IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

SNOMED CT-Based Standardized e-Clinical Pathways for Enabling Big Data Analytics in Healthcare

AYMAN ALAHMAR, (MEMBER, IEEE) AND RACHID BENLAMRI, (MEMBER, IEEE)
Lakehead University, Thunder Bay, ON P7B5E1 Canada

Corresponding author: Ayman Alahmar (e-mail: aalahmar@lakeheadu.ca).

ABSTRACT Automation of healthcare facilities represents a challenging task of streamlining a highly information-intensive sector. Modern healthcare processes produce large amounts of data that have great potential for health policymakers and data science researchers. However, a considerable portion of such data is not captured in electronic format and hidden inside the paperwork. A major source of missing data in healthcare is paper-based clinical pathways (CPs). CPs are healthcare plans that detail the interventions for the treatment of patients, and thus are the primary source for healthcare data. However, most CPs are used as paper-based documents and not fully automated. A key contribution towards the full automation of CPs is their proper computer modeling and encoding their data with international clinical terminologies. We present in this research an ontology-based CP automation model in which CP data are standardized with SNOMED CT, thus enabling machine learning algorithms to be applied to CP-based datasets. CPs automated under this model contribute significantly to reducing data missingness problems, enabling detailed statistical analyses on CP data, and improving the results of data analytics algorithms. Our experimental results on predicting the Length of Stay (LOS) of stroke patients using a dataset resulting from an e-clinical pathway demonstrate improved prediction results compared with LOS prediction using traditional EHR-based datasets. Fully automated CPs enrich medical datasets with more CP data and open new opportunities for machine learning algorithms to show their full potential in improving healthcare, reducing costs, and increasing patient satisfaction.

INDEX TERMS Clinical Pathway, Data Analytics, Decision Tree, Health Level 7, Length of Stay, Machine Learning, Semantic Web, SNOMED CT

I. INTRODUCTION

HEALTHCARE processes produce a large amount of data that has great potential for healthcare administrators, health policymakers, and big data researchers. However, a great portion of such data is not properly captured, missing in electronic format, and hidden inside paperwork and forms. It was the hope of Electronic Health Record (EHR) systems to store that vast amount of data in digital format. However, that target has never been achieved despite the fact that EHRs act as the central component of Health Information Systems (HIS) for decades. An important study on missing clinical and behavioral health data in a large EHR system revealed that EHRs inadequately capture various healthcare data such as diagnosis, visits, specialty care, hospitalizations, and medications [1]. The study concluded that missing data undermine many central functions of EHR and that "missing clinical information raises concerns about medical errors and research integrity" [1]. The authors stressed that "given the fragmentation of health care and poor EHR interoperability, information exchange and usability, priorities for further investment in health IT will need thoughtful reconsideration" [1]. This is not the only study regarding the vast amount of missing healthcare data in HIS. In [2], the authors presented multiple cases on how missing data in HIS is likely to result in medication errors and other patient harms. Missing data form obstacles in front of big data research in healthcare. In [3], the authors indicate that missing patients' data are prevalent in EHRs and are an impedance to utilizing machine learning for predictive and classification tasks in healthcare. For example, Length of Stay (LOS) prediction methods found in machine learning literature, operate without considering rehabilitation nursing interventions. This degrades the prediction accuracy of rehabilitation LOS. Although documented in papers, rehabilitation data are rarely captured

electronically in patient records [4].

A major source of missing data in healthcare institutions is paper-based forms and unstructured data. In [5], the authors list medical data written in an unstructured text format as one of the major sources of missing data in HIS. This is because EHRs are not designed to capture non-standardized data. In support of this analysis, and upon analyzing the literature, we found that an important reason for missing data in HIS is that a primary source of healthcare data is still paper-based and has not yet been fully automated. By this primary data source, we mean Clinical Pathways (CPs). CPs have been defined as optimal sequencing and timing of medical interventions by doctors, nurses, and other caregivers for a particular procedure or diagnosis, developed to minimize delays and resource utilization and to improve the quality of healthcare [6]–[8]. CPs appeared in healthcare first in the mid-1980s in the USA and then have spread all over the world [9]. The concept itself was not new because it has its roots in management theories that were proposed to improve the quality of business processes such as Critical Path Method (CPM), Program Evaluation and Review Technique (PERT), and Business Process Reengineering (BPR). These successful management theories were not applied in healthcare; thus, the concept of CP was an initiative to adopt effective management concepts in hospitals [10]–[12].

Despite the fact that CPs are becoming globally popular in hospitals as main components for patients' treatment and follow-up, CPs are still circulated in hospitals as paper-based documents and charts/tables. This forms a great barrier between CPs and their integration with today's automated hospitals. There have been several studies to automate CPs [13]–[19]. However, analyzing the relevant literature reveals that the common theme in all studies attempting to computerize CPs is that the main emphasis behind the computerization process was on how electronic CPs can support EHR systems. Therefore, the automation was not complete in the sense that CPs themselves were left with their unstructured nature (i.e., without standardization), and the resulting e-CPs were only partially automated and hidden behind EHRs.

The proper automation of CPs requires converting them to full digital entities that can work smoothly with all existing HIS, not only EHRs. This is crucial to reducing missing data in healthcare because CPs contain instructions on all interventions and procedures done on patients. Most missing data in healthcare (i.e., data not recorded digitally) exist due to the fact that CPs are not digitized.

The objective of this research is to study the effect of a new CP computerization framework on big data analytics in healthcare. In addition, the basic concepts of the new framework are briefly summarized here since its full details will be addressed in a different article dedicated mainly to the framework itself. The new SNOMED CT-based automation framework digitizes CPs and makes e-CP systems as central components in HIS. The new framework enables a detailed statistical analysis of CP interventions and maximizes data extraction from CPs. This contributes to decreasing data missingness and providing richer "CP-based" datasets for data analytics, as will be the focus of this research. Thus, we are motivated to hypothesize that the framework enables big data analytics and has positive effects on machine learning applications in healthcare. As an illustrative numerical example, we show the contribution of this work in improving the prediction accuracy of machine learning algorithms through experiments applied to stroke CP data to predict LOS of stroke patients in an acute rehabilitation facility. The experiments were applied to the CP data of real stroke patients in collaboration with the Regional Stroke Unit at Thunder Bay Regional Health Sciences Centre (TBRHSC), Ontario, Canada. We also present an example of time variation analysis of CP interventions.

The rest of the paper is organized as follows. Section 2 discusses CP automation and integration with HIS. Section 3 addresses the effect of SNOMED CT based e-clinical pathways on big data analytics in healthcare. In Section 4, we describe the experimental environment and discuss the experimental results. Finally, conclusions are drawn, and future research work is suggested in Section 5.

II. CP AUTOMATION AND INTEGRATION WITH HIS

As described above, e-CPs found in the literature have direct communication only with EHRs. They are not designed to be independent, fully-functioning systems. We view this as improper automation and integration of e-CPs since, in actual life, CPs are designed to be central healthcare components that produce data for all types of HIS commonly used in healthcare. This analysis of CP automation reveals two fundamental research challenges in the path of achieving complete automation of CPs and their proper integration with existing HIS. These challenges are summarized below.

