Identifying Chemotherapy Regimens in Electronic Health Record Data Using Interval-Encoded Sequence Alignment

Haider Syed[1,2(✉)] and Amar K. Das[1,3]

[1] Social Computing & Health Informatics Lab, Dartmouth College, Hanover, NH, 03755, USA
[2] Department of Computer Science, Dartmouth College, Hanover, NH, 03755, USA
[3] Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth Hanover, NH, 03755, USA
{haider.syed@cs.dartmouth.edu,amar.das@dartmouth.edu}

Abstract. Electronic health records (EHRs) play an essential role in patient management and guideline-based care. However, EHRs often do not encode therapy protocols directly, and instead only catalog the individual drug agents patients receive. In this paper, we present an automated approach for protocol identification using EHR data. We introduce a novel sequence alignment method based on the Needleman-Wunsch algorithm that models variation in treatment gaps. Using data on 178 breast cancer patients that included manually annotated chemotherapy protocols, our method successfully matched 93% of regimens based on the top score and had 98% accuracy using the top two scored regimens. These results indicate that our sequence alignment approach can accurately find chemotherapy plans in patient event logs while measuring temporal variation in treatment administration.

Keywords: Clinical guidelines · Plan recognition · Practice variation · Sequence alignment · Electronic health records · Cancer care

1 Introduction

The widespread adoption of electronic health records (EHRs) offers unprecedented opportunities to examine real-world treatment practices and outcomes. Many EHR systems, however, lack the internal ability to capture and monitor treatment plans, and may catalog the drug agents received by patients without encoding regimens directly. There is a need for robust computational methods to identify plan choice, adherence, and outcomes using EHR data, which can support research on quality of care, clinical practice variation and comparative effectiveness.

In this paper, we present a method for identifying cancer chemotherapy plans, which consist of patterns of single and multiple drug agents administered to patients in cyclical patterns known as treatment cycles. Physicians generally follow these chemotherapy protocols since they ensure standardized, evidence-based care with potentially toxic drugs. However, practice variations may result from delayed or dropped cycles.

J.H. Holmes et al. (Eds.): AIME 2015, LNAI 9105, pp. 143–147, 2015.
DOI: 10.1007/978-3-319-19551-3_17

A major open question in cancer care is determining which regimens are used in the current clinical practice and whether regimen variation affects efficacy. To address this challenge, we propose an autonomous treatment regimen identification method using known chemotherapy protocols. The approach models patient treatment histories and recommended treatment regimens as temporal sequences that encode the time interval of elapsed time between sequential events. Patient histories are compared to the treatment regimens using the proposed sequence alignment algorithm, which assigns a similarity score between the patient and protocol sequence. We present our method and evaluate it against a standard sequence alignment approach using institutional EHR data on chemotherapy administration.

2 Related Work

The use of sequence alignment methods to identify treatment regimens is a relatively new field of research. Lee et al. uses a local sequence alignment to detect HIV regimens of two sequential drug combinations [1], and a global alignment method without gaps to match patients to chemotherapy protocols with 68% accuracy [2]. Bourfa and Dankelman [3] use multi-sequence alignment to determine consensus procedural workflows in surgical activity logs, specifically for laparoscopy videos, and compare specific cases against the established consensus. Clinical workflow methods, such as [4, 5], have augmented classical edit distance to deal with temporal constraints.

3 Methods

Patient event logs were extracted, on November 1, 2014, from the EHR of Dartmouth-Hitchcock Medical Center, which was deployed on April 2, 2011. We selected all female patients who had invasive breast cancer diagnosed by the tumor registrar one year prior to November 1, 2014 and who had received NCCN (National Comprehensive Cancer Network, www.nccn.org) recommended chemotherapy agents for breast cancer based on medication administration records. We found 178 patients who met these criteria.

By reviewing the NCCN guidelines for breast cancer from 2011 through 2014, the author AD, who is a physician and has experience encoding protocols, identified 44 unique protocols. He manually matched the treatment sequences for the 178 patients to one of the 44 protocols. 115 patients received clearly identifiable NCCN recommended protocols while the rest received incomplete treatments or combinations of agents not recommended by NCCN guidelines. The patient and regimen data are encoded into event sequences, whereby chemotherapy agents are represented as single letter codes, with transition times between drug events appended to the letter codes as part of the proposed temporal extension for sequence alignment. Drugs given concurrently are cascaded sequentially, followed by the transition time.

