The Annotation Station: An Open-Source Technology for Annotating Large Biomedical Databases

OT Abdala, GD Clifford, M Saeed, A Reisner, G Moody, I Henry, RG Mark Harvard-MIT Division of Health Sciences & Technology, Cambridge MA, USA

Abstract

The authors present a new framework for annotating large databases of multi-channel clinical data such as MIMIC II. MIMIC II is an ICU database which includes both regularly sampled (but often discontinuous) high rate data (such as ECG and BP waveforms) and low resolution data (such as waveform derived averages, lab results, medication changes, fluid balances, and nurse-verified signal values) which are often sparse, asynchronous and irregularly sampled. Because of the extremely rich high-dimensional nature of MIMIC II medical data, we require a vast quantity of labeled data in order to test and validate ICU decision-support algorithms. MIMIC II presents a new annotation challenge which cannot be met by currently existing annotation structures due to the heterogeneous data types and unavailability of data. We have constructed a hardware/software configuration known as the "Annotation Station", a quad-monitor, time synchronized, viewing tool which displays all of this data in an organized fashion. The software gives the user the opportunity to produce annotations in a practicable format that serve the goals of the MIMIC II project. The annotation structure must apply to all the numeric signals in MIMIC as well as non-numeric data such as nursing notes, discharge summaries and patient histories. Furthermore, in order for the annotation framework to adequately represent the state of the patient to a human or machine, it must involve clinical coding using accepted medical lexicons and causal linkage of one annotation to another. This linkage is the basis of causal reasoning between significant events in different streams of the data. The annotations also include subjective expert assessments of a patient's hemodynamic state and trajectory. These assessments provide objective and subjective labels for assessing algorithms that track trends in the data with a view to producing intelligent alarms.

1. Introduction

Over the past 10 years, there has been a large ongoing effort at MIT to collect massive databases of Intensive Care Unit (ICU) patient stays. Initially, between 1994 and 1997, Mark and Moody [1, 2] collected data from the Beth Israel Deaconess Medical Center ICU and carefully selected 100 patients who were deemed to be representative of the ICU population. They then disseminated them in the research community. This database proved to be of great use in developing and validating algorithms related to heart arrhythmias and hemodynamic alarms based on ECG and ABP.

1.1. Current Data Collection Efforts

Following the success of this database, Mark *et al.* [3, 4] began an ongoing effort named MIMIC II which took advantage of new technological capabilities in high speed networking and massive data storage to record a richer set of data for a dramatically larger set of patients. MIMIC II now includes data for over 3000 patient stays.

There are several similar data collection efforts also ongoing in the medical research community. For example, the SIMON database at Vanderbilt University Medical Center (VUMC) [5] collects almost the same information as is in the MIMIC database. The IMPROVE data collection project [6] collected a significant amount of waveform and trend data at the Kuopio University Hospital in Finland. After the goals for the MIMIC II database are presented, a description of the extensions to the above data collection systems is then given.

1.2. Goals

There are two main functions we envisioned for the MIMIC II DB:

Firstly, a large, well characterized patient population with a vast array of associated data recorded is required. This will prove useful in clinical studies where a researcher can retrospectively examine the effects of various treatments and medications on patients with different histories and conditions and can hypothesize on how to improve therapy.

Secondly, a vast, annotated database will support intelligent alarm development and validation. In order to do this, we require annotations with several characteristics.

1. Coded identifiers of the patient state at the time of annotation. We use the UMLS [7] with a user specific abbreviation dictionary in order to standardize the concepts

Table 1. Data Types in MIMIC II

Signal	F_s Range	Data Type
Waveforms	125 Hz	8-bit
Trend Data	$\frac{1}{min}$	4-byte Float
CareVue Data	$\frac{1}{5min} \rightarrow \frac{1}{8hrs}$	4-byte Float
Text Data	$\frac{1}{8hrs} \rightarrow \frac{1}{\text{admission length}}$	ASCII text

included in these descriptions.
2. Identification of the onset and offset of medical conditions during the patient stay.
3. Links to the data streams included in MIMIC II which represent the fact based evidence of the presence of a particular condition.
4. Causal links between medical conditions in the patient record where the condition of the patient at one point in time or an element of the patient history affects a later state.
5. Subjective evaluations of the health of the patients organ systems. See Table 2 and Section 3.1 for further details.

