Data Reconstruction Based on Temporal Expressions in Clinical Notes

Zhikun Zhang
Shanghai Key Laboratory of Data Science, School of Computer Science
Fudan University
Shanghai, CHN
zhangzk17@fudan.edu.cn

Chunlei Tang
Brigham and Women's Hospital
Harvard Medical School
Boston, MA, USA
ctang5@partners.org

Joseph M. Plasek
Brigham and Women's Hospital
Harvard Medical School
Boston, MA, USA
jplasek@partners.org

Yun Xiong
Shanghai Key Laboratory of Data Science, School of Computer Science
Fudan University
Shanghai, CHN
yunx@fudan.edu.cn

Min-Jeoung Kang
Brigham and Women's Hospital
Harvard Medical School
Boston, MA, USA
mkang6@bwh.harvard.edu

Patricia C. Dykes
Brigham and Women's Hospital
Harvard Medical School
Boston, MA, USA
pdykes@bwh.harvard.edu

David W. Bates
Brigham and Women's Hospital
Harvard Medical School
Boston, MA, USA
dbates@bwh.harvard.edu

Li Zhou
Brigham and Women's Hospital
Harvard Medical School
Boston, MA, USA
lzhou@bwh.harvard.edu

***Abstract*—Learning representations of clinical notes poses challenges in handling complex content that necessitates preprocessing steps to make the data more suitable for data mining. An important issue, addressed here, is that of temporal expressions, where cues indicate the time when clinical events occur. We present a three-step data reconstruction algorithm for transforming similar clinical entities (e.g., symptoms, complications) into sequential data through unsupervised annotation of temporal expressions. First, the data reconstruction algorithm detects if an expression has temporal intent. Second, it decomposes and rewrites the expression into non-temporal sub-expression and temporal constraints. Finally, it clusters similar non-temporal sub-expressions by using unsupervised sentence embedding under the modified *K*-medoids paradigm. We experimented with our proposed algorithm on clinical notes associated with chronic obstructive pulmonary disease (COPD). Visualizing reconstruction results of cardiology reports for a longitudinal cohort of patients with COPD demonstrated that this algorithm is feasible.**

Keywords—Data reconstruction, cluster analysis, 'pulmonary disease, chronic obstructive', unsupervised learning

I. INTRODUCTION

In the medical field, a large amount of information (e.g., clinical findings and test result interpretations) is recorded in clinical notes that circulate between patients, physicians, and nursing professionals. Temporal expressions within these notes provide cues about relationships between clinical events for subsequent analysis tasks. However, learning temporal expressions and relations is challenging. They are often represented in various ways in clinical narratives concerning clinical entities (e.g., symptoms, diagnoses): based on a start time (e.g., a medication administration), qualitative constraint representation (e.g., days prior to death), duration-based representation (e.g., chronic obstructive pulmonary disease (COPD) stages, a hospital stay), and so forth [1]. The following are common practices in obtaining temporal expressions: one approach is to retrieve the temporal dimension of existing objects (e.g., the creation time for a specific clinical entity) and utilize this as a temporal component; another approach is to annotate all time-oriented information of task-specific entities via TimeML (http://www.timeml.org) [2] or similar markup languages to meet the requirements of temporal reasoning tasks. However, the former is too naive to effectively process detailed information on clinical entities; and the later depends on expensive and manual labels for diverse entities [3].

We envision abstracting a temporal task from a clinical expression is possible to the solution. The temporal abstraction task is defined as a set of time-stamped clinical entities. For example, viewing a temporal expression as two lower-level objects: point events (e.g., atrial fibrillation presents) and interval events (e.g., atrial fibrillation for a few seconds), and higher-level abstract objects called phases (e.g., atrial fibrillation has replaced sinus rhythm) based on these two lower-level objects. Processing should start with recognizing and annotating time and medical situations mentioned in the original text along with the identification and classification of these medical concepts. Our preliminary goal is to generate a set of time-stamped (or time-anchored) clinical entities, which happen to be represented in a sequential data format. In this task, we aim to reduce the need for labeling by automatically learning how to summarize a sequence of clinical events over time in a data-driven, unsupervised manner.

