

# Context-Sensitive Correlation of Implicitly Related Data: An Episode Creation Methodology

Roderick Y. Son, *Member, IEEE*, Ricky K. Taira, *Member, IEEE*, Hooshang Kangarloo, and Alfonso F. Cárdenas

**Abstract**—Episode creation is the task of classifying medical events and related clinical data to high-level concepts, such as diseases. Challenges in episode creation result in part because of data, in the patient record, only implicitly being associated with their respective episodes. Furthermore, traditional approaches have been limited to using feature-poor claims records to generate episodes. The accurate correlation of data to their episodes is valuable in health outcomes research to discern resource utilization with respect to medical conditions. This paper describes a combinatorial optimization approach for constructing episodes, which supports the incorporation of heterogeneous data types. Aspects of this approach include an episode model for characterizing the generation of data elements as a result of a process, a methodology for identifying the relationships between implicit processes and the data elements generated by the processes, a measure for evaluating candidate episode configurations, and an energy-minimization methodology for addressing episode creation. An implementation of this work, called Episode Creation Version 2 (EC2), has been applied on patient records with various episode types, which present with knee pain. EC2 demonstrated data element classification precision and recall scores of 78% and 82%, respectively. Significant improvements in precision and recall were observed over a traditional healthcare services approach.

**Index Terms**—Episode creation, probability model, simulated annealing.

## I. INTRODUCTION

IN THE medical arena, a patient's record is typically populated with a plethora of heterogeneous medical data elements including physician reports, imaging reports, laboratory results, pharmacy records, and claims records, all of which are generated as a result of one or more underlying medical problems, such as diseases or injuries. Unfortunately, relationships between the medical data and the corresponding medical problems are not always explicitly kept in the medical practice. The resolution of these relationships form data clusters called *episodes* [1]–[5]. Each episode, which represents an individual medical problem, contains all data that are generated during the process of care of that medical problem. Thus, episode creation is the task of automatically associating data elements contained within a patient's medical record into medical episodes.

Manuscript received August 9, 2004; revised January 8, 2005. Current version published September 4, 2008. This work was supported in part by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under Grant PO1-EB00216 and in part by the National Library of Medicine under Grant T15-LM07356.

R. Y. Son, R. K. Taira, and H. Kangarloo are with the Medical Imaging Informatics Group, University of California, Los Angeles, Los Angeles, CA 90024 USA (e-mail: rson@cs.ucla.edu; rtaira@mii.ucla.edu; hkangarloo@mii.ucla.edu).

A. F. Cárdenas is with the Computer Science Department, University of California, Los Angeles, Los Angeles, CA 90095 USA (e-mail: cardenas@cs.ucla.edu).

Digital Object Identifier 10.1109/TITB.2008.917901

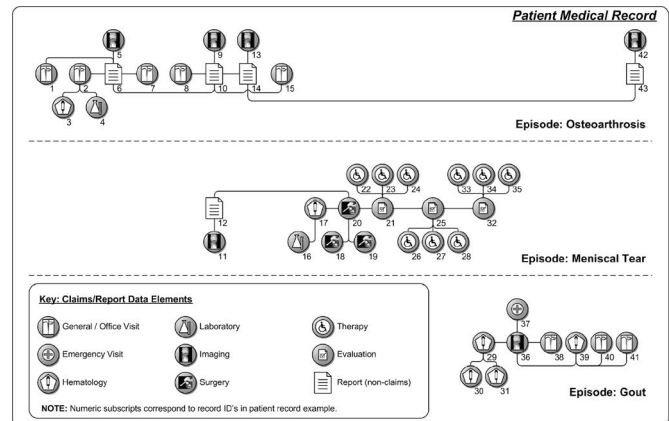


Fig. 1. Graphical representation of the patient record example. Nodes represent data elements in the record with time spanning from left to right. In the context of the cohesion metric, the edges represent a maximized pairwise correlation of data elements within episodes.

To illustrate the episode creation problem, consider the following clinical scenario of an example patient: a middle-aged female patient, who was recently diagnosed with and is currently being treated for osteoarthritis, enters the health care system with complaint of a knee pain due to trauma. A visit to the primary care physician and subsequent imaging study to diagnose the cause of the pain results in various medical documents being generated, including claims records and preliminary imaging reports, all of which are labeled with a diagnosis code corresponding to "knee pain." After reviewing the imaging study, a meniscal tear is identified, and the selected course of treatment is the surgical repair of the meniscus. Preoperative laboratory tests are taken, which generate additional medical data also labeled with a diagnosis code corresponding to "knee pain." Records of the knee arthroscopy and anesthesia are coded as "meniscal tear," and following postoperative therapies generate data coded as "knee pain." Fig. 1 illustrates a graphical representation of the data elements in this patient record. The nodes in the diagram correspond to data elements in the patient record.

Given this patient record, episode creation seeks to: 1) determine which episodes exist in the record and 2) associate each of the data elements within the record to its episode. However, episode creation is encumbered by several challenges, including the following: 1) the number of past episodes and types of episodes within a patient record are not explicitly known *a priori*; 2) the attributes of a medical data element may not definitively indicate the episode type with which it should be associated (e.g., a visit for joint pain can be observable in various

episodes, including a meniscal tear episode, an osteoarthritis episode, and a gout episode); and 3) multiple episodes can overlap in time (e.g., a meniscal tear episode occurring concurrently with an osteoarthritis episode).

One could suggest that the need for categorizing data based on medical episodes would dictate incorporating such information in the medical record data model and populating the field as each medical data element is made. However, because of the diagnostic nature of the medical practice, the identity of the episode that the data element is associated with is not always known [6]. To complicate matters, explicit links between medical data elements, such as a physician report to a laboratory result is typically not captured. For example, the fact that a physician made a request for an imaging study because of an earlier physical examination is not recorded. As a result, determination of the motivating medical problems that generated the observed medical findings in a patient's record and the correlation of the findings to their respective episode requires being able to infer relationships between medical data elements (i.e., implicit relationships).

Episode creation is of significant importance in the area of health outcomes research [5]. By accurately associating claims data to episodes, a higher fidelity picture can be garnered of how resources are utilized with respect to episodes. From this information, policies and protocols can be better formed and selected.

In this paper, we present a methodology to correlate medical data elements to their constituent episodes. This paper represents an expansion of work originally presented in [7]. Specifically, this paper presents a modified cost function, incorporating additional evidence to the original, and a more extensive evaluation in contrast to the prior work. The presented methodology utilizes a supervised learning approach to probabilistically characterize the processes that dictate the generation of data instances. From the probabilistic modeling, a combinatorial optimization approach is used to associate the optimal classification of data elements with their originating episodes.

## II. BACKGROUND

Approaches toward automatic episode creation have been explored using only medical claims records (i.e., billing statements) as the elements to construct the episodes [8]–[19]. Heterogeneous data types, such as physicians notes and imaging reports, which provide a more detailed understanding of the patient care, are not considered. Furthermore, these approaches utilize simplistic strategies to viewing the context of a patient record.

Wingert *et al.* provide a manual, iterative methodology for generating episodes of care from claims data [17]. Rules to discern episodes are manually generated and modified through iterative evaluation. However, manual methods for generating episodes are not scalable, given the difficulties of constructing a comprehensive rule set to fully characterize all the variances and subtleties characteristic of episodes. Rosen *et al.* proposed using clusters to group encounter types into gross categories, such as basic primary care, cancer care, and chronic care [14].

Hierarchy-based rules are used to guide the process of determining the cluster that a claims record should reside.

Lestina *et al.* provide a methodology for generating injury episodes from claims data by defining anatomical regions with respect to the diagnostic codes and defining “clear zones”—temporal thresholds based on statistical analysis—used for delineating distinct episodes [18]. Although this paper considers the probabilistic nature of time with respect to medical events, the episodes involved are related to injury, and the method used for isolating related events is overly simplistic for more complex episodes, such as episodes of disease. Their approach does not consider systemic diseases (e.g., diabetes), nor does it consider the interaction of different diseases impacting the same anatomical region.

Anderson *et al.* proposed, using the analysis of claims data based on diagnostic codes, a component of a medical billing record [19]. Diagnostic codes do not always signify a final diagnosis, but rather a hypothesis of the final diagnosis. Thus, each diagnostic code may not be conclusive of the true medical cause for a patient's treatment. Having nondefinitive diagnostic codes would result in potential misclassification because of rules insufficient to compensate for hypothetical diagnoses, or so complex, in an effort to compensate for all of the hypotheses, that they could be misconstrued.

