

Mining Time-dependent Patient Outcomes from Hospital Patient Records

R. Bharat Rao PhD¹, Sathyakama Sandilya PhD¹, Radu Niculescu², Colin Germond MD³, A. Goel¹

¹Data Mining, Siemens Corporate Research, Princeton, NJ

²Dept. of Computer Science, Carnegie Mellon University, Pittsburgh, PA

³North-Eastern Ontario Regional Cancer Center, 41 Ramsey Lake Road, Sudbury, ON Canada

ABSTRACT

We describe TMINER, a data mining framework that accurately infers missing clinical information by reasoning over the entire patient record. Hospitals collect computerized patient records (CPR's) in structured (database tables) and unstructured (free text) formats. Structured clinical data in the CPR's is often poorly recorded, and information may be missing about key outcomes and processes. For instance, for a population of 344 colon cancer patients, important clinical outcomes, such as disease state and its evolution, are stored only as unstructured data (doctors' dictations) in the CPR. Raw evidence (i.e., information extracted from CPR) is not a good predictor of disease state. Yet by combining this evidence in a principled fashion (using methods from uncertain and temporal reasoning), TMINER accurately infers disease state sequences for recurrence, a complex time-varying outcome, for these patients. These outcomes can now be added back into the CPR in structured form.

1. INTRODUCTION

The use of computerized patient records (CPR's) to store patient information is on the increase. A CPR contains many "data sources" each of which records patient information. Structured data sources, such as financial, lab, and pharmacy databases, record patient information in database tables. Information is also stored in unstructured data sources as free text (progress notes, radiology reports, etc.), as images (X-Ray, MR, CT, etc.), and even as waveforms (ECG). Often, key clinical findings are only stored within doctors' dictations, or even worse, in specialists reports that may not be included in the CPR.

Figure 1 (see next page) illustrates the lack of *structured* high-quality clinical data, particularly reliably recorded outcomes, which makes CPR's difficult to analyze. Yet, if CPR's can be structured, retrospective studies can complement traditional prospective outcomes studies, i.e., randomized clinical trials. These CPR's (although not randomized as in a trial) become a readily available source of patient data that can be mined. Further, we can validate guidelines as they are actually implemented in hospitals, as opposed to within a tightly monitored trial.

Our research focuses on automatically inferring structured clinical information (including patient outcomes that need

to be tracked as they vary over time) from existing CPR's.

This problem can be tackled in several different ways. One approach is to develop better methods for extracting information from unstructured data, for instance, natural language processing methods to analyze free text dictations. Another is to use very accurate domain knowledge (either from a domain expert, or by learning it from labeled data). A third is to improve data collection methods at the point of care, moving towards a structured electronic medical record. Our approach is different, and yet complementary to the above approaches.

In Section 2, we describe TMINER (for Temporal Miner), a data mining framework that *combines all available evidence in a principled fashion over time*. Our focus is not the extraction of information from individual parts of the CPR – we use readily available methods for that purpose – but to combine all available information. Furthermore, due to the very nature of medical data, TMINER must overcome the following challenges:

- Reason over information recorded at variable points in time; *not* just at evenly spaced time intervals.
- Information from the CPR may be unreliable: often due to errors in the extraction, but sometimes due to errors/missing data in the data source. TMINER must handle contradictory information about the same variable, and also deal with missing data.

Further, to facilitate rapid deployment for other diseases we require that any disease-specific knowledge used by TMINER be modular, simple, and easy to acquire.

Section 3 describes our study population: 344 colon cancer patients at NEORCC, a cancer care center in Ontario. NEORCC does not reliably record information about the most crucial of outcomes, recurrence of cancer. We also describe the high-level (medical) domain knowledge TMINER uses to automatically construct different models for each patient (i.e., the models vary with the data available for each patient).