Medical characterizations of the state of a patient are highly complex, involving many organ systems which are intricately linked. This is simplified into a scoring system where the annotator gives a quantitative opinion of the state and trajectory of the patient for each *disease process* at critical times during the patient stay. Essentially, the complex state of the patient is projected down into multiple simplified *disease process signals* where the progression of each evolving disease process is recorded over the stay of the patient. The reality of a medical diagnosis is more complex than this and the deterioration of one disease process can affect the onset of a new disease process. Therefore, through causal links, we allow a representation of which processes affect other processes, and when they do so. Our annotation structure provides the ability to extract both the cursory information of the annotator assessments (the disease process scores) as well as the detailed information. This includes:

- what specific condition existed (as indicated by which UMLS code was chosen) at each time point in the progression of each disease process,
- what disease processes are affecting each other, and
- what set of data is used as evidence for this hypothesis.

2. Data Description

There are various types of data in MIMIC II that come from different sources in different formats. Table 1 details the 4 main types of data in terms of format, source and sampling frequency, F_s. Waveforms include the electrocardiogram (ECG), arterial blood pressure (ABP), central venous pressure (CVP), and respiration signals. Trend Data includes unverified 1-minute averaged parameters such as heart rate, blood pressures, oxygen saturations, and cardiac output. The ICU Philips information database, CareVue [8], includes many parameters such as the CareVue versions of the Trend Data which are nurse-verified but less frequently recorded, intravenous drip medication rates, and fluid intake and discharge. The text data consists of notes that are entered by nurses at the end of their shifts, discharge summaries, and the results of some diagnostic tests such as X-rays. All data streams have missing data due to recording pauses, noise dropout, or change in frequency of sampling. This data has been de-identified according to HIPAA (Health Insurance Portability & Accountability Act) regulations as described by Douglass *et. al.* [9].

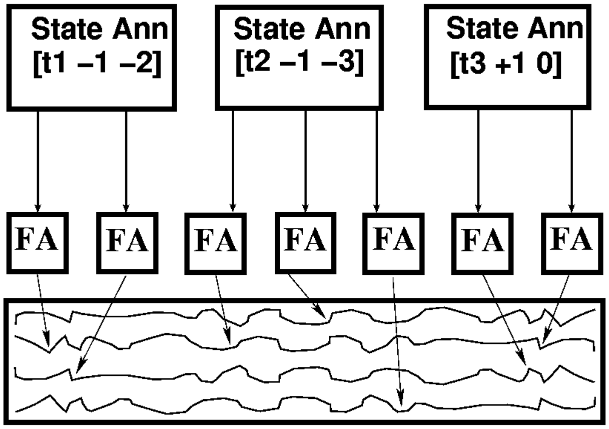

Figure 1. Annotation Scheme: This shows the three tiered approach to annotation. On the top tier, we see the /it-State Annotations, which have time values(t1, t2, t3), trajectory assessments(-1, -1, +1) and state assessments(-2, -3, 0) which represent meanings as detailed in Table 2

3. Methods

3.1. Annotation Methodology

In order to produce the web of causal reasoning we envision, a three-tiered approach was employed in the annotation structure. A schematic of this can be seen in Figure 1. At the bottom level, on the first tier, we have the data streams themselves. This level includes some itemization of what type of data is included in the stream (e.g. ABP, Hematocrit, Cardiac Output, or ECG). As clinicians review this data, they may want to flag particular pieces of data as interesting and relevant to a change in patient state. In our annotation scheme, this flagging creates what we call a *Flag Annotation* (FA), which is the second tier. This Flag Annotation automatically includes the time and data stream that was flagged. It also gives the annotators the op-

Table 2. Possible values and meanings for *State* and *Trajectory* values within State Annotations