Symmetry Health Data Systems has developed a new classification system to expedite episode creation [16]. Rather than using the standard diagnostic and procedure codes, a new set of codes were defined with episode tracking in mind. Because of the new classifications, episode creation appears to be more accurate. However, the system requires the adoption of a new standard for labeling medical activities. Like the previous approaches, this approach only utilizes claims data to construct episodes.

The various works in episode creation have been directed toward cost outcomes analysis. Thus, using only claims records, a patient's record is classified into medical problems of interest. Claims data are very useful for cost assessment; it is readily available because every encounter with the healthcare system is recorded for the purposes of payment. Moreover, the associated ICD9 diagnosis and CPT4 procedure codes are standardized to facilitate cost assessment analysis. Unfortunately, claims data alone are insufficient to address cost effectiveness analysis. Furthermore, literature has shown claims data, and in particular, the diagnosis codes contained in the records, to be insufficient for characterizing episodes [6]. Without utilizing other sources of data that are contained within a patient's record, the true motivation and cause of the medical intervention may be unclear. In particular, reports generated from imaging studies are one of the primary methods of objectively documenting the state of a medical condition. Thus, scaling a solution that can address heterogeneous data types is a necessary step toward evolving episode creation and obtaining a more accurate picture of a patient's healthcare.

Another shortcoming of current solutions is the dependence on rules. Rules tend to be elegant solutions for simple environments that have clear constraints. Unfortunately, medical episodes are not simple in nature, because looking at an

individual medical encounter without additional contextual information (i.e., the other elements surrounding a given encounter) may result in potential misclassification. An encounter that may indicate a diagnosis suggesting one possible episode may actually relate to a different episode as future medical procedures are administered. By considering the entire context, a more accurate classification is possible.

A third area of concern in current episode creation methodologies is the handling of time. Current solutions define an arbitrary window that is used to separate different episodes. However, such cutoff points are artificial. By incorporating a probabilistic component to time, such arbitrary solutions can be avoided, thus reducing false exclusion of encounters to an episode.

### III. COMBINATORIAL OPTIMIZATION STRATEGY

This paper hypothesizes that the classification of each data element in a patient record requires the consideration of the other data elements surrounding it in addition to itself, which motivates our global optimization strategy. Furthermore, we postulate that the ideal representation of a set of episodes for a patient record is the partitioning of the data into episodes such that the likelihood of the data elements being associated with the episodes is maximized. To this end, a simulated annealing strategy coupled with a cost function is presented.

#### A. Simulated Annealing

Given some patient record  $R$  (refer to Table I for notation details) containing a set of data elements  $\vec{d}_i$ , episode creation seeks to find a configuration  $C$ , a partitioning of the patient record into episodes  $\vec{e}_j$ . To address this task, the simulated annealing approach was selected as the global optimization method for episode creation because of its provably ensured convergence characteristics.

Simulated annealing is a knowledge-guided combinatorial technique, which seeks to find an optimal configuration by permuting through possible solutions using a cost function [20]. Per the simulated annealing framework, small perturbations to the configuration are iteratively proposed. Each configuration is evaluated by a cost function. Changes that improve the configuration are always accepted, and those that worsen the configuration are probabilistically accepted based on a Boltzmann distribution. A cooling schedule guides the iterative evaluation process. As the temperature is reduced, the probability of accepting a worse configuration is also decreased. The annealing approach allows us to find the global minima of our cost function without being stuck in a local minima. To map episode creation to the simulated annealing framework, the configuration and the cost function must be defined.

1) *Configuration*: In our episode creation framework, the simulated annealing *configuration* is a partitioning of the patient's set of data into labeled episode sets. The labels designate specific episode types (e.g., meniscal tear, osteoarthritis, and gout). In our currently presented framework, each data element is restricted to be a member of one and only one episode.

TABLE I  
SUMMARY OF TERMINOLOGY, NOTATION, AND TERMINOLOGY RELATIONSHIPS

Concept	Notation	Relationship	Comment
Data Element	$\vec{d}$	$\vec{d} \in \vec{e}, \vec{d} \in \vec{R}$	Structured data from sources such as Hospital Information System, Radiology Information System, and Picture Archiving and Communication System
Data Element Attribute	$d$	$d \in \vec{d}$	Attribute of a data element, such as an ICD-9 diagnosis code or CPT-4 procedure code from a claims record
Episode	$\vec{e}$	$\vec{e} \in C$	The underlying medical problem that dictates the generation of data elements, such as specific cases of meniscal tear, osteoarthritis, etc.
Episode Type	$\xi$	$\xi \in \Xi$	A classification for a set of episode instances that share a common underlying cause, such as "Meniscal Tear," "Osteoarthritis," and "Gout"
Episode Type Set	$\Xi$	—	The set of all episode types
Partitioned Patient Record (Configuration)	$C$	—	The environment within which episode instances exist; also implicitly associated with the patient record are all demographic information about a patient
Unpartitioned Patient Record	$R$	—	The unpartitioned patient record; the input of the episode creation system

The initial candidate configuration and succeeding configurations is probabilistically selected using the local model (discussed in Section IV). The local model functions as a context-free mechanism for associating data elements to episodes, without consideration of surrounding elements that coexist in the patient record.

2) *Configuration Evaluation*: A cost function  $K(C)$  is defined in this paper, which provides a relative measure of how well a candidate configuration represents a set of episodes. As the global minimum of the cost function is approached, the corresponding configuration should represent a more optimal solution. The cost function employed in this paper uses a multitier strategy for evaluating a configuration. The evaluation flow is shown in Fig. 2. The evaluation of cost begins by analyzing each candidate episode, a partition in the candidate configuration, using three metrics (local, cohesion, and evolution). These measures are combined to form a measure for each episode. The episode measures are then aggregated to generate a measure for the candidate configuration as a whole, the result of the function. In the following sections, the probability models used for the cost function, the training of the models, and how the models are incorporated into the cost function are described.

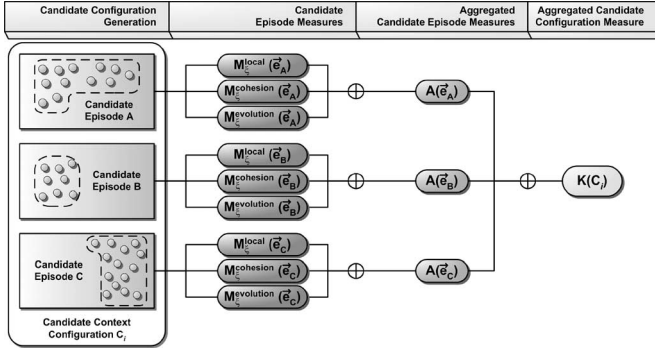


Fig. 2. Evaluation process flow for a candidate configuration.

#### IV. EPISODE PROBABILITY MODELS

The generation and evaluation of candidate configurations is driven by episode probability models. The models have been defined to capture characteristics pertaining to the observation of data as a result of an episode. The episode probability models support multiple types of patient-related data with various features.

Three probability models are employed to characterize episode types.

- 1) *Local model* characterizes the likelihood of an element being associated with a given episode type (used for generating candidate configurations and evaluating a candidate configuration).
- 2) *Cohesion model* characterizes the strength of affinity between pairs of elements within a given candidate episode (used in part to evaluate a candidate configuration).
- 3) *Evolutionary model* characterizes how far along the process of care the candidate episode instance has evolved (used in part to evaluate a candidate configuration).

Each probability model provides a different perspective for evaluating an episode. Combined, they are used in the energy-minimization approach to construct episodes.