- CP Automation: CPs are sometimes expressed in ambiguous local textual instructions. This not only makes them difficult to understand by other medical staff members but also makes them difficult to automate (with limited transferable electronic data). This usually creates a digital barrier or "digital divide" between CPs and HIS. Fig. 1 illustrates how paper-based or partially automated CPs create a digital divide between CPs and other HIS. e-CP research so far has ignored the presence of this digital divide, and most efforts were directed towards "programmatically" linking basic CP data with EHR systems while leaving CPs "as is," digitally invisible and far away from the digital age. This situation keeps CPs semantically non-operable with today's IT infrastructure.
- e-CP Integration with HIS: CP automation to produce computerized e-CPs forms only one aspect of CP inclusion in HIS. Another equally important consideration is: with what other HIS should e-CP be able to communicate? As mentioned in the introduction, by analyzing the literature so far, it is noticed that automated CPs were

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994286, IEEE Access

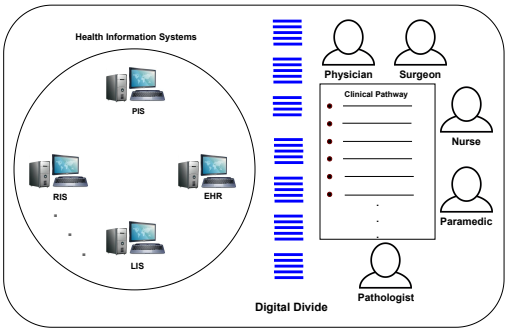

FIGURE 1. The Digital Divide caused by paper-based CPs.

positioned as side components that are created to support EHR systems. This positioning of CPs undervalues the real potential of CPs as the main healthcare plans for treatment and follow-up.

The following sub-sections give a summary of our CP automation framework with regard to the two considerations addressed above. We also show the importance of this research through a scenario related to detailed time variation analysis of CP interventions.

A. CP AUTOMATION

CPs are populated with data that are only partially transferred to other HIS. A key factor that is impeding the transfer of full CP data is that CPs are prepared in hospitals without attention to standardizing their medical terms. After a thorough review of CP research found in the literature and discussions with our domain experts at TBRHSC, it was clear that most CPs are currently developed using ambiguous local medical terms and abbreviations [20]–[29]. This situation makes CPs prone to human errors and forms a challenge to exchanging them across medical institutions. This also causes the loss of valuable CP data because existing HIS use standardized terminology systems in their encoding of medical terms. A solution for this, in our framework, is to encode CP data using internationally recognized medical reference terminology. For this objective, we selected the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) [30] to be the terminology system for encoding CP data because SNOMED CT is the most comprehensive and largest medical terminology in the world [31]. SNOMED CT encoding is realized by representing each CP data element by its equivalent SNOMED CT term and SNOMED CT ID (SCTID). For example, the ambiguous instruction "Complete Hx and PE" that is an element of a CP for open appendectomy [20], is converted in our framework into "Complete History and Physical Examination", where History and Physical Examination is a standard SNOMED CT term whose SCTID is 53807006. Table 1 gives more examples from a stroke CP. In addition to SNOMED CT encoding, rather than modeling CPs using traditional information systems, we modeled CPs using an ontology-based approach. The CP ontology is

TABLE 1. Encoding stroke CP Data using SNOMED CT Coding

Paper-based CP terms	Automated CP SNOMED CT terms
Intravenous	Intravenous injection (SCTID 43060002)
Neurovital signs	Taking neurologic vital signs (SCTID 82856007)

designed in consultation with domain experts and ontology research available in the literature [16], [17]. Fig. 2 shows the basic classes and relations in the main CP ontology that we developed using the Protege ontology editor and knowledge management system [32]. The ontology is integrated with a Java-based prototype CP management system that we developed based on our proposed framework. Disease-specific ontologies (e.g., ischemic stroke ontology) are then instantiated from the main CP ontology such that CP individuals (i.e., CP data) are SNOMED CT standardized. Fig. 3 shows part of the stroke ontology encoded with the proposed ontology-based modeling with SNOMED CT-based standardization.

Some advantages of the proposed modeling include [33]:

- Ontology, as a knowledge engineering modeling tool, allows for sharing and reuse of domain knowledge.
- An ontology defines the semantic of CP domain knowledge, which provides a shared understanding that can be communicated between heterogeneous applications in a machine-understandable way, thus facilitating semantic interoperability among e-CPs and HIS.
- Using SNOMED CT to represent CP data makes the integration of e-CPs with existing HIS smooth. This maximizes data extraction from CPs for less missing data and improved data analytics.
- Standardized CPs convert patients' CP traces (i.e., the path of each patient in the CP from admission to discharge) into path sequences of well-defined interventions with their start times and end times (e.g., $\{SCTID1, start\_time, end\_time\}$, $\{SCTID2, start\_time, end\_time\}$, etc.), which facilitates processing CP paths and performing useful statistical analyses on them, including time-based analytics (see Algorithm 1 in part C below).
- e-CPs with standardized data facilitate the retrieval of their contents for quality control. For example, the data of two CPs for the same disease used at two different hospitals can be retrieved digitally and compared by comparing their SNOMED CT terms and codes. This is a difficult task to realize with today's unstandardized, local text-based CPs. This also enables an e-CP system to act as a fully independent system, having its own CP functions (e.g., comparing CPs, exporting CP data to other systems, saving CP traces in a database of traces and applying data analytics on them).

B. E-CLINICAL PATHWAYS INTEGRATION WITH HIS

In order to address the proper communication level between e-CP systems and HIS, we first briefly consider the major

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994286, IEEE Access

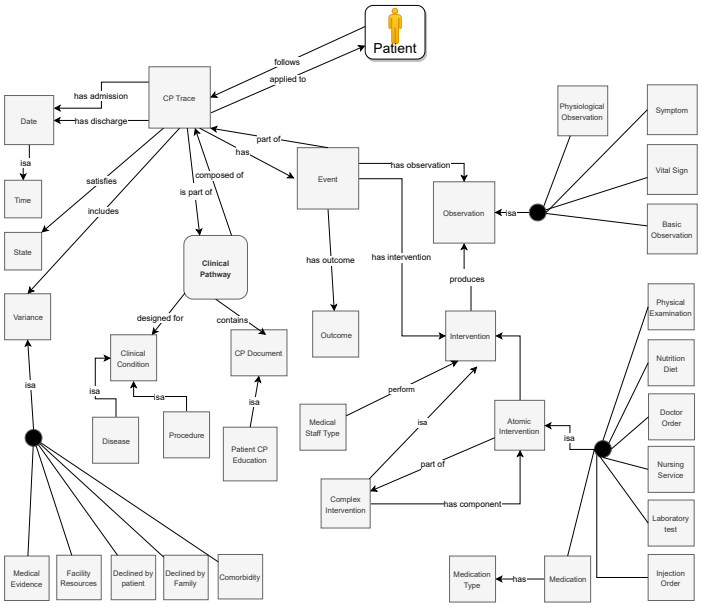

FIGURE 2. Basic classes and relations in the CP ontology.

