Conformance Checking Methodology Across Discharge Summaries and Standard Treatment Guidelines

by

raghavendra.ch , Kamalakar Karlapalem

in

ACM Transactions on Computing for Healthcare

Report No: IIIT/TR/2020/-1

Centre for Data Engineering
International Institute of Information Technology
Hyderabad - 500 032, INDIA
May 2020

Conformance Checking Methodology Across Discharge Summaries and Standard Treatment Guidelines

VEERA RAGHAVENDRA CHIKKA and KAMALAKAR KARLAPALEM, International Institute of Information Technology

Conformance checking of treatment plans in discharge summary data would facilitate the development of clinical decision support system, treatment plan quality assurance, and new treatment plan discovery. Conformance checking requires extraction of medical entities and relationships among them to form a computable representation of the treatment plan present in the discharge summary. We propose a workflow representation of patient's discharge summary that is referred to as workflow instance. We employ a multi-layer perceptron neural network to extract relationships between medical entities to construct the workflow instance. The aim of this work is to check the conformance of the workflow instance against standard treatment plan. Standard treatment plans are extracted from the treatment guidelines provided on healthcare websites such as WebMD, Mayo Clinic, and Johns Hopkins. For each disease, these guidelines are curated, aggregated, and represented as a workflow specification. We commend multiple measures to compute the conformance of workflow instance with workflow specification. We validate our conformance checking methodology using discharge summary data of three diseases, namely colon cancer, coronary artery disease, and brain tumor, collected from THYME corpus and MIMIC III clinical database. Our approach and the solution can be used by hospitals and patients to determine adherence, gaps, and additions to standard treatment plans. Further, our work can facilitate to identify common errors and goodness in actual enactment of treatment plans, which can further lead to refinement of standard treatment plans.

CCS Concepts: • **Computing methodologies** → *Artificial intelligence*; *Knowledge representation and reasoning*; *Natural language processing*; *Machine learning*;

Additional Key Words and Phrases: Treatment plan, discharge summaries, workflow instance extraction, workflow representation, conformance measure

ACM Reference format:
Veera Raghavendra Chikka and Kamalakar Karlapalem. 2020. Conformance Checking Methodology Across Discharge Summaries and Standard Treatment Guidelines. *ACM Trans. Comput. Healthcare* 1, 3, Article 13 (May 2020), 19 pages.
https://doi.org/10.1145/3377328

1 INTRODUCTION

A treatment plan defines the necessary therapeutic interventions for a patient's medical problem. It includes guidelines for each particular diagnosis. Healthcare providers follow several standard treatment plans for various

This work was partially supported by R&D Centre, Hitachi India Pvt. Ltd.
Authors' addresses: V. R. Chikka, International Institute of Information Technology Hyderabad (IIITH), Gachibowli, Hyderabad, Telangana, India – 500032; emails: raghavendra.ch@research.iiit.ac.in, ch.raghavendra66@gmail.com; K. Karlapalem, International Institute of Information Technology Hyderabad (IIITH), Gachibowli, Hyderabad, Telangana, India – 500032; email: kamal@iiit.ac.in.

2637-8051/2020/05-ART13 $15.00
https://doi.org/10.1145/3377328

diseases, such as acute and chronic heart failure [28] and pulmonary hypertension [14]. Consistent execution of such standard treatment plans is essential for the best evidence of treatment success. Research [42] conducted on 10,000 patients with inflammatory breast cancer showed that under-utilization of trimodality treatment (chemotherapy, surgery, and radiation therapy) have a negative impact on patient survival. Survival rates are highest among patients who have undergone trimodality treatment than the combination of chemotherapy plus surgery, radiation therapy plus surgery, or surgery alone. Similar cohort studies on hypertension [3] and osteoporosis [6] reported the under-treatment of recommended care. Another study [46] suggests an effective implementation of a standardized post-resuscitates protocol to improve survival rates among cardiac arrest patients, thus emphasizing the importance of following standard treatment plans. In cases like these, determining deviations in the treatment process helps to caution healthcare providers for immediate corrective measures. The advantage of such an analysis is twofold: first, to monitor the quality assurance of the treatment provided, and second, to discover new treatment plans for different co-morbidities, which would further aid in improving the standard treatment plans based on expert physicians' best practices rather than on traditional systematic trials [21].

The objective of this study is to systematically determine the conformance of discharge summary's treatment plan with the standard treatment guidelines. The discharge summary is one of the primary documents used for storing and retrieving information about a patient's hospitalization [31]. A discharge summary contains a patient's medical information that includes the reason for patient's admission, physical findings (symptoms), laboratory tests, treatments, and responses to treatments. It acts as a main source of information about the treatment provided to the patient for further continuity of care [5]. Digitization of health records and documents have generated a large number of discharge summaries. Analyzing such a large number of free-text discharge summaries manually is a strenuous task. Hence, there is a requirement for machine-readable and processable representation for discharge summary information. Some of the representations include temporally abstracted event sequences [5], ICD procedural code sequences [34, 57], and comprehensive event sequences (comprised of diagnosis, lab tests, medications, etc.) [53]. But these sequence representations are ineffective in capturing parallel or overlapping medical entities that occur at the same time. To overcome this limitation, Wang et al. [54] developed a graph-based representation where nodes indicate medical entities and edges indicate temporal difference among the entities. Enhancing this idea, we propose a graph representation for the discharge summary, referred to as workflow instance, orchestrated with nodes as medical entities and edges as semantic relations (*associated with*, *administered for*, and *shows*) among the entities. Thus, for the medical entities with semantic types of medical problems, tests, and treatments, we have a total of eight relationships: Test Associated with Test (TeATe), Treatment Associated with Treatment (TrATr), Problem Associated with Problem (PAP), Test Associated with Treatment (TeATr), Treatment Administered for Problem (TrAP), Test Administered for Problem (TeAP), Treatment Shows Problem (TrSP), and Test Shows Problem (TeSP).

Standard treatment guidelines for disease/disorders are well specified by various health institutes, such as Mayo Clinic [27], WebMd [56], and Johns Hopkins [18]. For each disease, medical entities, such as symptoms, tests, and treatments, can be extracted from each institute's guidelines. These medical entities collected from each institute are aggregated and manually curated to represent in the form of graphical representation that is referred to as workflow specification of the disease.

A workflow is a cumulative, concise, and explicit version of a complex set of activity sequences. The concept of workflow is analogous to a *process* in business process modeling [12]. The usual notion of conformance in process modeling is that an executable process with low-level details should hold conformance with the abstract process with high-level granularity [26]. Similarly, in our work, the workflow instance of a discharge summary is considered as a trace of a possible execution corresponding to a workflow specification. Therefore, by adopting the concepts from business/process models, we define various conformance measures for comparing the workflow instance with the specification [12].

In summary, the novelty of this work lies in automatically extracting the semantic relations to construct the workflow representation, and the methodology for the conformance checking problem.

The major contributions of our work are as follows:

(1) We build a multi-layer perceptron neural network (MLPNN) model using novel lexical, context, and similarity features to automatically extract medical relations from the discharge summary and to form *workflow instance*.
(2) We generate a human-curated *workflow specification* for a disease by gathering standard treatment guidelines from various web sources.
(3) We propose graph-based conformance measures to evaluate the conformance of workflow instance with the corresponding workflow specification of a standard treatment plan.

Scope of the work. Our study is based on relation extraction, machine learning techniques, and the comparative evaluation aspect of the workflow. Our focus is mainly on establishing the conformance methodology and showing the results based on the datasets that are publicly available. In this approach, any of our machine learning modules can be replaced by other best techniques to improve the qualitative performance of the system.

The rest of the article is organized as follows. In Section 3, we describe the overall architecture of our system. Next, in Section 3.2, we describe an MLPNN to extract a workflow instance of the discharge summary. In Section 3.3, we propose conformance measures to check conformance of a workflow instance with its respective workflow specification/treatment plan. In Section 4, we elaborate conformance checking methodology on a case study. Finally, in Section 5, we illustrate our experiments on three diseases with real discharge summaries and conclude our findings and discussions.

