ClinicalTime: Identification of Patients with Acute Kidney Injury using Temporal Abstractions and Temporal Pattern Matching

Daniel Capurro, MD, PhD\textsuperscript{1,2}, Mario Barbe, MD\textsuperscript{1}, Claudio Daza\textsuperscript{1}, Josefa Santa María\textsuperscript{1}, Javier Trincado\textsuperscript{1}, Ignacio Gomez\textsuperscript{1};

\textsuperscript{1}Pontificia Universidad Católica de Chile, Santiago, Chile; \textsuperscript{2}University of Washington, Seattle, WA.

Abstract

Introduction: The rising cost of providing healthcare services creates an extreme pressure to know what works best in medicine. Traditional methods of generating clinical evidence are expensive and time consuming. The availability of electronic clinical data generated during routine patient encounters provides an opportunity to use that information to generate new clinical evidence. However, electronic clinical data is frequently marred by inadequate quality that impedes such secondary uses. This study provides a proof-of-concept and tests the classification accuracy of ClinicalTime—a temporal query system—to identify patient cohorts in clinical databases.

Methods: we randomly selected a sample of medical records from the MIMIC-II database, an anonymized database of intensive care patients. Records were manually classified as having an acute kidney injury or not according to the AKIN criteria. Those records were then blindly classified using ClinicalTime to represent the AKIN criteria. Classification accuracy was measured. Results: ClinicalTime correctly classified 88% of all patients, with a sensitivity of 0.93 and specificity of 0.84. Its performance was superior to simply using ICD-9 codes, which correctly classified 66% of all patients. Conclusions: ClinicalTime, a temporal query system, is a valid method to add to the currently available ones to identify patient phenotypes in patient databases and, thus, improving our ability to re-use routinely collected electronic clinical data for secondary purposes.

Introduction

The rising cost of providing health services (1) generates an intense pressure to identify which interventions are effective and which ones are not. The current evidence based medicine paradigm—that evidence needs to emerge from randomized controlled trials and systematic reviews of such trials—is not always feasible given the high costs and sometimes limited applicability of such studies (2,3). The increasing adoption of electronic health records is generating massive amounts of routinely collected clinical data that could be useful to produce ‘real-world’ evidence of what works. This, a healthcare system that is able to generate knowledge through expedited analysis of the data it generates has been labeled a learning healthcare system (4).

This attractive vision of a healthcare system that can generate knowledge from routinely collected clinical data is still hampered by multiple issues, including problems with inadequate data quality. Clinical data is primarily collected for individual patient care and not for secondary knowledge generation. Electronic clinical data required for quality improvement and clinical research is frequently stored as free-text or other unstructured formats which severely limit its reusability. For example, we recently conducted a study that analyzed the availability of structured clinical data for quality improvement and comparative effectiveness research across six hospitals. We demonstrated that in average, 75% of all data elements were not available as structured and computable database fields (5).

The most commonly used strategy to overcome this lack of availability of structured clinical data—a key component of data quality—is to manually abstract clinical data from individual patient records. Manual abstraction of electronic data defies the purpose of capturing the data electronically in the first place. Other approaches include the use of diagnostic, procedure and medication codes to identify patient traits, natural language processing to extract meaning from clinical narratives, and combinations of the above. Despite the significant advances produced in this field in the last decade, these methods still require notable efforts to develop and test. This leaves space for designing new methods to add to the currently available tools to identify patient traits within electronic clinical databases.

In this paper we describe, as a proof-of-concept, the utilization and testing of ClinicalTime, a system to identify patients based on temporal patterns of clinical events that does not require clinical knowledge abstraction in advance. We test its performance against manual abstraction by clinicians and against commonly used ICD-9 codes.
Methods

We used ClinicalTime, a temporal abstraction and temporal pattern matching system that uses a subset of the Knowledge Based Temporal Abstraction (KBTA) framework developed by Shahar et al. (6,7). An initial description of this system is available in a previous publication (8), a complete description will be provided in an upcoming publication. Briefly, ClinicalTime allows researchers to define their cases based on temporal intervals and their temporal relationships using a custom temporal query language. Once the pattern of intervals and relationships is created, the system abstracts all relevant temporal intervals from a generic relational database and then searches for patients that match the pre-defined pattern. The main difference between ClinicalTime and a full implementation of the KBTA framework, is that ClinicalTime does not require clinical contexts to abstract temporal intervals but relies on the clinical knowledge possessed by researchers at the time they describe the pattern of interest. For instance, if a researcher uses ‘Fever’ as one of the intervals included in the pattern, he or she must define, at query time what temperatures meet the Fever criteria and, given the clinical context in which the data was collected, which temperature readings can be interpolated to build a single interval. This allows ClinicalTime to be used in different settings without having to extract and represent domain-specific clinical knowledge to abstract relevant intervals.