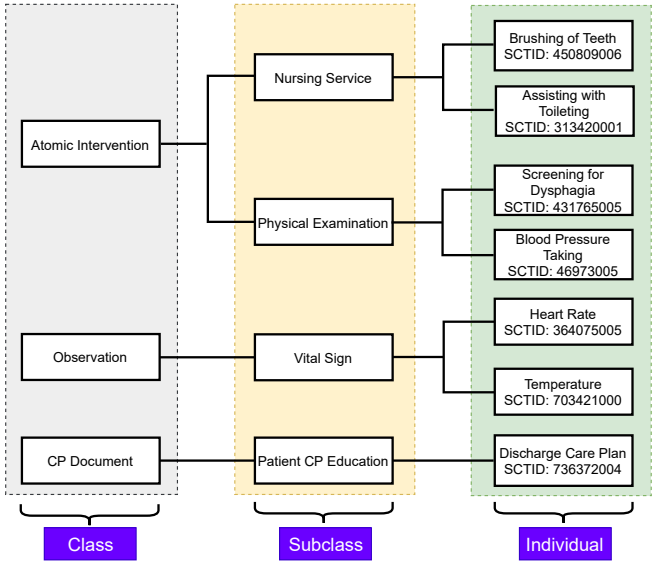

FIGURE 3. Part of the CP ontology for stroke.

subsystems of HIS and the data they contain [34]. Then we consider e-CPs and analyze the relationship between CP data and HIS.

1) Electronic Health/Medical Record Systems (EHR/EMR):

An EMR is a digital version of patients' paper charts. It is used by single practice clinics and small hospitals for their local records. Typically, an EMR contains the medical history of the patients, diagnoses, and treatments. An EHR can be viewed as a "large-scale" EMR that stores more data and facilitates the sharing of health records across different institutions. Modern systems are capable of playing the roles of both EMR and EHR since they offer options to keep patients' data local inside the institution or sharing the data with larger systems [4].

2) Laboratory Information Systems (LIS):

LIS are software systems with features that support modern laboratory operations and informatics. The main functions of LIS include recording, managing, and storing clinical laboratory data for patients. LIS have traditionally been most adept at sending laboratory test orders to lab instruments, tracking orders, and recording lab test results. In addition, LIS support the operations of public health institutions and

their labs by managing and reporting critical data concerning immunology and infection [35].

3) Radiology Information Systems (RIS):

RIS are the core systems for the e-management of imaging departments and are critical to the efficient workflow of radiology practices. The main functions of RIS include scheduling of patients, managing resources of radiology departments, image performance tracking, and distribution of results. A central component of RIS is the radiology PACS (Picture Archiving and Communication System) which provides storage and easy access to medical images from various sources (e.g., computed tomography (CT), medical ultrasound, X-ray, magnetic resonance imaging (MRI), computed radiography (CR), etc.) [34].

4) Pharmacy Information Systems (PIS):

PIS provide functions to maintain the organization and supply of drugs. A PIS can be a separate system for pharmacy usage, or it can be coordinated with inpatient hospital order entry systems. PIS are used to increase patient safety, report drug usage, and track costs. Outpatient PIS have a strong emphasis on medication labeling, drug warnings, and instructions for administration. The effective and safe dispensing of pharmaceutical drugs is the most important function of PIS. During the dispensing process, PIS prompt pharmacists to verify that the medication they have filled is for the correct patient, contains the right quantity and dosage, and displays accurate information on the prescription label [4].

5) e-Clinical Pathways Systems (e-CPS):

The concept of applying CPs in hospitals was a novel initiative to adopt successful management practices in healthcare. Therefore, since their introduction to healthcare institutions, the main objective of CPs was to coordinate and "manage" healthcare processes as central components. CPs contain all the interventions required to treat the patients; thus, within CPs lies the very heart of medical planning and treatment, including cost and quality factors in healthcare. The considerations above suggest that CPs were designed to produce all types of data in healthcare described above (e.g., EHR data, LIS data, etc.). Fig. 4 makes this point clear by illustrating how CPs generate data for all types of HIS discussed above. Two CPs for Diabetes Mellitus and Carotid Artery Disease are illustrated. As shown by the arrows in the figure, both CPs include order instructions that result in data that need to be transferred to all types of HIS. Thus, computerized CP management systems should be designed and positioned such that they are "centralized" (i.e., positioned at the centre of HIS) and allowed to communicate with all types of HIS, not only EHRs as the common theme in CP systems found in the literature. Besides using ontological modeling and SNOMED CT-based standardization, this positioning and high communication level of e-CPs can be achieved by equipping CP management systems with Health Level 7 (HL7) messaging functionality to communicate with existing HIS (Fig. 5). HL7 consists of a set of international standards for the transfer of clinical data between software applications [36]. This is achieved through standard, machine-readable HL7 messages. The generation of standard HL7 messages can be automated through application programs in high-level languages such as the Java-based HL7 Application Programming Interface (API) toolkit [37]. Fig. 6 shows an illustration of an HL7 observation result message to communicate the result of human immunodeficiency virus status using SNOMED CT encoding.

C. TIME VARIATION ANALYSIS OF CP INTERVENTIONS

Fully automated and digitized CPs open new opportunities for the application of data and statistical analyses on CPs. This helps improve the CP performance and optimize hospital resources. Examples can cover a wide range of applications. We present here an example scenario related to time variation analysis of CP interventions. In today's busy medical work environments, time management in hospitals is important to control costs and save lives. Wasted time may deprive other patients of having the required healthcare service due to a lack of medical staff members. This may cause the death of some patients. e-CP systems in which the start time and end time of each intervention are recorded can provide great help in discovering inefficient practices related to time management. For this objective, we developed the CP time analytics algorithm shown in Algorithm 1. CP traces of all patients are saved on an external file called Patient_CP_interventions_file (see Algorithm 1). The output of the algorithm is the maximum time, minimum time, and average time of each intervention of a CP considering all patients who went through that CP within a certain period of time (e.g., one year). Fig. 7 illustrates the graph of a sample output related to Screening for Dysphagia for ischemic stroke patients (SCTID 431765005). In ischemic stroke CP, one of the medical interventions is "Screening for dysphagia" because stroke often causes a swallowing disorder called dysphagia [38]. The average time for screening is around 30 minutes. A long time for screening (e.g., 57 minutes, as shown in the figure), makes the testing room more occupied and deprive other patients of having this test on time. In this particular case, solutions to control the time could be either by preparing the test room before the arrival of patients, analyzing the cases of a long time for better control, or by providing more training for nurses on efficient practices in dysphagia screening. Time variation analysis of CP interventions for all CPs in a hospital helps healthcare managers in improving healthcare provision and optimizing hospital resources. Such detailed CP analysis is not possible with today's unstructured paper-based CPs.

