SPECIAL SERIES: INFORMATICS TOOLS FOR CANCER RESEARCH AND CARE

original reports abstract

Developing an FHIR-Based Computational Pipeline for Automatic Population of Case Report Forms for Colorectal Cancer Clinical Trials Using Electronic Health Records

Nansu Zong, PhD[1]; Andrew Wen, MS[1]; Daniel J. Stone, BS[1]; Deepak K. Sharma, MS[1]; Chen Wang, PhD[1]; Yue Yu, PhD[1]; Hongfang Liu, PhD[1]; Qian Shi, PhD[1]; and Guoqian Jiang, MD, PhD[1]

PURPOSE The Fast Healthcare Interoperability Resources (FHIR) is emerging as a next-generation standards framework developed by HL7 for exchanging electronic health care data. The modeling capability of FHIR in standardizing cancer data has been gaining increasing attention by the cancer research informatics community. However, few studies have been conducted to examine the capability of FHIR in electronic data capture (EDC) applications for effective cancer clinical trials. The objective of this study was to design, develop, and evaluate an FHIR-based method that enables the automation of the case report forms (CRFs) population for cancer clinical trials using real-world electronic health records (EHRs).

MATERIALS AND METHODS We developed an FHIR-based computational pipeline of EDC with a case study for modeling colorectal cancer trials. We first leveraged an existing FHIR-based cancer profile to represent EHR data of patients with colorectal cancer, and then we used the FHIR Questionnaire and QuestionnaireResponse resources to represent the CRFs and their data population. To test the accuracy of and overall quality of the computational pipeline, we used synoptic reports of 287 Mayo Clinic patients with colorectal cancer from 2013 to 2019 with standard measures of precision, recall, and F1 score.

RESULTS Using the computational pipeline, a total of 1,037 synoptic reports were successfully converted as the instances of the FHIR-based cancer profile. The average accuracy for converting all data elements (excluding tumor perforation) of the cancer profile was 0.99, using 200 randomly selected records. The average F1 score for populating nine questions of the CRFs in a real-world colorectal cancer trial was 0.95, using 100 randomly selected records.

CONCLUSION We demonstrated that it is feasible to populate CRFs with EHR data in an automated manner with satisfactory performance. The outcome of the study provides helpful insight into future directions in implementing FHIR-based EDC applications for modern cancer clinical trials.

JCO Clin Cancer Inform 4:201-209. © 2020 by American Society of Clinical Oncology

INTRODUCTION

It has been increasingly recognized that the common data collection and management methods used by the oncology clinical trial community are laborious, imprecise, and expensive, which greatly hinders the implementation of novel trial design and the achievement of study integrity and reproducibility for trial. For example, rapid data inputs are required for adaptive design, especially when the adaptation of the trial depends on accurate individual patients' outcome data status, which can be obtained quickly. The case report forms (CRFs) are questionnaires specifically used by researchers in clinical trial research to collect information about each participating patient.[1] The development and population of the CRFs play a significant role in the selection of participants.[2] One trend is to design the electronic data capture (EDC)-oriented electronic forms using model-driven solutions, such as FHIRForm,[3] Research Electronic Data Capture,[4] and OpenClinica.[5] However, the rapid growth in the scale of medical data and collaborations between different medical facilities for developing clinical trials creates new challenges. These challenges include integrating data across different EHR systems as well as ensuring the data required by the protocol and study-specific hypothesis is attributed in an efficient way.[4,6,7] There is a benefit of having a standard for the format as well as a clear value definition for data and responses of CRFs to facilitate EDC, which allows plug and play functionality for any developed CRFs, because they will be interuseable between the different institutions as long as data are generated in that standardized format.[8,9]

There are a few efforts to provide a standardized data model for the secondary use of electronic health

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

Accepted on February 4, 2020 and published at ascopubs.org/journal/cci on March 5, 2020: DOI https://doi.org/10.1200/CCI.19.00116

This study has been reviewed and approved by the Mayo Clinic Institutional Review Board.

CONTEXT

Key Objective

The objective of this study was to examine the existing cancer model based on a design of pipeline to harmonize with real-world electronic health records (EHRs) for the automation of the case report forms (CRFs) population for cancer clinical trials.

Knowledge Generated

We demonstrated it is feasible to populate CRFs with Fast Healthcare Interoperability Resources (FHIR)-based EHR data in an automated manner with high performance. We observed limited information loss in the extract, transform, load process to generate a standard-based pathology-report data representation, because there was a similar performance for the CRF population with the standardized representation versus raw pathology data.

