Matching Data Fragments with Imperfect Identifiers from Disparate Sources

Michael B Craig[1,2], Benjamin E Moody[1,2], Sherman Jia[1,2], Mauricio C Villarroel[1,2], Roger G Mark[1,2]

[1]Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, USA
[2]Massachusetts Institute of Technology, Cambridge, MA, USA

Abstract

The Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-II) Database includes waveforms and derived parameters from bedside monitors, clinical data from an ICU information system, and data from other hospital laboratories and archives, for thousands of patients. These data come from devices under separate domains that often do not retain detailed information regarding relationships between parameters. We developed software for matching data fragments with incomplete and sometimes incorrect identifiers. We found that names, medical record numbers, waveform times and durations, and ICU admission and discharge records were most helpful when available; however, physiological data can also be used in some circumstances. Rule-based normalization and text edit-distance metrics are used in addition to a visual verification tool for patients whose records cannot be assembled automatically. Thus, a majority of the available waveform recordings are matched to patients in the clinical database.

1. Introduction

The Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-II) Database contains both comprehensive clinical data, stored as a relational database, and bedside monitor-derived waveform data from thousands of Intensive Care Unit (ICU) patients.[1] Both the clinical data and the waveform data require significant curation before they can be made available for biomedical research; here, we focus on curation of the waveform data.

The patient-waveform data are originally recorded on commercially produced devices, the purpose of which is simply transmission to and display at nurses' stations, within the hospital, in real time. As archival collection was not intended for these devices, spurious and inconsistent data are common; especially, the lack of temporal markers makes retrospective analysis a challenge. Computationally intensive curation of the raw data is performed offline, before the waveforms are included into the MIMIC-II database.

Additionally, as the clinical data and the waveform data come from separate computer systems, under administratively distinct domains at the same hospital, a definitive relationship between the waveform data and their respective patients is not provided. We describe our process for rebuilding that relationship. As much as possible, we use automatic procedures to generate *matches* between waveforms and patients; however, we have developed GUI-based software tools to assist in manual matching, and in verification of automatic matches. While studies can be done using the waveform data alone, when properly matched to patients in the clinical database (DB), significant research opportunities appear ([2], [3], [4]).

2. Data collection

The waveform data consist of ECG, ABP, and PAP signals sampled at 125 Hz with 10- or 12-bit resolution, varying in length from minutes to hours, or days. In addition to these raw signals, there are low-resolution (1 Hz) "trend" signals that are calculated from these. The types of trends include heart rate, respiratory rate, non-invasive BP, CVP, PAWP, cardiac output, and both PAP and ABP. Finally, the clinical DB includes once-per-hour recordings of the same trend types; these data are entered by nurses monitoring the waveforms in real time, but may be (and often are) modified by them.

2.1. Data-collection generations

The waveform data come from two generations of collection. In the first, up to four waveform channels are collected simultaneously from each patient. These waveform data are mostly well behaved: most are time-aligned with their respective hourly clinical-DB recordings. Moreover, the large majority of waveforms are tagged with medical record numbers (MRNs); however, those that are not have no other direct patient-identifying information.

The second generation contains up to eight simultaneous waveform channels per patient, but these data are less well behaved: they require significant curation in order to be useful as complete, continuous waveforms; their time stamps are unreliable, and are often significantly time-

shifted (by up to several hours) compared to the hourly clinical-DB recordings; and just over half are tagged with MRNs. In this generation, though, the majority of waveforms are tagged with patient names, which are very important as patient identifiers.

2.2. Waveform recording

Patients in the ICU are attached with multiple leads to *bedside monitors*, which have network connections to a centralized *database server*, which feeds the waveform data in real time to the nurses' stations. The *archiving agent* is a relatively low-powered PC attached directly to the database server, acquiring batches of waveform data in soft-real time.[1]

Along with the waveform data, *export logs* containing human- and machine-readable metadata are produced. The export logs mark each waveform with a unique *Case ID* and – depending on the generation – patient names, MRNs, and ICU names.

2.3. Waveform curation

The archiving agent provides two files for each patient waveform: the "raw" file, which contains the high-resolution waveforms and 1 Hz trends, and the "flat" file, which contains a single, theoretically exact replica of one waveform from the raw file. The raw file contains the relevant data, but does not contain any reliable time stamps: the archiving agent receives the raw file in contiguous *chunks*, and it only stamps it with the time of its arrival. The flat file, in contrast, contains generally accurate time stamps from the bedside monitor itself.

Processing occurs in several steps. After conversion into WDFB format [5], the flat-file reference waveform – though it is typically scaled and clipped – should match one of the signals in the raw file. Since no reliable marker of which raw-file signal the flat-file signal matches to is available, a brute-force procedure searches the latter for locations where a raw-file signal's chunk agrees with it. The first significant change in the raw chunk – where the signal goes "up" or "down" an appreciable amount – is used to latch onto the flat-file signal; if the rest of the chunk agrees, with the same computed scale factor for the flat-file signal, then that chunk can be properly time-aligned. Even a very permissive matching, though, does not always work: guesses about the alignment, based on the alignment of nearby segments, must sometimes be made. The resulting WFDB record consists of the properly time-aligned raw chunks.

[1]The archiving agent was not part of the original commercial design.

3. Matching fragments

Next, the waveforms are matched against their respective patients in the clinical DB. To perform this matching, several identifiers – including the waveforms' export-log metadata, physiological information from the waveform trends, and patient information from the clinical DB – are used.

3.1. Physiological matching

The first generation of waveforms consists of 2957 records with unique Case IDs. Of these, 80% are tagged with MRNs: these have been directly matched against patient records in the clinical DB. The remaining 607 have undergone a "bottom-up" matching process: a short list of potential patient matches is produced for each Case ID, and then a comparison of physiological data is presented to a human user to aid in selecting the best among these matches.