III. EFFECT OF SNOMED CT BASED E-CPS ON BIG DATA ANALYTICS IN HEALTHCARE

Data analytics algorithms prefer rich datasets. They work better when data are as much complete as possible. Missing data have always formed a challenge in the face of getting

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994286, IEEE Access

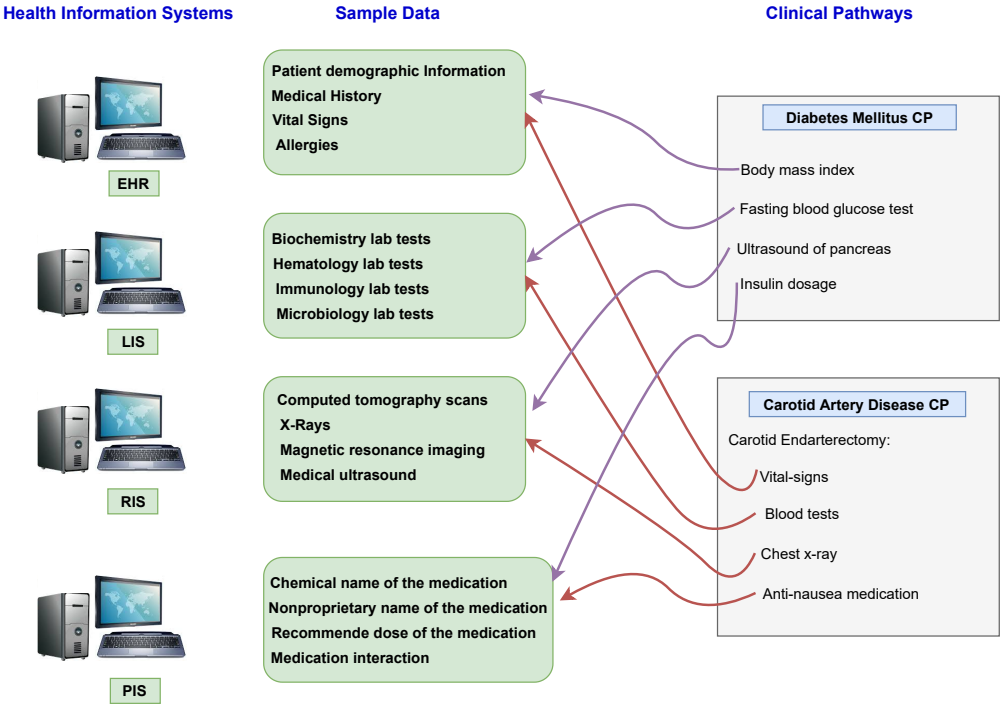

FIGURE 4. Clinical Pathways produce data for all types of Health Information Systems.

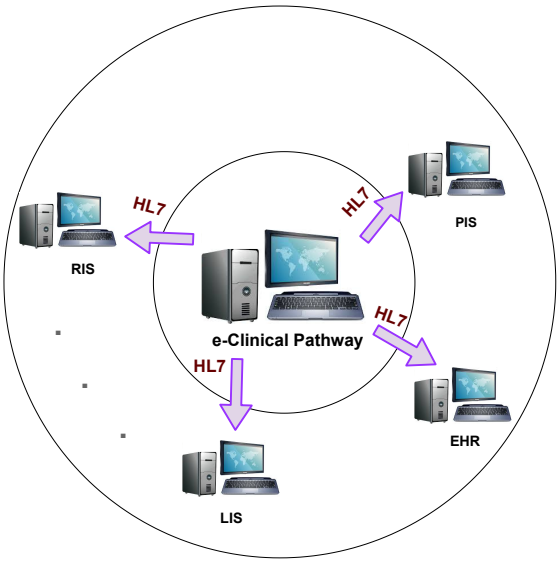

FIGURE 5. HL7 enables e-CPs to communicate with all existing HIS.

good classification and prediction results through machine learning algorithms. This is more critical in the field of healthcare data analytics because patients' outcomes are sensitive to data collected in hospitals. CPs are intended to be one of the most important sources of patients' data. Since the framework addressed in this research has the objective of generating computerized CPs that are fully encoded with SNOMED CT terms, then this framework contributes to reducing missing data and supplying rich CP-based datasets. This is achieved by making all CP data digitally visible and communicable with existing HIS (through SNOMED CT encoding, ontology-based modeling, and HL7 messages). We illustrate this contribution by machine learning experiments from the domain of hospital Length of Stay (LOS) prediction. LOS refers to the number of days that an inpatient stays in a hospital. LOS has long been a crucial metric of hospital efficiency and quality of care. The uncertainty of LOS increases costs and makes it difficult for hospitals to optimize their scheduling process [39]. The clinical and financial consequences of long LOS have made LOS as one of the most observed measures in healthcare systems [40]. LOS predictions that are related to rehabilitation CPs (i.e., CPs applied to patients whose hospital stays are in rehabilitation facilities) suffer from the fact that many rehabilitation interventions are not stored in EHRs. This makes EHR-based datasets yield less accurate LOS predictions. For example, stroke patients are initially treated in hospitals in emergency departments and stroke units, and after that, they move to acute rehabilitation facilities and follow the rehabilitation CP during their hospitalization. The challenge with rehabilitation CPs is that they contain many nursing care interventions. We refer to such interventions as "soft" interventions in this paper. By soft interventions, we mean interventions like assisting with toileting, etc. Although such interventions are specified on CPs, actually performed on patients, and documented on papers, they are rarely recorded in EHRs compared with what we term as "hard" CP interventions like

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994286, IEEE Access

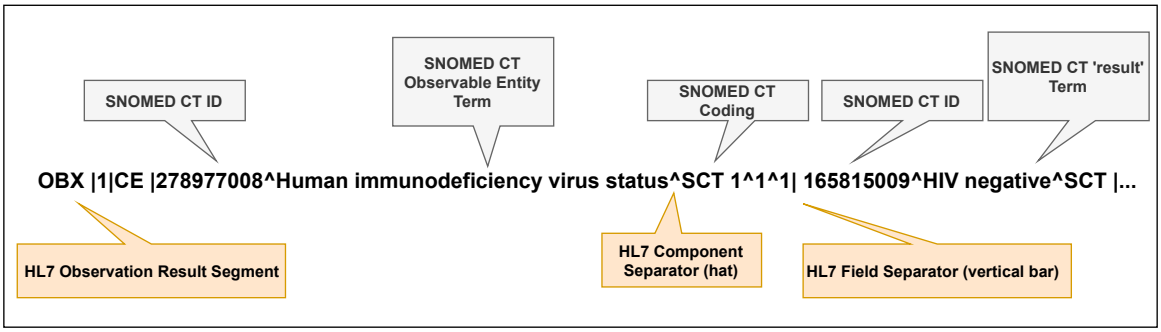

FIGURE 6. Illustration of an HL7 observation result message.

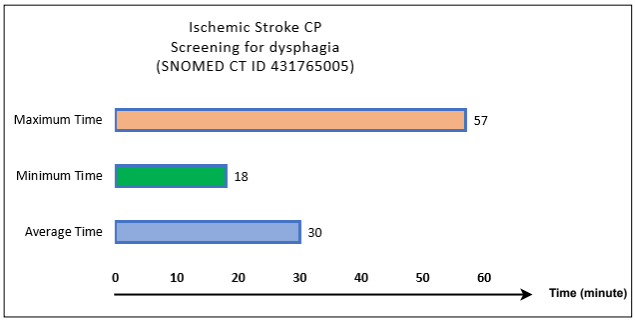

FIGURE 7. CP Intervention time analytics based optimization.