Relevance

With the FHIR-based CRF population pipeline prototyped in this study, the data collection, transformation, and quality assurance process became streamlined and generalizable to support further adaptation of other cancer types.

record (EHR) data, such as Informatics for Integrating Biology and the Bedside,[10] the Observational Health Data Sciences and Informatics Common Data Model,[11] and Fast Healthcare Interoperability Resources (FHIR).[12] Notably among these, FHIR is emerging as the next-generation standards framework for exchanging electronic health care data. FHIR defines data formats and elements, known as resources, as well as messages to exchange medical records. FHIR provides a standard data communication method that directly delivers discrete data elements, such as Patient, Diagnosis, Procedure, and Medication, rather than the traditional document-centric methods, and enables data to be quickly transitioned and easily parsed by analytics platforms.[8,12] The importance of standardization for cancer phenotypic data has been increasingly recognized by the cancer research informatics community. For example, the Clinical Data Interchange Standards Consortium has published a number of therapeutic area standards for cancers.[13] The Royal College of Pathologists of Australasia (RCPA)/HL7 Australia[14] has released the cancer profiles for structured colorectal and prostate reports.

To meet the EDC requirements for cancer clinical trials, the existing models need to be carefully examined and harmonized with real-world EHR and cancer trial data. In particular, the examination of the capability of FHIR in EDC applications for effective cancer clinical trials needed to be conducted. The objective of the study was to design, develop, and evaluate an FHIR-based method that enables the automation of the CRF population for cancer clinical trials using real-world EHRs. As such, we developed a computational pipeline with two corresponding efforts conducted: (1) FHIR-based data representation for Mayo Clinic patients with colorectal cancer, based on an existing cancer profile, the Australian Colorectal Cancer Profile[14]; and (2) FHIR-based CRFs' representation and population. As a proof of concept, we conducted a case study for modeling cancer phenotypic data from Mayo Clinic pathologic reports and populated the CRFs designed for colorectal cancer trials.[15]

MATERIALS AND METHODS

Data

Mayo Clinic's Unified Data Platform (UDP)[16] is a clinical data warehouse that provides a combined view of heterogeneous data across multiple data sources, including Epic Clarity, through effective data orchestration. UDP provides access to all the information on patients with colorectal cancer. In practice, we only extract the surgical and pathologic information in this study. For surgical information, we collected surgical reports for obtaining cancer-specific data required by clinical trials. For pathologic information, we mainly used a semistructuralized form known as a synoptic report, as well as the original pathologic report as a supplementary source. A synoptic report is an internal effort by the Mayo Clinic since 2013 to enforce compliance with the College of American Pathologists' protocols[17] on exactly what data elements must be included and generally have roughly templated values[18] when documenting certain cancers within pathology reports. The protocol of clinical data access was approved by the Mayo Clinic Institutional Review Board.

FHIR-Based Standardization and Tools

The resources defined by FHIR cover a wide range of concepts that are clinically related (eg, Clinical) and supportive (eg, Foundation, Base, Financial, and Specialized). Recognized as classes, those concepts and their subconcepts (ie, subclasses) can better interpret and facilitate the use of resources. For example, the two resources *Observation* and *DiagnosticReport* belong to the subconcept *Diagnostic* under the concept *Clinical* to classify the resources. To model the EHR data, a resource is described with a set of attributes. Each attribute is limited to be valued with certain predefined data types, such as "*string*," "*dateTime*," and "*Reference*," "*CodeableConcepts*," and "*code*." The popular clinically related terminologies, such as SNOMED CT,[19] Logical Observation Identifiers Names and Codes (LOINC),[20] and International Classification of Diseases–9 and –10,[21] are adopted as preferred vocabularies.

HL7 application programming interface (HAPI)-FHIR[22] is an open-source Java library implementation of the FHIR specification for data modeling, parsing, and management. HAPI-FHIR supports the following tools: (1) parser and encoder to convert between the source data model and FHIR-based data model, and (2) communicate between the client application and server.

Australian Colorectal Cancer Profile

RCPA has initiated an effort for adopting the use of structured cancer reporting and the use of FHIR to facilitate the data exchange.[14] As mentioned, RCPA has released the cancer profiles for structured colorectal and prostate reports. The colorectal cancer protocol used in this study is based on the structured colorectal cancer profile released by RCPA.[23] A logical model that captures the concepts and defines the value sets for the published protocol is formed and is further be represented with FHIR. The atomic data items in the report are mainly represented as the FHIR resources *DiagnosticReport* and *Observation*.

Automatic Population of CRFs

We proposed an FHIR-based method that enables the standard representation of data elements from the pathologic report for cancer trials and functionalizes the automation of the data population of CRFs to support real-world cancer trials. There are three main steps, as shown in Figure 1: (1) cancer data preparation, (2) FHIR-based data profiling, and (3) FHIR-based CRF data population.