Section 4 assesses TMINER's results for inferring recurrence in 2 ways: for classification accuracy (did the patient recur), and for temporal sequence accuracy (if the patient recurred, did TMINER correctly estimate the disease-free survival time). We compare TMINER with "ground truth", as established by a colon cancer specialist after retrospective review of all 344 patient charts. (The

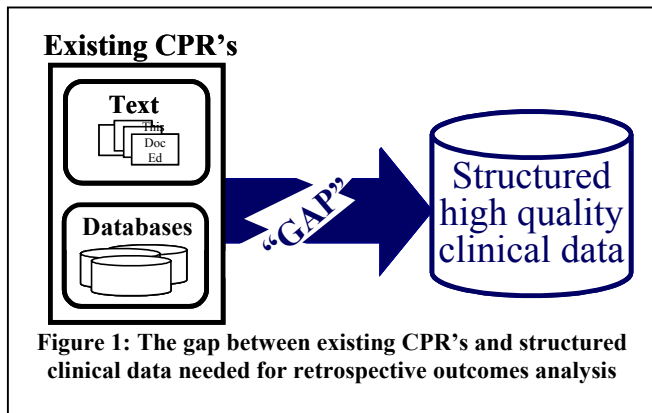


Figure 1: The gap between existing CPR's and structured clinical data needed for retrospective outcomes analysis

chart has information that is not available to TMINER.) Initial examination of the cases where TMINER's results differed from the specialist's has turned up 2 patients, where the specialist changed his retrospective diagnosis of recurrence. We also demonstrate that TMINER can accurately infer recurrence even when the extracted evidence is a relatively poor predictor. Section 5 reviews related work. We conclude in Section 6 with thoughts for future research.

2. METHODS

We begin by providing an analogy that describes the spirit in which TMINER performs inference.

2.1 Analogy: French Medical Student

A French medical student who has some knowledge about cancer is provided with cancer patient CPR's. The CPR's contain transcribed English dictations and pharmacy data. The student's task is to classify which patients have had a recurrence, and if they have, determine when it occurred.

Unfortunately his English is poor, though he does know some key medical words and a few of the drug names. However, he cannot rely purely on the presence of some key words, such as *metastases*, in the dictation, because he knows that physicians often make negative statements ("Patient is free of evidence of metastases.").

How might the student best carry out his task?

The student can collect all relevant evidence from the CPR – without trusting any single piece of evidence – and combine it to reconcile any disparities. He can use his

knowledge about the treatment of cancer – for instance, on noting that a patient had a liver resection, the student can conclude that the patient (probably) previously had a recurrence. TMINER works in a similar fashion as illustrated in Figure 2.

2.2 Problem Definition

Let S be a continuous time random process taking values in Σ that represents the state of the system. Let $T = \{t_1, t_2, \dots, t_n\}$, where $t_i < t_{i+1}$, be the n "times of interest" when S has to be inferred. Let S_i refer to the sample of S at time $t_i \in T$. Let V be the set of variables that depend upon S .

Let O be set of all (probabilistic) observations for all variables, $v \in V$. Let O_i be the set of all observations "assigned" to $t_i \in T$; i.e., all observations about variables, $v \in V$, that are relevant for this time-step t_i . Similarly, let $O_i^j(v)$ be the j -th observation for variable v assigned to t_i . Let $seq = \langle S_1, S_2, \dots, S_n \rangle$ be a random variable in Σ^n ; i.e., each realization of seq is a state sequence across T .

GOAL: Estimate the most likely state sequence, seq_{MAP} , (the maximum *a posteriori* estimate of seq) given O .

$$seq_{MAP} = \arg \max_{seq} P[seq | O]$$

2.3 Extraction: Observations from data

TMINER extracts information, o_i , from every data source in a uniform format called *probabilistic observations*. Each o_i is drawn entirely from a single piece of information in a data source (e.g., from a phrase in a sentence, or a row in a database table), and hence is assumed to be inherently undependable. The observation {"Recurrent", "12/17/01", <T=0.1, F=0.0>}, states that the Boolean variable "Recurrent" has an associated distribution over all possible values that can be taken by "Recurrent". The probabilities do not have to add up to 1.0; any remainder (here 0.9) is assigned to *unknown*, and is smoothed over T/F, based upon the (time-dependent) a priori distribution.

Extraction from Structured data: TMINER communicates with all databases via JDBC, Java's built-in interface to relational databases. Executing a query (e.g., retrieve drug administered) is expressed as a probabilistic observation.