X-rays and surgical procedures [4]. As noticed by nurses and domain experts, in stroke patients, soft interventions have an effect on LOS prediction because patients who need more nursing services show long LOS. Thus, we hypothesize that data mining algorithms that work on datasets missing nursing care interventions yield less accurate LOS prediction results. Although terminology systems (like SNOMED CT and ICD-10) have advanced in recent years and have included standardized terms and codes for nursing care tasks, current CPs in use at hospitals still present most soft interventions as unstructured text with local terms. This is a major reason for the soft interventions being missed in EHR-based datasets (i.e., datasets obtained from EHRs without capturing all data on CPs). The framework outlined in this research contributes to recording soft interventions by means of standardized e-CPs that can transfer all their data to HIS. This results in CP-based datasets that are richer in data. To illustrate this experimentally, we performed machine learning experiments on the prediction of rehabilitation LOS using a stroke CP and CP-based dataset of stroke patients from TBRHSC, Ontario, Canada. The objective is to compare LOS prediction results between the CP-based dataset that includes nursing services and the same dataset without nursing services (EHR-based dataset) using the same machine learning algorithm on both datasets. Dataset preparation and experiments were performed with the help of stroke domain experts from the Regional Stroke Unit at TBRHSC, who assisted in the study with their knowledge on stroke patients' rehabilitation. The dataset contains rehabilitation admission data for 500 patients who had stroke secondary to Carotid Artery Disease (CAD). CAD is a chronic vascular disease characterized by the formation of plaques in the wall of the carotid artery, causing stenosis and impairing the flow of blood to the brain. In the case of plaque rupture, a clot of blood may form and detach, then move with blood to smaller brain vessels, potentially leading to an ischemic stroke [41]. In the dataset, each patient record contains several characteristics such as demographic data (age, gender, and ethnicity), disease history (sleep apnea, atrial fibrillation, diabetes, and hypertension), length of stay, medical history and habits (e.g., previous carotid artery intervention, alcohol consumption, smoking, type of stenosis, speech and language disorder), and the CP-based required nursing care interventions (e.g., assisting with toileting, brushing teeth, combing of hair, assistance with shaving). The LOS values were between four and nine days. The median LOS was five days. The objective of the data mining model is to predict short versus long LOS. We used the median LOS as the threshold dividing long vs. short LOS. Thus, patient records with LOS less than or equal to five days were labeled as short LOS while records greater than five days were labeled as long LOS. The initial dataset was imbalanced due to having more short than long LOS. This was the reason for considering the median (rather than the average) LOS as the threshold dividing short vs. long LOS since the median is the preferred measure for the central tendency for skewed data [42]. Furthermore, to avoid the drawbacks of dealing with an imbalanced classification problem, we used under-sampling techniques to train the classification model on balanced classes. Our LOS prediction example is a binary classification problem that is non-linear in nature, and this makes decision tree-based methods very suitable for this problem because they are successful in dealing with non-linear classification [43]. Furthermore, many researchers have reported that decision tree methods are successful in LOS prediction [44]–[47]. In general, decision tree algorithms use entropy-based methods to form tree nodes. This is done by selecting the most informative attributes based on two measures: entropy and information

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994286, IEEE Access

Algorithm 1: CP time-based data analytics

input : *Patient_CP_interventions_file (contains the sequences of: intervention_SCTID, intervention_startTime, and intervention_endTime for each patient)*
output: *max_duration[SCTID]: Maximum duration for each intervention (SCTID) in the CP*
output: *min_duration[SCTID]: Minimum duration for each intervention (SCTID) in the CP*
output: *avg_duration[SCTID]: Average duration for each intervention (SCTID) in the CP*

Data Structures:
SCTID: A variable to store an intervention's SCTID
intervention_duration[SCTID]: Hash-table with list of all durations for each intervention (SCTID)
total_duration[SCTID]: List to store total durations of each intervention (SCTID)
begin
  foreach *record* ∈ *Patient_CP_interventions_file* **do**
    foreach *intervention* ∈ *record* **do**
      SCTID← *intervention_SCTID*
      if *(SCTID* ∉ *intervention_duration[SCTID])* **then**
        Add *SCTID* to *intervention_duration[SCTID]*
      end
      intervention_duration[SCTID] ← *intervention_endTime - intervention_startTime*
    end
  end
  foreach *record* ∈ *intervention_duration[SCTID]* **do**
    total_duration[SCTID]=0
    max_duration[SCTID] = min_duration[SCTID] = duration [0] (duration of 1st intervention in the record)
    for *i*← *0* **to** *length of record* **do**
      total_duration[SCTID]←*total_duration[SCTID] + duration[i]*
      if *(duration[i] > max_duration[SCTID])* **then**
        max_duration[SCTID] ← *duration[i]*
      end
      if *(duration[i] < min_duration[SCTID])* **then**
        min_duration[SCTID] ← *duration[i]*
      end
    end
    avg_duration[SCTID] ← *total_duration[SCTID] / length of record*
  end
end

gain, as follows.

- Entropy (H) measures the impurity of a category or class (X), as shown in equation (1).

$$H_X = -\sum_{\forall x \in X} P(x) log_2 P(x) \tag{1}$$

  where $P(x)$ is the probability of label x in X [43].
- Information gain measures the purity of an attribute based on the conditional entropy determined by equation (2) below.

$$H_{Y|X} = -\sum_{\forall x \in X} P(x) \sum_{\forall y \in Y} P(y \mid x) log_2 P(y \mid x) \tag{2}$$

  where $H_{Y|X}$ is the conditional entropy for each attribute (X) relative to base entropy (Y) which is the entropy of the output variable, LOS in our case. The information gain of an attribute X is defined as the difference between the base entropy and the conditional entropy of the attribute, as shown in equation (3).

$$InfoGain_X = H_Y - H_{Y|X} \tag{3}$$

  Information gain compares the degree of purity of the upper node (parent node) before a split with the degree of purity of the lower node (child node) after a split. At every split, an attribute (or predictor) with the highest information gain is considered as the most informative attribute and is chosen for the split [43].

Among the most commonly used decision tree learning algorithms are ID3 (Iterative Dichotomiser 3), C4.5, and C5.0. ID3 algorithm has the drawback that it may construct a complex and deep tree that causes overfitting leading to poor prediction results. The C4.5 algorithm is an improved ID3 algorithm that addresses the overfitting problem in ID3 by using the technique of pruning to simplify the decision tree. Pruning is done by removing the tree nodes and branches that do not provide additional information [43]. C5.0 algorithm offers a number of improvements over C4.5, including faster processing and more efficient memory usage [48]. Another commonly used decision tree algorithm is CART (Classification And Regression Trees); however, preliminary experiments on our datasets showed that C5.0 has a slightly better overall performance than CART. Based on the above analysis, we adopted the C5.0 algorithm in our simulation experiments. As will be detailed below in the results section, experimental evaluation demonstrates that LOS prediction with fully automated CPs that result in rich datasets (CP-based datasets) outperforms traditional LOS predictions based on EHR-based datasets.