A maximum entropy approach was selected to generate the aforementioned probability models. This classifier framework was selected because of its ability to handle conditionally dependent features (unlike with naive Bayes) and its flexibility in handling interdependencies (unlike decomposable models) [21]. To review, the maximum entropy classifier is based on the log-linear form. The classifier provides a framework that constrains the estimated distribution to exactly match the expected frequency of features within a training set (i.e., maximizes the uncertainty of a distribution) [20], [22]. Constraints are imposed on a maximum entropy model by defining feature functions  $\tilde{f}_i(\vec{x}, c)$ . The general equation of the maximum entropy classifier is

$$P(c | \vec{x}) = \frac{1}{Z(\vec{x})} \prod_i \alpha_i^{\tilde{f}_i(\vec{x}, c)}$$

$$Z(\vec{x}) = \sum_c \prod_j \alpha_j^{\tilde{f}_j(\vec{x}, c)} \quad (1)$$

where

$\vec{x}$  context; the observed set of data;

$c$  class;

$$\tilde{f}_i(\vec{x}, c) = \begin{cases} 1, & \text{if } (\vec{x}, c) \text{ satisfies a certain constraint} \\ 0, & \text{otherwise} \end{cases}$$

$\alpha_i$  weight corresponding to feature  $\tilde{f}_i$ ;

$$Z(\vec{x}) \text{ normalization factor to ensure } \sum_{\vec{x}} p(c | \vec{x}) = 1.$$

The maximum entropy classifier presumes that a training set of the form  $T = \{(\vec{x}_1, c_1) \cdots (\vec{x}_n, c_n)\}$  exists. To implement the probability model, a set of binary features  $\tilde{f}_i(\vec{x}, c)$  must also be defined for each episode model, with the aforementioned constraints. Through training, the weights  $\alpha_i$  are derived given the set of constraints. The modeling framework has the dual property of: 1) assuming nothing that is not observed in the training data and 2) providing the maximum-likelihood distribution given the training data and their features. Computationally, the classifier is calculated using the method of Lagrange undetermined multipliers with iterative scaling numerical methods [20]. The details of the three models, which follow the maximum entropy framework, are described next.

##### A. Local Model

The local model provides a simple, context-free classification of the elements within a patient record. It characterizes the likelihood of an element being related to any given episode type prior to the consideration of other elements in the same patient record. The local model is expressed as  $P_{\text{local}}(\xi | \vec{d})$  where  $\xi$  represents the episode type to which the element may be related and  $\vec{d}$  represents an element, such as a structured report instance, in the patient record.

Features used for the classifier are defined by a domain expert. Example of a feature used in the local model includes

$$\tilde{f}_i(\vec{d}, \xi) = \begin{cases} 1, & \text{if } d_{\text{anatomy}} = \text{leg} \\ & \text{and } \xi = \text{meniscal tear} \\ & \text{where } d_{\text{anatomy}} \in \vec{d} \\ 0, & \text{otherwise} \end{cases}$$

where  $d_{\text{anatomy}}$  represents an attribute of data element  $\vec{d}$ .

Given an element and its associated attribute values, such as an element from a transformed claims report with diagnostic and procedure codes, a likelihood of the report being associated with any episode type is provided. As mentioned before, the local model is also used in the initial configuration and succeeding augmented configurations to be evaluated are generated for the simulated annealing process.

##### B. Cohesion Model

The cohesion model measures the likelihood of a pair of data elements being associated with an episode type. The attributes of the element pair are incorporated as features in the maximum entropy classifier. As illustrated in Fig. 3, the model considers multiple granularity of features.

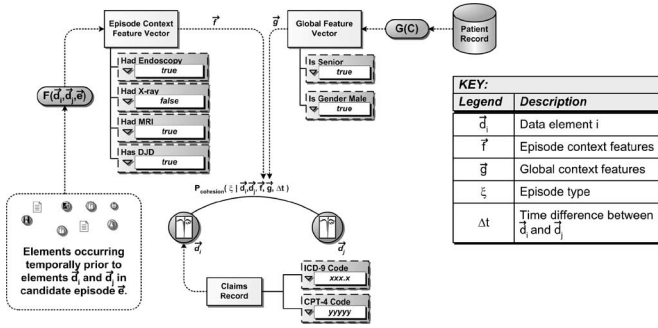


Fig. 3. Cohesion between element pairs are dependent on features including those of the elements, episode context, patient characteristics, as well as the time delta between elements.

- 1) *Element-specific features*: A localized examination of the data elements (e.g., procedure and diagnostic codes from a claims record).
- 2) *Episode context features*: A context for what has been observed in the episode relative to the elements (e.g., a positive meniscal tear finding; a specific surgical event for repairing the meniscus).
- 3) *Global context features*: Information that pertains to the episode but is not observed directly during the process of the episode (e.g., the age category and gender of the patient).

In addition to these feature values, the temporal distance between the two elements contributes to the measure of cohesiveness. Time differences are partitioned into multiple categories based on observed distributions from the training set and serve as additional features for the classifier (i.e., same day, within seven days, within half a month, ...).

The cohesion probability, as shown in (2), reflects the likelihood that the elements  $\vec{d}_i, \vec{d}_j \in \vec{e}$  are in an episode of type  $\xi$  where the function  $L_{\Xi}(\vec{e})$  specifies the episode type  $\xi$  that the episode instance  $\vec{e}$  represents,  $\vec{f}_{\text{cohesion}}$  is the episode's local context features in binary vector format,  $\vec{g}$  is the global context features also in a binary vector format, and  $\Delta t$  is the time difference between  $\vec{d}_i$  and  $\vec{d}_j$  measured in days. The set  $\Xi$  denotes the set of all possible episode types

$$P_{\text{cohesion}}(L_{\Xi}(\vec{e}) | (\vec{d}_i, \vec{d}_j, \vec{f}_{\text{cohesion}}, \vec{g}, \Delta t)). \quad (2)$$

The cohesion probability model follows the maximum entropy approach shown in (1).

Like the features in the local model, the features employed in the cohesion model are domain expert defined. An example of features for the cohesion model include

$$\tilde{f}_{i-1}(\vec{x}, \xi) = \begin{cases} 1, & \text{if } d_{\text{procedure}I} = \text{office visit} \\ & \text{and } d_{\text{procedure}J} = \text{imaging} \\ & \text{and } \Delta t = 0 \\ & \text{and } \xi = \text{osteoarthritis} \\ & \text{where } d_{\text{procedure}I} \in \vec{d}_i \\ & \text{and } d_{\text{procedure}J} \in \vec{d}_j \\ 0, & \text{otherwise} \end{cases}$$

$$\tilde{f}_i(\vec{x}, \xi) = \begin{cases} 1, & \text{if } f_{\text{hadEndoscopy}} = \text{true} \\ & \text{and } f_{\text{isAnatomyKnee}} = \text{true} \\ & \text{and } \xi = \text{meniscal tear} \\ & \text{where } f_{\text{hadEndoscopy}} \in \vec{f}_{\text{cohesion}} \\ & \text{and } f_{\text{isAnatomyKnee}} \in \vec{f}_{\text{cohesion}} \\ 0, & \text{otherwise} \end{cases}$$

$$\tilde{f}_{i+1}(\vec{x}, \xi) = \begin{cases} 1, & \text{if } g_{\text{isGenderMale}} = \text{true} \\ & \text{and } \xi = \text{osteoarthritis} \\ & \text{where } g_{\text{isGenderMale}} \in \vec{g} \\ 0, & \text{otherwise} \end{cases}$$

where  $\vec{x} = (\vec{d}_i, \vec{d}_j, \vec{f}_{\text{cohesion}}, \vec{g})$ .

A function  $F_{\text{cohesion}}(\vec{d}_i, \vec{d}_j, \vec{e})$  is the feature extraction function that populates the episode context features  $\vec{f}_{\text{cohesion}}$  for some candidate episode  $\vec{e}$  given the two elements  $\vec{d}_i$  and  $\vec{d}_j$  being evaluated. A simplified definition that may be used for  $F_{\text{cohesion}}$  is shown in (3), which aggregates all elements that are temporally during or prior to the element pair in consideration

$$\vec{f}_{\text{cohesion}} = F_{\text{cohesion}}(\vec{d}_i, \vec{d}_j, \vec{e}) = \bigvee_{\vec{d} \in \vec{e}} \vec{d} \quad (3)$$

where

$$\vec{e} = \{\vec{d} | \vec{d} \in \vec{e},$$

$$d_{\text{time stamp}} \leq \max(d_{\text{time stamp}I}, d_{\text{time stamp}J}),$$

$$d_{\text{time stamp}} \in \vec{d}, d_{\text{time stamp}I} \in \vec{d}_i, d_{\text{time stamp}J} \in \vec{d}_j\}$$

The operator symbol  $\bigvee$  represents an aggregation function, analogous to the bitwise OR function, that is used to merge a set of data into a single representation, a feature vector. For example, if the elements are represented as binary feature vectors, a bitwise-OR operation is done across the elements that have occurred prior or during the time of the element pair.

A function  $G(C)$  is defined that populates the global characteristics  $\vec{g}$  for some patient record  $C$ . Specifically, the function extracts pertinent demographics information such as the patient's age and gender. For notation simplicity,  $C$  is not included as an input to the probability model, although it is implied when incorporating  $\vec{g}$ .

