1.00	1.00	1.00	1.00
Chest X-ray result	Unknown, no metastasis, metastasis	1.00	1.00	1.00	1.00	1.00	1.00
Mammography date	dd/mm/yyyy	1.00	1.00	1.00	1.00	1.00	1.00
Mammography results	Clear, suspicious lesion, clear, unknown	0.98 (0.92 to 0.98)	0.98	0.99	0.98	0.99	0.99
MRI breast date	dd/mm/yyyy	1.00	1.00	1.00	1.00	1.00	1.00
MRI breast max size	Length × width × height	0.98 (0.94 to 0.98)	0.98	1.00	0.98	1.00	1.00
MRI breast lymph node	Node absent, node present	0.98 (0.94 to 1.0)	0.98	0.99	0.98	0.99	0.99
U/S breast biopsy date	dd/mm/yyyy	0.98 (0.92 to 0.98)	0.98	0.99	0.98	0.99	0.99
U/S breast biopsy result	Length × width × height	1.00	1.00	1.00	1.00	1.00	1.00
U/S breast biopsy lymph node	Node absent, node present	0.98 (0.94 to 1.0)	0.98	0.99	0.98	0.99	0.99
Surgery							
Axillary node dissection date	dd/mm/yyyy	0.98 (0.94 to 0.9)	0.98	0.99	0.98	0.99	0.99
Axillary node dissection nodes removed	Integer	0.98 (0.94 to 1.0)	0.98	1.00	0.98	1.00	1.00
Axillary node dissection positive nodes	Integer, unknown	0.98 (0.94 to 1.0)	0.98	0.99	0.98	0.99	0.99
Sentinel node biopsy date	dd/mm/yyyy	1.00	1.00	1.00	1.00	1.00	1.00
Sentinel node biopsy nodes removed	Integer, unknown	1.00	1.00	1.00	1.00	1.00	1.00
Sentinel node biopsy positive nodes	Integer, unknown	0.98 (0.92 to 0.98)	0.98	0.99	0.98	0.99	0.99
Surgery date	yyyy/mm	0.88 (0.78 to 0.95)	0.88	0.88	0.88	0.88	0.88
Surgery side	Left, right	0.89 (0.82 to 0.95)	0.89	0.95	0.89	0.95	0.93
Surgery procedure	Mastectomy, modified radical mastectomy, breast-conserving surgery	0.89 (0.82 to 0.95)	0.89	0.96	0.89	0.96	0.95

(Continued on following page)

TABLE 1. Evaluation of Natural Language Processing Compared With Manual Chart Review (Continued)

Variable	Possible Values	F1 (95% CI)	Sensitivity/ Recall	Specificity	Positive Predictive Value/ Precision	Negative Predicative Value	Overall Accuracy
Comorbidities							
AF	Yes/no	0.8 (0.5 to 1.0)	0.67	1.00	1.00	0.99	0.99
CAD	Yes/no	0.89 (0.57 to 1.0)	0.80	1.00	1.00	0.98	0.99
COPD	Yes/no	0.8 (0.5 to 1.0)	0.67	1.00	1.00	0.99	0.99
DM	Yes/no	1.00	1.00	1.00	1.00	1.00	1.00
HTN	Yes/no	1.00	1.00	1.00	1.00	1.00	1.00
Stroke	Yes/no	0.86 (0.5 to 1.0)	0.75	1.00	1.00	0.99	0.99
Pathology							
Diagnostic biopsy date	dd/mm/yyyy	0.98 (0.92 to 0.98)	0.98	0.98	0.98	0.98	0.98
ER biomarker	Negative, positive, unknown	1.00	1.00	1.00	1.00	1.00	1.00
PR biomarker	Negative, positive, unknown	1.00	1.00	1.00	1.00	1.00	1.00
HER2 biomarker	Negative, positive, unknown	0.88 (0.72 to 1.0)	0.88	0.96	0.88	0.96	0.94
Tumor side	Bilateral, left, right, unknown	0.94 (0.86 to 1.0)	0.94	0.97	0.94	0.97	0.96
Tumor site	Central, others, unknown	0.94 (0.86 to 1.0)	0.94	0.97	0.94	0.97	0.96
Tumor type	Ductal, others, DCIS, mixed, lobular, LCIS, unknown	1.00	1.00	1.00	1.00	1.00	1.00
Clinical examination nodes found	No, yes, unknown	0.85 (0.74 to 0.94)	0.85	0.93	0.85	0.93	0.90
Clinical examination size of primary tumor	cm^3	0.88 (0.77 to 0.93)	0.88	0.99	0.88	0.99	0.98
Complications							
Myocardial infarction	Yes/no	1.00	1.00	1.00	1.00	1.00	1.00
Sepsis	Yes/no	1.00	1.00	1.00	1.00	1.00	1.00
Stroke	Yes/no	NA	NA	0.96	0.00	1.00	0.96
Venous thromboembolism	Yes/no	0.57 (0.25 to 1.0)	1.00	0.94	0.40	1.00	0.95
Febrile neutropenia	Yes/no	NA	NA	NA	NA	NA	NA
Hypercalcemia	Yes/no	NA	NA	NA	NA	NA	NA
Metastasis							
Metastasis date	dd/mm/yyyy	0.87 (0.78 to 0.94)	0.87	0.87	0.87	0.87	0.87
Metastasis site	Bone, brain, liver, lungs, unknown	0.96 (0.91 to 1.0)	0.96	0.99	0.96	0.99	0.99