Extraction from Free Text: TMINER strips document

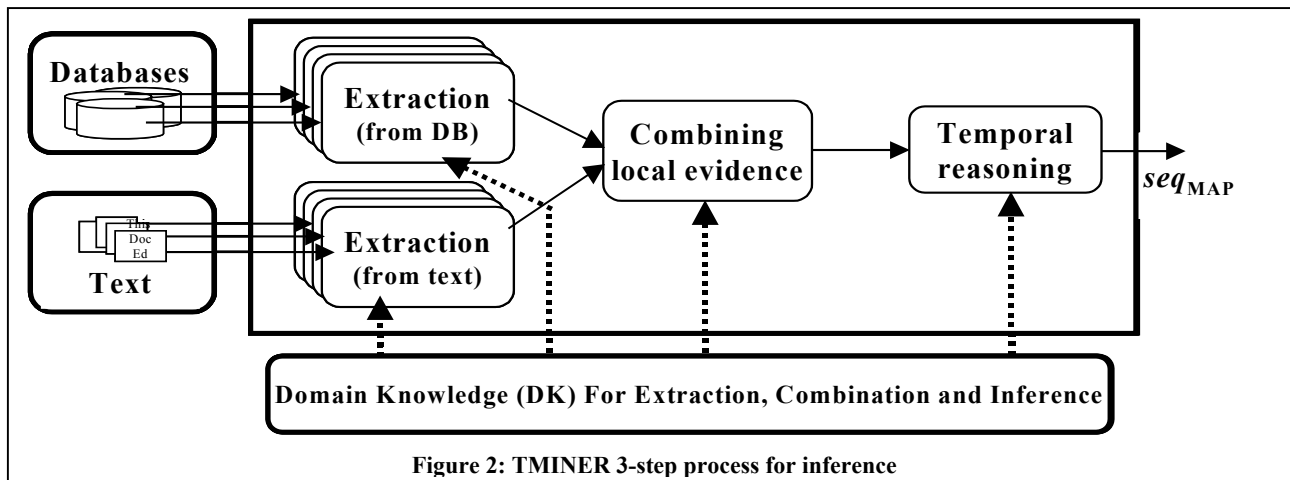


Figure 2: TMINER 3-step process for inference

headers/footers, and tokenizes free text. We extract information from the token stream via phrase spotting, an easy-to-implement method from computational linguistics. Phrase spotting is about as simple as it sounds. A phrase-spotting rule is applied within a single sentence. The rule:

$[metastasis \ \& \ malignant] \Rightarrow \{“Recurrent”, \langle T=0.5 \rangle\}$

states that if the 2 words (actually aliases) in the rule are found in a sentence, a probabilistic observation about recurrence should be generated. TMINER also has compound rules to detect “negation” and “imprecision”, which modify the probabilities in existing observations.

TMINER can clearly benefit from using better Natural Language Processing methods [1] than phrase spotting. Of direct relevance is the analysis of doctors’ dictations by Chapman [2] which identifies the 7 most common uses of negation in doctors’ dictations. Augmenting our aliases with a general lexical reference [3] or a medical language dictionary (SNOMED) should improve performance.

2.4 Combination + Inference

The primary focus of our interest is estimating what happened to the patient across \mathbf{T} , the duration of interest. The estimation of the MAP state sequence can be done in two steps, the first of which is combination of observations at a fixed point in time and the second is the propagation of these inferences across time.

Each (smoothed) o_i is in the form of an *a posteriori* probability of a variable given the small context that it is extracted from. All observations, $\mathbf{O}_i^j(v)$, about a variable for a single time t_i are combined into one assertion in a straightforward manner by using Bayes’ theorem:

$$P[v_i | O_i^1(v_i), \dots, O_i^k(v_i)] \propto P[v_i] \cdot \prod_{j=1}^k P[O_i^j(v_i) | v_i] \propto \frac{\prod_{j=1}^k P[v_i | O_i^j(v_i)]}{P[v_i]^{k-1}}$$

At every $t_i \in \mathbf{T}$, we model the relationships among \mathbf{S}_i and \mathbf{V} using a Bayesian Network [4]. Because we model the state process as being Markov and the state as being causative (directly or indirectly) of all the variables that we observe, we have the following equation:

$$P[seq | O] \propto P[S_0] \cdot \prod_{i=2}^n P[S_i | S_{i-1}] \cdot \prod_{i=1}^n P[O_i | S_i] \\ \propto \prod_{i=2}^n \frac{P[S_i | S_{i-1}]}{P[S_i]} \cdot \prod_{i=1}^n P[S_i | O_i]$$