### C. Evolutionary Model

The third model employed for evaluating a candidate configuration is the evolution model. This model characterizes the clinical progress of a given episode instance. Thus, the resulting measure of a candidate episode that only reflects the early stages of an episode type should be less than that of a sequence that represents an episode in a more advanced stage. In essence, the evolutionary measure emphasizes coverage of an episode for a sequence. The evolutionary probability is defined as

$$P_{\text{evolution}}(L_{\Xi}(\vec{e}) | (\vec{f}_{\text{evolution}}, \vec{g})). \quad (4)$$

This probability model, which follows the structure of (1), uses two types of features: 1) the episode context features  $\vec{f}_{\text{evolution}}$  and 2) the global context features  $\vec{g}$  used in the cohesion probability model. Unlike the cohesion model, where the episode

context features are populated dependent on the elements that occur during or prior to the pair of elements that are being evaluated, the evolutionary model utilizes all elements within the candidate episode to populate the feature vector.

An example of domain expert defined features for the evolutionary model include

$$\tilde{f}_i(\vec{x}, \xi) = \begin{cases} 1, & \text{if } f_{\text{hadMRI}} = \text{true} \\ & \text{and } \xi = \text{meniscal tear} \\ & \text{where } f_{\text{hadMRI}} \in \vec{f}_{\text{evolution}} \\ 0, & \text{otherwise} \end{cases}$$

$$\tilde{f}_{i+1}(\vec{x}, \xi) = \begin{cases} 1 & \text{if } f_{\text{hadEndoscopy}} = \text{true} \\ & \text{and } \xi = \text{meniscal tear} \\ & \text{where } f_{\text{hadEndoscopy}} \in \vec{f}_{\text{evolution}} \\ 0, & \text{otherwise} \end{cases}$$

where  $\vec{x} = (\vec{f}_{\text{evolution}}, \vec{g})$ .

Given the two example features, training of the probability model may discover that a meniscal tear episode is more likely observed having both an MRI examination and an endoscopy procedure than one with just an MRI procedure only.

The function  $F_{\text{evolution}}(\vec{e})$  is the feature extraction function that populates the episode features  $\vec{f}_{\text{evolution}}$  for some candidate episode  $\vec{e}$

$$\vec{f}_{\text{evolution}} = F_{\text{evolution}}(\vec{e}) = \bigvee_{\vec{d} \in \vec{e}} \vec{d}. \quad (5)$$

A possible definition for  $F_{\text{evolution}}$  is shown in (5), which aggregates all elements that occur during or prior to the element pair in consideration. Since the elements are represented as binary feature vectors, a bitwise-or operation is done across the elements that have occurred in the episode.

## V. TRAINING PROBABILITY MODELS

Now that the local, cohesion, and evolutionary probability models have been described, the training process is described to construct the models. All three models are based on a maximum entropy classifier [(1)]; thus, a labeled training set of the form  $(\vec{x}, c)$  must be constructed for each of the models. A training data  $T$  consists of a set of patient records that function as the context within which episodes are observed. The organization of the training data is presented as

$$T = \{C_1, C_2, \dots, C_m\} \quad (6)$$

$$C = \{\vec{e}_\xi \mid \vec{e}_\xi \neq \emptyset, \xi \in \Xi\} \quad (7)$$

$$\vec{e} = \{\vec{d}_1, \vec{d}_2, \dots, \vec{d}_n\} \quad (8)$$

where

$\vec{e}_\xi$  episode instance of type  $\xi$ ;

$m$  number of patient records (contexts) in training set;

$n$  Number of elements in episode example data set;

For the purposes of learning models for specific episode types, the episode specific training set  $T_\xi$  is defined. It contains the set of episode instances of type  $\xi$  across all contexts within the main

training set  $T$ . From the fundamental training data, training sets for the local, cohesion, and evolutionary models can be derived

$$T_\xi = \{x \mid \mathcal{L}_\Xi(\vec{e}) = \xi, \vec{e} \in C, C \in T\} \quad \forall \xi \in \Xi. \quad (9)$$

During the training process of the three models, the generalized iterative scaling (GIS) technique [23], a simple hill-climbing algorithm, is used to compute the feature weights  $\alpha_j$  for each of the models, which follow form of (1). The GIS algorithm requires a set of training data of the form  $(\vec{x}, c)$ , where  $\vec{x}$  represents the input data from which the features are extracted and  $c$  represents the classification for the given example. The specific training data used for each of the probability models will now be discussed in the following sections.

### A. Local Model

The training of the local model probability  $P_{\text{local}}(\xi \mid \vec{d})$  (see Section IV-A), the model that facilitates context-free analysis of an individual element, is a simple process based on a set of labeled training data. A local probability distribution is constructed by collecting all of the elements from the training set and their respective episode types  $\xi$ , as has been defined in  $T_\xi$ . To map the local model training collection to the form  $(\vec{x}, c)$  for the GIS technique, the following set is generated:

$$(\vec{x}, c) \in \{(\vec{d}, \xi) \mid \vec{d} \in T_\xi, \xi \in \Xi\}. \quad (10)$$

Given that the training set has been generated, the corresponding probability distribution for the local probability model is constructed through iterative scaling to provide a set of maximum entropy probabilities of the form  $P_{\text{local}}(\xi \mid \vec{d})$ .

### B. Cohesion Model

The training of the cohesion model  $P_{\text{cohesion}}(\xi \mid (\vec{d}_i, \vec{d}_j, \vec{f}_{\text{cohesion}}, \vec{g}, \Delta t))$  (see Section IV-B), the model used for evaluating the likelihood of a pair of elements coexisting within an episode, is also based on a labeled training set. In the case of the cohesion model, the training set is generated as

$$(\vec{x}, c) \in \{((\vec{d}_i, \vec{d}_j, F_{\text{cohesion}}(\vec{d}_i, \vec{d}_j), G(C), \Delta t), \mathcal{L}_\Xi(\vec{e})) \mid \vec{d}_i, \vec{d}_j \in \vec{e}, \vec{e} \in C, C \in T\}. \quad (11)$$

Given that the training set has been generated, the corresponding probability distribution for the cohesion probability model is constructed through iterative scaling to provide a set of maximum entropy probabilities of the form  $P_{\text{cohesion}}(\xi \mid (\vec{d}_i, \vec{d}_j, \vec{f}_{\text{cohesion}}, \vec{g}, \Delta t))$ .

### C. Evolutionary Model

The training of the evolutionary model  $P_{\text{evolution}}(\xi \mid (\vec{f}_{\text{evolution}}, \vec{g}))$  (see Section IV-C), the model used for evaluating the likelihood of a set of elements as a whole comprising an episode is also based on a labeled training set. In the case of the evolutionary model, the training set is generated as

$$(\vec{x}, c) \in \{(F_{\text{evolution}}(\vec{e}), G(C)), \mathcal{L}_\Xi(\vec{e}) \mid \vec{e} \in C, C \in T\}. \quad (12)$$

Given that the training set has been generated, the corresponding probability distribution for the evolutionary probability model is constructed through iterative scaling to provide a set of maximum entropy probabilities of the form  $P_{\text{evolution}}(\xi | (\vec{f}_{\text{evolution}}, \vec{g}))$ .

## VI. CONFIGURATION COST FUNCTION

### A. Episode Measure Components

Characterizing the “goodness” of a candidate episode is based on inspecting the candidate set from three perspectives, a local perspective that considers how well the elements individually correspond to an episode, a cohesion perspective that focuses on how strong the affinity is between elements within the episode, and an evolutionary perspective, which measures how far into the episode process the candidate episode is.

1) *Local Episode Measure*: The local episode measure utilizes the evidence of how data elements are classified in a given episode without consideration of other data elements in the patient record. The local measure is

$$M_{\xi}^{\text{local}}(\vec{e}) = 1 - \sum_{\vec{d} \in \vec{e}} \frac{P_{\text{local}}(\xi | \vec{d})}{|\vec{e}|}. \quad (13)$$

As the elements in a candidate episode better represent the episode type  $\xi$ , the value of  $M_{\xi}^{\text{local}}(\vec{e})$  decreases to 0. As the elements more poorly represent the episode type  $\xi$ , the value increases to 1.

2) *Cohesion Episode Measure*: The second measure is an inspection of the candidate using the cohesion model and focuses on the strongest way of bonding elements to each other within a given episode based on the episode’s type. The strength of the bonds between elements is modeled by the cohesion probability model, which characterizes the likelihood of two elements coexisting in an episode of some type.