2 RELATED WORK

Workflow instance. Generating workflow instance from the discharge summary constitutes extracting medical entities and relationships among them. We extract medical entities by using machine learning techniques developed in our previous work that has shown comparable performance with state-of-the-art systems [44]. In this work, we focus on medical relation extraction. Identifying semantic relations among medical entities is a challenging task involving the extraction of context and structure of the text [2]. Hence, mere bag-of-words and co-occurrence techniques are not effective to detect these relations [2]. Significant prior research has focused on pattern-based methods where these structures are manually analyzed to build patterns for relations [1]. Even though such patterns were able to achieve good precision, they were not sufficient to identify all patterns for good recall. A recent I2b2 challenge [50] encouraged the research on identifying relationship types between medical problems, tests, and treatments. Most of participants in the challenge used a support vector machine (SVM) classifier for identification of the relationship type among the entities [11, 35, 41]. Among these, the SVM classifier [41] has been shown to be effective for this relation extraction task. Motivated by the recent success of deep learning techniques, we experimented with neural network models trained using novel similarity features in identifying the relationships.

Workflow specification. There are various well-known representations for treatment guidelines, such as GLIF [32] and PROforma [47]. GLIF is a common representation format that facilitates guideline sharing across organizations. These guidelines encompass appropriate usage of specific technologies, surgical procedures, and tests as part of clinical care [32]. PROforma is more focused on designing guideline modeling language for decision support systems and guidelines. However, each of these representations were developed to incorporate different kinds of knowledge addressing particular modeling challenges [13]. In this work, we create a workflow-based representation with semantic relations that allows us to check conformance of a discharge summary's treatment plan. Other systems that are relevant to our study include PRODIGY [37] and GUIDE [38]. PRODIGY is a clinical decision system that advises on prescriptions when provided with a problem header, patient scenario,

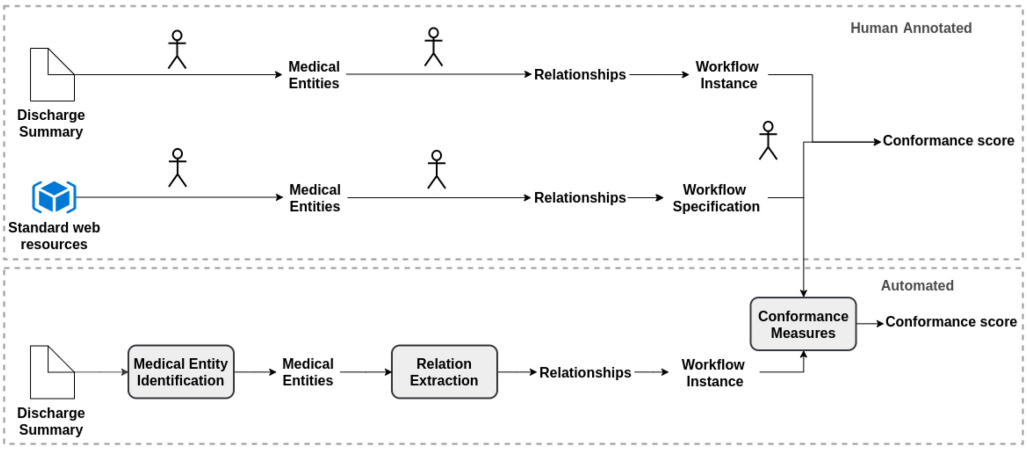


Fig. 1. Conformance checking methodology. Top: Human-annotated process where the standard resources and discharge summaries are manually processed to identify medical entities and relationships. The respective relationships are connected to form workflow representations that are then compared to get the conformance score. Bottom: The automated process. The discharge summary is automatically processed by using *Medical Entity Identification* and *Relation Extraction* to extract medical entities and relationships to get *a workflow instance*. *Conformance Measures* are used to automatically compute conformance of workflow instance with manually created workflow specification.

and therapy. GUIDE is a clinical care workflow management system for effective utilization of organization resources such as supporting the ward, pharmacy, imaging, laboratories, and external resources [38]. These tools are built to address different problems and cannot be compared with our conformance checking methodology, where we use publicly available treatment guidelines and discharge summaries.

Conformance checking. Even though several studies [3, 6, 33] were focused on computing the percentage of population who have not received a specific therapeutic procedure, the conformance checking on a patient's whole treatment plan is a less explored problem. Our work closely aligns with CareGap [16, 52], which correlates the physician's treatment decisions with the treatment guidelines provided by a clinical decision support system for adult soft tissue sarcoma patients. In CareGap, the treatment plan is mentioned as a single sentence, such as "Wide excision surgery and adjuvant radiotherapy pre- or post- operation." In addition, CareGap is dependent on the clinical decision support system in recommending a treatment plan of only one disease with specific attributes. In our work, we considered the treatment plan of a patient's entire visit by using the discharge summary. Our workflow representation can incorporate various symptoms, tests, and treatments and their relationships to any disease, and in general, our solution can be easily extended to other diseases.

3 CONFORMANCE CHECKING

We developed a conformance checking methodology to process discharge summary and standard resources data to create workflow specification and workflow instance, which are then compared to derive conformance score. The methodology, which includes *human-annotated* and *automated* components, is shown in Figure 1. In human-annotated processing, to ease manual effort, as a first step each document is processed by a MetaMap biomedical tool [4] to extract entities that are then human curated to get the ground-truth entities. These entities are connected by relationships to form workflow representation (detailed in Section 4). In automated processing, the discharge summary is passed through a *Medical Entity Identification* module to generate medical entities that are then processed by *Relation Extraction* to generate relationships that are connected to form a workflow

instance. This instance is compared with human-curated workflow specification to compute a conformance score using *Conformance Measures*. Each of these modules are elaborated in the following sections.

3.1 Medical Entity Identification

Medical Entity Identification is analogous to the traditional Named Entity Recognition (NER) task in the medical domain. In a sequence labeling problem like NER, the goal is to predict the sequence of output labels for a given sequence of the input text. A typical approach for the natural language NER task is to use BIO tagging to tag the named entity. In BIO tagging, *B*, *I*, and *O* tags refer to *Beginning* of a named entity, *Intermediate* words of the named entity, and *Others*, respectively. We used the popular sequence labeling machine learning model Conditional Random Fields (CRF) for tagging the text. CRF is modeled using features such as term features, part-of-speech (POS) tags, phrase tags, prefix, suffix, and UMLS features. CRF uses BIO tagging to automatically identify medical entities of semantic types: symptom, test, treatment, and medication. Our CRF model has achieved comparable performance with state-of-the-art techniques as described in our previous work [44]. As a next step, these medical entities are passed to the Relation Extraction module.

3.2 Relation Extraction

The Relation Extraction module identifies the relationship types (from eight pre-defined classes: TrATr, TrAP, TrSP, TrATe, TeAP, TeSP, TeATe, and PAP) among the medical entities. We used a neural network–based approach to automatically identify these relationships.

3.2.1 Multi-Layer Perceptron Neural Network. Motivated with the recent success of neural networks in many NLP applications, we explored a multi-layer neural network to extract relationships (edges) of workflow instance. MLPNN learns a function $f : R^n \rightarrow R^c$, where n is the number of features and c is the number of output class labels[1] [10]. Each layer computes $h = tanh(Wx + b)$, where $h \in R^k$ is the output, $W \in R^{nxk}$ is the weight matrix, $x \in R^n$ is the input features, $tanh$ is the activation function, $b \in R^k$ is the bias, and n and k are input and output dimensions of layer.

Output layer. For the output layer, the Softmax activation function is used to compute probability estimates of the output labels:

$$p_j = \frac{e^{\hat{y}_j}}{\sum_{i=1}^{C} e^{\hat{y}_i}}, j = 1 \ldots C, \tag{1}$$

where $\hat{y}_j$ is the output of the last layer and C is the total number of class labels.