Data reconstruction is, most commonly, used to process real time series (e.g., from wireless sensor networks [4-5]), in order

This work was partially funded by the CRICO/Risk Management Foundation of the Harvard Medical Institutes Incorporated, Partners Innovation Fund, the National Natural Science Foundation of China Projects No. U1636207, and the Shanghai Science and Technology Development Fund No.19511121204, 16JC1400801, and Suzhou Science and Technology Bureau Technology Demonstration Project (SS201712, SS201812).

978-1-7281-1867-3/19/$31.00 ©2019 IEEE

to solve the information loss issue. There is no highly cited literature related to data reconstruction on document summarization [6-7]. We propose a novel data reconstruction algorithm for creating effective representations of clinical notes. As far as we know, our approach is among the first to transform free-text data into sequential data by considering clinical entities using machine learning. Sequential data is easy to be used in downstream analytical activities, including time distribution visualization, sequence mining, pattern discovery, classification, and prediction [8-9]. Examples of sequential data include: DNA, protein, customer purchase history, web surfing history, etc. Sequential data can be thought of as "an extension of record data, where each record has as a time associated with it [9]," of which one or more attributes of sequential data have relationships that involve a discrete-time order. Making predictions with sequential data occurs in a variety of ways, such as predicting the next value for a given input sequence. We argue that annotating and processing valid temporal expressions in clinical notes is critical for achieving an appropriate reconstruction. The aim of this paper is to establish the feasibility of our data reconstruction algorithm for use in the clinical domain.

II. The Proposed Three-Step Data Reconstruction Algorithm

The architecture of the proposed data reconstruction algorithm has three steps: first, detect if an expression has temporal intent when scanning an entire sentence; second, decompose and rewrite the expression into non-temporal sub-expressions and temporal constraints; and third, cluster similar non-temporal sub-expressions into clinical entities using sentence embedding clustering.

A. Temporal Expression Annotation

An annotation in a clinical document is identified as a temporal expression if it contains any of the following [10]: (a) explicit time expressions (e.g., dates, times); (b) implicit temporal signals (i.e., cue words for temporal relations); (c) ordinal words (e.g., 'first').

Explicit temporal expressions occur in entirely-specified or under-specified forms of date time (e.g., 'August 1, 2019, 9:00 am', '2019/08/01', '9:00 am', 'noon'), or durations (e.g., 'after two years', 'every Tuesday'). Common approaches to identify temporal expressions via natural language processing (NLP) follow a rules-based approach (e.g., regular expression matching) that utilize a markup language like TimeML [2]. In this study, our annotations follow the TimeML TIMEX3 standard [2]. TimeML has five grammatical categories [2]: (1) nouns like 'today', 'Thursday', (2) noun phrases like 'the morning', (3) adjectives like 'current', (4) adverbs like 'recently', and (5) adjective or adverb phrases like 'two weeks ago'. Currently, TimeML cannot be used to treat a Prepositional Phrase or a clause of any type.

Consider the following cardiology examples:

- **Example 1:** *When compared with ECG of 18-JUL-YYYY 10:41, (unconfirmed) no significant change was found. Confirmed by X MD on 7/22/YYYY 17:18.*
- **Example 2:** *Subsequently, a pharmacological stress test was performed with IV adenosine infusion after which sestamibi was injected IV at peak drug effect."*

TimeML's recognition capacity on Example 1 is adequate. In Example 2, no explicit date is mentioned. The phrase "after which" refers to an event (*after IV adenosine infusion*). TimeML could detect this phrase, but does not properly disambiguate it to a normalized date. The temporal preposition "with" is a cue as well, and words like "subsequently" are also used in temporal contexts. Thus, TimeML and similar markup languages still struggle with representing the implicit temporal conditions exemplified in Example 2.

Temporal signals mark textual elements that denote implicit temporal relations among temporal expressions, such as 'after' or 'during'. We extend the TimeML scheme to include cues when a clinical entity is mentioned implicitly, such as 'after which'. We referenced Allen's 13 temporal relations [11] between time intervals for temporal reasoning which include 'equal', 'start', 'finish', 'meet', 'compared to', and their inverses. We also consider ordinals (e.g., 'last', 'previous'). These are frequent in expressions when entities can be chronologically ordered, such as "the previous tracing heart rate is faster."

We define a temporal expression as follows:

- **Definition 1** (*Temporal Expression*) [10, 12]. A temporal expression is a three-tuple $t_{e_i} = \langle e_i, t_i, n_i \rangle$, where e_i is the expression itself as it occurs in the textual clincial notes, t_i represents the type of temporal expression, and n_i is the normalized value.