Assessment	Value	Example Meaning
State	[0]	normal, stable
State	[-1]	mild instability (detectable CHF, hypovolemia, early sepsis)
State	[-2]	moderate instability
State	[-3]	severely decompensated; cardio -genic shock, pulmonary edema, pulmonary embolus, septic shock
Trajectory	[-1]	deterioration
Trajectory	[0]	no change since last assessment
Trajectory	[1]	improvement

portunity of qualifying the tagged piece of data with a free text description of what they believe is occurring in this data stream alone. Annotators are encouraged to use Shu's [10] coding scheme to attempt to assign a UMLS code to this text description. Annotators also have the opportunity to add a more detailed free-text description of the flagged event.

Once all the relevant data evidence is flagged in the second tier, annotators make judgments about when there are significant changes in patient state. They then create the third tier, termed the *State Annotations* where annotators indicate their assessment of the state of the current disease process and its trajectory since the last assessment. This should be done according to the guidelines in Table 2 In order to justify their observations, they may then link in all of the Flag Annotations that they consider relevant to this State Annotation as well as all previous State Annotations that play a role in bringing about the state that is currently being annotated. For each disease process, there must be at least two State Annotations, one that represents the onset of the disease process, and one that represents the offset. Each annotation in a disease process must link to the previous annotation in the disease process, as the previous state obviously has great impact on the state at the next time step. For the case where more than one disease process per patient is being annotated, we need to be able to quickly and reliably identify which annotations are relevant to which disease process. This involves the creation of a list of Disease Processes that we will annotate, and the assignment of one of these disease processes labels to each State Annotation. This will result in a number of disease processes for which we have a train of state annotations in a patient record. This *Disease Process Signal* could then be quickly and correctly extracted from these State Annotations, thus satisfying our original goal.

3.2. Using the Annotation Station

In order to annotate the MIMIC II database, annotators must be able to efficiently review patient records, then to quickly note their beliefs about the patient state. In the next two sections, a brief description of how the Annotation Station provides the annotator with the ability to do this is presented.

The Annotation Station is a four screen Java client which is attached to a data server serving the MIMIC II database on a Postgres database. The four screens display an Annotation Window, an Information Viewer, a Trend Viewer, and a Waveform Viewer. The choice of view in the Annotation Station is driven by the idea that annotators require the ability to view a patient stay at many different levels of granularity.

The first screen that the annotator looks at is the information viewer. At the top of this screen are a series of timelines. Currently, this includes what we call an *Orienter*, the main driving timeline, and an alarms timeline. There is a well-defined Timeline Java interface such that timelines with new information can easily be added if needed. The function of the Orienter is to show the region of the patient stay that is currently being viewed. It shows the admission date, the release date, and ticks at every midnight in between these two times. The main timeline below the orienter shows created annotations and nurse's notes, so the annotator has a frame of reference. The main timeline is the driver for the whole system. Actions, such as zooming and scrolling, on the timeline can (optionally) cause all pieces of data on the annotation station to update themselves based on the timeline's new view. Also, performing these actions on the data presented in the annotation station can (again optionally) cause the timeline to update itself based on the new view selected by the annotator. Below this main timeline, there is an alarm timeline, which plots the alarms generated over the patient stay. The annotator's current region of interest is shaded in the orienter so that it exactly matches the region that is shown on any timeline.

Further down on the Information viewer are several types of information including a nurse's note pad where the annotator can sequentially review the nurse's notes. Also, all data described as CareVue data is displayed in text boxes and placed on the information viewer. These boxes enumerate all of the available signals and indicate the nearest previous in time. The annotator also has the option of plotting a time series of all of these values on a signal panel.

The signal panel can plot the annotator's choice of up to 10 signals simultaneously. These plots are zoom-able and scrollable. There are also types of data that cause incorporation of several other data streams into one plot. For example, the unverified systolic diastolic and mean arterial blood pressures are displayed together simultaneously with a resolution of one minute. In addition, the asynchronous nurse verified values are overlaid to give the annotator an alternative source of the same information. We also display the seven waveforms, when available, on the wave panel. All of these plots can optionally control the active time range the annotation station is focused on if the annotator wishes it to be synchronized.