Utilizing the minimum spanning tree (MST) algorithm [22], the maximal way of linking the elements together within a candidate episode is computed. For a graph  $G = (\tilde{V}, \tilde{E})$ , where  $\tilde{V}$  is a set of vertices,  $\tilde{E}$  is a set of edges between pairs of vertices, and weight function  $w(\tilde{u}, \tilde{v})$ , the task is to find a path in the graph that minimizes  $w(\tau) = \sum_{(\tilde{u}, \tilde{v}) \in \tau} w(\tilde{u}, \tilde{v})$  where  $\tau$  is a spanning tree. From the pairwise element probabilities defined in the cohesion model, a set of vertices and weighted edges are formed. The vertices are represented by the elements and the edge weights are based on the aforementioned probabilities.

Given a set of elements and a set of cohesion probabilities describing the transition from one element instance to another, the MST technique is used to maximize the set of transition edges, to span the elements. The transition edge weights, the weights applied to the MST are the difference of unity and the original edge probability, given as

$$w(\vec{d}_i, \vec{d}_j) = 1 - P_{\text{cohesion}}(\xi | (\vec{d}_i, \vec{d}_j, \vec{f}_{\text{cohesion}}, \vec{g}, \Delta t)). \quad (14)$$

Once the MST is computed, the episode cohesion measure  $\mu$  for the candidate episode  $\vec{e}$  of type  $\xi$  is computed according to

$$M_{\xi}^{\text{cohesion}}(\vec{e}) = \frac{\sum_{\varepsilon \in \text{MST}(\tilde{V}, \tilde{E}, w)} w(\varepsilon)}{|\tilde{V}| - 1} \quad (15)$$

where  $\text{MST}(\tilde{V}, \tilde{E}, w)$  represents the set of edges that form the MST for the graph consisting of the elements as vertices and the cohesion between elements as edges. Note that this measure results in a more cohesive candidate episode having a lower measure value. A visualization of a possible set of spanning trees is shown in Fig. 1 with the lines between the data elements representing the spanning tree edges.

3) *Evolutionary Episode Measure*: The evolutionary measure focuses on how far along the process of a given episode a candidate data set represents. Thus, the resulting measure of a sequence that only shows the early stages of an episode should be higher than that of a sequence that represents an episode in a more advanced stage. This measure is modeled with the evolutionary probability model

$$M_{\xi}^{\text{evolution}}(\vec{e}) = 1 - P_{\text{evolution}}(\xi | (\vec{f}_{\text{evolution}}, \vec{g})). \quad (16)$$

The resulting evolutionary measure is shown in (16).

### B. Aggregated Episode Measure

The aggregate episode measure incorporates the three dimensions of evaluating an episode, local, cohesion, and evolution, into a single measure. A weighted average is used

$$A(\vec{e}) = \gamma_{\text{local}} M_{\xi}^{\text{local}}(\vec{e}) + \gamma_{\text{cohesion}} M_{\xi}^{\text{cohesion}}(\vec{e}) + \gamma_{\text{evolution}} M_{\xi}^{\text{evolution}}(\vec{e}). \quad (17)$$

During the evaluation of this work,  $\gamma_{\text{local}} = \gamma_{\text{cohesion}} = \gamma_{\text{evolution}}$ .

### C. Configuration Measure

The energy cost function for a candidate configuration  $C$  of episodes aggregates the episode measures to give an overall measure of goodness

$$K(C) = \frac{1}{Z} \sum_{\vec{e} \in C} |\vec{e}| A(\vec{e}), \quad \text{where} \quad Z = \sum_{\vec{e} \in C} |\vec{e}| \quad (18)$$

Given a configuration of  $C$ , the final cost is a weighted average of the individual episode measures, as shown in (18). The weighting is guided by the number of elements associated with each episode. The normalization factor  $Z$  is the number of data elements in the patient record.

As the configuration  $C$  better represents an optimal partitioning of the data elements into episodes, the cost function  $K(C)$  decreases to 0. Otherwise,  $K(C)$  increases to 1.

## VII. PRELIMINARY IMPLEMENTATION AND EVALUATION

An implementation of the proposed episode creation methodology, called Episode Creation Version 2 (EC2), was developed in Java. Presented in this section includes the details of the labeled medical data set employed in the evaluation and the baseline approach used to compare with EC2. A ten fold cross-validation was conducted to compare precision and recall performance scores. A robust linear regression was also employed to better understand the performance behavior of EC2.

TABLE II  
EPISODE LABELS

Episode Code	Episode Name	Description
BU	bursitis	inflammation of a bursa
CACLMT	compound cruciate ligament and meniscal tear	tears of both the cruciate ligament(s) and meniscus
CLS	cruciate ligament sprain	sprain of the anterior, collateral or posterior cruciate ligament
CLT	cruciate ligament tear	sprain of the anterior, collateral or posterior cruciate ligament
CO	contusion	direct, blunt, compressive force to a muscle
CP	chondromalacia patellae	softening of the articular cartilage of the knee-cap
FX	fracture	broken bone
G	gout	peripheral arthritis resulting from the deposition of sodium urate crystals in one or more joints
KR	knee replacement	post-operative knee replacement encounters; knee replacement as a result of severe osteoarthritis
MT	meniscal tear	frayed, radial, circumferential, flap, parrot-beak or bucket-handle tear of the meniscus
N	“normal”	idiopathic pain; no identified cause of pain and of short episode duration
OA	osteoarthritis	noninflammatory degenerative joint disease characterized by degeneration of the articular cartilage, hypertrophy of bone at the margins and changes in the synovial membrane
OM	osteomyelitis	bacterial infection of bone and bone marrow in which the resulting inflammation can lead to a reduction of blood supply to the bone
OP	osteoporosis	abnormal loss of bony tissue resulting in fragile porous bones
PAT	patellar tendonitis	inflammation and degeneration of the tendon that connects the patella to the tibia
RA	rheumatoid arthritis	inflammatory disease that causes pain, swelling, stiffness, and loss of function in the joints

### A. Patient Data Set

The evaluation was performed on a set of 468 de-identified patient records from the Harris Family Medical Center in Melbourne, FL. Each patient included at least one data element that was labeled as a “knee pain” event (i.e., ICD-9-CM code of “719.46”). Each patient record consists of two types of data elements: claims records and structured imaging reports. Each data element was labeled by a domain expert with respect to the episode with which it is associated; these labels were utilized during training. The list of episode labels that were used in this evaluation are shown in Table II. The episode types were selected based on consultation with domain experts and inspection of the training set to determine what level of granularity (e.g., “joint pain” [gross granularity] versus “osteoarthritis” [fine granularity]) was possible to characterize episodes.

In total, 8145 data elements (7586 claims records and 559 structured imaging reports) from 468 patient records were labeled. The structured imaging reports were semi automatically generated from free-text reports with support from developed natural language processing tools [24], [25]. Free-text reports were structured using several natural language processing tools [24], [25], coupled with a manually generated keyword list, constructed by a domain expert, used to identify utterances of pertinent negative and pertinent positive findings, and conditions. A structured report would thus consist of a set of binary features, such as `subluxation.negative=true` or `meniscus.tear=false`, representing the existence or absence of a given keyword/phrase. From the structured data set, the episode creation probability models were learned.

### B. Results

An identical tenfold cross-validation was done on both the baseline strategy and the implementation of the proposed methodology called EC2. The data set was divided into ten folds with each fold consisting of 46 or 47 randomly selected patient records. During each iteration of the tenfold cross-validation, one fold was used as a test set and the remaining nine folds were used as the training set. Both approaches were tested using the same folds and compared to truth, the original labeling of the data elements.

The baseline approach was an implementation of the “clean” period strategy to episode creation. Data elements are delimited into episodes based on temporal gaps (i.e., a period of time in which no data elements are observed). Adjacent knee pain visits are considered associated with the same episode if the difference between the respective service dates are less than a specified length of time, called the clean period; a standard 90-day clean period was selected for this study [18], [26], [27]. Thus, the first data element of an episode is not adjacent to a preceding data element within the prior 90 days; the last data element of an episode is not adjacent to a successive data element within the following 90 days. The labeling of the generated episodes utilized the evolution model, described in Section III, for classifying the episodes.

The comparison of the two methodologies were based on precision and recall of the data elements for each episode type. For a given episode type, precision measures the proportion of the data elements identified as being associated with the episode type that are correct. Recall measures the proportion of the data elements associated with the episode type that were retrieved.