Needleman-Wunsch Algorithm

The Needleman-Wunsch algorithm (NW) is a global sequence alignment algorithm (NW) [18]. The method guarantees the optimal alignment for a given scoring schema,

which quantifies similarity for all possible event pairings. The alignment can include event gaps, which carry a scoring penalty; the alignment score is the sum of scores for aligned events between the sequences. We use a scoring scheme that assigns exact matching events a score of 1 and mismatching events a score of -1; gaps are penalized 0.5. More complex and continuous scoring schemes can also be used. Multiple drugs administered concomitantly, such as TC, are treated as single events; partial matches of the constituent agents are treated as mismatches; therefore, TC and T receive a score of -1. This scoring choice signifies that missing agents in drug combinations indicate different treatment recommendations.

NW ignores timing between events and aligns sequences encoded without timing. For the NW method, we encoded treatment sequences that had the same series of treatment events but different intra-event distances as a single type of sequence since NW cannot distinguish between sequences with different intra-event interval lengths.

Interval-Encoded Needleman-Wunsch Algorithm

The Interval-encoded Needleman-Wunsch Algorithm (IENW) algorithm uses temporally-ordered events and the interval time between events to incorporate timing information into NW without modifying the algorithm itself; instead, temporal information is handled using a novel scoring scheme. We introduce the idea of *temporal concepts* or *temporal events* whereby a drug or drug combination followed by its' associated transition time are defined to be a single temporal event. For example, we encode four bi-weekly cycles of concurrent Docetaxel (T) and Cyclophosphamide (C) as TC.14 TC.14 TC.14 TC.14, which contains a single temporal event, TC.14. Aligning the sequence to a hypothetical patient sequence of TC.21 TC.21 TC.21 TC.21, the algorithm scores TC.14 and TC.21. By contrast, the NW algorithm ignores the timing and scores TC and TC directly. Since all medication administration events in the data set are time stamped with dates, we have chosen to measure the time distance between events at the granularity of days.

The IENW uses the static NW scoring scheme and a user-defined temporal penalty that accounts for temporal variations. For events $A.t_A$ and $B.t_B$ where A and B are drug events and t_A and t_B are their respective transition times, we define the temporal scoring function, S_t:

$$S_t(A.t_A, B.t_B) = \begin{cases} S(A,B) - T_P \dfrac{|t_A - t_B|}{\max(t_A, t_B)} = 1 - T_P \dfrac{|t_A - t_B|}{\max(t_A, t_B)}, & \forall A \equiv B \\ S(A,B) = -1.1, & \forall A \neq B \end{cases}$$

where S(A, B) is the static similarity score between events A and B used in the traditional NW algorithm, TP is a heuristic representing the maximum temporal penalty that can be imposed on the score for temporal discrepancies, we set Tp to 0.3. In cases where drug events match, a temporal penalty dependent on the percentage difference between the timing of the events in question, is applied. IENW requires the scoring scheme to hold scores for every event-timing combination that can be encountered in sequences; for event X and transition times [0, N], the scoring scheme must hold scores for X.0, X.1

through X.N. For K events, the IENW holds scores for K(N + 1) events while the static-NW scoring scheme holds scores for K events.

4 Experiments and Results

For the 115 patients that were manually matched to recommended protocols, the NW and IENW were used to score the patients against protocols; the scores were normalized by dividing the raw score by the number of events in the protocol sequences. Results against the manual annotations show that the NW and IENW correctly identify a patient's protocol with 93% accuracy using the top-scoring protocol as the prediction, and 98% using the two matches as the prediction. Using Lee's [2] gapless global sequence alignment algorithm has a 54% accuracy.

5 Discussion

Our proposed IENW method identifies the treatments of breast cancer patients based on known protocols with high accuracy. Although the accuracies of the IENW and NW are the same, the standard NW algorithm does not compare sequences using inter-event timings, and thus could not distinguish the correct regimens in 24 cases, which we rated as a tie. By contrast, IENW treats such protocols as distinct regimens and can successfully match patients to the exact protocol. The patient histories misclassified by both NW and IENW did not complete the drug cycles at the end of the regimens, so they were matched to similar, shorter protocols. For patients that were correctly classified by the IENW, the range of scores was 5 to 100, the mean score was 77 and the median score was 89.