First, pathologic and surgical reports are extracted from UDP. The original unstructured reports are converted to structured reports that better describe the data with the schema (ie, data elements) by natural language processing (NLP) tools. To allow this study to focus primarily on the automated data population of CRFs in an interoperable manner, in practice, instead of directly applying NLP tools to obtain the structured information from pathologic reports, we used the semi-structured synoptic reports that represent the pathologic reports with a structured presentation. From the synoptic reports, a total of 21 data elements, such as tumor site, tumor size, and surgical margins, are directly extracted and used to populate the data model. For the data elements that cannot be covered by synoptic reports, we also use the original pathologic reports to complement the capture of additional information based on simple NLP-based methods. Similarly, we applied the same NLP-based methods for pathologic reports to process surgical reports.

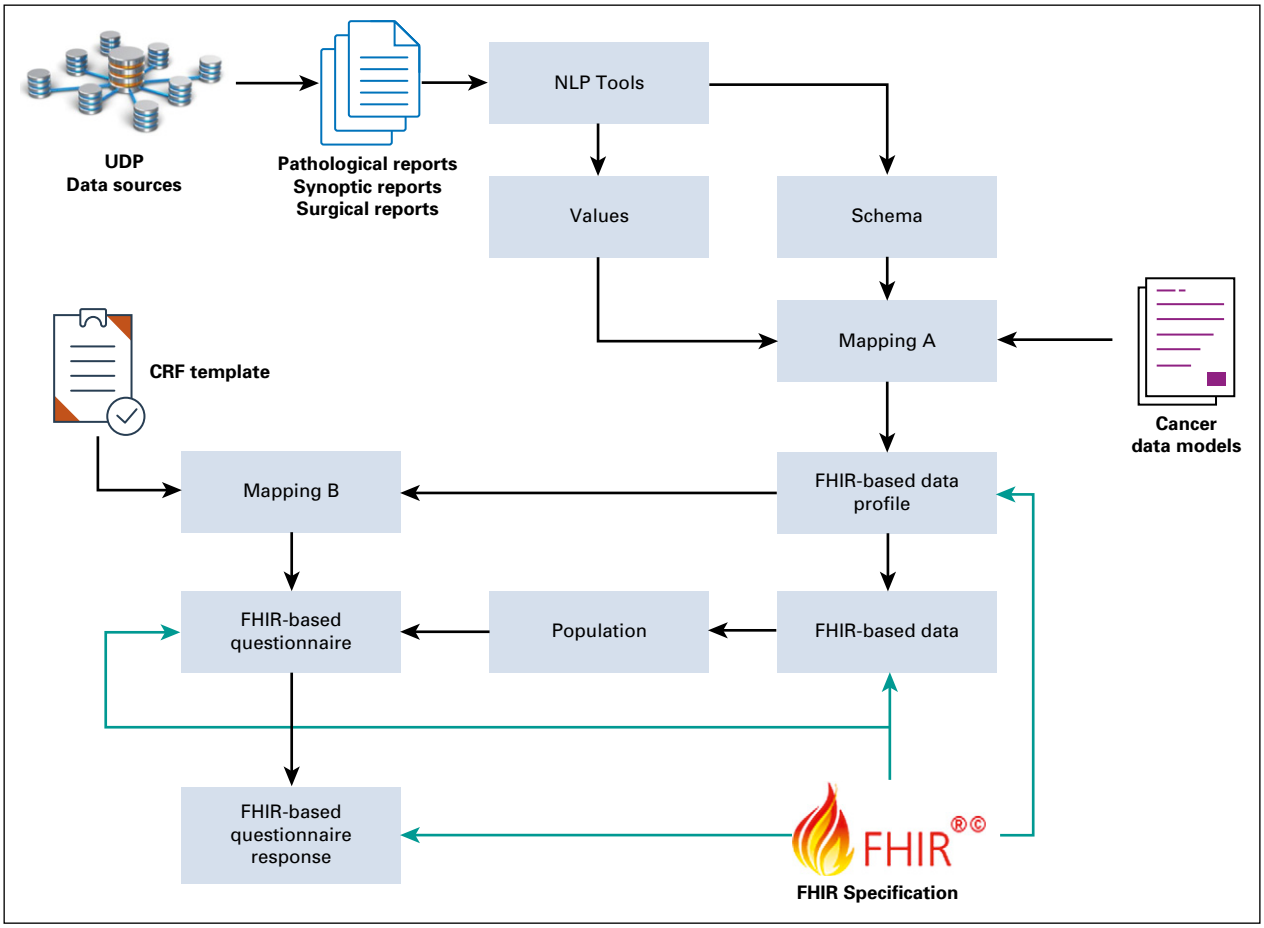

FIG 1. A Fast Healthcare Interoperability Resources (FHIR)–based pipeline for automatic data population of CRFs. CRF, case report form; NLP, natural language processing; UDP, Unified Data Platform.