The results of precision and recall for baseline and EC2 are shown in Tables III and IV, respectively. The tables are sorted by whether the episode type consisted of information-rich structured reports and coded claims data, or claims data alone. The tables were secondarily sorted by the number of episode instance examples that existed in the labeled data set. The column *Diff* represents the difference between baseline and EC2 with a positive value showing improvement for EC2 and a negative value reflecting improvement toward baseline. The columns TP+FP (in precision) and TP+FN (in recall)



TABLE III  
PRECISION RESULTS FOR THE BASELINE AND PROPOSED EPISODE CREATION STRATEGIES

Episode	# Episode Instances	EC2		Baseline		Has Report	% Diff.	P-value
		Precision	TP+FP	Precision	TP+FP			
OA	166	67.24%	2027	65.73%	1608	TRUE	1.51%	0.3383
MT	107	77.42%	2662	74.28%	1932	TRUE	3.15%	0.0135
N	88	64.75%	244	12.89%	1606	TRUE	51.86%	<.0001
CLS	47	82.85%	519	38.43%	890	TRUE	44.42%	<.0001
CP	41	90.46%	304	40.92%	606	TRUE	49.54%	<.0001
BU	37	78.65%	267	36.93%	417	TRUE	41.72%	<.0001
CLT	23	71.99%	632	63.97%	272	TRUE	8.02%	0.0162
CO	21	60.00%	85	74.00%	100	TRUE	-14.00%	0.0426
OM	16	94.74%	57	NaN	0	TRUE	NA	NA
PAT	11	77.27%	44	91.30%	46	TRUE	-14.03%	0.0664
FX	8	92.31%	26	92.59%	27	TRUE	-0.28%	0.9687
KR	8	68.80%	266	NaN	0	TRUE	NA	NA
CACLMT	6	35.38%	732	100.00%	104	TRUE	-64.62%	<.0001
OP	96	97.55%	286	48.78%	572	FALSE	48.78%	<.0001
G	15	92.31%	39	100.00%	35	FALSE	-7.69%	0.0939
RA	12	93.85%	130	99.13%	115	FALSE	-5.28%	0.0282

TABLE IV  
RECALL RESULTS FOR THE BASELINE AND PROPOSED EPISODE CREATION STRATEGIES

Episode	# Episode Instances	EC2		Baseline		Has Report	% Diff.	P-value
		Recall	TP+FN	Recall	TP+FN			
OA	166	85.67%	1591	66.44%	1591	TRUE	19.23%	<.0001
MT	107	86.34%	2387	60.12%	2387	TRUE	26.23%	<.0001
N	88	59.62%	265	78.11%	265	TRUE	-18.49%	<.0001
CLS	47	62.14%	692	49.42%	692	TRUE	12.72%	<.0001
CP	41	77.03%	357	69.47%	357	TRUE	7.56%	0.0224
BU	37	47.30%	444	34.68%	444	TRUE	12.61%	0.0001
CLT	23	55.29%	823	21.14%	823	TRUE	34.14%	<.0001
CO	21	30.91%	165	44.85%	165	TRUE	-13.94%	0.0091
OM	16	66.67%	81	0.00%	81	TRUE	66.67%	<.0001
PAT	11	46.58%	73	57.53%	73	TRUE	-10.96%	0.1851
FX	8	57.14%	42	59.52%	42	TRUE	-2.38%	0.8248
KR	8	50.97%	359	0.00%	359	TRUE	50.97%	<.0001
CACLMT	6	66.07%	392	26.53%	392	TRUE	39.54%	<.0001
OP	96	98.24%	284	98.24%	284	FALSE	0.00%	1.0000
G	15	64.29%	56	62.50%	56	FALSE	1.79%	0.8445
RA	12	93.13%	131	87.02%	131	FALSE	6.11%	0.0983

represent the number of classifications observed for each approach (TP = “true positive”; FP = “false positive”; FN = “false negative”); the columns correspond to the denominators of the precision and recall proportions. The column *P-value* represents the likelihood that the difference between the baseline and EC2 results is statistically significant, with values tending toward 0.0000 representing very strong evidence of statistically significant differences and values tending toward 1.0000, which represents little to no statistical significance in differences [28].

The *P-value* for the precision and recall results was computed using the *z*-test. The computations are dependent on the proportions (i.e., precision or recall) and the number of observations (i.e., TP+FP or TP+FN) of the two methods being compared. The *z*-test is an asymptotic result and requires a sufficiently large number of observations to compare two proportions, which was satisfied in this analysis.

In the healthcare environment, *P-value* of less than 0.05 is considered statistically significant [28]. In this paper, a further refined interpretation is used. A *P-value* of less than 0.001 shows very strong evidence, less than 0.01 shows strong evidence, 0.05 shows moderate evidence, 0.10 shows weak evidence and

greater than 0.10 shows no evidence of statistical significance [29].

1) *Precision Results*: The precision results for the baseline and EC2 implementations are shown in Table III. Out of the 16 episode types, 8 show precision differences favoring EC2; six favor baseline. Two episode types did not have baseline precision data because none of the episode clusters generated by the baseline strategy was classified as KR or OM.

Very strong statistically significant differences (*P-value* of <0.0001) in precision are observed. Five episode types favor EC2; only one favored baseline. For episode types of larger episode instance example sizes (exceeding 20 episode examples), and incorporating both claims data and information-rich structure imaging report data, EC2 performed better in precision than baseline in all but one episode. For episode types of smaller episode instance example sizes (below 20 episode examples) or incorporating only claims data, the baseline strategy performed better than EC2 in all but one case. However, only one showed very strong indications of statistical significance.

2) *Recall Results*: The results of recall for the baseline and EC2 implementations are shown in Table IV. EC2 performs better than baseline in all but five cases, one of which demonstrated

TABLE V  
CONFIDENCE INTERVAL (C.I.) AND ESTIMATED AVERAGE OF DATA ELEMENT CLASSIFICATION PRECISION AND RECALL RESULTS FOR ALL EPISODES WHEN COMPARING EC2 WITH BASELINE

ALL EPISODES	EC2	Baseline	Difference
<b>Prec. - Avg.</b>	78.70%	68.71%	9.99%
<b>Prec. - C.I.</b>	(77.82%, 79.58%)	(67.95%, 69.47%)	<b>NO OVER-LAP</b>
<b>Recall - Avg.</b>	82.40%	67.06%	15.34%
<b>Recall - C.I.</b>	(81.57%, 83.23%)	(66.10%, 68.02%)	<b>NO OVER-LAP</b>

no difference. Seven episode types show very strong evidence of statistically significant improvement favoring EC2 with recall improvements ranging from 13% to 67%. Only one episode type showed a very strong statistical significance of improvement favoring baseline with a recall score of 19%.

For episode types of larger episode instance example sizes (exceeding 20 episode examples), and incorporating both claims data and information-rich structure imaging report data, EC2 performed better in recall than baseline in six of eight episodes. Two episode types favored baseline; both showed evidence of statistical significance. The remaining six episode types favoring EC2 showed recall improvements of around 8% and above, and all were shown to have strong evidence of statistical significance. For episode types of smaller episode instance example sizes (below 20 episode examples) or incorporating only claims data, the EC2 strategy also performed better than baseline in five of eight episode types. The three episode types not favoring EC2 were statistically insignificant.

3) *Estimated Averages and Confidence Intervals for Precision and Recall:* To compare the overall performance for the two approaches, the estimated variance-weighted average and the 95% confidence interval for the precision and recall results of the baseline and EC2 approach were computed. The variance-weighted average, which requires a sufficiently large number of observations (satisfied in this study), is a robust averaging scheme that is not affected by extreme values [28]. The confidence interval provides a measure of how precise the estimated average is. Overlapping confidence intervals reflect that the differences between two averages is not significant. Both the estimated average and confidence interval were computed using FastPro [28].

Three cases were evaluated that varied based on the labeled data set characteristics: all episode types, episode types that incorporated structured imaging reports, and episode types that incorporated reports as well as having an episode instance example size of 20 or more episode instances. The results for the 95% confidence interval and estimated averages for the three cases are shown in Tables V–VII. In all cases, EC2 performed better than baseline. Average precision improvements ranged from 9% to 31% favoring EC2 and average recall improvements ranged from 15% to 25%. The largest differences favoring EC2 are seen when looking at episode types that had episode instance examples exceeding 20 and incorporated report data.