Loss function. Cross entropy is used as the loss function. For our multi-class classification, cross entropy loss is given as

$$loss = -\sum_{j=1}^{C} y_j \log(p_j), \tag{2}$$

where y_j is a binary indicator; it is equal to 1 for correct label j and 0 otherwise. C is the total number of class labels.

Training. To train our model, we used a stochastic gradient descent algorithm to optimize our loss function. The neural network parameters are updated during training using back propagation. In the next section, we discuss the features used to train our model.

3.2.2 Features Used for Relation Extraction. We used three kinds of features for building our machine learning model: lexical, context, and similarity.

Lexical features. Lexical features are created from a word string of relation arguments. These features include the word string of relation arguments, an individual word and lemma of the words forming relation arguments,

[1]Output class labels include eight pre-defined relationship types and one NULL relationship. The NULL relation implies that there is no relationship between the pair of entities.

Table 1. Features for the Sentence "*Colonoscopy* Showed a *malignat mass* in the Proximal Ascending Cecum"

Type	Feature	Example
Lexical features	Word string of relation arguments	colonoscopy, (cannot match "malignat mass" because of misspelling)
	Individual word	colonoscopy, mass
	Lemma	colonosco-, malign-, mass
	Concept types	Test, symptom
Context features	Word in between relation arguments	showed, a
	POS tags	VB (verb), DT (determiner)
	Chunk tags	B-VP(verb phase), B-NP (noun phase)
	Two words preceding first argument	—
	Two words succeeding first argument	showed, a
	Two words preceding second argument	showed, a
	Two words succeeding second argument	in, the
	Word sequence in-between	showed a
	POS tag sequence in-between	VB DT
	Chunk tag sequence in-between	B-VP B-NP
Similarity features	Levenshtein distance similarity	colonoscopy, mass, malignat (matches the word "malignant" since their Levenshtein distance is 1)

Note: We intentionally misspelled the term *malignant* to show the utility of the similarity feature.

and concept types. Table 1 shows example features from the sentence "Colonoscopy showed a malignat mass in the proximal ascending cecum" for entities "colonoscopy" and "malignant mass." For these entities, the lexical features are "colonoscopy," "mass," "colonosco," "malign," and so forth.

Context features. Context features deal with word string contents in between the relation arguments. Context features include any word, POS tag, chunk tag, words preceding and succeeding the relation arguments, any concept type used in between the relation arguments, word sequence, POS tag sequence, and chunk tag sequence. Only top frequent sequences are considered as features that have occurred more than once in the training dataset. The number of top words, POS tags, and chunk sequences identified are 187, 197, and 221, respectively. The shortest dependency path between the pair of entities is used as the feature. In the case of inter-sentential relations, dependency trees of two sentences are combined to form a single tree [48]. Dependency of entities across these sentences is identified using the shortest path in the newly constructed tree. Table 1 shows example context features based on neighboring word strings "showed a" and "in the."

Similarity features. The preceding lexical and context features are susceptible to misspellings and minor lexical variations in the word strings. To overcome this, the Levenshtein distance similarity metric is used. Levenshtein distance is the number of additions, deletions, and substitutions needed to change one string to another. Empirically, we choose a Levenshtein distance of 2 to compute the similarity. From Table 3 (shown later), in the Similarity Features row, even though the word *malignat* is misspelled in the given sentence, the Levenshtein distance helps us find the nearest matching word *malignant* with less than two transformations. However, this metric works better only when the string length is greater than 4. For example, the word *mass* may match words such as *as* or *ass*, which is not correct.

The lexical and context features were used as defined in existing studies [41]. The similarity features are the novel features introduced in this work.

3.3 Conformance Measures

The workflow specification created from various sources contains near exhaustive information about the treatment plan of a particular disease, in that it contains many of the possible symptoms, tests, treatments, and medications, whereas a workflow instance contains only information about current clinical events based on the condition of the patient for a particular duration of time from entry to discharge. Given a discharge summary, conformance checking involves the comparison of the workflow instance of a patient with the workflow specification of the disease of which the patient is diagnosed. The workflow specification for particular a patient is identified by extracting the disease terms from our previous work [44]. The number of disease terms present in the discharge summary is used to assign the treatment plan workflow specification of the disease for the discharge summary.

The conformance measure reveals how well events in a workflow instance are in compliance with the workflow specification. We experimented with various measures for checking conformance. Our conformance measures are mainly inspired by a precision metric ($|a \cap b|/|a|$), where a and b are sets. They signify the precision with which workflow instance W_i adheres to workflow specification W_s. The higher measure implies that the workflow instance conforms to the treatment plan used in the standard specifications. We cannot use a recall metric for our problem because as the workflow specification is very large compared to a workflow instance, recall is always near zero.

3.3.1 Node Measure. The node elements are compared considering syntactic and semantic aspects of natural language. The syntactic aspect includes the edit distance of the two strings (Levenshtein distance) [25], where the semantic aspect focuses on the meaning of the words of two labels using WordNet and UMLS [7, 22]. Word edit distance and character edit distance are used to quantify the similarity of the elements. These are further enhanced by stemming and stopword removal techniques. A major challenge with discharge summaries is handling the synonyms of medical terms. For example, the medical term *cancer* is also represented using names such as *adenocarcinoma*, *carcinoma*, and *ademos*. A biomedical ontology such as UMLS is used to handle synonyms, where a cluster of similar medical terms is represented with a concept. For example, all medical terms of the concept *cancer* are clustered and represented with a concept unique identifier (CUI). CUIs along with text values are used to compare the elements. Given two graphs $g_i = (V_i, E_i, L_i)$ and $g_s = (V_s, E_s, L_s)$, V_i and V_s are nodes, E_i and E_s are edges, and L_i and L_s are labels of the corresponding graphs. The conformance score of g_i is defined as

$$CS_{node} = \frac{|g_i \cap g_s|}{|g_i|}, \tag{3}$$

where $|g_i \cap g_s|$ denotes the common nodes in g_i and g_s, and $|g_i|$ and $|g_s|$ denote the number of nodes in g_i and g_s, respectively.

3.3.2 Graph Measure. Similarity based on common nodes and edges is represented as the graph measure [29]. Let $CE(g_i, g_s)$ be the number of common edges in g_i and g_s; the conformance score based on the total number of matched nodes and edges is defined as

$$CS_{graph} = \frac{1}{2}\left(\frac{|g_i \cap g_s|}{|g_i|} + \frac{|CE(g_i, g_s)|}{|E_i|}\right), \tag{4}$$

where $|g_i \cap g_s|$ denotes the common nodes in g_i and g_s, $|CE(g_i, g_s)|$ denotes number of common edges, E_i is the number of edges of g_i, and $|g_i|$ and $|g_s|$ denote the number of nodes in g_i and g_s, respectively.

3.3.3 Trace Measure. A trace is defined as the sequence of events in the execution of workflow specification. In conformance checking, a trace of W_i has to be mapped with the traces of workflow specification to quantify their similarity [15]. However, a workflow specification can have a large number of traces, which makes it difficult to compute the longest common subsequence for each trace. Thus, we define a metric considering the sequences up to three consecutive entities—that is, a window of size 3. Since we are using an overlapping and connected sub-sequence window approach, we can achieve the sequence order present in the workflow

instance. Considering the complexity of long traces, we believe that these triplet (window of size 3) sequences instead of long traces would be a reasonable assumption in the conformance measure. Let $CT(g_i, g_s)$ represent the common three consecutive nodes in g_i and g_s, and let T_i be the number of triplets in g_i; the conformance score based on the total number of matched triplets is defined as

$$CS_{trace} = \frac{1}{3}\left(\frac{|g_i \cap g_s|}{|g_i|} + \frac{|CE(g_i, g_s)|}{|E_i|} + \frac{|CT(g_i, g_s)|}{|T_i|}\right). \tag{5}$$

3.3.4 Common Maximal Subgraph Measure. Common maximal subgraph similarity is developed in Bunke and Shearer [8] and Labriji et al. [23]. Let Common maximal subgraph, denoted by $CMS\ (g_i, g_s)$, is the largest common subgraph between g_i and g_s. The conformance score based on the maximal subgraph is defined as

$$CS_{cms} = \frac{|CMS(g_i, g_s)|}{|g_i|}. \tag{6}$$

4 HUMAN-ASSISTED CONFORMANCE CHECKING

4.1 Data Preparation

Workflow specification of a disease is extracted from the treatment guidelines provided by healthcare web sources such as Mayo Clinic [27], WebMD [55], and the National Comprehensive Cancer Network [17]. To create workflow instances, we collected discharge summaries of three diseases, namely colon cancer, coronary artery disease (CAD), and brain tumor, from the THYME corpus [45] and the MIMIC III clinical database [19]. Our dataset is composed of 300 discharge summaries with 100 each belonging to the three diseases.