Second, the Australian Colorectal Cancer Profile (ACP) was used as the data model to describe and standardize the data extracted from the pathologic reports. In practice, a small subset of 21 data elements from ACP was extracted to map the element from the synoptic report manually. For example, the element tumor site from the synoptic report is mapped to *Colorectal.preAnalytic.tumourLocation* and *Colorectal.macro.tumourSite*. The complete list of mapping is presented in Table 1. We adopted the mappings of data elements between ACP and FHIR (defined in the ACP) to enable the FHIR-based representation of data. For example, *Colorectal.macro.tumourSite* can be represented with the FHIR resource *Observation* with the value *LOINC#33725-3 (Tumor site)*. The detailed mapping table and conversion script can be found at http://hl7.org.au/fhir/rcpa/cmap.html#summary. The data conversion was based on standard string processing, such as tokenization, stemming (ie, lemmatization), and dictionary look-up to represent the raw values with standardized terminology.

Last, in step 3, a CRF is represented with an FHIR resource *Questionnaire,* and the answers from each patient are generated via a mapping between the question items and the data elements of ACP. For example, "Primary Site(s) of a patient" is mapped to *Colorectal.preAnalytic.tumourLocation* and *Colorectal.macro.tumourSite*. The detailed mappings for nine questions are listed in Table 2. Note that, for questions 7 and 8

TABLE 1. Mapping Table A: Synoptic Report to Colorectal Cancer Profile

ID	Cancer Model Element	Synoptic Report Element	Sample Value
1	Colorectal.macro.otherMacroComments	Specimen	Sigmoid colon, rectum
2	Colorectal.preAnalytic.typeOfOperation	Procedure	Rectosigmoid colectomy (low anterior resection)
3	Colorectal.preAnalytic.tumourLocation/ Colorectal.macro.tumourSite	Tumor site	Cecum
4	Colorectal.macro.maxTumourDiameter	Tumor size	4- × 3.5- × 0.6-cm tumor bed (rare microscopic foci within)
5	Colorectal.macro.tumourPerforation	Macroscopic tumor perforation	Not identified.
6	Colorectal.macro.intactnessOfMesorectum	Macroscopic intactness of mesorectum	
7	Colorectal.micro.tumourType	Histologic type	Signet-ring cell carcinoma
8	Colorectal.micro.histologicalGrade	Histologic grade	High grade (poorly differentiated to undifferentiated)
9	Colorectal.micro.maxDegreeLocalInvasion	Microscopic tumor extension	Tumor invades through muscularis propria into pericolorectal soft tissue, but does not extend to serosal surface
10	Colorectal.micro.proximalOrDistalResectionMargins/ Colorectal.micro.involvedMargins/ Colorectal.micro.marginsMicroClearance / Colorectal.micro.nonperitonealisedCircumMargin	Surgical margins	All margins negative for tumor, 2 cm from radial
11	Colorectal.micro.neoadjuvantTherapy	Treatment effect	
12	Colorectal.micro.lymphNodeInvolvement / Colorectal.micro.venousSmallVesselInvasion / Colorectal.micro.intramuralVeinInvasion / Colorectal.micro.extramuralVeinInvasion / Colorectal.micro.smallVesselInvasion	Lymphovascular invasion	Not identified
13	Colorectal.micro.perineuralInvasion	Perineural invasion	Not identified
14	Colorectal.micro.extramuralTumourDeposits	Tumor deposits	Indeterminate
15	Colorectal.micro.polypDetails	Type of polyp tumor arises from	None identified
16	Colorectal.synthesisOverview.tumourStagingSystem	Pathologic staging (*AJCC Cancer Staging Manual*, 7th edition)	
17	Colorectal.synthesisOverview.tumourStageT	Primary tumor	pT3
18	Colorectal.synthesisOverview.tumourStageN	Regional lymph nodes	pN1c
19	Colorectal.micro.lymphNodesDetails.numExamined	Number examined (total)	38
20	Colorectal.micro.lymphNodesDetails.numPos	Number involved (total)	4
21	Colorectal.micro.histoConfDistMetastases / Colorectal.micro.histoConfDistMetastasesSite / Colorectal.synthesisOverview.tumourStageM	Distant metastasis	

Abbreviations: AJCC, American Joint Committee on Cancer; ID, identifier.

regarding the procedure information of a patient, we extended the ACP with an extension defined in the FHIR resource *Observation* to represent procedure information. We used the FHIR resource "*QuestionnaireResponse*" to represent answers to the questions in the CRF, as shown in our example in Figure 2.

Creation of Mappings

We created two sets of mappings, as shown in Tables 1 and 2: mapping A and mapping B.