4) *Correlation of Training Set Characteristics to F-Score:* To better understand the behavior of the episode creation methodology, an analysis was performed to determine what

TABLE VI  
CONFIDENCE INTERVAL (C.I.) AND ESTIMATED AVERAGE OF DATA ELEMENT CLASSIFICATION PRECISION AND RECALL RESULTS FOR EPISODE TYPES THAT INCORPORATED STRUCTURED IMAGING REPORTS WHEN COMPARING EC2 WITH BASELINE

REPORTS	EC2	Baseline	Difference
<b>Prec. - Avg.</b>	72.98%	63.87%	9.11%
<b>Prec. - C.I.</b>	(71.94%, 74.02%)	(63.01%, 64.73%)	<b>NO OVER-LAP</b>
<b>Recall - Avg.</b>	77.24%	52.17%	25.07%
<b>Recall - C.I.</b>	(76.25%, 78.23%)	(50.97%, 53.37%)	<b>NO OVER-LAP</b>

TABLE VII  
CONFIDENCE INTERVAL (C.I.) AND ESTIMATED AVERAGE OF DATA ELEMENT CLASSIFICATION PRECISION AND RECALL RESULTS FOR EPISODE TYPES THAT INCORPORATED REPORTS AS WELL AS HAVING A TRAINING SET SIZE OF 20 OR MORE EPISODE INSTANCES WHEN COMPARING EC2 WITH BASELINE

REPORTS & SIZE > 20	EC2	Baseline	Difference
<b>Prec. - Avg.</b>	76.20%	44.36%	31.84%
<b>Prec. - C.I.</b>	(75.00%, 77.40%)	(43.20%, 45.52%)	<b>NO OVER-LAP</b>
<b>Recall - Avg.</b>	78.58%	53.81%	24.77%
<b>Recall - C.I.</b>	(77.47%, 79.69%)	(52.45%, 55.17%)	<b>NO OVER-LAP</b>

factors caused EC2 to perform better on some episode types and less so on others. A robust linear regression analysis (i.e., linear regression with Huber–White standard error estimates [30]) was employed to identify potential correlations between the characteristics of the domain expert labeled data set for each episode type and the overall performance of EC2 on each episode type.

The precision and recall scores of each episode type was combined as an  $F$ -score, shown in (19), to represent the overall performance on a given episode type. The  $F$ -score is a single measure that combines precision and recall [31]

$$F\text{-score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (19)$$

Three possible factors for each episode type were considered that may influence the performance of the presented methodology: 1) the number of episode examples in the data set; 2) the average number of data elements within an episode instance, and 3) the proportion of data elements for a given episode type that were nondescript “knee pain” claims records. Nondescript “knee pain” data elements (i.e., claims records coded as ICD-9-CM = “719.46”) are considered in this regression because they can be associated with most episode types considered in this evaluation. The characteristics of the data set employed in the regression and the associated  $F$ -scores are shown in Table VIII.

In this robust linear regression, only episodes that included report data were considered. This ensured that the analysis was not biased by lack of structured report availability. To verify that the coefficients computed by the robust regression represented maximum-likelihood estimates, the Wilk-Shapiro normality test was performed to confirm that the residuals were independent and normally distributed. The resulting robust regression using Stata is shown in Table IX [30].

TABLE VIII  
COMPUTED F-SCORES BASED ON PRECISION/RECALL SCORES AND CHARACTERISTICS OF THE DOMAIN EXPERT LABELED DATA SET FOR EACH EPISODE TYPE (NUMBER OF EPISODE INSTANCES OF EACH EPISODE TYPE, EXISTENCE OF REPORTS, AVERAGE NUMBER OF DATA ELEMENTS PER EPISODE INSTANCE, AND PROPORTION OF “KNEE PAIN” DATA ELEMENTS FOR A GIVEN EPISODE TYPE)

Episode	F-score	# Episode Instances	Has Report	# D.E. per Episode	Prop. of Knee Pain D.E.
OA	75.35%	166	TRUE	9.5843	0.3394
MT	81.64%	107	TRUE	22.3084	0.3071
N	62.08%	88	TRUE	3.0114	0.9774
CLS	71.02%	47	TRUE	14.7234	0.1705
CP	83.21%	41	TRUE	8.7073	0.1513
BU	59.07%	37	TRUE	12.0000	0.4167
CLT	62.54%	23	TRUE	35.7826	0.1580
CO	40.80%	21	TRUE	7.8571	0.4000
OM	78.26%	16	TRUE	5.0625	0.0123
PAT	58.12%	11	TRUE	6.9091	0.1316
FX	70.59%	8	TRUE	5.2500	0.0238
KR	58.56%	8	TRUE	44.8750	0.1448
CACLMT	46.09%	6	TRUE	65.3333	0.2908
OP	97.90%	96	FALSE	2.9583	0.0000
G	75.79%	15	FALSE	3.7333	0.0179
RA	93.49%	12	FALSE	10.9167	0.0000

TABLE IX  
ROBUST LINEAR REGRESSION TO IDENTIFY POTENTIAL CORRELATION BETWEEN THE F-SCORE AND CHARACTERISTICS OF THE DOMAIN EXPERT LABELED DATA SET

F-SCORE	Coef.	Std. Err.	t	P-value
# Episode Inst.	0.003786	0.0003490	10.85	<.0001
# D.E. per Episode	0.000067	0.0004948	0.13	0.8970
Prop. of Knee Pain D.E.	-0.927048	0.0755940	-12.26	<.0001
Intercept	0.693051	0.0191313	36.23	<.0001

As demonstrated by the robust linear regression, the number of episode instances is positively significant to the *F*-score (i.e., the larger the number of episode examples, the better the EC2 performs). Also, the proportion of “knee pain” data elements for a given episode type is negatively significant to the *F*-score (i.e., the larger the proportion of nondescript “knee pain” data elements, the worse the EC2 performs). The average size of the episode instances did not demonstrate any correlation to *F*-score.

## VIII. CONCLUSION

The test results of the EC2 strategy indicate significant improvements in both precision and recall relative to the baseline. As seen in the earlier results, for episode types with a large training set size and that also incorporated reports, EC2 performed significantly better than baseline in both precision and recall. As the training set size and the presence of documents decreased, EC2 performed less well.

These improvements are significant to health outcomes research, one of the primary applications of episode creation. By more accurately classifying claims data to their originating episodes, EC2 can generate a higher fidelity picture of how resources are utilized with respect to an episode. In turn, the future development of strategies for optimal utilization of healthcare services can be facilitated.

The improvements favoring EC2 in overall precision and recall can be attributed in part to the incorporation of both the cohesion and evolutionary probability models in characterizing an episode. The cohesion model assists in the linking of data

elements that are difficult to classify with data elements that are more definitive in their episode origin. Likewise, the evolutionary model aids in identifying groups of data elements that collectively represent an episode. Furthermore, EC2 is more adept at handling scenarios where episodes may be temporally overlapping, unlike the baseline approach.

In particular, EC2 performed well with episode types such as osteoarthritis, meniscal tear, cruciate ligament sprain, chondromalacia patella, bursitis, and cruciate ligament tear; these episode types favored EC2 in both precision and recall. The “normal” episode type showed a higher precision but lower recall when comparing EC2 to baseline. Only in the case of contusion did baseline perform better in both precision and recall. The remaining episode types, all of which either did not incorporate reports or were of small training set size, showed improvements in precision at the cost of recall or vice versa.

When the training size is smaller, baseline performs better, particularly in recall. This observation may be an indication in those cases of the insufficient training examples to learn the evolution probability model. Unlike the local or cohesion models, whose number of training examples are a function of the number of data elements in each episode instance in the training set, the evolution model is solely dependent on the number of labeled episode instances. Thus, the learned model from a small number of episode instance are likely to be poor. EC2 also suffers when only claims data is used in the training set to represent an episode type. This observation is representative of the lack of sufficient characterizing features associated with claims data to sufficiently learn cohesion and evolution models.

As observed in the robust linear regression, EC2 performs better with larger episode instance training examples. Also, EC2’s performance is degraded by the increasing proportion of vague “knee pain” data elements, which can be associated with a myriad of conditions. The reduction of the uncertainty associated with these data elements can be better addressed by incorporating additional sources of data, such as physician’s notes, which can then better disambiguate the origin of these elements.

Future work includes additional evaluation of the contributions of the cohesion and evolutionary probability models. Also,

the refining of the cost function by incorporating hierarchical level maximum entropy classifier to aggregate the various measures used in this paper is being explored. The evaluation only utilized a simple average of the measures to aggregate the measures together for the final cost. In addition, the EC2 system is being tested on a second training set that pertains to a different class of episodes (e.g., disease entities that present with sciatica or lumbago).