This equation connects the *a posteriori* probability of *seq* (any sequence of samples of the state process across time) given all observations, to $P(\mathbf{S}_i | \mathbf{O}_i)$, the temporally local *a posteriori* probability of the state given the observations for each time instant. Essentially, we string together the

temporally local Bayesian Networks by modeling each state sample, \mathbf{S}_i , as the the cause of the next sample, \mathbf{S}_{i+1} . TMINER is described in greater detail in [5].

3. NEORCC COLON CANCER PATIENTS

Our study focuses on Stage III colon cancer patients. Colon cancer patients are grouped into 4 stages, based on the extent of disease at the time of diagnosis. The survival for each stage is different and is worse as the stage increases [6]. All references to patients hereafter, will be assumed to be Stage III colon cancer patients.

3.1 Data Sources at NEORCC

At NEORCC, the principal source of *structured* patient data is OPIS [7], an oncology patient relational database. OPIS contains data about patient demographics, staging, diagnosis date, and chemotherapy. However, OPIS does not reliably record information about the most basic outcome, recurrence of cancer (and date thereof). NEORCC has stored information in OPIS from 1988.

The principal source of unstructured patient data is another database that stores “doctors’ dictations” (free text that has been transcribed). These dictations include oncologists’ progress notes (from patient visits to the oncologist, the cancer specialist), and notes from visits to non-specialists, such as nutritionists. NEORCC began transcribing and storing dictations in early 1994.

The data in our study is based on a snapshot of both data sources taken in January 2001. We restrict our study to 344 patients for whom at least one dictation has been stored. The earliest and latest a patient in our study began treatment was 1985 and November 2000 respectively. Patients averaged 10.8 dictations, with a maximum of 53 dictations. 18 patients had 1 dictation, and 54 patients had 3 or fewer dictations. Many dictations had no recurrence information (at least, based on phrase spotting extraction).

3.2 TMINER Domain Knowledge

Each step in the process for defining domain knowledge (DK) for TMINER requires little or no fine-tuning.

1. *Identify state S*. Here \mathbf{S} is a single variable, “disease-state”. The only transition of interest is if/when the patient recurs – i.e., from State “0” (disease-free) to State “1” (recurrent). The legal sequences are 0^*1^* .
2. *Identify V, and its data sources*: “Recurrent” and “CEA” (a test result for a tumor marker) are drawn from doctors’ dictations. From OPIS we track if chemotherapy is provided, and its intent. (This is only recorded from 1998 onwards.)
3. *Define extraction DK*: The queries to extract chemo from OPIS, and the rules to extract “Recurrent” and “CEA” from text. We defined 9 phrase-spotting rules, and set the probabilities induced by every rule

to 0.5 (see example rule in Section 2.3). Two general compound rules detect “negation” and “imprecision”.

4. *Define T*, and a mechanism to assign observations to $t_i \in T$: As we want to infer when the doctor diagnoses recurrence, each visit is defined a time of interest. Observations between visits are assigned to the next visit. (Note that **T** and *n* are patient-dependent.)
5. *Identify probabilistic dependencies*: A Bayesian network encoding the dependencies between **S** and **V** is used to compute the distributions of **S_i** given **O_i**.
6. *Define State Transitions*: A standard survival curve for colon cancer patients [9] is used to model dwell times and the transition probabilities.

4. RESULTS

An oncologist reviewed the patient charts for these patients (this took the doctor about 3 months) to establish “ground truth”. We randomly chose 138 (40%) patients to be in the “training” set. We reviewed the notes on half the training set to arrive at phrase-spotting rules, and used the other half of the training set to adjust potential gross errors in the conditional probabilities in the model (in consultation with the specialist). After making all such adjustments, we ran TMINER over the remaining 60% (206) patients. *All tables and figures show results over the entire data set; test set results are reported in the text.*

4.1 Adjustments to “Ground Truth”

TMINER initially produced 7 classification errors on the training set. On review, we found that 2 of these were not classification errors, but in fact misclassifications by our specialist. The results presented here are after correcting the ground truth for these mistakes. We have not yet undertaken a similar analysis of the errors on the test set.