The selected case study involved the identification of patients with acute kidney injury during an intensive care unit (ICU) stay as defined by the Acute Kidney Injury Network (AKIN) (9). We selected this condition because it is frequently encountered in the ICU and should allow a significant number of cases in a relatively small random sample of patient records; it is often overlooked in critical care patients despite its impact in morbidity and mortality; and it has a well-defined pattern of temporal events amenable to be represented as temporal intervals. To meet the AKIN definition of acute kidney injury a patient must have one of the following:

- Serum creatinine increase of 0.3 mg/dL or 50% increase from baseline, within 48 hours;
- Urinary output less than 0.5 mL/Kg/hour for 6 hours or more.

The system was tested against the Multiparameter Intelligent Monitoring in Intensive Care research database (MIMIC - II) (10). This database contains more than 30,000 intensive care unit episodes including demographic information, laboratory results, clinical notes, medications, and physiologic measurements. The data has been anonymized and it is freely available for research purposes. For this particular project we used patient laboratory data to obtain serum creatinine values and the input/output table to identify obtain hourly urinary output. In addition, manual classification used tables providing information from clinical notes, diagnostic and procedure codes, discharge summaries, and radiology reports.

Two clinicians manually classified 100 randomly selected records. After independently reviewing each patient’s laboratory results, clinical notes, radiology reports, procedure and diagnostic codes, they classified patients according to whether they met or not the AKIN criteria mentioned above. They could classify patients as ’yes’, ‘no’ or ‘unsure’. Although the AKIN criteria are purely numeric—and thus, theoretically unambiguous—we decided to allow some interpretation of inconsistent results. For example, a patient with a very low baseline serum creatinine (0.2 mg/dL) and multiple serum creatinine measurements in one day, may have a transient elevation to 0.3 mg/dL which would constitute a 50% increase and, thus meet the criteria, but if no reference to a kidney issue contained in his or her clinical notes, no reduction in urinary output, no nephrotoxic medications, no evidence of hypotension or other potential injurers, it could be considered an artifact or due to intrinsic variability of creatinine measurements (11). Discrepancies and records annotated as ‘unsure’ were manually reviewed and assigned a final classification.

After manually annotating this sample, we built a query using ClinicalTime that represented the following patterns of time intervals as shown in Figure 1.

![Figure 1](image)

**Figure 1**: patterns of clinical time intervals used to search for patients meeting the AKIN acute kidney injury criteria. Pattern a) involves two temporal intervals of serum creatinine related by a temporal relation (Before), a change in creatinine with a magnitude (0.3 mg/dL or 50% from baseline), a maximum distance (<48 hours) and a direction (elevation). Pattern b) involves only one temporal interval of urinary output of type ‘moving window’ in which ClinicalTime checks whether the condition was met in any 6-hour time-window.
Currently, ClinicalTime does not have a graphical user interface but relies on a simple query language to express a pattern of temporal intervals and their relations which has been described in detail in (8). Briefly, it enables a researcher to define a query consisting of two clinical temporal intervals (Interval1, Interval2) and a temporal relation that connects them. ClinicalTime abstracts from a relational database the temporal intervals that match both intervals and produces two temporal interval lists. Relations are such that they represent all possible temporal relations as defined by Allen in (12). The result of a query is another list of temporal intervals that be used as input for the next query, thus, allowing the creation of nested queries. The final output is a list of patients that match the pattern of interest.

ClinicalTime classified each record as ‘yes’ or ‘no’ and we built a 2 x 2 contingency table to calculate precision measures. In addition, we identified records coded with ICD-9 codes indicating acute kidney injury. All comparisons were made using the manual annotation by clinicians as the gold standard.

We used Cohen’s kappa to assess agreement between clinicians when manually annotating clinical records and McNemar’s X² to compare sensitivities and specificities of ClinicalTime and ICD-9 codes against manual annotation. A P-value <0.05 was considered statistically significant.