IV. RESULTS AND DISCUSSION

We implemented the experiments using "RStudio" integrated development environment for R programming language [49], [50]. To compare the results with and without full CP data, we performed identical processing and experiments on two

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994286, IEEE Access

datasets. The first dataset does not include the CP nursing services (EHR-based dataset), while the second dataset includes the CP nursing services (CP-based dataset). The datasets were split into training/testing sets. In order not to generalize the results from a single split, we conducted experiments with 70:30 and 80:20 training/testing split ratios. Furthermore, we have diversified the performance metrics by including multiple major common metrics, including the area under the receiver operating characteristic curve (AUROC), accuracy, sensitivity, specificity, and precision, as shown in equations (4), (5), (6) and (7). In the equations, TP, TN, FP, and FN stand for True Positive, True Negative, False Positive, and False Negative, respectively.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{6}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

Fig. 8 and 9 show the experimental results. As shown in the figures, the performance of the prediction model that includes the CP nursing services is better than the model without the nursing services in terms of the considered metrics. The results show better AUROC for CP-based dataset ($\approx$ 88% and 93%) compared to EHR-based dataset ($\approx$ 78% and 84%) for split ratios 70:30 and 80:20, respectively. The most commonly reported measure of a classifier is the accuracy because accuracy evaluates the overall efficiency. The results show better accuracy for the CP-based dataset ($\approx$ 85% and 92%) compared to EHR-based dataset ($\approx$ 77% and 85%) for split ratios 70:30 and 80:20, respectively. The results show better performance for the CP-based dataset in terms of sensitivity and equal performance with the EHR-based dataset in terms of specificity. Sensitivity assesses the effectiveness of the classifier on the positive/minority class. In our experiments, this is the class of patients with long LOS. Thus, the CP-based dataset yields better long LOS prediction performance. Specificity, on the other hand, measures the effectiveness of predicting negative cases (short LOS in our experiments). Since fewer nursing services are related to short LOS, then it is reasonable that both datasets show equal specificity, i.e., equal prediction performance on patients with less nursing services. Precision is also called positive predictive value. The obtained results show that the CP-based dataset gives better precision under both training/testing split ratios.

The metrics mentioned above are the most used performance measures for such classification problems. However, since our datasets are slightly imbalanced, we decided to investigate more metrics that consider imbalanced datasets.

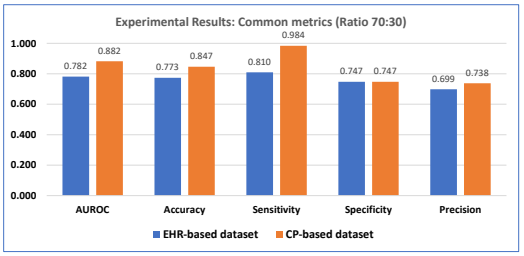

FIGURE 8. Experimental results: Common metrics, 70:30 split ratio.

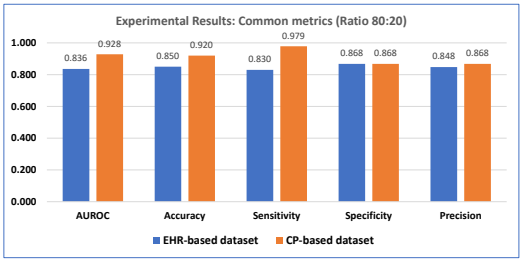

FIGURE 9. Experimental results: Common metrics, 80:20 split ratio.

This helps in generalizing the results by considering performance metrics that combine the previous metrics to account for imbalanced datasets. Therefore, we considered the Balanced Accuracy and Geometric Mean (G-mean), which are common imbalance-oriented performance metrics [51], as shown in equations (8) and (9).

$$\begin{aligned}\text{Balanced Accuracy} &= \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \\ &= \frac{1}{2}(sensitivity + specificity)\end{aligned} \tag{8}$$

$$\begin{aligned}\text{G-mean} &= \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \\ &= \sqrt{sensitivity \times specificity}\end{aligned} \tag{9}$$

Fig. 10 and 11 show the experimental results considering the imbalance-oriented metrics. The balanced accuracy is the average between the sensitivity and the specificity, which measures the average accuracy obtained from both the majority and minority classes. "This quantity reduces to the traditional accuracy if a classifier performs equally well on either classes. Conversely, if the high value of the traditional accuracy is due to the classifier taking advantage of the distribution of the majority class, then the balanced accuracy will decrease compared to the accuracy" [51].

Our results show that both the traditional accuracy and the balanced accuracy have close values for both datasets, with the CP-based dataset showing better performance. This is an indication of the good performance of both classifiers on the majority and minority classes with the CP-based dataset

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2994286, IEEE Access

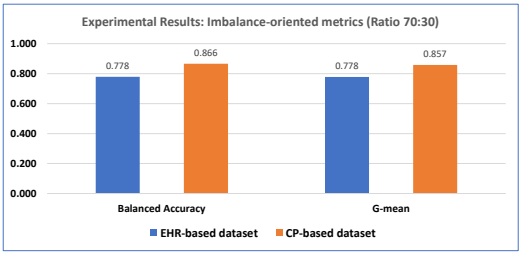

FIGURE 10. Experimental results: Imbalance-oriented metrics, 70:30 split ratio.

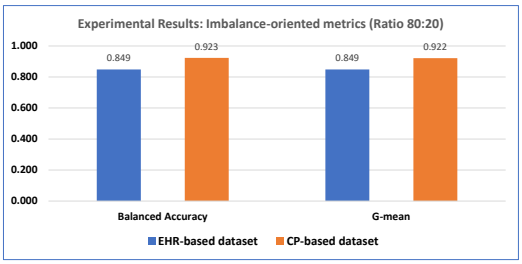

FIGURE 11. Experimental results: Imbalance-oriented metrics, 80:20 split ratio.

yielding improved performance. G-Mean is a metric suitable for imbalanced datasets because it measures the balance between classification performances on both the majority and minority classes [51]. Our results show higher G-mean values for the CP-based dataset ($\approx$ 86% and 92%) than the G-mean values of the EHR-based dataset ($\approx$ 78% and 85%), for both 70:30 and 80:20 split ratios, respectively.

As shown in the analysis above, experimental results support the hypothesis that complete CP standardization and automation reduces data missingness in healthcare, resulting in rich datasets and contributing to improving the performance of data analytics algorithms. It is worth mentioning that there are other factors than nursing services that affect stroke patients' LOS (e.g., comorbidity, diabetes, etc.). However, such data are common to both datasets in our experiments; thus, the only differentiating data are nursing services available on the clinical pathway.

V. CONCLUSION

CPs are crucial components of healthcare systems and deserve more studies regarding their automation and complete integration with HIS. In this research, we presented the summary of a CP automation framework based on three major components: SNOMED CT standardization of all CP data, ontology-based modeling, and HL7 communication. Our model enables detailed statistical analyses on CP data that help in optimizing hospital resources and providing better healthcare services for patients. Regarding big data analytics in healthcare, our research hypothesis is that the problem of missing data in HIS can be addressed by fully digitizing and integrating CPs with HIS. This helps in generating rich datasets and improving the performance of data analytics algorithms in healthcare. To test this hypothesis, we standardized and converted a hospital stroke CP into an e-CP that was integrated with a prototype e-CP system. The data of real 500 stroke patients were collected and simulated based on the standardized stroke e-CP in collaboration with the Regional Stroke Unit at TBRHSC, Ontario, Canada. Next, we conducted machine learning experiments on the resulting datasets. The experiments were from the domain of predicting rehabilitation LOS.

The experimental results show that LOS prediction with the dataset that includes the required CP nursing services outperforms LOS prediction using the dataset that does not include the nursing services. Five common performance metrics were evaluated to reach this conclusion: AUROC, accuracy, sensitivity, specificity, and precision. Furthermore, since the datasets were imbalanced, we supported our analysis by considering two additional combined performance measures that account for the imbalance nature of the LOS problem, namely, the balanced accuracy and G-mean. The experimental results using the imbalance-oriented performance metrics confirmed that LOS prediction using the CP-based dataset outperforms the prediction using the EHR-based dataset. The results can be justified by the fact that patients who are affected more negatively by the stroke incidence require more nursing services in stroke rehabilitation and thus show longer LOS. Therefore, datasets that include the full healthcare data available on CPs can enable and guide machine learning algorithms more accurately towards improved predictive results.