Our goal for workflow extraction from these documents has two challenges: (1) extraction of medical entities/terms from the text and (2) identifying relationships between the entities. Medical entities from these documents are first extracted by using a MetaMap biomedical tool [4] for initial annotations. These initial entities are human curated to get the ground-truth entities of semantic types: problem, test, and treatment (treatment/medication). These entities are connected using eight pre-defined relationship types. Entities with the same semantic type are connected using the *associated with* relationship type, hence the relationships: TeATe, TrATr, and PAP. This relationship type is also used to connect test and treatment by TeATr. Entities with the semantic types of problem and test/treatment are connected using the *administered for* and *shows* relationship types: TrAP, TeAP, TrSP, and TeSP. These relationships implicitly incorporate temporal ordering of medical entities. For example, the *administered for* relationship type implies that the *problem* occurred before *test/treatment* and the *shows* relationship type implies the other way round. In the case of the relationship type *associated with*, the entity that is mentioned earlier is considered to be as the one that occurred earlier. These individual relationship pairs may span within the sentence or across the multiple consecutive sentences. If the individual relationships are not connected with one or the other pairs, this results in gaps in the workflow. Hence, assuming that the medical entities in a discharge summary of one particular disease are related, we fill any gaps by connecting the individual relationship pairs to the nearest entities with the *associated with* relationship type to form a workflow.

Conformance scores are annotated based on comparing the workflow instance created from the discharge summary and the corresponding workflow specification of the disease. Each discharge summary is manually annotated by a conformance score on a scale from 0 to 1, where 0 = no conformance, 0.5 = partial conformance, and 1 = total conformance.

The annotation procedure for each of the preceding tasks is a two-step process as presented in Styler et al. [45]:

(1) *Annotation phase*: Two annotators were provided with discharge summaries for annotating medical entities and relationships of workflow instances and conformance scores with respect to workflow specification. Although the kappa inter-annotation agreement among annotators is approximately 0.70 in each of the tasks, we passed the annotations through a verification phase where the conflicts were resolved.

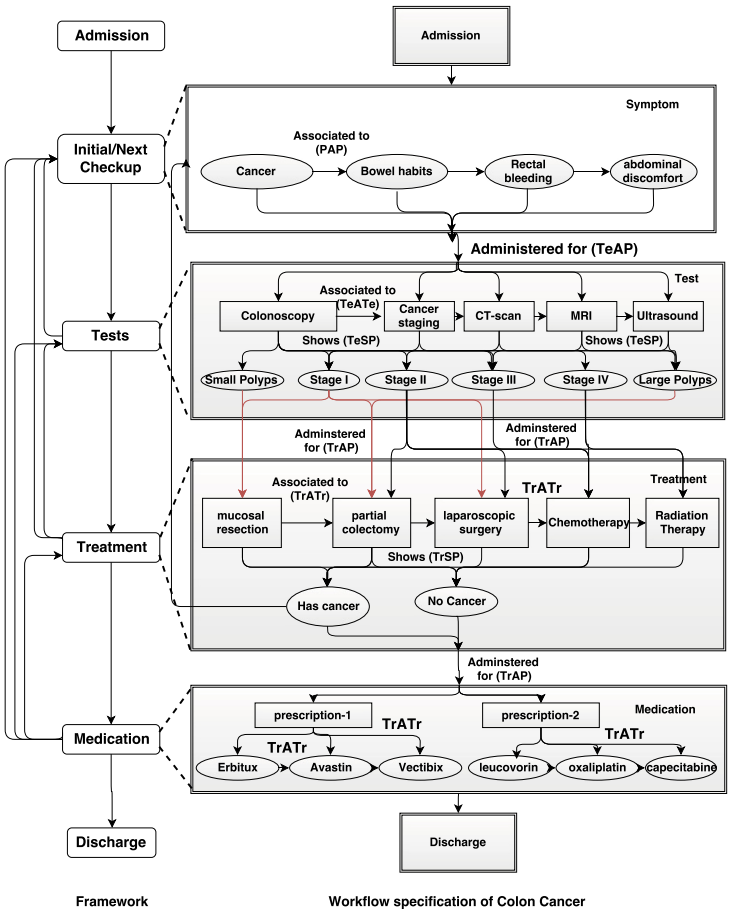


Fig. 2. Workflow Specification. Left: A high-level treatment plan framework with *activities* shown in text boxes and arrows show the flow of activities. Each activity is composed of low-level events shown in the large rectangular boxes on the right side. Right: Treatment plan of colon cancer in the workflow representation. The rectangular boxes represent test/treatment/medication, and ellipses represent sign/symptom/outcome. These entities are connected using pre-defined relationships.

(2) *Verification phase:* In the verification phase, a third annotator performed adjudication on the annotations to produce ground-truth annotations.

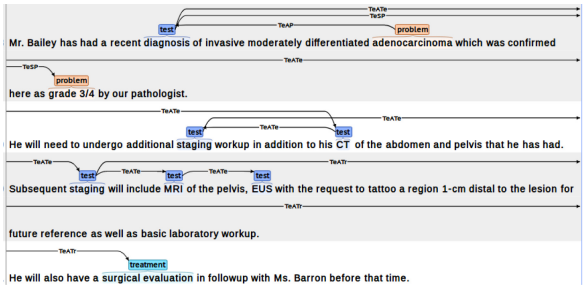
These ground-truth annotations are used to evaluate relation extraction and conformance measures as given later in Section 5.

4.2 Colon Cancer Example

4.2.1 Workflow Specification. The workflow specification of a disease is modeled based on standard treatment guidelines available on standard web sources. For each disease, we manually extract medical entities and their relationships from treatment guidelines and model them as a workflow specification. Figure 2 shows a sample workflow specification of colon cancer.[2] A *high-level specification* of the patient's treatment activities, shown on

[2]These workflow figures are only for representational purposes. Our solution for relation extraction and conformance checking are performed at a much lower level—that is, on medical entities and relationships.

Table 2. Extraction of the Workflow Instance from Discharge Summary Snippet (Left) Sample Text of Discharge Summary (Right) Workflow Instance with Medical Entities and Relations

Discharge Summary Snippet	*Workflow Instance*
Mr. Bailey has had a recent diagnosis of invasive moderately differentiated adenocarcinoma which was confirmed here as grade 3/4 by our pathologist. He will need to undergo additional staging workup in addition to his CT of the abdomen and pelvis that he has had. Subsequent staging will include MRI of the pelvis, EUS with the request to tattoo a region 1-cm distal to the lesion for future reference as well as basic laboratory workup. He will also have a surgical evaluation in followup with Ms. Barron before that time.	

the left, includes *Admission*, *Initial/Next Checkup*, *Tests*, *Treatment*, *Medication,* and *Discharge*. A treatment plan can be a sequence of activities (*Admission* → *Initial Checkup* → *Tests* → *Treatments* → *Medication* → *Discharge*). However, in reality, tests or treatments can be conducted multiple times for chronic conditions. Such scenarios can be captured by back links that show the possibility of flow from *Treatments* → *Tests*, *Tests* → *Initial Checkup*, or within *Tests* itself. In the right half of Figure 2 is a workflow snippet of *Colon Cancer diagnosis*. The initial checkup is the first activity carried out after admission, which is mostly concerned with the condition of the patient. A few *Symptoms* of *colon cancer* are *cancer*, *bowel habits*, *rectal bleeding*, and *abdominal discomfort*. *Tests* are conducted based on these *symptoms* from the *initial checkup*. *Tests* activity contains diagnostic procedures such as *Colonoscopy*, *Cancer staging*, *CT-scan*, *MRI*, and *ultrasound*. One or more of these tests are performed to diagnose the patient based on colon cancer symptoms. The preceding tests may reveal any of the *findings*: stage I, stage II, stage III, stage IV, small polyps, and large polyps. Therapeutic procedures are conditioned upon the findings of *Tests*, such as stage and severity of the cancer. A few cases are listed next:

- For findings, *stage I* and *small polyps*, the *therapeutic procedure* of mucosal resection is carried out.
- If *stage I* has *large polyps*, laparoscopic surgery is performed.
- For stages II, III, and IV, one or more treatments from laparoscopic surgery, partial colectomy, chemotherapy, and radiation therapy are performed.