Mapping A (Table 1) provides a map for the information from the structure of the synoptic report to the ACP. For example, pathologic information, such as tumor site and microscopic tumor extension, is obtained from a synoptic report and is mapped to the corresponding element in ACP. Mapping B (Table 2) provides a map from the ACP FHIR model to the CRF questions, which is used to populate the CRFs represented by FHIR "*QuestionnaireResponse*." In this study, we used the colorectal cancer adjuvant on a study form used in the study by Alberts et al,[15] which contains nine primary questions: "Q1 Primary Site(s)," "Q2 Was there bowel obstruction," "Q3 Was there bowel perforation," "Q4 Disease extent," "Q5 Number of lymph nodes examined," "Q6 Number of positive lymph nodes," "Q7 Surgery date," "Q8 Type of procedure," and "Q9 Comments." Please note, to completely populate the CRF with data, we also used the original pathologic reports and surgical reports as the supplementary data sources for the corresponding mappings to address the four questions, Q2, Q7, Q8, and Q9. Because the values for the answers are restricted, such as "Yes" for "Was there bowel obstruction" and "Open/ Laparoscopic" for "Type of procedure," we were able to design an extraction strategy based on the simple rules to obtain the answers.

Experiments

To test our pipeline, we extracted 1,037 colorectal cancer synoptic reports of 287 Mayo Clinic patients from UDP dated from 2013 to 2019. We obtained the data from a UDP search based on the colorectal cancer–related International Classification of Diseases–9 codes filtered in compliance with Mayo Clinic's policies of research authorization. We ran our pipeline successfully against 1,037 synoptic reports and populated them into the FHIR-based cancer profile instances. Our pipeline further populated the FHIR Questionnaire resource (represents the answers to the CRF) with data points from both the FHIR cancer profile instances and structured data.

TABLE 2. Mapping Table B: Colorectal Cancer Profile to Case Report Form

ID	Form Question	Value	Element	Source
1	Primary site(s)	Cecum, transverse colon, sigmoid colon, ascending colon, splenic flexure, hepatic flexure, descending colon	Colorectal.preAnalytic.tumourLocation/ Colorectal.macro.tumourSite	Synoptic report
2	Was there bowel obstruction?	Yes, no	Colorectal.preAnalytic.clinicalObstruction	Pathologic report
3	Was there bowel perforation?	Yes, no	Colorectal.macro.tumourPerforation	Synoptic report
4	Disease extent	Tumor invades submucosa (T1), tumor invades muscularis propria (T2), tumor invades through the muscularis propria into the subserosa, or into nonperitonealized pericolic or perirectal tissue (T3), tumor directly invades or is adherent to other organs or structures and/or involves the visceral peritoneum (T4), primary tumor cannot be assessed (TX)	Colorectal.micro.maxDegreeLocalInvasion/ Colorectal.synthesisOverview.tumourStageT	Synoptic report
5	Number of lymph nodes examined	Integer	Colorectal.micro.lymphNodesDetails.numExamined	Synoptic report
6	Number of positive lymph nodes	Integer	Colorectal.micro.lymphNodesDetails.numPos	Synoptic report
7	Surgery date (date primary tumor removed; < 56 days before random assignment)	Date	Colorectal.surgical.	Surgical report
8	Type of procedure	Open, laparoscopic	Colorectal.surgical.description	Surgical report
9	Comments	Text	Colorectal.synthesisOverview.overarchingComment	Pathologic report

Abbreviation: ID, identifier.

```
{
  "resourceType": "Bundle",
  "entry": [
    {
      "resource": {
        "resourceType": "QuestionnaireResponse",
        "item": [
          {
            "text": "Primary Site(s)",
            "answer": [
              {
                "valueCoding": {
                  "system": "https://snomed.info/sct",
                  "code": "34402009",
                  "display": "Rectum"
                }
              }
            ]
          },
          {
            "text": "Was there bowel obstruction?",
            "answer": [
              {
                "valueCoding": {
                  "system": "https://snomed.info/sct",
                  "code": "2667000",
                  "display": "No"
                }
              }
            ]
          },
          {
            "text": "Was there bowel perforation?",
            "answer": [
              {
                "valueCoding": {
                  "system": "https://snomed.info/sct",
                  "code": "2667000",
                  "display": "No"
                }
              }
            ]
          },
          {
            "text": "Disease Extent",
            "answer": [
              {
                "valueCoding": {
                  "system": "https://snomed.info/sct",
                  "code": "395707006",
                  "display": "Tumor invades through the muscularis propria
                  into the subserosa, or into non-peritonealized pericolic
                  or perirectal tissues (T3)"
                }
              }
            ]
          },
          {
            "text": "Number of Lymph Nodes Examined",
            "answer": [
              {
                "valueQuantity": {
                  "value": 38
                }
              }
            ]
          },
          {
            "text": "Number of Positive Lymph Nodes",
            "answer": [
              {
                "valueQuantity": {
                  "value": 4
                }
              }
            ]
          },
          {
            "text": "Type of Procedure",
            "answer": [
              {
                "valueString": "Open"
              }
            ]
          },
          {
            "text": "Surgery Date (date primary tumor removed;
            < 56 days prior to randomization)",
            "answer": [
              {
                "valueDate": "2019-**-**"
              }
            ]
```


FIG 2. An example of the Fast Healthcare Interoperability Resources–based case report form populated with data.