Abbreviations: AF, atrial fibrillation; CAD, coronary artery disease; COPD, chronic obstructive pulmonary disorder; CT, computed tomography; DCIS, ductal carcinoma in situ; DM, diabetes; ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; HTN, hypertension; LCIS, lobular carcinoma in situ; MRI, magnetic resonance imaging; PR, progesterone receptor.

cohort in our platform may not be representative of national or global populations. Patients report outcome data on a voluntary basis, with adoption at around 70% and with some disruption associated with the COVID-19 pandemic. Our hospital system did not collect individual-level socioeconomic data, so data on marginalization are based on neighborhood-level estimates, although our ability to link to the most granular census data (district areas) minimizes the risk of ecological fallacy when using the index as an individual-level proxy.[49]

We created an automated, longitudinal, prospective data and analytics platform for breast cancer at a regional cancer center. This platform combines principles of data warehousing with NLP to provide the integrated, timely, meaningful, high-quality, and actionable data required to establish a learning health system.

AFFILIATIONS

[1]Centre for Data Science and Digital Health, Hamilton Health Sciences, Hamilton, Canada
[2]Institute for Health Policy Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
[3]Division of Cardiology, Department of Medicine, Faculty of Health Sciences, McMaster University, Hamilton, Canada
[4]Population Health Research Institute, Hamilton Health Sciences, Hamilton, Canada
[5]Pentavere Research Group, Toronto, Canada
[6]Hamilton Health Sciences, Hamilton, Canada
[7]Escarpment Cancer Research Institute, Hamilton Health Sciences, Hamilton, Canada
[8]Department of Oncology, Faculty of Health Sciences, McMaster University, Hamilton, Canada
[9]Department of Surgery, Faculty of Health Sciences, McMaster University, Hamilton, Canada

CORRESPONDING AUTHOR

Jeremy Petch, PhD, MA, 207-175 Longwood Rd, S Hamilton, ON L8P 0A1; e-mail: Jeremy.petch@utoronto.ca.

SUPPORT

Supported by a grant from Roche Canada.

AUTHOR CONTRIBUTIONS

Conception and design: Jeremy Petch, Christopher Pettengell, Steven Aviv, Peter Sztur, Mark Levine
Administrative support: Peter Sztur
Collection and assembly of data: Jeremy Petch, Joel Kempainnen, Christopher Pettengell, Steven Aviv, Bill Butler, Jessica Bogach, Alexandria Allard-Coutu, Peter Sztur
Data analysis and interpretation: Jeremy Petch, Joel Kempainnen, Christopher Pettengell, Steven Aviv, Greg Pond, Ashirbani Saha, Jessica Bogach, Jonathan Ranisau, Mark Levine
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Jeremy Petch
Research Funding: Roche Canada
Uncompensated Relationships: Mutuo Health

Joel Kempainnen
Research Funding: Roche Canada (Inst), Roche Canada (Inst)

Christopher Pettengell
Employment: Pentavere Research Group Inc
Leadership: Pentavere Research Group Inc

Steven Aviv
Employment: Pentavere
Leadership: Pentavere
Stock and Other Ownership Interests: Pentavere

Greg Pond
Employment: Roche Canada
Stock and Other Ownership Interests: Roche Canada
Honoraria: AstraZeneca
Consulting or Advisory Role: Takeda, Merck, Profound Medical

Peter Sztur
Employment: Apotex Inc

Jonathan Ranisau
Consulting or Advisory Role: Cardea Health

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

We thank Duane Bender, Julie Bosworth, Paul Brown, Tim Dietrich, and Ted Scott for their contributions to this study.