A total of five possible types are used in t_i, including: date, time, duration, signal, and ordinal. The normalized value represents the semantics of a temporal expression, which is specified in the markup language of TimeML.

In summary, the goal of this stage is to extract carefully every temporal expression e_i and to correctly assign the type and value attributes t_i and n_i, respectively.

B. Contextual Expression Decomposition

In this stage, our goal is to decompose and rewrite a composite temporal expression into one or more non-temporal sub-expressions (returning the clinical entities), and one or more temporal sub-expressions (returning the temporal constraints). The constraints may be applied to time scopes associated with results of any two non-temporal sub-expressions.

Example 1 can be intuitively decomposed into three sub-expressions: 1.1 "*When compared with ECG of 18-JUL-YYYY 10:41*," 1.2 "*(unconfirmed) no significant change was found*," and 1.3 "*Confirmed by X MD on 7/22/YYYY 17:18*." Two basic rules this process applies include:

- **Rule 1:** The signal word separates the non-temporal and temporal sub-expressions, acting as a pivot for decomposition.
- **Rule 2:** Each sub-expression needs to have an entity and a relation (generally represented using verbs).

Example 2 should be decomposed into: 2.1 "*Subsequently*," 2.2 "*a pharmacological stress test was performed with IV adenosine infusion*," and 2.3 "*after which sestamibi was injected IV at peak drug effect*." This decomposition applies an additional rule:

- **Rule 3:** If a subsequent sub-question lacks an entity or its relation, then borrow the missing component from a prior sub-expression.

C. Sentence Embedding Clustering

This study aims to reconstruct clinical notes in an unsupervised way so as to not require any labeled training data. To achieve this, we chose the *K*-medoids clustering method to obtain similar non-temporal sub-expressions. *K*-medoids is a variant of *K*-means and aims to improve accuracy by reducing sensitivity to the outliers. Under the *K*-medoids paradigm, we adapted Arora et al.'s [11] unsupervised sentence embedding as a similarity measure. We modified the traditional *K*-medoids algorithm to merge any two clusters with center points that are very close. The final number of clusters thus depends upon calculated results which may result in fewer clusters than the specified *K*. Take Example 1 as an example, the duration is the time interval between 1.1 and 1.3. After having calculated the similarity of non-temporal sub-expression as 1.2, the number of clusters of similar ECG diagnosis was easy to obtain. Thus, a reconstruction result of "no significant change was found: 4 days." Pseudocode statements 1 to 3 are sentence embedding steps, and statements 4 to 5 are clustering steps.

Input: $S = \{S_1, S_2, \cdots, S_n\}$ $i\epsilon\ [1, \cdots, n]$ //set of sentences (each S_i consists of multiple Words W_{ij} $j\epsilon\ [1, \cdots, m]$), parameter a, K // number of desired clusters

Output: $C = \{C_1, C_2, \cdots, C_K\}$ // set of clusters, in which each data point is a vector v_{S_i} embedded by a sentence S_i

1. **for** all sentence S_i **do**
   for all words W_{ij} in S_i **do** calculate the frequency of W_{ij} and the length of sentence S_i, then estimate the proportion of a word W_{ij} emitted in the sentence S_i as $p(W_{ij})$ **done**;
   $v_{S_i} \leftarrow \frac{1}{|S_i|}\sum_{W_{ij}\epsilon S_i}\frac{a}{a + p(W_{ij})}$
   done
2. form a matrix X whose columns are v_{S_i}, and let u be its first singular vector
3. **for** all sentence S_i **do** $v_{S_i} \leftarrow v_{S_i} - uu^{\mathrm{T}}v_{S_i}$ **done**
4. assign initial values for C as medoids (i.e., select K random vectors from the matrix X as the central vectors)
5. **repeat**
   assign each sentence v_{S_i} to the clusters which has the closest mean (that is the cost of vector); calculate new mean for each cluster;
   if the distance between medoids of any two clusters is lower than the average distance within one of the two clusters **then** merge the two clusters
   until there is no exchange of sentences between clusters and no merges among clusters

III. Experimental Results

We designed an experiment to evaluate our proposed algorithm cardiology ECG associated with chronic obstructive pulmonary disease (COPD).