Every piece of data in the annotation station can be used as a marker for the creation of a Flag Annotation. For Flag Annotations, the source of the data is the selected data stream. State Annotations can be created with time anchors from data streams, but State Annotations do not have a source in data, except indirectly through its Flag Annotation links.

4. Problems

It is easy to imagine a situation where, over several annotators, we have time-dispersed annotations which do not match-up well with each other. Essentially, we would have no basis of comparison, no basis of measuring inter-annotator agreement, and we would create the very difficult job of merging these time-dispersed annotations. Therefore, we propose that annotations be created on a CIA (Criteria for Instigation of Annotation). These are objective criteria that must be met in order to begin annotating a disease process. This guarantees a consistent onset for disease processes, which provides tells us which disease process from one annotator corresponds with which disease process of another annotator.

The UMLS is a medical lexicon that codes physical objects, concepts, or ideas such as medications, anatomical structures, interventions or treatments. The lexicon also has a hierarchy of relationships between these concepts which attempts to categorize and organize them. Due to the complex nature of the possible combinations of all of these elements of the database, there are often several ways of coding the same medical idea in the UMLS with no way of easily mapping between them. It is for this reason that it was decided that the annotators must select out of a pre-condensed list of possible disease processes. Annotators can additionally add more detail if they believe the predefined category does not fully describe the disease process which is being annotated. However, at least one of the standard disease processes must be selected. This gives us a useful limited UMLS subset from which to develop and train causal reasoning algorithms.

5. Conclusions

Using the Annotation Station, we have already annotated 20 records for hemodynamic stability. Over the next year, we intend to annotate 200 cases to be posted on PhysioNet [2]. It is also intended that a version of this system will be made available within ICUs to facilitate quickly reviewing a patient's history to improve continuity of care in an environment where clinicians change shifts.

6. Acknowledgments

This publication was made possible by Grant Number R01 EB001659 from the National Institute of Biomedical Imaging and Bioengineering.

References

[1] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 2000 (June 13);101(23):e215–e220. Circulation Electronic Pages: http://circ.ahajournals.org/cgi/content/full/101/23/e215.

[2] http://www.physionet.org/.

[3] Saeed M, Lieu C, Raber G, Mark R. MIMIC II: A massive temporal ICU patient database to support research in intelligent patient monitoring. Computers in Cardiology 2002; 29:641–644.

[4] Mark RG. Integrating data, models and reasoning in critical care, 2003. National Institute of Biomedical Imaging and Bioengineering Proposal R01 EB001659.

[5] Dawant B, Uckun S, Manders E, Lindstrom D. The simon project: Model-based signal analysis and interpretation in intelligent patient monitoring. IEEE Engineering in Medicine and Biology 1993;12(4):82–91.

[6] Korhonen I, Ojaniemi J, Niieminen K, Van Gils M, Heikela A, Kari A. Building the improve data library. IEEE Engineering in Medicine and Biology Magazine Nov/Dec 1997; 16(6):25–32.

[7] Umls knowledge sources, 16th edition - July release: 2004ab documentation http://www.nlm.nih.gov/research/umls/umlsdoc.html.

[8] http://www.medical.philips.com/main/products/ patient_monitoring/products/carevue/.

[9] Douglass M, Clifford G, Reisner A, Moody G, Mark R. Computer-assisted deidentification of free text in the MIMIC II database. Computers in Cardiology 2004;M6.2.

[10] Shu J, Clifford G, Saeed M, Long W, Moody G, Szolovits P, Mark R. An open-source, interactive java-based system for rapid encoding of significant events in the icu using the unified medical language system. Computers in Cardiology 2004;S41.6.

Address for correspondence:

Omar Abdala
Laboratory for Computational Physiology
Harvard-MIT Division of Health Sciences & Technology
Rm E25-505, 45 Carleton St.,
Cambridge MA 02142 USA
ota@mit.edu