Results

The initial random selection of 100 medical records were manually annotated by two clinicians using information from laboratory results, diagnostic codes, clinical notes, radiology reports and discharge summaries. The initial agreement between the two annotators was 93.3% (Kappa 0.85, P<0.0001). Sources of disagreement were adequately recorded as ‘unsure’ by all annotators. ClinicalTime’s classification accuracy was measured against a second batch of randomly selected clinical records.

Overall, 48% of patients met at least one of the AKIN criteria for acute kidney injury. ICD-9 codes correctly classified 67% of all patients and ClinicalTime correctly classified 88% of all patients presenting acute kidney injury. ClinicalTime had a higher sensitivity than ICD-9 codes for identifying patients with acute kidney injury (0.93 vs. 0.33, p<0.00001) but a lower specificity (0.84 vs 0.98, p=0.04).

| Table 1: comparison of classification accuracies |
|-----------------------------------------------|
| ICD-9 codes vs. Manual Annotation | ClinicalTime vs. Manual Annotation |
| Sensitivity (recall) [95% CI] | 0.33 [0.20 – 0.48] | 0.93 [0.82 – 0.98] |
| Specificity [95% CI] | 0.98 [0.88 – 1.00] | 0.84 [0.71 – 0.93] |
| Positive likelihood ratio [95% CI] | 16.65 [2.29 – 120.83] | 5.98 [3.15 – 11.35] |
| Negative likelihood ratio [95% CI] | 0.69 [0.56 – 0.83] | 0.07 [0.02 – 0.22] |
| True positive rate (precision) [95% CI] | 0.94 [0.69 – 1.00] | 0.85 [0.72 – 0.93] |
| True negative rate [95% CI] | 0.60 [0.48 – 0.71] | 0.84 [0.70 – 0.92] |

We analyzed the sources of disagreement between manual annotations and ClinicalTime in the 11 patients where we found them. Four (36%) were due to cases in which the patient did have an elevation in creatinine that met the AKIN criteria but clinicians considered them an artifact due to intrinsic variability of the serum creatinine measurement. In these patients, the baseline serum creatinine level ranged from 0.3 to 0.4 mg/dL and the observed elevation in serum creatinine was 0.2 mg/dL in every case; the use of clinical narratives were key to the final classification of these patients. In four additional cases, the system failed to classify the patients correctly since there was incomplete data regarding urinary output; again, these patients were finally classified using information from clinical narratives.

Considering the high true positive rate of ICD-9 codes, we explored whether using an acute kidney injury ICD-9 code to patients classified as negative by ClinicalTime would improve our classification accuracy. This slightly decreased sensitivity to 0.92 [95% CI: 0.80 – 0.97] and increased specificity to 0.86 [95% CI: 0.72 – 0.94].
Discussion

The overall prevalence of acute kidney injury, using the AKIN criteria in the studied sample, was concordant with previous publications, with a prevalence of 54% (13). Overall, classification accuracy of ClinicalTime was higher than what could be obtained by using ICD-9 codes for the condition. ICD-9 codes’ low sensitivity may be surprising but this is consistent with what has been reported by others (14).

The improved accuracy, when compared to ICD-9 codes, was mostly determined by a higher sensitivity (recall) but at the expense of a somehow lower specificity and, hence, increasing false positives from 6% to 15%. This provides an additional tool, with comparable classification accuracy, to the currently existing methods used to identify patient cohorts in large clinical databases. However, building an interval-based query should take significantly less time than building a natural language processing engine to identify a new condition and be more intuitive for non-experts than constructing a complex temporal SQL database query, especially in future iterations of ClinicalTime when a graphical user interface will be available.

In this study, the main source of false positives using ClinicalTime was patients with a low baseline serum creatinine (range 0.3 to 0.4 mg/dL) who experienced increases larger than 50% from that baseline. Adding ICD-9 codes to ClinicalTime did not significantly alter the proportion of false positives. In those cases, manual classification was resolved by extracting information from clinical narratives (clinical notes, discharge summaries), which highlights the importance of clinical documentation in correctly identifying a patient’s phenotype and the need to combine methods to automatically classify clinical records.