## REFERENCES

- [1] J. A. Solon, J. J. Feeney, S. H. Jones, R. D. Rigg, and C. G. Sheps, "Delineating episodes of medical care," *Amer. J. Public Health*, vol. 57, pp. 401–408, Mar. 1967.
- [2] S. J. Kilpatrick, "The distribution of episodes of illness—A research tool in general practice," *J. R. Coll. Gen. Pract.*, vol. 25, no. 158, pp. 686–690, Sep. 1975.
- [3] G. L. Stoddart and M. L. Barer, "Analyses of demand and utilization through episodes of medical service," in *Health Economics, and Health Economics*, Amsterdam, The Netherlands, North Holland, 1981, pp. 149–170.
- [4] L. G. Kessler, D. M. Steinwachs, and J. R. Hankin, "Episodes of psychiatric care and medical utilization," *Med. Care*, vol. 20, no. 12, pp. 1209–1221, Dec. 1982.
- [5] M. C. Hornbrook, A. V. Hurtado, and R. E. Johnson, "Health care episodes: Definition, measurement and use," *Med. Care Rev.*, vol. 42, no. 2, pp. 163–218, 1985.
- [6] J. Strausberg, H. Lang, U. Obertacke, and F. Rauhut, "Classifications in routine use: Lessons from icd-9 and icpm in surgical practice," *J. Amer. Med. Inf. Assoc.*, vol. 8, no. 1, pp. 92–100, 2001.
- [7] R. Y. Son, R. K. Taira, A. A. T. Bui, A. F. Cardenas, and H. Kangarloo, "A context-sensitive methodology for automatic episode creation," in *Proc. AMIA Annu. Fall Symp.*, 2002, pp. 707–711.
- [8] R. Schneeweiss, R. A. Rosenblatt, D. C. Cherkin, C. R. Kirkwood, and G. Hart, "Diagnostic clusters: A new tool for analyzing the content of ambulatory medical care," *Med. Care*, vol. 21, no. 1, pp. 105–122, 1983.
- [9] R. A. Rosenblatt, R. Schneeweiss, D. C. Cherkin, and G. Hart, "Inpatient diagnostic clusters: Analyzing hospital care in family practice," *J. Family Pract.*, vol. 18, no. 1, pp. 93–101, 1984.
- [10] R. Schneeweiss, D. C. Cherkin, G. Hart, D. A. Revicki, E. V. Dunn, H. L. Tindall, and R. A. Rosenblatt, "Diagnosis clusters adapted for icd-9-cm and icd-9-cm," *J. Family Pract.*, vol. 22, no. 1, pp. 69–72, 1986.
- [11] B. Starfield, J. Weiner, L. Mumford, and D. Steinwachs, "Ambulatory care groups: A categorization of diagnoses for research and management," *Health Serv. Res.*, vol. 26, no. 1, pp. 53–74, Apr. 1991.
- [12] J. P. Weiner, B. H. Starfield, D. M. Steinwachs, and L. M. Mumford, "Development and application of a population-oriented measure of ambulatory care case-mix," *Med. Care*, vol. 29, no. 5, pp. 452–472, May 1991.
- [13] D. G. Cave, "Analyzing the content of physicians' medical practices," *J. Ambulatory Care Manage.*, vol. 17, no. 3, pp. 15–36, 1994.
- [14] A. K. Rosen and A. Mayer-Oakes, "Episodes of care: Theoretical frameworks versus current operational realities," *J. Qual. Improvement*, vol. 25, no. 3, pp. 111–128, Mar. 1999.
- [15] D. J. Brailer and E. A. Kroch, "Member risk adjustment for ambulatory episodes of care," *Health Care Manage.*, vol. 2, pp. 125–136, 1999.
- [16] D. K. Dang, "Computer-implemented method for profiling medical claims," U.S. Patent 5,835,897 Nov. 1998.
- [17] T. D. Wingert, J. E. Kralowski, T. J. Lindquist, and D. J. Knutson, "Constructing episodes of care from encounter and claims data: Some methodological issues," *Inquiry*, vol. 32, no. 4, pp. 430–443, 1995.
- [18] D. C. Lestina, T. R. Miller, and G. S. Smith, "Creating injury episodes using medical claims data," *J. Trauma: Injury, Infection Crit. Care*, vol. 45, no. 3, pp. 565–569, 1998.
- [19] G. Anderson, E. P. Steinberg, J. Whittle, N. R. Powe, S. Antebi, and R. Herbert, "Development of clinical and economic prognoses from medicare claims data," *J. Amer. Med. Assoc.*, vol. 263, no. 7, pp. 967–972, 1990.
- [20] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. 2nd ed. New York: Wiley, 2001.
- [21] A. Ratnaparkhi, "Maximum entropy models for natural language ambiguity resolution" Ph.D. dissertation, Univ. of Pennsylvania, Pennsylvania, PA, 1998.
- [22] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley Interscience, 1991.
- [23] J. N. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *Ann. Math. Stat.*, vol. 43, pp. 1470–1480, 1972.
- [24] R. K. Taira and S. G. Soderland, "A statistical natural language processor for medical reports," in *Proc. AMIA Annu. Fall Symp.*, 1999, pp. 970–974.
- [25] R. K. Taira, S. G. Soderland, and R. M. Jakobovits, "A statistical natural language processor for medical reports," in *Proc. Int. Conf. Math. Eng. Technol. Med. Biol. Sci.*, 2000, pp. 497–503.
- [26] E. Keeler, W. Manning, and K. Wells, "The demand for episodes of mental health services," *J. Health Econ.*, vol. 7, pp. 369–392, 1988.
- [27] E. Keeler and J. Rolph, "The demand for episodes of treatment in the health insurance experiment," *J. Health Econ.*, vol. 7, pp. 337–367, 1988.
- [28] B. Dawson and R. G. Trapp, *Basic and Clinical Biostatistics*. New York: McGraw-Hill, 2001.
- [29] A. H. Kuiper, "P-value as a measuring tool and decision procedure for the goodness-of-fit test," *J. Appl. Stat.*, vol. 15, pp. 131–135, 1988.
- [30] Stata, Stata Corp., College Station, TX, 2001.
- [31] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.



**Roderick Y. Son** (M'99) received two B.Sc. degrees in computer science and applied mathematics, the M.A. degree in applied mathematics, the M.S. and Ph.D. degrees in computer science from the University of California, Los Angeles (UCLA), Los Angeles, in 1995, 1997, 1999, and 2005, respectively. He is currently with the UCLA Medical Imaging Informatics Group. His current research interests include classification of medical documents and natural language processing, specifically in the area of coreference resolution.



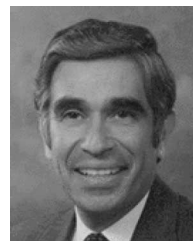
**Ricky K. Taira** (M'97) received the B.Sc. degree in electrical engineering and the Ph.D. degree in biomedical physics from the University of California, Los Angeles (UCLA), in 1982 and 1988, respectively.

He is currently a Professor in the Department of Radiological Sciences, David Geffen School of Medicine, UCLA. He was engaged in research on development of picture archive and communication systems (PACS), and medical knowledge bases (the KMeD project). His current research interests include natural language processing of medical corpora and formal ontological representations of disease entities.



**Hooshang Kangarloo** received the M.D. degree from Tehran University, Tehran, Iran, in 1970.

He is currently a Professor of pediatrics and radiological sciences in the David Geffen School of Medicine, University of California, Los Angeles (UCLA), where he is a Professor of bioengineering in the Henry Samueli School of Engineering and Applied Sciences. He is also the Director of the UCLA Medical Imaging Informatics Group, and the Co-Director of the UCLA Biomedical Informatics Center. His current research interests include telemedicine, healthcare process modeling and evaluation, and imaging informatics.



**Alfonso F. Cardenas** received the B.Sc. degree from San Diego State University, San Diego, CA, and the M.S. and Ph.D. degrees in computer science from the University of California, Los Angeles (UCLA), in 1969.

In 1972–1975, he was a Visiting Scientist/Consultant to IBM Corporation. He is currently a Professor in the Computer Science Department, UCLA. His current research interests include database management, distributed heterogeneous and multimedia (text, image/picture, voice) systems, information systems planning and development methodologies, software engineering, and legal and intellectual property issues.