The complete digitization and automation of e-CPs is a key contribution towards capturing all CP data. This allows detailed statistical analysis on CP contents and patients' treatments, as well as proved to be important for big data analytics in healthcare.

Currently, only EHR-based datasets are available for researchers. Obtaining CP-based datasets is a challenge since most CPs in use today are paper-based, and CP-based datasets can only be obtained by collaborating directly with cooperating hospitals. One of our future research directions is to collaborate with multiple hospitals to obtain more and larger CP-based datasets.

We consider this research as a starting initiative towards the complete automation of CPs, hoping to encourage more research in this area to achieve better CP automation, improved healthcare outcomes, enhanced patient satisfaction, and healthier society.

ACKNOWLEDGMENT

We want to thank Dr. Kofi Darko for sharing his knowledge on stroke CP and LOS. We would also like to thank the nurses and staff members of the Regional Stroke Unit, who provided expertise and shared data that greatly assisted the research and improved its quality. This study was approved by the both Lakehead University and TBRHSC Research Ethics Boards.

• • •

REFERENCES

[1] J. M. Madden, M. D. Lakoma, D. Rusinak, C. Y. Lu, and S. B. Soumerai, "Missing clinical and behavioral health data in a large electronic health record (EHR) system," Journal of the American Medical Informatics Association, vol. 23, no. 6, pp. 1143–1149, 2016.

[2] J. Leviss and P. Charney, HIT or miss: lessons learned from health information technology implementations. AHIMA Press Chicago, 2013.

[3] J. Codella, H. Sarker, P. Chakraborty, M. Ghalwash, Z. Yao, and D. Sow, "eXITs: An Ensemble Approach for Imputing Missing EHR Data," in 2019 IEEE International Conference on Healthcare Informatics (ICHI). IEEE, 2019, pp. 1–3.

[4] P. A. Potter, A. G. Perry, P. Stockert, A. Hall, B. J. Astle, and W. Duggleby, Canadian Fundamentals of Nursing. Elsevier Health Sciences, 2018.

[5] H. Hegde, N. Shimpi, A. Panny, I. Glurich, P. Christie, and A. Acharya, "Mice vs ppca: Missing data imputation in healthcare," Informatics in Medicine Unlocked, vol. 17, p. 100275, 2019.

[6] X. Wang, S. Su, H. Jiang, J. Wang, X. Li, and M. Liu, "Short-and long-term effects of clinical pathway on the quality of surgical non-small cell lung cancer care in china: an interrupted time series study," International Journal for Quality in Health Care, vol. 30, no. 4, pp. 276–282, 2018.

[7] R. J. Coffey, J. S. Richards, C. S. Remmert, S. S. LeRoy, R. R. Schoville, and P. J. Baldwin, "An introduction to critical paths," Quality Management in Healthcare, vol. 14, no. 1, pp. 46–55, 2005.

[8] S. D. Pearson, S. F. Kleefield, J. R. Soukop, E. F. Cook, and T. H. Lee, "Critical pathways intervention to reduce length of hospital stay," The American journal of medicine, vol. 110, no. 3, pp. 175–180, 2001.

[9] K. Zander, K. A. Bower, and M. Etheredge, "Nursing case management: blueprints for transformation," Boston: New England Medical Center Hospitals, pp. 1–128, 1987.

[10] E. Rooney, "Developing care pathways–lessons from the steele review implementation in e ngland," Gerodontology, vol. 31, pp. 52–59, 2014.

[11] G. Schrijvers, A. van Hoorn, and N. Huiskes, "The care pathway: concepts and theories: an introduction," International journal of integrated care, vol. 12, no. Special Edition Integrated Care Pathways, 2012.

[12] R. S. Russell and B. W. Taylor, Operations and supply chain management. John Wiley & Sons, 2017.

[13] S. R. Abidi and S. S. R. Abidi, "An ontological modeling approach to align institution-specific clinical pathways: Towards inter-institution care standardization," in 2012 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS). IEEE, 2012, pp. 1–4.

[14] S. R. Abidi, S. S. R. Abidi, L. Butler, and S. Hussain, "Operationalizing prostate cancer clinical pathways: An ontological model to computerize, merge and execute institution-specific clinical pathways," in Workshop on Knowledge Management for Health Care Procedures. Springer, 2008, pp. 1–12.

[15] A. Daniyal, S. R. Abidi, and S. S. R. Abidi, "Computerizing clinical pathways: ontology-based modeling and execution." in MIE, 2009, pp. 643–647.

[16] J. Liu, Z. Huang, X. Lu, and H. Duan, "An ontology-based real-time monitoring approach to clinical pathway," in 2014 7th International Conference on Biomedical Engineering and Informatics. IEEE, 2014, pp. 756–761.

[17] Y. Ye, Z. Jiang, X. Diao, D. Yang, and G. Du, "An ontology-based hierarchical semantic modeling approach to clinical pathway workflows," Computers in biology and medicine, vol. 39, no. 8, pp. 722–732, 2009.

[18] Z. Hu, J.-S. Li, T.-S. Zhou, H.-Y. Yu, M. Suzuki, and K. Araki, "Ontology-based clinical pathways with semantic rules," Journal of medical systems, vol. 36, no. 4, pp. 2203–2212, 2012.

[19] H.-Q. Wang, T.-S. Zhou, Y.-F. Zhang, L. Chen, and J.-S. Li, "Research and development of semantics-based sharable clinical pathway systems," Journal of medical systems, vol. 39, no. 7, p. 73, 2015.

[20] A. L. Hilario, J. D. H. Oruga, M. P. B. Turqueza, and D. V. Hilario, "Utilization of clinical pathway on open appendectomy: A quality improvement initiative in a private hospital in the philippines," International journal of health sciences, vol. 12, no. 2, p. 43, 2018.

[21] P. Ibeziako, K. Brahmbhatt, A. Chapman, C. De Souza, L. Giles, S. Gooden, F. Latif, N. Malas, L. Namerow, R. Russell et al., "Developing a clinical pathway for somatic symptom and related disorders in pediatric hospital settings," Hospital pediatrics, vol. 9, no. 3, pp. 147–155, 2019.

[22] N. A. Patel, R. A. Bly, S. Adams, K. Carlin, S. R. Parikh, J. P. Dahl, and S. Manning, "A clinical pathway for the postoperative management of hypocalcemia after pediatric thyroidectomy reduces blood draws," International journal of pediatric otorhinolaryngology, vol. 105, pp. 132–137, 2018.

[23] S. Ovaere, I. Boscart, I. Parmentier, P. J. Steelant, T. Gabriel, J. Allewaert, H. Pottel, F. Vansteenkiste, and M. D'Hondt, "The effectiveness of a clinical pathway in liver surgery: a case-control study," Journal of Gastrointestinal Surgery, vol. 22, no. 4, pp. 684–694, 2018.