REFERENCES

1. Olsen LA, Aisner D, McGinnis JM. The Learning Healthcare System: Workshop Summary [Internet]. The Learning Healthcare System 1–354, 2007 https://pubmed.ncbi.nlm.nih.gov/21452449/
2. McGinnis JM, Fineberg Hv, Dzau VJ: Advancing the learning health system. N Engl J Med 385:1-5, 2021
3. Levine MN, Alexander G, Sathiyapalan A, et al: Learning health system for breast cancer: Pilot project experience. JCO Clin Cancer Inform 3:1-11, 2019
4. Palakshappa D, Miller DP, Rosenthal GE: Advancing the learning health system by incorporating social determinants. Am J Manag Care 26:e4-e6, 2020
5. Enticott JC, Melder A, Johnson A, et al: A learning health system framework to operationalize health data to improve quality care: An Australian perspective. Front Med (Lausanne) 8:1824, 2021

6. Sheikh A, Anderson M, Albala S, et al: Health information technology and digital innovation for national learning health and care systems. Lancet Digital Health 3:e383-e396, 2021
7. Swords DS, Scaife CL: Granular neighborhood-level socioeconomic data: An opportunity for a different kind of precision oncology? Am J Surg 222:8-9, 2021
8. Liaw ST, Guo JGN, Ansari S, et al: Quality assessment of real-world data repositories across the data life cycle: A literature review. J Am Med Inform Assoc 28:1591-1599, 2021
9. Beauchemin M, Weng C, Sung L, et al: Data quality of chemotherapy-Induced nausea and vomiting documentation. Appl Clin Inform 12:320-328, 2021
10. Shepheard J: Clinical coding and the quality and integrity of health data. Health Inf Manag J 49:3-4, 2020
11. Shiri A, Thornton GM: Canada's Health Data Repositories: Challenges of Organization, Discoverability and Access [Internet]. Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI. 2019 https://journals.library.ualberta.ca/ojs.cais-acsi.ca/index.php/cais-asci/article/view/1111
12. Jones RD, Krenz C, Griffith KA, et al: Governance of a learning health care system for oncology: Patient recommendations. JCO Oncol Pract 17:e479-e489, 2021
13. Kurian AW, Mitani A, Desai M, et al Breast cancer treatment across health care systems: Linking electronic medical records and state registry data to enable outcomes research. Cancer 120:103-111, 2022
14. Weber SC, Seto T, Olson C, et al: Oncoshare: Lessons learned from building an integrated multi-institutional database for comparative effectiveness research. AMIA Annu Symp Proc 970, 2012
15. Hazlehurst BL, Kurtz SE, Masica A, et al: CER Hub: An informatics platform for conducting comparative effectiveness research using multi-institutional, heterogeneous, electronic clinical data. Int J Med Inform 84:763-773, 2015
16. Trifiletti DM, Showalter TN: Big data and comparative effectiveness research in radiation oncology: Synergy and accelerated discovery. Front Oncol 5:274, 2015
17. Richesson RL, Horvath MM, Rusincovitch SA: Clinical research informatics and electronic health record data. Yearb Med Inform 9:215-223, 2014
18. Seneviratne MG, Seto T, Blayney DW, et al: Architecture and implementation of a clinical research data warehouse for prostate cancer. EGEMS (Wash DC) 6:13, 2018
19. Evans RS, Lloyd JF, Pierce LA: Clinical use of an enterprise data warehouse. AMIA Annu Symp Proc 189, 2012
20. Danciu I, Cowan JD, Basford M, et al: Secondary use of clinical data: The Vanderbilt approach. J Biomed Inform 52:28-35, 2014
21. Wang X, Liu L, Fackenthal J, et al: Towards to an Oncology Database (ONCOD) using a data warehousing approach. AMIA Summits Translational Sci Proc 105:2012, 2012
22. Foran DJ, Chen W, Chu H, et al: Roadmap to a comprehensive clinical data warehouse for precision medicine applications in oncology. Cancer Inform 16, 2017 https://pubmed.ncbi.nlm.nih.gov/28469389/
23. Choi IY, Park S, Park B, et al: Development of prostate cancer research database with the clinical data warehouse technology for direct linkage with electronic medical record system. Prostate Int 1:59-64, 2013
24. Boehm KM, Khosravi P, Vanguri R, et al: Harnessing multimodal data integration to advance precision oncology. Nat Rev Cancer 22:114-126, 2022
25. Tuominen S, Uusi-Rauva K, Blom T, et al: Real-world data on diffuse large B-cell lymphoma in 2010–2019: Usability of large data sets of Finnish hospital data lakes. Future Oncol 18:1103-1114, 2022