In an ideal post-procedure scenario, cancer can either be completely removed or partially removed. If the cancer cells are completely removed (*no cancer*), the patient would be *discharged* with the *prescriptions* and *suggestions* represented by activity *Medication*. If the cancer cells are not removed completely (*has cancer*), then the workflow continues from activity *Initial/Next Checkup*, which is represented by the separate sub-workflow following the treatment plan.

4.2.2 Workflow Instance of a Discharge Summary Diagnosed for Colon Cancer. A workflow instance consists of all events that happen in the clinical time line. Table 2 shows an example discharge summary snippet and the

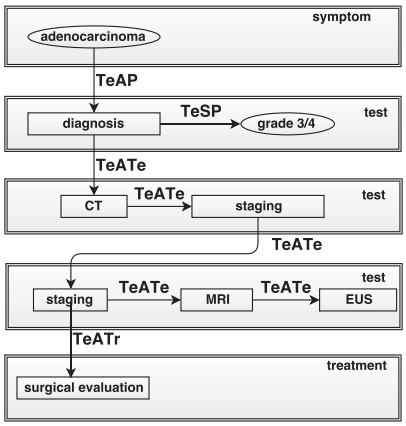


Fig. 3. Graphical form of the workflow instance of Table 2. Each box shows medical entities of the same head semantic type as given in Figure 2. Tests of each sentence are presented in separate boxes for better representation. However, comparison of the workflow instance with workflow specification happens at the entity and relationship level.

workflow instance containing medical terms and relations among them. Even though the occurrence of future or follow-up procedures is uncertain, they are included in the workflow instance to check whether the planned treatment adheres to workflow specification. Figure 3 shows the graphical representation of the workflow instance. This instance explains that the patient has been *diagnosed* with *adenocarcinoma*, which resulted in *grade 3/4*. The patient will further undergo a sequence of *diagnostic procedures*, such as *CT*, *staging*, *MRI*, and *EUS*, followed by treatment, *surgical evaluation*. Figure 3 is shown only for representational purposes. However, medical entities and their relationships are the atomic components of the workflow instance that are further used.

4.2.3 Mapping the Workflow Instance to Workflow Specification. Mapping of the workflow instance to workflow specification starts with the first event, *adenocarcinoma*. The problem *adenocarcinoma* maps to its synonym *cancer* of workflow specification. *Diagnosis* is a generic term for the high-level entity *test*, which resulted in *grade 3/4*, implying *stage III/IV*. Then, the workflow continues with a sequence of tests connected by the *associated with* relation (TeATe). This set of tests, including *CT*, *MRI*, *staging*, and *EUS* is mapped to *CT-scan*, *MRI*, *cancer staging*, and *ultrasound*, respectively. Finally, the flow is from *tests* to *treatment*, where *surgical evaluation* refers to the treatment *laparoscopic surgery*. Thus, the given workflow instance completely conforms with the workflow specification of *Colon cancer* with a conformance score of 1.0.

5 RESULTS

5.1 Workflow Instance Relation Extraction

5.1.1 Experimental Setup. The main challenge in our work is to automatically identify the relationships among entities present in discharge summaries. We are mainly focused on entities with semantic types: medical problem, test, and treatment (treatment plus medication). In this work, we have used a total of eight relationships between the semantic types: TeATe, TrATr, PAP, TeATr, TrAP, TeAP, TrSP, and TeSP. In our experiments, we assume that the medical terms and their semantic types are already known as in other existing works [40, 43]. Our dataset is composed of 300 discharge summaries collected from the THYME corpus [45] and the MIMIC III clinical database [19]. Statistics of the distribution of relationship types in our dataset is given in Table 3.

To have a comparative analysis on a highly class imbalanced dataset, performance metrics such as precision, recall, and *F*-score were used. Standard classification metrics of precision (P), recall (R), and *F*-score (F) were used to evaluate workflow instance relation extraction. The whole dataset is randomly divided into training (70%) and

Table 3. Data Statistics of the Whole Dataset: Relationship Types and Number of Instances

Relationship Types	Total	Train	Test
ALL	3,929	2,559	1,370
TrAP	525	354	171
TrSP	453	296	157
TrATr	1205	819	386
TeAP	98	50	48
TeSP	240	144	96
TeATe	197	147	150
TeATr	239	132	107
PAP	872	617	255

Table 4. Comparison of Different Classification Models for Identifying All Relationship Types Among Medical Entities (Problem, Test, and Treatment)

Classifier	Precision	Recall	*F*-Score
Co-occurrence (baseline)	0.53	**0.90**	0.67
MLP NN	**0.8**	0.76	**0.77**
SVC (linear, C=0.1)	0.77	0.73	0.74
SVC (linear, C=0.025)	0.8	0.68	0.72
SVC (rbf, c=1000)	0.8	0.66	0.7
Decision tree	0.66	0.67	0.65
Nearest neighbors	0.49	0.55	0.49

Note: The baseline system has given better recall, and MLP NN has given good performance with a high precision and *F*-score.

testing (30%) using train_test split [36]. We used the Scikit-learn library [36] to implement the machine learning classifiers.

5.1.2 Comparative Performance of MLPNN with Other Classifiers. Relation extraction among medical entities is formulated as a multi-class classification task [11]. We implemented a baseline system based on a co-occurrence approach that has shown value on relation extraction tasks in the literature [51]. We experimented with several well-known classifiers based on SVM, decision trees, and *k*-nearest neighbor. Each classifier's hyper-parameter values were empirically chosen based on fivefold cross validation on the training dataset. Table 4 shows the comparative results of different classifiers, namely MLPNN [10], SVM [9], decision tree [39], and nearest neighbor classifier [20]. Each classifier is modeled using all of the features. Results show that the co-occurrence baseline system had higher recall than any other approach, as it considers all possible relations that can exist among the entities [51]. However, MLPNN outperformed all systems on precision and *F*-score. MLPNN showed 3% improvement over the best accuracy of SVM (cost-weight $C = 0.1$) with statistical significance of a one-way analysis of variance (ANOVA) test p-value < 0.05.

5.1.3 Class-Wise Performance of MLPNN with All Features. Table 5 summarizes the class-wise results of MLPNN modeled based on all of the features explained in Section 3.2.2. We see that the relationships TrATr, TeATe, and PAP have high accuracy, as the similar semantic-type neighbor entities are grouped together in the construction of workflow. The relationship TeAP has shown relatively low accuracy because of a low number of samples (refer to Table 3).

Table 5. Class-Wise Performance of the MLPNN Classifier with All Features

Relationship Types	Precision	Recall	*F*-Score
All	0.78	0.76	0.77
TrAP	0.63	0.66	0.64
TrSP	0.83	0.7	0.76
TrATr	0.79	0.9	0.84
TeAP	0.65	0.23	0.34
TeSP	0.83	0.75	0.78
TeATe	0.87	0.9	0.88
TeATr	0.66	0.5	0.57
PAP	0.87	0.81	0.84

Note: A high *F*-score was noted for relationship types with higher numbers of samples.