A. Experiment 1: Data Reconstruction of COPD Cardiology Reports

COPD is a progressive lung disease where the airways in the lungs are damaged [13]. It is the third leading cause of mortality in the United States, affecting an estimated 14.7 million diagnosed patients [14]. The quality of life of the patients with COPD deteriorates as the disease progresses. Unfortunately, there is currently no cure for COPD. Thus, gathering similar COPD conditions within a stage across a patient population may be insightful to understand COPD progression. In addition, prior research has demonstrated that there is a relationship between COPD and cardiovascular morbidity [15].

We retrieved 1,029,363 free-text cardiology reports corresponding to 13,918 unique COPD patients who had received care at the Partners Healthcare network and died between 2011 and 2017. The average time span of a patient using the cardiology data source was 740.8 days, with a max span of 2,459 days. We utilized the section of abnormal electrocardiogram (ECG) findings. To obtain more clinical entities, we extracted 30,363 ECG paragraphs from reports generated 90 days prior to death. This study was approved by the Partners Institutional Review Board (IRB). We utilized a time distribution chart to visualize our reconstruction results (see Fig. 1). We obtained the same quantity (i.e., 30,363) of reconstructed ECG findings as sequential data. Table I shows the number of clusters (i.e., 10) and example results of sentence embedding clustering. We found that most of the clusters have two opposing polarities of sentiment. This occurs because sentence embedding is a similarity measure and thus cannot be used for sentiment analysis.

Comparing the proportion at the same duration concerning one ECG finding (see Fig. 1), the time interval between two ECGs with the first "no significant change was found" can be properly extended to about 20 days. In Fig. 1 and 2, the results indicate there exists a possible relationship between QT prolongation and polymorphic tachycardia that conforms to common clinical observations [16] because the duration constraints of "QT has lengthened" and "atrial fibrillation has replaced sinus rhythm" are very similar and commonly last around 3 months. Specifically, within mentions that contained a 70-days duration constraint, about 98.4% of patients who had QT prolongation were also diagnosed with atrial fibrillation.

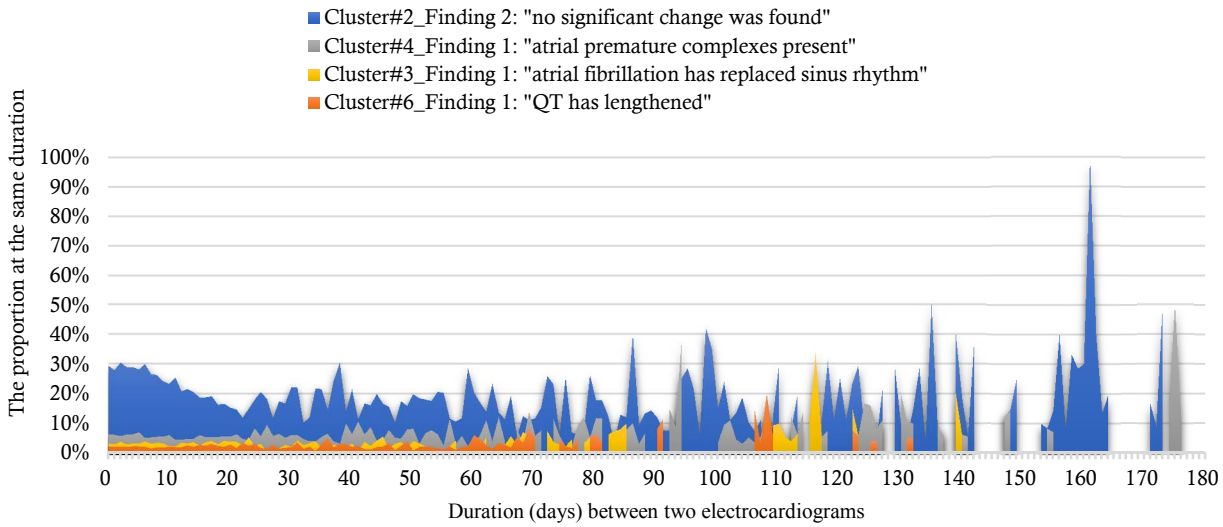

Fig. 1 A time distribution chart visualizing ECG notes reconstruction. Y-axis indicates the proporation of a cluster finding constrained within a specific duration (x-axis) found in the reports.