Although there are ways to further improve the classification accuracy adding the presence or absence of specific procedures, medications and other clinical variables, we wanted to demonstrate the utility of temporal abstractions to identify patients with the condition of interest and, in particular ClinicalTime’s performance against well-known phenotyping strategies such as ICD-9 coding. In future releases of the system, researchers will be able to build intervals of non-numeric attributes to identify phenotypes such as the ‘presence of a urinary catheter’, ‘presence of a central line’ or ‘mechanical ventilation’ which should enable them to identify patients with more complex conditions for which ICD-9 or other codes may be even less accurate or simply not available.

Limitations and future work

We chose kidney injury as our case study due to its prevalence and the availability of clearly defined ICD-9 codes to identify it. However, the use of such a simple case study like this one may not completely reflect the possibilities of a temporal query system for more complex conditions. Additional evaluations of ClinicalTime using more complex case definitions will be tested in the future.

In addition, using other non-temporal attributes such as removing patients with previously known chronic kidney failure in dialysis, or complementing it with diagnostic or procedural codes could have easily improved the algorithm, but we decided to limit the evaluation to a purely temporal based query system to isolate the system’s value.

Conclusions

This study provides a proof-of-concept of the classification accuracy of ClinicalTime, a temporal interval abstraction and query system in the identification of patients in large clinical data repositories. The time and knowledge required building a temporal interval pattern and query should be significantly shorter than what is needed for other methods such as natural language processing or complex SQL queries.

Despite performing better than other methods, perfect classification still relies on human interpretation of clinical narratives. This highlights the ongoing need to use a combination of methods to improve the automatic classification of large patient datasets.

Acknowledgements

This project was funded by CONICYT through FONDECYT grant Nº 11130577

References

1. Percent Annual Increase in National Health Expenditures (NHE) per Capita vs. Increase in Consumer Price
1. Index (CPI), 1980-2009 [Internet]. Kaiser Family Foundation. 2011 [cited 2012 Feb 24]. pp. 1–1. Available from: http://facts.kff.org/chart.aspx?ch=212

2. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials. JAMA. Am Med Assoc; 2003;290(12):1624.

3. Rothwell PM. External validity of randomised controlled trials: “To whom do the results of this trial apply?.” The Lancet. 2005 Jan;365(9453):82–93.

4. Olsen L, Aisner D, McGuinness MJ, editors. The Learning Healthcare System. Washington DC: The National Academies Press; 2007.

5. Capurro D MD, PhD, Yetisgen M PhD, van Eaton E MD, Black R, Tarczy-Hornoch P MD. Availability of Structured and Unstructured Clinical Data for Comparative Effectiveness Research and Quality Improvement: A Multi-Site Assessment. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2014 Dec 31;2(1).

6. Medical temporal-knowledge discovery via temporal abstraction. 2009;2009:452–6. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=20351898&retmode=ref&cmd=prlinks

7. Shahar Y. A framework for knowledge-based temporal abstraction. Artificial intelligence. 1997 Jan 1.

8. Capurro D. Secondary Use of Electronic Clinical Data: Barriers, Facilitators and a Proposed Solution. 2013.

9. Ricci Z, Cruz DN, Ronco C. Classification and staging of acute kidney injury: beyond the RIFLE and AKIN criteria. Nature Publishing Group. Nature Publishing Group; 2011 Mar 1;7(4):201–8.

10. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L-W, Moody G, et al. Multiparameter Intelligent Monitoring in Intensive Care II: A public-access intensive care unit database*. Critical Care Medicine. 2011 May;39(5):952–60.

11. Joffe M, Hsu C-Y, Feldman HI, Weir M, Landis JR, Hamm LL. Variability of Creatinine Measurements in Clinical Laboratories: Results from the CRIC Study. Am J Nephrol. 2010;31(5):426–34.

12. Allen J. An interval-based representation of temporal knowledge. Proc 7th International Joint Conference on Artificial Intelligence, Vancouver, Canada. 1981;:221–6.

13. Case J, Khan S, Khalid R, Khan A. Epidemiology of Acute Kidney Injury in the Intensive Care Unit. Critical Care Research and Practice. 2013;2013(4):1–9.

14. Waikar SS. Validity of International Classification of Diseases, Ninth Revision, Clinical Modification Codes for Acute Renal Failure. Journal of the American Society of Nephrology. 2006 Jun 1;17(6):1688–94.