Table 6. Performance Evaluation of Lexical (L), Context (C), and Similarity (S) Features Using Feature Ablation

Features Used	Precision	Recall	*F*-Score	Difference
All features (L+C+S)	0.78	0.76	0.77	—
-Lexical (C+S)	0.6	0.54	0.56	**−0.21**
-Context (L+S)	0.73	0.7	0.71	**−0.06**
-Similarity (L+C)	0.76	0.72	0.74	**−0.03**

Note: Each feature has shown improvement in the performance of the model.

5.1.4 Contribution of Each Feature. To examine the contribution of each feature, we ablate each feature from the model. Table 6 summarizes the results with feature ablation. The first row shows the performance of the MLPNN model trained with all of the features. The remaining rows show the result with the removal of each feature. The Difference column in Table 6 shows the significance of lexical, context, and semantic features on the model by 21%, 6%, and 3%, respectively.

5.2 Conformance Score

The conformance score is computed between the discharge summary and the standard treatment plan. A discharge summary is represented as a workflow instance with medical entities as nodes and relationships as edges. Each workflow instance is compared with the corresponding disease's workflow specification (standard treatment plan) and manually labeled with a ground-truth conformance score on a scale from 0 to 1 explained in Section 4.1. We automatically compute conformance scores for each workflow instance based on the measures provided in Section 3.3. These conformance measures are validated by finding their correlation with ground-truth scores [30] using the Spearman rank correlation coefficient (Spearman rho) and the Pearson product-moment correlation coefficient (PPMCC) [24]. The Spearman rho finds the monotonicity in the ranking order of scores but does not measure the difference in the magnitude. PPMCC measures the linear dependence of metric values with manual scores. These coefficients imply the correlation of conformance measures with that of ground-truth scores on discharge summaries. Table 7 summarizes the results of these measures on three diseases. Results show that the trace measure shows slightly better evaluation in relative ordering of discharge summaries. The node measure, graph measure, and trace measure show high Pearson coefficients for CAD, colon cancer, and brain tumor, respectively. This leads to the fact that no unique correct measure exists for conformance and that one

Table 7. Spearman Rho and PPMCC Coefficients on the Workflow Instances of Three Diseases Indicating the Correlation of Our Conformance Measures with Manually Annotated Scores

Disease	Conformance Measure	Evaluation Metric	
		Spearman Rho	Pearson Coe
Colon cancer	Node measure	0.661	0.633
	Graph measure	0.768	**0.713**
	Trace measure	**0.776**	0.712
	CMS measure	0.443	0.379
CAD	Node measure	0.698	**0.708**
	Graph measure	0.728	0.707
	Trace measure	**0.729**	0.703
	CMS measure	0.501	0.452
Brain tumor	Node measure	0.628	0.685
	Graph measure	0.735	0.705
	Trace measure	**0.767**	**0.710**
	CMS measure	0.452	0.386

Note: All coefficient values are statistically significant with p-value < 0.01.

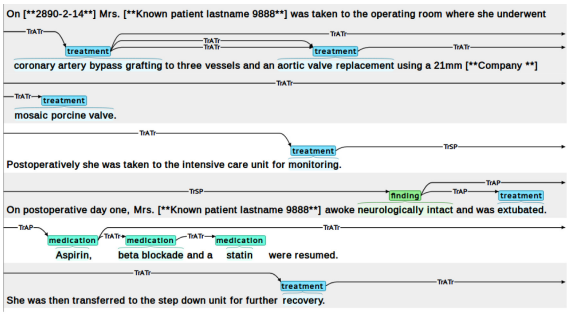
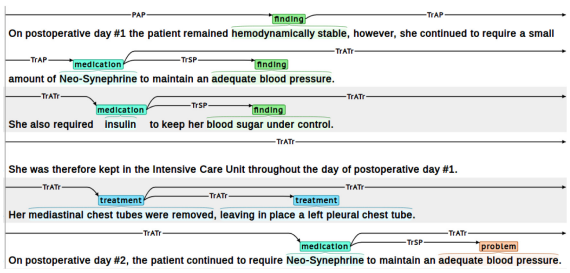
of the measures can be picked based on the application [49]. Low correlation of the CMS measure implies that common maximal sub-graph might not essentially contain the significant medical entities in the treatment plan.

6 DISCUSSION

Our neural network approach that incorporates lexical, context, and similarity features generates relationships among the medical entities to give a workflow representation from free text and outperformed traditional models. Performance comparison with other works is not possible because of the difference in the datasets. Thus, we have experimented with other traditional baseline classifiers such as SVM and decision trees. Specifically, our method using MLPNN achieved an F-score greater than 0.77 compared to SVM and decision trees with corresponding F-scores of 0.74 and 0.65, respectively. The novelty of this approach is not only in improving performance over traditional baseline models but rather in using the novel similarity features that work better in identifying relationships.

The significance of conformance scores is illustrated using two different patient scenarios. Table 8 shows the discharge summary snippets and their corresponding workflow instances of two patients with ground-truth conformance scores ($CS_{groundtruth}$) of 1 and 0.5, respectively. In the first row, a patient diagnosed with CAD underwent *Coronary Artery bypass graft* and *aortic valve replacement* using *mosaic porcine valve*. The different conformance measures for the workflow instance are shown in the first cell. For each measure, a different threshold can be chosen empirically to estimate a low or high conformance score of a discharge summary. The second row shows the diagnosis of a diabetic patient for CAD with a conformance score of 0.5. The patient had to continue with the medications *Neo-synephrine* and *insulin* for maintaining *blood pressure* and *sugar under control*, respectively. Here, the terms *insulin* and *sugar under control* are specific to diabetes. Workflows of such specific conditions have resulted in a lower conformance score with the specification (compared to the first row). Identifying such less conformance patients has two advantages: first, to help a doctor in identifying such co-morbidity cases and pay focused attention, and second, these co-morbidity cases can be leveraged in formulating co-morbid specific treatment plans. For example, a CAD treatment plan specific to a patient with diabetes can be suggested for future patients. Thus, the proposed conformance measures would allow us to effectively identify patient records that would help in modifying the treatment plan used by healthcare providers.

Table 8. Extraction of the Workflow Instance from Discharge Summary Snippet (Left) Sample Text of Discharge Summary (Right) Workflow Instance with Medical Entities and Relations

Discharge Summary Snippet	***Workflow***
On [**2890-2-14**] Mrs. [**Known patient lastname 9888**] was taken to the operating room where she underwent coronary artery bypass grafting to three vessels and an aortic valve replacement using a 21mm [**Company **] mosaic porcine valve. Postoperatively she was taken to the intensive care unit for monitoring. On postoperative day one, Mrs. [**Known patient lastname 9888**] awoke neurologically intact and was extubated. Aspirin, beta blockade and a statin were resumed. She was then transferred to the step down unit for further recovery. ($CS_{node} = 0.823, CS_{graph} = 0.778, CS_{trace} = 0.698, CS_{cms} = 0.142, CS_{groundtruth} = 1.0$)	
On postoperative day #1 the patient remained hemodynamically stable, however, she continued to require a small amount of Neo-Synephrine to maintain an adequate blood pressure. She also required insulin to keep her blood sugar under control. She was therefore kept in the Intensive Care Unit throughout the day of postoperative day #1. Her mediastinal chest tubes were removed, leaving in place a left pleural chest tube. On postoperative day #2, the patient continued to require Neo-Synephrine to maintain an adequate blood pressure. ($CS_{node} = 0.568, CS_{graph} = 0.489, CS_{trace} = 0.405, CS_{cms} = 0.055, CS_{groundtruth} = 0.5$)	

Note: Conformance measures of discharge summaries are provided under each snippet.