TABLE I. RESULT CLUSTERS OVER 30,363 ECG PARAGRAPHS

Cluster	***Record Number***	***Example***
#1	2059	Finding 1: *"aberrant conduction is now present"* Finding 2: *"aberrant conduction is no longer present"*
#2	10,920	Finding 1: *"significant changes have occurred"* Finding 2: ***"no significant change was found"***
#3	1591	Finding 1: ***"atrial fibrillation has replaced sinus rhythm"*** Finding 2: *"sinus rhythm has replaced Atrial fibrillation criteria for septal infarct are no longer present"*
#4	2742	Finding 1: ***"atrial premature complexes present"*** Finding 2: *"atrial premature complexes no longer detected"*
#5	5730	Finding 1: *"premature atrial complexes are now present the axis shifted right T wave inversion no longer evident in Lateral leads"* Finding 2: *"inverted T waves have replaced nonspecific T wave abnormality in lateral leads"*
#6	518	Finding 1: ***"QT has lengthened"*** Finding 2: *"QT has shortened"*
#7	2549	Finding 1: *"wide QRS tachycardia has replaced Sinus rhythm"* Finding 2: *"junctional rhythm has replaced Wide QRS rhythm"*
#8	1237	Finding 1: *"vent. rate has decreased by 59 BPM"* Finding 2: *"vent. rate has increased by 5 BPM and AV synchrony is improved"*
#9	2961	Finding 1: *"current undetermined rhythm precludes rhythm comparison, needs review"* Finding 2: *"previous ECG has undetermined rhythm, needs review Borderline criteria for lateral infarct are no longer present"*
#10	56	Finding 1: *"NSC (i.e., neurogenic stress cardiomyopathy)"* Finding 2: *"NSR (i.e., normal sinus rhythm)"*

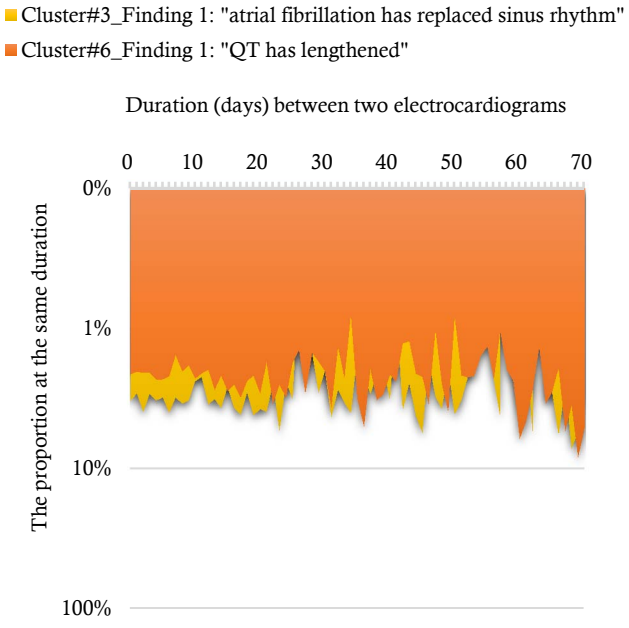

Fig. 2 a part of Fig. 1 visualizing two clusters (i.e., Cluster #3 and #6) having a similar distribution of duration constraints.

IV. DISCUSSION AND CONCLUSIONS

We developed a novel data reconstruction scheme that has three stages. First, it detects if an expression has temporal intent. Second, it decomposes and rewrites the expression into non-temporal sub-expression and temporal constraints. Finally, it clusters similar non-temporal sub-expressions by using unsupervised sentence embedding under the *K*-medoids paradigm. Our main finding is that it is feasible to utilize our proposed algorithm for unsupervised labeling of clinical notes.

Our data reconstruction algorithm for temporal clinical expression captured phrase or sentence embeddings in a way that was feasible to address several gaps in natural language processing functionality. An interesting line of efforts is that text mining is not limited to methods of complex neural network models. The best practices on data preparation and integration processes should include data format conversions, which provide enhanced data analysis capabilities. Those include time distribution visualization, sequence mining, pattern discovery, classification, prediction, and so forth.

From this reconstruction, we successfully developed a simple time distribution visualization of ECG findings. As shown in Fig. 1, a simple visualization of the time distribution of ECG findings brought significant data insights. The ability to track data from the patient to note level allows providers and health service researchers to examine the effects of care like never before. The time interval between two ECGs on "no significant change was found" allowed us to determine, for example, if the use of an ECG truly reduces costs in a real-world setting and, if so, determines its cost-effectiveness. Fig. 2 indicates an inevitable relationship between QT prolongation and polymorphic that conforms to clinical observations and previous literature.