26. Mayo CS, Kessler ML, Eisbruch A, et al: The big data effort in radiation oncology: Data mining or data farming?. Adv Radiat Oncol 1:260-271, 2016
27. Petch J, Batt J, Murray J, et al: Extracting clinical features from dictated ambulatory consult notes using a commercially available natural language processing tool: Pilot, retrospective, cross-sectional validation study. JMIR Med Inform 7:e12575, 2019
28. Gauthier M-P, Law JH, Le LW, et al: Automating access to real-world evidence. JTO Clin Res Rep 3:100340, 2022
29. Banerjee I, Li K, Seneviratne M, et al: Weakly supervised natural language processing for assessing patient-centered outcome following prostate cancer treatment. JAMIA Open 2:150-159, 2019
30. Hernandez-Boussard T, Kourdis PD, Seto T, et al: Mining electronic health records to extract patient-centered outcomes following prostate cancer treatment. AMIA Annu Symp Proc 876-882, 2017 /pmc./articles/PMC5977629/
31. Gori D, Banerjee I, Chung BI, et al: Extracting patient-centered outcomes from clinical notes in electronic health records: Assessment of urinary Incontinence after radical prostatectomy. EGEMS (Wash DC) 7:43, 2019
32. Lindvall C, Deng C-Y, Agaronnik ND, et al: Deep learning for cancer symptoms monitoring on the basis of electronic health record unstructured clinical notes. JCO Clin Cancer Inform 6, e2100136, 2022
33. Kim BJ, Merchant M, Zheng C, et al: Second prize: A natural language processing program effectively extracts key pathologic findings from radical prostatectomy reports. J Endourology 28:1474-1478, 2014
34. Savova GK, Tseytlin E, Finan S, et al: DeepPhe - a natural language processing system for extracting cancer phenotypes from clinical records. Cancer Res 77:e115, 2017
35. Zhang X, Zhang Y, Zhang Q, et al: Extracting comprehensive clinical information for breast cancer using deep learning methods. Int J Med Inform 132:103985, 2019
36. Zhou S, Wang N, Wang L, et al: CancerBERT: A cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. J Am Med Inform Assoc 29:1208-1216, 2022
37. Hernandez-Boussard T, Blayney DW, Brooks JD: Leveraging digital data to inform and improve quality cancer care. Cancer Epidemiol Biomarkers Prev 29:816-822, 2020
38. Abernethy AP, Gippetti J, Parulkar R, et al: Use of electronic health record data for quality reporting. JCO Oncol Pract 13:530-534, 2017
39. Morin O, Vallières M, Braunstein S, et al: An artificial intelligence framework integrating longitudinal electronic health records with real-world data enables continuous pan-cancer prognostication. Nat Cancer 2:709-722, 2021
40. Nagels J, Wu S, Anderson D, et al: Image exchange in Canada: Examples from the province of Ontario. J Digit Imaging 2022 https://pubmed.ncbi.nlm.nih.gov/35386033/
41. Reis T, Faria I, Serra H, et al: Barriers and facilitators to implementing a continuing medical education intervention in a primary health care setting. BMC Health Serv Res 22:638, 2022
42. Tappen RM, Wolf DG, Rahemi Z, et al: Barriers and facilitators to implementing a change initiative in long-term care utilizing the INTERACT™ quality improvement program. Health Care Manag (Frederick) 36:219-230, 2017
43. Greenberg A, Kramer S, Welch V, et al: Cancer care Ontario's computerized physician order entry system: A province-wide patient safety innovation. Healthc Q 9:108-113, 1132
44. Hui D, Bruera E: The Edmonton symptom assessment system 25 years later: Past, present, and future developments. J Pain Symptom Manage 53:630-643, 2017

45. van Ingen T, Matheson FI: The 2011 and 2016 iterations of the Ontario marginalization index: Updates, consistency and a cross-sectional study of health outcome associations. Can J Public Health 113:260-271, 2022
46. Glassman B: Multidimensional Deprivation in the United States: 2017. American Community Survey Report [Internet], 2019 www.census.gov/acs
47. Finlayson SG, Subbaswamy A, Singh K, et al: The clinician and dataset shift in artificial intelligence. N Engl J Med 385:283-286, 2021
48. Feng J, Phillips Rv, Malenica I, et al: Clinical artificial intelligence quality improvement: Towards continual monitoring and updating of AI algorithms in healthcare. NPJ Digit Med 5:66, 2022
49. Flora M, Moloney G, van Ingen T: User Guide: 2016 Ontario Marginalization Index [Internet]. Toronto, 2022 https://www.publichealthontario.ca/-/media/documents/o/2017/on-marg-userguide.pdf