Further, we evaluated our solution by considering discharge summaries with a conformance score less than or equal to 0.5 and a conformance score greater than 0.5. These two sets of discharge summaries gave a few interesting patterns. The therapeutic procedure *hemicolectomy* (removal of part of colon) is often followed by *anastomotic leak* → *emergency surgery and ileostomy* → *blood transfusion* (DocID: ID015_clinic_043). In addition,

most common co-morbidities of the diseases have been identified. For example, the *Coronary Artery Disease* medical entity often occurs with *diabetes*, *high cholesterol*, and *hypertension*, and, *Brain tumor* has co-morbidities such as *eye blurring*, *aphasia*, and *urinary dysfunction*. In some cases, a discharge summary can contain diagnosis of multiple disorders. For example, a discharge summary (DocId: ID197_clinic_579) contains procedures (e.g., *colonoscopy*, *Appendectomy*, and *Cystoprostatectomy*) that are related to *colon cancer*, *appendix*, and *prostate gland*. By analyzing the workflow instances of discharge summaries, we found new diagnostic procedures of *colon cancer* such as *proctocolectomy* and *anoscopy*, which are not present in the treatment guidelines extracted from web sources. Populating such knowledge from discharge summaries can in turn improve the standard treatment plans.

Our study has some limitations. First, diseases with co-morbidities involve high variability of treatment processes that are often complicated. It is not straightforward to implement the treatment guidelines of individual disease directly. Such scenarios require specialized treatment plans for a specific set of co-morbidities. Second, in this work, all procedures are considered to be at the same level of detail. For example, *chemotherapy* and its drug regimes, such as *FOLFOX* and *FOLFIRI* are considered as separate procedures. However, in general, these drug regimes are part of the therapeutic procedure *chemotherapy*. Third, we only used publicly available web sources treatment guidelines to model workflow specification. Given the flexibility of workflow representation, any new treatment guidelines can be easily incorporated in our solution.

However, our study has important applications: first, concise representation of discharge summaries in a computable form (workflow) allows clinicians to check the patient history records in a shorter time frame; second, discovery of new treatment plans for different co-morbidities; and third, in a real-world setting, an individual health record can be documented for every 72 hours of patient stay. Provided that the individual health record is properly documented, the conformance score alerts the doctor to take note of the patient case if it deviates from the standard treatment plan. Our work is a proof of concept that demonstrates the importance of conformance checking for quality healthcare.

7 CONCLUSION

In this article, we addressed an important and challenging problem of determining the conformance of the patient's discharge summary to standard treatment plan. Our study was based on public domain data that was available as discharge summaries and treatment plans extracted from reputed health sites. The key contributions are (i) modeling the conformance problem as a problem of determining whether a workflow instance conforms to its workflow specification, (ii) determining workflow specification from treatment plans, (iii) extracting the workflow instance from the discharge summary using various classifiers with novel additional features, and (iv) attributing the conformance score for a discharge summary to its treatment plan. Our experimental results on three diseases and 300 discharge summaries show the viability of our solution. The aim of our work is a project wherein patients can upload their text discharge summary and get a conformance score with respect to standard treatment plans. The user will also be able to determine variation in treatments from standard plans. Thus, medical providers manage conflicts with patients in agreeing to a treatment plan. The key idea of workflow specification and its conformance to the workflow instance is powerful, and can be applied to other domains such as maintenance work and project management.

7.1 Data Availability

The data used in this work is available from the authors upon reasonable request.

7.2 Code Availability

Source code of this work is available from the authors upon request.

REFERENCES

[1] Asma Ben Abacha and Pierre Zweigenbaum. 2011. Automatic extraction of semantic relations between medical entities: A rule based approach. *Journal of Biomedical Semantics* 2, 5 (2011), S4.

[2] Sophia Ananiadou, Sampo Pyysalo, Jun'ichi Tsujii, and Douglas B. Kell. 2010. Event extraction for systems biology by text mining the literature. *Trends in Biotechnology* 28, 7 (2010), 381–390.

[3] Susan E. Andrade, Jerry H. Gurwitz, Terry S. Field, Michael Kelleher, Sumit R. Majumdar, George Reed, and Robert Black. 2004. Hypertension management: The care gap between clinical guidelines and clinical practice. *American Journal of Managed Care* 10, 7 Pt 2 (2004), 481–486.

[4] Alan R. Aronson. 2006. *MetaMap: Mapping Text to the UMLS® Metathesaurus®*. Bethesda, MD: NLM, NIH, DHHS.

[5] Iyad Batal, Hamed Valizadegan, Gregory F. Cooper, and Milos Hauskrecht. 2013. A temporal pattern mining approach for classifying electronic health record data. *ACM Transactions on Intelligent Systems and Technology* 4, 4 (2013), 63.

[6] L. Bessette, L.-G. Ste-Marie, S. Jean, K. S. Davison, M. Beaulieu, M. Baranci, J. Bessant, and J. P. Brown. 2008. The care gap in diagnosis and treatment of women with a fragility fracture. *Osteoporosis International* 19, 1 (2008), 79–86.

[7] Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research* 32, Suppl. 1 (2004), D267–D270.

[8] Horst Bunke and Kim Shearer. 1998. A graph distance metric based on the maximal common subgraph. *Pattern Recognition Letters* 19, 3–4 (1998), 255–259.

[9] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2, 3 (2011), 27.

[10] Ronan Collobert and Samy Bengio. 2004. Links between perceptrons, MLPs and SVMs. In *Proceedings of the 21st International Conference on Machine Learning*. ACM, New York, NY, 23.

[11] Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. 2010. NRC at i2b2: One challenge, three practical tasks, nine statistical systems, hundreds of clinical records, millions of useful features. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*.

[12] Marc Ehrig, Agnes Koschmider, and Andreas Oberweis. 2007. Measuring similarity between semantic business process models. In *Proceedings of the 4th Asia-Pacific Conference on Conceptual Modelling, Volume 67*. 71–80.

[13] Ariel Farkash, J. T. Timm, and Zeev Waks. 2013. A model-driven approach to clinical practice guidelines representation and evaluation using standards. *Studies in Health Technology and Informatics* 192 (2013), 200–204.

[14] N. Galie, N. M. Hoeper, A. Torbicki, J. L Cachiery, J. A. Barbera, M. Beghetti, et al.; ESC Committee for Practice Guidelines (CPG). 2009. Guidelines for the diagnosis and treatment of pulmonary hypertension: The Task Force for the Diagnosis and Treatment of Pulmonary Hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS), endorsed by the International Society of Heart and Lung Transplantation (ISHLT). *European Heart Journal* 30, 20 (2009), 2493–2537.

[15] Kerstin Gerke, Jorge Cardoso, and Alexander Claus. 2009. Measuring the compliance of processes with reference models. In *Proceedings of OTM Confederated International Conferences "On the Move to Meaningful Internet Systems."* Springer, 76–93.

[16] Esther Goldbraich, Zeev Waks, Ariel Farkash, Marco Monti, Michele Torresani, Rossella Bertulli, Paolo Giovanni Casali, and Boaz Carmeli. 2015. Understanding deviations from clinical practice guidelines in adult soft tissue sarcoma. *Studies in Health Technology and Informatics* 216 (2014), 280–284.

[17] National Comprehensive Cancer Network. 2015. NCCN Guidelines. Retrieved March 28, 2020 from https://www.nccn.org/professionals/physician_gls/default.aspx.

[18] Johns Hopkins. 2015. Johns Hopkins Medicine. Retrieved March 28, 2020 from https://www.hopkinsmedicine.org/healthlibrary/.

[19] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, et al. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3 (2016), 160035.

[20] James M. Keller, Michael R. Gray, and James A. Givens. 1985. A fuzzy *k*-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics* 4 (1985), 580–585.

[21] Vinod Khosla. 2014. "20 Percent Doctor Included" & Dr. Algorithm: Speculations and Musings of a Technology Optimist. In March 28, 2020 from https://www.khoslaventures.com/20-percent-doctor-included-speculations-and-musings-of-a-technology-optimist.

[22] Adam Kilgarriff and Christiane Fellbaum. 2000. WordNet: An electronic lexical database. *Language* 76, 3 (Sept. 2000), 706.

[23] Amine Labriji, Salma Charkaoui, Issam Abdelbaki, Abdelouhaed Namir, and El Houssine Labriji. 2017. Similarity measure of graphs. *International Journal of Recent Contributions from Engineering, Science & IT* 5, 2 (2017), 42–56.