Limitation of our study include that this work was based on a single organization, Partners Healthcare, using a single language, English, and therefore might not be generalizable to other organizations that use different documentation policies, procedures, or documentation systems.

ACKNOWLEDGMENT

The authors would like to thank Meihan Wan for assisting with method development.

REFERENCES

[1] L. Zhou, G. Hripcsak, "Temporal reasoning with medical data--a review with emphasis on medical natural language processing," J Biomed Inform, Apr. 2007, vol. 40, no. 2, pp. 183-202.

[2] TimeML Working Group, "Guidelines for Temporal Expression Annotation for English for TempEval 2010," August 14, 2009. Avariable from: http://www.timeml.org/tempeval2/tempeval2-trial/guidelines/timex3guidelines-072009.pdf [Accessed July 2019].

[3] E. Apostolova, T. Velez, "Toward Automated Early Sepsis Alerting: Identifying Infection Patients from Nursing Notes," arXiv preprint arXiv:1809.03995, 2018.

[4] L. Tian, L G. Li, C. Wang, "A Data Reconstruction Algorithm Based on Neural Network for Compressed Sensing," In Proceedings of 2017 Fifth International Conference on Advanced Cloud and Big Data, IEEE, 2017, August, pp. 291-295.

[5] Z. Chen, L. Chen, G. Hu, W. Ye, J. Zhang, G. Yang, "Data reconstruction in wireless sensor networks from incomplete and erroneous observations," IEEE Access, 2018, vol. 6, pp. 45493-45503.

[6] Z. He, C. Chen, J. Bu, C. Wang, L. Zhang, D. Cai, X. He, "Document summarization based on data reconstruction," In Proceedings of Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012, pp. 620-626.

[7] M. Gambhir, V. Gupta, V, "Recent automatic text summarization techniques: a survey," Artificial Intelligence Review, 2017, vol. 47, no. 1, pp. 1-66.

[8] Z. Xing, J. Pei, E. Keogh, "A brief survey on sequence classification," ACM Sigkdd Explorations Newsletter, 2010, vol. 12, no. 1, pp. 40-48.

[9] P. Tan, M. Steinbach, A. Karpatne, V. Kumar, "Introduction to data mining," Pearson Education India, 2nd ed., 2019.

[10] Z. Jia, A. Abujabal, R. S. Roy, J. Strötgen, G. Weikum, "TEQUILA: Temporal Question Answering over Knowledge Bases," In Proceedings of the 27th ACM International Conference on Information and Knowledge Management. ACM, 2018, pp. 1807-1810.

[11] J. F. Allen, "Maintaining knowledge about temporal intervals," In Readings in qualitative reasoning about physical systems, Elsevier, 1990, pp. 361-372.

[12] J. Strötgen, M. Gertz, "HeidelTime: High quality rule-based extraction and normalization of temporal expressions," In Proceedings of the 5th International Workshop on Semantic Evaluation. Association for Computational Linguistics, July 2010, pp. 321-324.

[13] R. M. Shavelle, D. R. Paculdo, S. J. Kush, D. M. Mannino, D. J. Strauss. Life expectancy and years of life lost in chronic obstructive pulmonary disease: findings from the NHANES III Follow-up Study. International journal of chronic obstructive pulmonary disease, 2009(4): 137.

[14] American Lung Association. Trends in COPD (Chronic Bronchitis and Emphysema): Morbidity and mortality; March 2013. Avariable from: http://www.lung.org/assets/documents/research/copd-trend-report.pdf [Accessed July 2019].

[15] R. G. Barr, D. A. Bluemke, F. S. Ahmed, J. J. Carr, P. L. Enright, E. A. Hoffman, R. Jiang, S. M. Kawut, R. A. Kronmal, J. A.C. Lima, E. Shahar, L. J. Smith, K. E. Watson, "Percent emphysema, airflow obstruction, and impaired left ventricular filling," New England Journal of Medicine, 2010, vol. 362(3): pp. 217-227.

[16] D. Darbar, P. A. Harris, A. Hardy, A. Frye-Anderson, B. White, K. J. Norris, D. M. Roden, "Marked steepening of QT restitution following cardioversion of atrial fibrillation," Heart Rhythm, 2004, vol. 1, pp. S192.