Figure 2 shows an example of a populated FHIR-based CRF. The form is organized with FHIR resource *Bundle*, which contains an FHIR resource *QuestionnaireResponse* that represents the responses as the entry. Each question in CRF is considered a question item, which includes the original question and the filled answers. For example, Q1 with an answer "Rectum" is represented in the first item. The values are represented by the standardized values of coding systems. In the same example, "Rectum" is encoded as "34402009" from SNOMED CT.[19] We applied diverse data formats to support the presentation of values, such as quantity (eg, "Q6"), string (eg, "Q8"), date (eg, "Q7") and text (eg, "Q9").

We designed two tasks to evaluate the accuracy of the ACP conversion in the module and the overall quality of the automatically populated CRFs.

Task 1: ACP-based data conversion. We randomly selected 200 records, with any sensitive information removed that could be used to identify the patient, to evaluate how precisely the values of all the data elements are converted. The results were reviewed by two subject matter experts (N.Z. and G.J.) majoring in medical informatics and metadata harmonization. The κ score of the two experts for interagreement was 0.90. The reviewers marked the true positive (TP), true negative (TN), false positive (FP), false negative (FN) defined as follows:

- If the value is "present" and the pipeline correctly parsed it, we labeled it as TP.
- If the value is "present" and the pipeline wrongly parsed it, we labeled it as FN.
- If the value is "absent" and the pipeline correctly recognized it, we labeled it as TN.
- If the value is "absent" and the pipeline wrongly asserted a value, we labeled it as FP.

Then, the accuracy of conversion could be evaluated for each element on the basis of the counts of those measures by reviewers.

Task 2: CRF data population. We randomly selected 100 deidentified records to evaluate the overall quality of the populated CRFs. The results were reviewed by two medical informatics experts (Y.Y. and G.J.); the interagreement κ score was 0.97. The reviewers were required to complete the answers for each question on the basis of an investigation of the given original synoptic reports. The answers annotated by the reviewers were considered the gold standard, and the precision, recall, and F1 score were calculated correspondingly.

RESULTS

Task 1: ACP-Based Data Conversion

As Table 3 shows, most of the values mapped perfectly to the corresponding concepts as defined in the cancer profile

for all the elements (average accuracy, 0.99) when tumor perforation was excluded (0.09). Looking into the reasons for the low performance of tumor perforation, we found that a mapping rule was missing for the values of the tumor perforation, resulting in most of the FNs recorded. Other FNs were caused by unexpected data formats or data values, such as a one-dimensional value instead of three-dimensional values for *maxTumourDiameter* or *ileum*, *rectosigmoid* colon used for tumor site.

Task 2: CRF Data Population

As Table 4 shows, those elements populated on the basis of the source of structured sources, such as surgical date and type, and bowel obstruction, received perfect precision (1.0), recall (1.0), and F1-score (1.0) values. The elements populated on the basis of the synoptic report received the average F1 score of 0.95.

DISCUSSION

In this study, we used the FHIR standard as a unified framework for automating the CRFs data population for cancer clinical trials. We performed a case study and implemented a computational pipeline focusing on the population of an on-study CRF form for a colorectal cancer trial. We demonstrated that it is feasible to populate CRFs with EHR data in an automated manner, as can be seen by the high performance shown in Table 4, based on the adoption of the FHIR as a standard data access mechanism. Please note, the purpose of the experiment and the collected performance metrics is to demonstrate the reliability of our proposed method for automatically generating the CRFs from raw data. Because the mappings and NLP-based string processing play an important role, despite the performance metric being intended to cover our whole pipeline, the results were influenced by the performance of the NLP methods used. The reason for choosing FHIR can be summarized as follows: (1) FHIR is one of the most popular data standards in the medical field, with all major EHRs adopting FHIR for health care data exchange; and (2) FHIR is a messaging standard that allows information to be captured as it is generated. As opposed to the Observational Health Data Sciences and Informatics Common Data Model and Integrating Biology and the