[24] I. S. Lanig, P. W. New, A. S. Burns, G. Bilsky, J. Benito-Penalva, D. Bensmail, and M. Yochelson, "Optimizing the management of spasticity in people with spinal cord damage: a clinical care pathway for assessment and treatment decision making from the ability network, an international initiative," Archives of physical medicine and rehabilitation, vol. 99, no. 8, pp. 1681–1687, 2018.

[25] C. R. Shubert, M. L. Kendrick, E. B. Habermann, A. E. Glasgow, B. J. Borah, J. P. Moriarty, S. P. Cleary, R. L. Smoot, M. B. Farnell, D. M. Nagorney et al., "Implementation of prospective, surgeon-driven, risk-based pathway for pancreatoduodenectomy results in improved clinical outcomes and first year cost savings of us1 million," Surgery, vol. 163, no. 3, pp. 495–502, 2018.

[26] V. Wylde, W. Bertram, A. D. Beswick, A. W. Blom, J. Bruce, A. Burston, J. Dennis, K. Garfield, N. Howells, A. Lane et al., "Clinical and cost effectiveness of the star care pathway compared to usual care for patients with chronic pain after total knee replacement: study protocol for a uk randomised controlled trial," Trials, vol. 19, no. 1, p. 132, 2018.

[27] M. Li, J. Zhang, T. J. Gan, G. Qin, L. Wang, M. Zhu, Z. Zhang, Y. Pan, Z. Ye, F. Zhang et al., "Enhanced recovery after surgery pathway for patients undergoing cardiac surgery: a randomized clinical trial," European Journal of Cardio-Thoracic Surgery, vol. 54, no. 3, pp. 491–497, 2018.

[28] W. Kamil, O. Y. Hian, S. Mohd-Said, S. L. A. Zainuddin, H. Ramli, E. Noor, R. Ayob, A. F. A. Aziz, A. Ismail, S. Sulong et al., "Development of clinical pathway for non-surgical management of chronic periodontitis," Malaysian Journal of Public Health Medicine, vol. 2018, no. Specialissue1, pp. 26–32, 2018.

[29] "Clinical pathways, thunder bay regional health sciences centre," https://tbrhsc.net/, 2019, accessed: 2019-10-23.

[30] "SNOMED CT, SNOMED International," https://www.snomed.org/snomed-ct, 2020, accessed: 2020-01-15.

[31] "Canada health infoway," https://infoway-inforoute.ca/en/, 2019, accessed: 2019-10-20.

[32] M. A. Musen, "The protégé project: a look back and a look forward," AI Matters, vol. 1, no. 4, pp. 4–12, 2015. [Online]. Available: https://doi.org/10.1145/2757001.2757003

[33] H. S. Pinto and J. P. Martins, "Ontologies: How can they be built?" Knowledge and information systems, vol. 6, no. 4, pp. 441–464, 2004.

[34] K. A. Wager, F. W. Lee, and J. P. Glaser, Health care information systems: a practical approach for health care management. John Wiley & Sons, 2017.

[35] The Laboratory, Health, and Science Informatics Encyclopedia, "Laboratory Information System," 2020. [Online]. Available: https://www.limswiki.org/

[36] T. Benson and G. Grieve, Principles of health interoperability: SNOMED CT, HL7 and FHIR. Springer, 2016.

[37] "University Health Network, HAPI FHIR," http://hapifhir.io, 2019, accessed: 2019-07-08.

[38] "Stroke association," http://www.strokeassociation.org, 2020, accessed: 2020-01-20.

[39] A. Alahmar, E. Mohammed, and R. Benlamri, "Application of data mining techniques to predict the length of stay of hospitalized patients with diabetes," in 2018 4th International Conference on Big Data Innovations and Applications (Innovate-Data). IEEE, 2018, pp. 38–43.

[40] M. Shulan and K. Gao, "Revisiting hospital length of stay: what matters?" The American journal of managed care, vol. 21, no. 1, pp. e71–7, 2015.

[41] G. M. Karageorgos, I. Z. Apostolakis, P. Nauleau, V. Gatti, R. Weber, E. S. Connolly, E. C. Miller, and E. E. Konofagou, "Arterial wall mechanical inhomogeneity detection and atherosclerotic plaque characterization using high frame rate pulse wave imaging in carotid artery disease patients in vivo," Physics in Medicine & Biology, vol. 65, no. 2, p. 025010, 2020.

[42] F. J. Gravetter and L. B. Wallnau, Statistics for the behavioral sciences. Cengage Learning, 2016.

[43] E. E. Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. Wiley, 2015.

[44] P. Liu, L. Lei, J. Yin, W. Zhang, W. Naijun, and E. El-Darzi, "Healthcare data mining: Prediction inpatient length of stay," in 2006 3rd International IEEE Conference Intelligent Systems. IEEE, 2006, pp. 832–837.
[45] L. Breiman, Classification and regression trees. Routledge, 2017.
[46] K. B. Highland, M. R. Cole, S. P. Bilir, J. Pruett, and W. Drake, "Decreasing length of stay and pulmonary arterial hypertension-related hospitalizations with macitentan using a decision tree model in a medicare population," in A68. WOW: PHARMACOLOGICAL TREATMENT OF PULMONARY HYPERTENSION. American Thoracic Society, 2017, pp. A2286–A2286.
[47] M. A. Rahman, B. Honan, T. Glanville, P. Hough, and K. Walker, "Using data mining to predict emergency department length of stay greater than 4 hours: Derivation and single-site validation of a decision tree algorithm," Emergency Medicine Australasia, 2019.
[48] M. Kuhn and K. Johnson, Applied predictive modeling. Springer, 2013, vol. 26.
[49] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2019. [Online]. Available: https://www.R-project.org/
[50] RStudio Team, RStudio: Integrated Development Environment for R, RStudio, Inc., Boston, MA, 2019. [Online]. Available: http://www.rstudio.com/
[51] J. Akosa, "Predictive accuracy: a misleading performance measure for highly imbalanced data," in Proceedings of the SAS Global Forum, 2017, pp. 2–5.

AYMAN D. ALAHMAR received his Bachelor's Degree, Master's Degree, and Ph.D. Degree in Mechanical Engineering with a specialization in Computer Information Systems from Middle East Technical University, Ankara, Turkey. He then received his MBA Degree from Lakehead University, Ontario, Canada. After years of academic and industrial experience as an Associate Professor and Serial Entrepreneur, his passion for software engineering and artificial intelligence led him to pursue a Ph.D. Degree in Software Engineering at Lakehead University, where he is currently a Ph.D. Candidate and Researcher at the Artificial Intelligence and Data Science Lab. His current research interests include Automation and Digitization of Clinical Pathways, Health Informatics, Software Engineering, and Artificial Intelligence.

RACHID BENLAMRI is a Professor of Software Engineering at Lakehead University - Canada. He received his Master's degree and a Ph.D. in Computer Science from the University of Manchester - UK in 1987 and 1990, respectively. He is the head of the Artificial Intelligence and Data Science Lab at Lakehead University. He supervised over 80 students and postdoctoral fellows. He served as a keynote speaker and general chair for many international conferences. Professor Benlamri is a member of the editorial board for many refereed international journals. His research interests are in the areas of Artificial Intelligence, Semantic Web, Data Science, Ubiquitous Computing, and Mobile Knowledge Management.