4.2 Results: Classification

Table 1 compares TMINER classification (the final state of seq_{MAP}), against “ground truth” and also with outcomes recorded in OPIS. (The recurrence information in OPIS is not used by TMINER – NEORCC wanted an independent estimate of recurrence.) On just the test set, TMINER had similar false positive (4%) and negative (13%) rates. Even though TMINER is optimized to infer the *state evolution for each patient*, it estimates the final state very

	Pred Recur rent	False +ve	Pred Non Recur	False -ve	False +ve rate	False -ve rate
Truth	99	0	245	0	0	0
TMINER	94	7	250	12	3%	12%
OPIS	67	1	277	33	<1%	33%

Table 1: Classification Error on 344 patients

well.

We note that 4 of TMINER’s classification errors were unavoidable as the patient had recurred, was resected and became disease-free even before any data was available to TMINER. The oncologist’s classification of recurrence is based on the patient chart (which includes handwritten notes from before 1994, and external specialists reports).

4.3 Results: Sequence Prediction

Figure 3 displays a histogram of the distribution of errors in the time of recurrence (when seq_{MAP} predicts 0→1) for 87 patients that TMINER correctly declared recurrent. The X-axis is the number of visits that our prediction missed by (“3E” is “3 visits early”, and “2L” is “2 visits late”). TMINER predicted the exact time of recurrence for 54 recurrent patients. Including the 238 correct sequence predictions for non-recurrent patients, TMINER predicts 85% of all patient sequences with complete accuracy

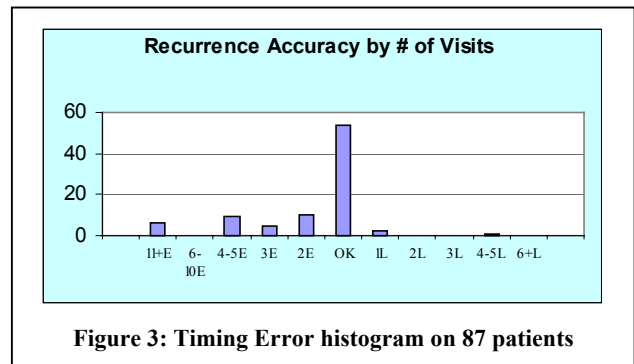


Figure 3: Timing Error histogram on 87 patients

(84% of the test set sequences are predicted perfectly.)

Table 2 demonstrates the efficacy of temporal inference to estimate patient state and the necessity of something more than just temporally local document-based analysis. We collected statistics about evidence of recurrence in dictations. Table 2:Row 1 shows that for 3735 documents, nearly 60% (2219) say nothing about recurrence; at least, as per phrase spotting. An inference engine that only uses a text document to declare the patient’s current state will make no declaration 60% of the time, and be expected to make a mistake around 18% of the time when there is evidence in the note. Whereas TMINER is able to “fill in the gaps” using information

	Correct	No Evidence	Wrong	Error
Raw Evidence	1248	2219	268	18%
TMINER Correct	1204	2081	215	6%
TMINER Error	4%	6%	20%	

Table 2: Visits correctly classified for all 344 patients

about the patient across time and thus labels 94% of the

visits with the correct state. Table 2:Row 3 shows that even when the documents have missing or misleading evidence, TMINER can achieve highly accurate results. (Note that some patients receive chemotherapy in other hospitals. OPIS has chemo observations with intent for only 9% of our patients, making it also a poor predictor in isolation.)

5. REVIEW OF RELATED WORK

Our work draws heavily on earlier work on Bayesian networks and graphical models (see [4] for an overview). Here we briefly list some temporal reasoning methods in medicine (see [8] for a extensive bibliography) that are similar to TMINER in some aspects; [5] provides greater detail about how these methods relate to our research.

Ngo et al [9] describe a temporal probabilistic reasoning method via context-sensitive model construction. Bellazi et al [10] describe a system that uses a Dynamic BN to analyze the blood glucose level of a patient over a time interval. Kayaalp et al [11] use structured information to predict probabilities of survival for ICU patients. Other related research [12][13][14] deals with representing temporal data and enforcing temporal integrity.