TABLE 3. Accuracy of Conversion of FHIR-Based Cancer Profile

ID	Element of FHIR-Based Cancer Profile	TP	TN	FP	FN	Accuracy
1	Colorectal.macro.otherMacroComments	200	0	0	0	1.0
2	Colorectal.preAnalytic.typeOfOperation	200	0	0	0	1.0
3	Colorectal.preAnalytic.tumourLocation/ Colorectal.macro.tumourSite	195	0	0	4	0.98
4	Colorectal.macro.maxTumourDiameter	180	7	1	5	0.97
5	Colorectal.macro.tumourPerforation	9	8	0	182	0.09
6	Colorectal.macro.intactnessOfMesorectum	24	176	0	0	1.0
7	Colorectal.micro.tumourType	189	11	0	0	1.0
8	Colorectal.micro.histologicalGrade	190	9	0	0	1.0
9	Colorectal.micro.maxDegreeLocalInvasion	187	7	0	0	1.0
11	Colorectal.micro.neoadjuvantTherapy	191	7	0	0	1.0
12	Colorectal.micro.lymphNodeInvolvement / Colorectal.micro.venousSmallVesselInvasion / Colorectal.micro.intramuralVeinInvasion / Colorectal.micro.extramuralVeinInvasion / Colorectal.micro.smallVesselInvasion	200	0	0	0	1.0
13	Colorectal.micro.perineuralInvasion	190	8	0	0	1.0
14	Colorectal.micro.extramuralTumourDeposits	193	4	0	0	1.0
15	Colorectal.micro.polypDetails	0	134	0	17	0.89
16	Colorectal.synthesisOverview.tumourStagingSystem	1	0	0	0	1.0
17	Colorectal.synthesisOverview.tumourStageT	190	0	0	0	1.0
18	Colorectal.synthesisOverview.tumourStageN	199	0	0	0	1.0
19	Colorectal.micro.lymphNodesDetails.numExamined	195	0	0	0	1.0
20	Colorectal.micro.lymphNodesDetails.numPos	190	0	0	0	1.0
21	Colorectal.micro.histoConfDistMetastases / Colorectal.micro.histoConfDistMetastasesSite / Colorectal.synthesisOverview.tumourStageM	200	0	0	0	1.0

Abbreviations: FHIR, Fast Healthcare Interoperability Resources; FN, false negative; FP, false positive; ID, identifier; TN, true negative; TP, true positive.

TABLE 4. Precision, Recall, and F1 Score of the Population of Case Report Forms

ID	Form Question	Validation Source	Precision	Recall	F1 Score
1	Primary site(s)	Synoptic report	0.93	0.96	0.95
2	Was there bowel obstruction?	Synoptic report (diagnosis)	1.0	1.0	1.0
3	Was there bowel perforation?	Synoptic report	0.96	0.96	0.96
4	Disease extent	Synoptic report	0.97	0.94	0.95
5	Number of lymph nodes examined	Synoptic report	0.93	0.93	0.93
6	Number of positive lymph nodes	Synoptic report	0.95	0.95	0.95
7	Surgery date (date primary tumor removed; < 56 days before random assignment)	Surgical report	1.0	1.0	1.0
8	Type of procedure:	Surgical report	1.0	1.0	1.0
9	Comments	Synoptic report (comment)	1.0	1.0	1.0

Abbreviation: ID, identifier.

Bedside, which typically are the secondary use of EHR data requiring batched queries, the mechanism of FHIR makes it easier to populate the data model and keep it updated. With the FHIR-based CRF population pipeline prototyped in this study, the data collection, transformation, and quality assurance process became streamlined and can be replaced.

Despite the reliability of the proposed framework, as demonstrated in this study, there are a number of limitations in this study. First, we just used a single on-study CRF with limited data elements included. Although this study is a proof of concept, a single CRF is certainly not enough to evaluate the coverage of the existing cancer profiles. In the future, we will look into more CRFs to fully understand the data element requirements that should be able to inform the enhancement of cancer profile development. Second, we adopted the ACP to cover most of the elements from synoptic reports. However, our target CRF asks for more data elements than those covered. We argue that in the future, a cancer profile with a comprehensive list of data elements from both structured and unstructured data should be developed. Third, most of the cancer-specific phenotypic data are largely embedded in the unstructured clinical narratives. In this study, we used the synoptic reports Mayo Clinic has implemented.[17] Although we have argued that the use of synoptic reports can greatly reduce the complexity for the pipeline implementation, as well as cover many high-quality data elements, we realize the necessity of using advanced NLP tools, because the structured or semi-structured reports (eg, synoptic reports) are not always available. We have developed an FHIR-based clinical-data normalization tool that enables the extraction of structured information from unstructured medical notes,[24,25] which can be adapted. In addition, more sophisticated tools like MedKAT[26] and DeepPhe[27] need to be examined for processing the narratives in diverse notes to extend the coverage of more questions. Last, the mapping rules and evaluation methods developed in this study were based on the consensus from our study team and done on a small scale as a proof of concept. In the future, a more sophisticated and rigorous method and evaluation (eg, a community-based consensus) will need to be designed and conducted.

AFFILIATION

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, MN

CORRESPONDING AUTHOR

Guoqian Jiang, MD, PhD, Department of Health Sciences Research, Harwick Building, Mayo Clinic Minnesota, Rochester, MN 55901; e-mail: jiang.guoqian@mayo.edu

SUPPORT

Supported in part by a National Institutes of Health Big Data to Knowledge (BD2K) grant (Grant No. U01 HG009450; G.J.).