The IENW approach has several limitations. It cannot support continuous or an infinite number of transition times. As we derive timing between drugs using encounter dates, we only have discrete times available. Another shortcoming of the IENW is that it always uses the timing appended to an event to compute the temporal penalty; however, when gaps are introduced, the total timing across the gaps should also be used in computing the penalty. Future work will present an algorithm that overcomes this problem.

Identifying treatment plans based on medical events has been an active area of research in artificial intelligence in medicine for the past two decades. Tu et. al. [7] proposed generated therapy advice using a patient's treatment information. Probabilistic-topic modeling [8, 9, 10] and process mining approaches [9, 10] can abstract patterns from events logs and summarize clinical pathways but cannot match clinical pathways to recommended chemotherapy protocols. Bhatia and Levy [11] create a chemotherapy-plan detection method for EHR data that was found to perform poorly on complex patterns. Therapy-plan recognition for data containing intervals has also been studied extensively [12, 13]. Other work has also been done [14, 15]. Our approach is distinct from prior work in using a novel temporal sequence alignment method that measures variation from recommended treatment protocols based on timing and completeness of drug administration.

6 Conclusions

Automated chemotherapy-protocol identification is an important challenge as adherence to clinical guidelines ensures standardized and optimal-care for patients. In the clinical setting, the specific regimen patients receive is not always recorded in their medical record or may be mis-recorded. Abstracting the regimen from the patient's event log manually can be time-consuming and challenging. However, this information is pivotal when deciding what treatments patients actually receive in clinical practice. We have shown that using an interval-encoded sequence alignment as an approach for identifying chemotherapy regimens can provide high accuracy.

References

1. Lee, W.N., Das, A.K.: Local alignment tool for clinical history: temporal semantic search of clinical databases. In: AMIA Annu. Symp. Proc., pp. 437–441 (2010)
2. Lee, W.N.: Evaluating clinical practice variation using a knowledge-based temporal sequence alignment framework. Ph.D. Thesis. Stanford University: U.S.
3. Bouarfa, L., Dankelman, J.: Workflow mining and outlier detection from clinical activity logs. J. Biomed. Inform. 45(6), 1185–1190 (2012)
4. Combi, C., Gozzi, M., Oliboni, B., Juarez, J., Marin, R.: Temporal similarity measures for querying clinical workflows. Artificial Intelligence in Medicine 46, 37–54 (2009)
5. Montani, S., Leonardi, G.: Retrieval and clustering for supporting business process adjustment and analysis. Information Systems 40, 128–141 (2014)
6. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48(3), 443–453 (1970)
7. Tu, S.W., Musen, M.A.: Episodic refinement of episodic skeletal-plan refinement. International Journal of Human–Computer Studies 48, 475–497 (1998)
8. Huang, Z., Dong, W., Ji, L., Gan, C., Lu, X., Duan, H.: Discovery of clinical pathway patterns from event logs using probabilistic topic models. J. Biomed. Inform. 47, 39–57 (2014)
9. Huang, Z., Lu, X., Duan, H.: On mining clinical pathway patterns from medical behaviors. Artif. Intell. Med. 56, 35–50 (2012)
10. Van der Aalst, W., Weijters, T., Maruster, L.: Workflow Mining Discovering Process Models from Event Logs. IEEE Trans. Knowl. Data Eng. 16(9), 1128–1142 (2004)
11. Hares, B., Levy, M.: Automated plan-recognition of chemotherapy protocols. In: AMIA Annu. Symp. Proc., pp. 108–114 (2011)
12. Batal, I., Fradkin, D., Harrison, J., Moerchen, F., Hauskrecht, M.: Mining Recent Temporal Patterns for Event Detection in Multivariate Time Series Data. In: Proceedings of Knowledge Discovery and Data Mining (KDD), Beijing, China (2012b)
13. Sacchi, L., Larizza, C., Combi, C., Bellazi, R.: Data mining with temporal abstractions: learning rules from time series. Data Mining and Knowledge Discovery (15) (2007)
14. Combi, C., Franceschet, M., Peron, A.: Representing and Reasoning about Temporal Granularities. Journal of Logic and Computation 14(1), 51–77 (2004)
15. Juárez, J.M., Guil, F., Palma, J.T., Marín, R.: Temporal similarity by measuring possibilistic uncertainty in CBR. Fuzzy Sets and Systems 160(2), 214–230 (2009)