6. SUMMARY

Our goal is to extract outcomes and other variables in a structured fashion from existing CPR's with minimal manual intervention. We have described a general framework for performing Bayesian inference on random processes that are sampled at arbitrary time instants by integrating information from structured and unstructured data sources. TMINER uses simple domain that is easy to solicit and requires little fine-tuning (e.g., we set all rule probabilities arbitrarily to 0.5, and approximate the Stage III survival curve [6] with an exponentially decaying curve). We have demonstrated that even with missing patient information and misleading evidence, TMINER successfully extracts structured information for a complex time-varying medical outcome, namely recurrence.

We plan to incorporate better natural language processing methods [1][2] into our work; new extraction methods can be "plugged into" TMINER. We plan to investigate the potential for learning the model parameters from data [15]. For this application, it made sense to define the times of interests to coincide with the visits; however, TMINER can infer state at any time in the process, and indeed can be used to predict future values of the state. Also, TMINER can be extended to handle non-stationary Markov processes. Currently, we are applying TMINER to ICU AMI diabetic patients to analyze the impact of two different treatment guidelines. In another hospital, we will analyze the use of expensive drugs in diabetic neuropathy. We are considering TMINER for automatic verification of government safety standards[16], and

possibly for early detection of bio-terrorism events from temporal patterns in unstructured patient records.

REFERENCES

- [1] Manning, C.D., Schutze, H. Foundations of statistical natural language processing. Cambridge, MIT Press.
- [2] Chapman, W., Bridewell W., Hanbury P., Cooper, G., Buchanan, B.G. Evaluation of negation phrases in narrative clinical reports. Proc AMIA Symp 2001.
- [3] Fellbaum, C., WordNet: An electronic lexical database. MIT Press, May 1998.
- [4] Jensen, F.V. An introduction to Bayesian networks. UCL Press, 1996.
- [5] Rao, R.B., Sandilya, S., Niculescu, R., Germond, C. Temporal inference from structured & unstructured data sources, Siemens Corporate Research Technical Report Series, SCR-02-TR-720, Princeton, Feb 2002.
- [6] Sobin, L.H., Wittekind, C., TNM classification of malignant tumours, 5th ed. Wiley-Liss, New York, 97.
- [7] Cancer Care Ontario. OPIS, The oncology patient information system, Version 4.3. Toronto, 1999.
- [8] Combi, C., Shahar, Y. Temporal reasoning and temporal data maintenance in medicine: issues and challenges. Computers in Biology and Medicine 27(5), 1997, 353-368.
- [9] Ngo, L., Haddawy, P., Krieger, R.A., Helwig, J. Efficient temporal probabilistic reasoning via context-sensitive model construction. Computers in Biology and Medicine, 1997.
- [10] Bellazzi, R., Larizza, C., De Nicolao, G., Riva, A., Stefanelli, M. Mining biomedical time series by combining structural analysis and temporal abstractions. JAMIA, 1998, vol. 5, 160-164.
- [11] Kayaalp, M., Cooper, G. F., Clermont, G. Predicting ICU mortality: A comparison of stationary and nonstationary temporal models. Proc AMIA Symp., 2000, 418-422.
- [12] Horn, W., Miksch, S., Egghart, G., Popow, C., Paky, F. Effective data validation of high frequency data: time-point, time-interval, and trend-based methods. Computers in Biology and Medicine, 1997.
- [13] Kahn, M., Fagan, L., Tu, S. Extensions to the time-oriented database model to support temporal reasoning in expert medical systems. Methods of Information in Medicine, 1991, 30, 4-14.
- [14] Larizza, C., Moglia, A., Stefanelli, M. M-HTP: A System for monitoring heart transplant patients. Artificial Intelligence in Medicine, 1992, 4, 111-126

4/3/2003 11:14 AM

[15] Heckerman, D. A tutorial on learning with Bayesian networks. Microsoft Research MSR-TR-95-06, 1996.

[16] Joint Commission on Accreditation of Healthcare Organizations (JCAHO): <http://www.jcaho.org>.