AUTHOR CONTRIBUTIONS

Conception and design: Nansu Zong, Andrew Wen, Hongfang Liu, Guoqian Jiang
Financial support: Guoqian Jiang
Collection and assembly of data: Nansu Zong, Andrew Wen, Daniel J. Stone, Guoqian Jiang
Data analysis and interpretation: Nansu Zong, Andrew Wen, Daniel J. Stone, Deepak K. Sharma, Chen Wang, Yue Yu, Qian Shi, Guoqian Jiang
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).

Qian Shi
Stock and Other Ownership Interests: Amgen, Johnson & Johnson
Consulting or Advisory Role: Yiviva
Research Funding: Celgene (Inst), Roche (Inst)

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

We thank Grahame Grieve for his guidance on the access of the Australian colorectal cancer profile.

REFERENCES

1. Grimes DA, Hubacher D, Nanda K, et al: The Good Clinical Practice guideline: A bronze standard for clinical research. Lancet 366:172-174, 2005
2. Bellary S, Krishnankutty B, Latha MS: Basics of case report form designing in clinical research. Perspect Clin Res 5:159-166, 2014
3. Eapen BR, Costa A, Archer N, et al: FHIRForm: An open-source framework for the management of electronic forms in healthcare. Stud Health Technol Inform 257:80-85, 2019
4. Harris PA, Taylor R, Thielke R, et al: Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 42:377-381, 2009
5. Gessner S, Storck M, Hegselmann S, et al: Automated transformation of CDISC ODM to OpenClinica. Stud Health Technol Inform 243:95-99, 2017
6. El Emam K, Jonker E, Sampson M, et al: The use of electronic data capture tools in clinical trials: Web-survey of 259 Canadian trials. J Med Internet Res 11:e8, 2009
7. Nahm ML, Pieper CF, Cunningham MM: Quantifying data quality for clinical trials using electronic data capture. PLoS One 3:e3049, 2008
8. Mandel JC, Kreda DA, Mandl KD, et al: SMART on FHIR: A standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 23:899-908, 2016
9. Rea S, Pathak J, Savova G, et al: Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. J Biomed Inform 45:763-771, 2012
10. 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 19:181-185, 2012
11. Stang PE, Ryan PB, Racoosin JA, et al: Advancing the science for active surveillance: Rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med 153:600-606, 2010
12. Bender D, Sartipi K: HL7 FHIR: An agile and RESTful approach to healthcare information exchange. Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems. 2013:326-331, 2013.
13. CDISC: Therapeutic areas. 2019. https://www.cdisc.org/standards/therapeutic-areas
14. ADHA/HL7 Australia: HL7 Australia Implementation Guide. 2014. http://fhir.hl7.org.au/fhir/rcpa/index.html
15. Alberts SR, Sargent DJ, Nair S, et al: Effect of oxaliplatin, fluorouracil, and leucovorin with or without cetuximab on survival among patients with resected stage III colon cancer: S randomized trial. JAMA 307:1383-1393, 2012
16. Kaggal VC, Elayavilli RK, Mehrabi S, et al: Toward a learning health-care system–knowledge delivery at the point of care empowered by big data and NLP. Biomed Inform Insights 8:13-22, 2016 (suppl 1).
17. College of American Pathologists: Cancer Protocol Templates. 2019. https://www.cap.org/cancerprotocols
18. Srigley JR, McGowan T, Maclean A, et al: Standardized synoptic cancer pathology reporting: A population-based approach. J Surg Oncol 99:517-524, 2009
19. Donnelly K: SNOMED-CT: The advanced terminology and coding system for eHealth. Stud Health Technol Inform 121:279-290, 2006
20. McDonald CJ, Huff SM, Suico JG, et al: LOINC, A universal standard for identifying laboratory observations: A 5-year update. Clin Chem 49:624-633, 2003
21. World Health Organization: International classification of diseases (ICD). 2006. http://www.who.int/classifications/icd/en/
22. University Health Network: HAPI-FHIR library homepage. http://hapifhir.io/
23. Royal College of Pathologists of Australasia: Cancer protocols. https://www.rcpa.edu.au/Library/Practising-Pathology/Structured-Pathology-Reporting-of-Cancer/Cancer-Protocols
24. Hong N, Wen A, Shen F, et al: Integrating structured and unstructured EHR data using an FHIR-based type system: A case study with medication data. AMIA Jt Summits Transl Sci Proc 2017:74-83, 2018
25. Hong N, Wen A, Mojarad MR, et al: Standardizing heterogeneous annotation corpora using HL7 FHIR for facilitating their reuse and integration in clinical NLP. AMIA Annu Symp Proc 2018:574-583, 2018
26. Coden A, Savova G, Sominsky I, et al: Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model. J Biomed Inform 42:937-949, 2009
27. 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-e118, 2017