[24] Ann Lehman, Norm O'Rourke, Larry Hatcher, and Edward Stepanski. 2005. *JMP for Basic Univariate and Multivariate Statistics: A Step-by-Step Guide*. SAS Institute Inc., Cary, NC.

[25] Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10 (1966), 707–710.

[26] Axel Martens. 2005. Consistency between executable and abstract processes. In *Proceedings of the 2005 IEEE International Conference on e-Technology, e-Commerce, and e-Service (EEE'05)*. IEEE, Los Alamitos, CA, 60–67.

[27] Mayo Clinic. 2015. Diseases and Conditions. Retrieved March 28, 2020 from http://www.mayoclinic.org/diseases-conditions.
[28] John J. V. McMurray, Stamatis Adamopoulos, Stefan D. Anker, Angelo Auricchio, Michael Böhm, Kenneth Dickstein, Volkmar Falk, et al. 2012. ESC guidelines for the diagnosis and treatment of acute and chronic heart failure 2012. *European Journal of Heart Failure* 14, 8 (2012), 803–869.
[29] Mirjam Minor, Alexander Tartakovski, and Ralph Bergmann. 2007. Representation and structure-based similarity assessment for agile workflows. In *Proceedings of the International Conference on Case-Based Reasoning (ICCBR'07)*, Vol. 7. 224–238.
[30] Hans Moen, Juho Heimonen, Laura-Maria Murtola, Antti Airola, Tapio Pahikkala, Virpi Terävä, Riitta Danielsson-Ojala, Tapio Salakoski, and Sanna Salanterä. 2014. On evaluation of automatically generated clinical discharge summaries. In *Proceedings of the 2nd European Workshop on Practical Aspects of Health Informatics (PAHI'14)*. 101–114.
[31] Jennifer S. Myers, C. Komal Jaipaul, Jennifer R. Kogan, Susan Krekun, Lisa M. Bellini, and Judy A. Shea. 2006. Are discharge summaries teachable? The effects of a discharge summary curriculum on the quality of discharge summaries in an internal medicine residency program. *Academic Medicine* 81, 10 (2006), S5–S8.
[32] Lucila Ohno-Machado, John H. Gennari, Shawn N. Murphy, Nilesh L. Jain, Samson W. Tu, Diane E. Oliver, Edward Pattison-Gordon, Robert A. Greenes, Edward H. Shortliffe, and G. Octo Barnett. 1998. The guideline interchange format: A model for representing guidelines. *Journal of the American Medical Informatics Association* 5, 4 (1998), 357–372.
[33] S. Pathare, A. Brazinova, and I. Levav. 2018. Care gap: A comprehensive measure to quantify unmet needs in mental health. *Epidemiology and Psychiatric Sciences* 27, 5 (2018), 1–5.
[34] Debprakash Patnaik, Patrick Butler, Naren Ramakrishnan, Laxmi Parida, Benjamin J. Keller, and David A. Hanauer. 2011. Experiences with mining temporal event sequences from electronic medical records: Initial successes and some challenges. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, 360–368.
[35] J. D. Patrick, D. H. M. Nguyen, Y. Wang, and M. Li. 2010. I2b2 challenges in clinical natural language processing 2010. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data.*
[36] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (Oct. 2011), 2825–2830.
[37] Ian N. Purves. 1998. PRODIGY: Implementing clinical guidance using computers. *British Journal of General Practice* 48, 434 (1998), 1552.
[38] Silvana Quaglini, Mario Stefanelli, Giordano Lanzola, Vincenzo Caporusso, and Silvia Panzarasa. 2001. Flexible guideline-based patient careflow systems. *Artificial Intelligence in Medicine* 22, 1 (2001), 65–80.
[39] J. Ross Quinlan. 1986. Induction of decision trees. *Machine Learning* 1, 1 (1986), 81–106.
[40] Bryan Rink, Sanda Harabagiu, and Kirk Roberts. 2011. Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association* 18, 5 (2011), 594–600.
[41] Kirk Roberts, Bryan Rink, and Sanda Harabagiu. 2010. Extraction of medical concepts, assertions, and relations from discharge summaries for the fourth i2b2/VA shared task. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data.*
[42] Natasha M. Rueth, Heather Y. Lin, Isabelle Bedrosian, Simona F. Shaitelman, Naoto T. Ueno, Yu Shen, and Gildy Babiera. 2014. Underuse of trimodality treatment affects survival for patients with inflammatory breast cancer: An analysis of treatment and survival trends from the National Cancer Database. *Journal of Clinical Oncology* 32, 19 (2014), 2018.
[43] Sunil Kumar Sahu, Ashish Anand, Krishnadev Oruganty, and Mahanandeeshwar Gattu. 2016. Relation extraction from clinical texts using domain invariant convolutional neural network. arXiv:1606.09370.
[44] Mihir Shekhar, Veera Ragahvendra Chikka, Lini Thomas, Sunil Mandhan, and Kamalakar Karlapalem. 2015. Identifying medical terms related to specific diseases. In *Proceedings of the ICDM Workshop on Biological Data Mining and Its Applications in Healthcare (BioDM'15).*
[45] William F. Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C. de Groen, Brad Erickson, et al. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics* 2 (2014), 143–154.
[46] Kjetil Sunde, Morten Pytte, Dag Jacobsen, Arild Mangschau, Lars Petter Jensen, Christian Smedsrud, Tomas Draegni, and Petter Andreas Steen. 2007. Implementation of a standardised treatment protocol for post resuscitation care after out-of-hospital cardiac arrest. *Resuscitation* 73, 1 (2007), 29–39.
[47] David R. Sutton and John Fox. 2003. The syntax and semantics of the PRO forma guideline modeling language. *Journal of the American Medical Informatics Association* 10, 5 (2003), 433–443.
[48] Kumutha Swampillai and Mark Stevenson. 2011. Extracting relations within and across sentences. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'11)*. 25–32.
[49] Tom Thaler, Philip Hake, Peter Fettke, and Peter Loos. 2014. Evaluating the evaluation of process matching techniques. In *Proceedings of Multikonferenz Wirtschaftsinformatik*. 1600–1612.
[50] Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association* 18, 5 (2011), 552–556.
[51] Karin M. Verspoor, Go Eun Heo, Keun Young Kang, and Min Song. 2016. Establishing a baseline for literature mining human genetic variants and their relationships to disease cohorts. *BMC Medical Informatics and Decision Making* 16, 1 (2016), 68.

[52] Zeev Waks, Esther Goldbraich, Ariel Farkash, Michele Torresani, Rossella Bertulli, Nicola Restifo, Paolo Locatelli, Paolo Casali, and Boaz Carmeli. 2013. Analyzing the "careGap": Assessing gaps in adherence to clinical guidelines in adult soft tissue sarcoma. *Students in Health Technology and Informatics* 186 (2013), 46–50.

[53] Fei Wang, Noah Lee, Jianying Hu, Jimeng Sun, and Shahram Ebadollahi. 2012. Towards heterogeneous temporal clinical event pattern discovery: A convolutional approach. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, 453–461.

[54] Fei Wang, Namyoon Lee, Jianying Hu, Jimeng Sun, Shahram Ebadollahi, and Andrew F. Laine. 2013. A framework for mining signatures from event sequences and its applications in healthcare data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 2 (2013), 272–285.

[55] WebMD. 2015. Health A-Z. Retrieved March 28, 2020 from http://www.webmd.com/a-to-z-guides/common-topics/default.htm.

[56] WebMD. 2010. Diagnosing Brain Cancer. Retrieved March 28, 2020 from https://www.webmd.com/cancer/brain-cancer/brain-cancer-diagnosis.

[57] Illhoi Yoo and Min Song. 2008. Biomedical ontologies and text mining for biomedicine and healthcare: A survey. *Journal of Computing Science and Engineering* 2, 2 (Jun 2008), 109–136.

Received December 2018; revised November 2019; accepted December 2019