For a given Case ID, the list of potential matches is created as follows: all patients are culled except for those with an ICU stay whose time interval overlaps that of the Case ID's waveform; and those for which for which a majority of trend types "matched" those available for the patient in the clinical DB (i.e. either both were available, or both not).[2] Culling by both of these identifiers produced an average potential-match list length of forty patients per Case ID.

Next, the 1 Hz waveform trends were compared directly to their corresponding potential hourly trend data in the clinical DB by (a) applying a 10-minute median filter to the waveform trend, (b) computing the absolute values of the differences between this filtered waveform and the clinical trend data, and (c) taking the median of those differences. For each such median, a confidence rating was calculated by comparing it to similarly computed medians for a fixed set of 100 previously, randomly selected, and verified matches: the number of these 100 medians that are larger than a potential match's median – for each trend type – gives the per-trend confidence rating.

A final confidence rating takes into account all of the types of trends available, along with the number of data points used for comparison, as well as an *importance weight* based on the relative perceived reliability of the trend types. For example, as respiratory-rate trends tend to be noisy, they are given less weight than heart rate trends, which are calculated directly from ECG signals. The full list of importance weights is shown in table 1.

The formula for the overall confidence rating is:

$$R_{overall} = \frac{\sum_{i=1}^{n} R_{trend,i} \cdot W_{trend_i} \cdot P_{trend_i}}{\sum_{i=1}^{n} W_{trend,i} \cdot P_{trend_i}}$$

[2]Systolic and diastolic ABP were not used here, because the information for their availability was not present at the time.

Table 1. Importance Weights for Trend Types

Trend Type	Importance Weight
Heart Rate	1.5
Resp. Rate	0.5
NBP sys./dias./mean	0.33 each
PAP sys./dias./mean	0.5 each
CVP	0.5
ABP sys./dias./mean	0.5 each
PAWP	1
CO	1

given the per-type confidence rating $R_{trend,i}$ for each trend i, the importance weight $W_{trend,i}$, and the number of waveform-trend points used, $P_{trend,i}$.

Given the confidence rating-ranked list of potential matches, a human user was able to verify which, if any, match is valid for each previously unmatched Case ID. Furthermore, The GUI-based tool used to enable this was also adapted to verify previous matches (based on MRN) that had unusually low overall confidence ratings. The results are given in section 4.

3.2. Matching with discrete identifiers

Physiological matching has so far been found not to be feasible on the second generation. This may largely be a result of the fact that waveform times from this generation are not well synchronized with the hourly trend data in the clinical DB.[3] Instead, direct matches are attempted using patient names and MRNs, with confirmation provided by agreement on ICU identity and time-interval overlap between the waveforms and the ICU stays.

3.3. Matching with MRNs

In the data collected for MIMIC-II, when MRNs are available, they may encounter two problems: a typo produces a well formed, but incorrect MRN for a patient; or the MRN is not well formed (i.e. not a seven-digit number).

Damerau-Levenshtein distance[6] is used to compare two MRNs. This is an example of an "edit distance" metric, which is the minimum number of "edits" performed on one string to transform it into the other. For Damerau-Levenshtein distance, an "edit" is either a character insertion, a character deletion, a change of a single character in place, or a swapping of two different adjacent characters. A simpler edit-distance metric is Levenshtein distance, which incorporates only the first three of these operations.[7] Damerau-Levenshtein distance can be more useful than plain Levenshtein distance when trying to account for human-originating typos, because it treats a character swap on the same level as other typos; thus, we use it to compare MRNs.

[3]More advanced techniques, such as using cross-correlation to synchronize between the two sets of data, have been considered but not yet applied.

3.4. Matching with names

Unlike an MRN, one person's name can be written in numerous different ways, all of which may be valid, but some of which are more likely to be encountered than others. Similarly, one written name may legitimately identify two (or more) different people.

Regular expression-based normalization is attempted, to extract a first name, middle name(s), last name, and suffix.[8] This attempted normalization takes account of the usual patterns observed, separately, in the clinical DB, and in the waveforms' export logs. Ambiguity may still occur: e.g., if only one name is available, it may be a first or last name; first or last names be written out of order; or middle names may be falsely distinguished from parts of first or last names.

After normalization, a set of three matching functions of the form $names_i : (Name, Name) \rightarrow Boolean$, along a scale from "strong" to "weak" – returning $true$ for names believed to match to the same patient – are used. Stronger functions require more similarity between names to return $true$; weaker functions are more lax. Techniques employed include: swapping or rotating first, middle, and last names, before computing Levenshtein distance over the concatenation; expanding first initials to match full names; matching a single name against either of the other names; and allowing larger Levenshtein distances between first names than last names. The three functions, $names_w$ (weak), $names_m$ (medium), and $names_s$ (strong) use combinations of such techniques that are relatively specific to the export logs and clinical data in question.

3.5. Matching with multiple identifiers

Full matching is performed by exploring the space of MRN and name identifiers in a piecewise fashion, using time-interval overlaps to confirm or reject potential matches. The entire process is performed inside of a relational database.

The procedure for Case IDs with MRNs is shown in figure 1. As with the three name-matching functions described, we have two MRN-matching functions: $MRNs_e$ (matching two MRNs exactly) and $MRNs_s$ (matching two MRNs strongly); as well as three time-interval matching functions, with different overlap constraints: $times_w$ (weak), $times_m$ (medium), and $times_s$ (strong); and one ICU-identity matching function: $ICUs_e$ (matching two ICU unit names exactly). In this figure, a rectangular box

represents a join, in the relational-database sense, with the list of patients in the clinical DB;[4] and a hexagonal box represents a pure filter, which simply removes potential matches.

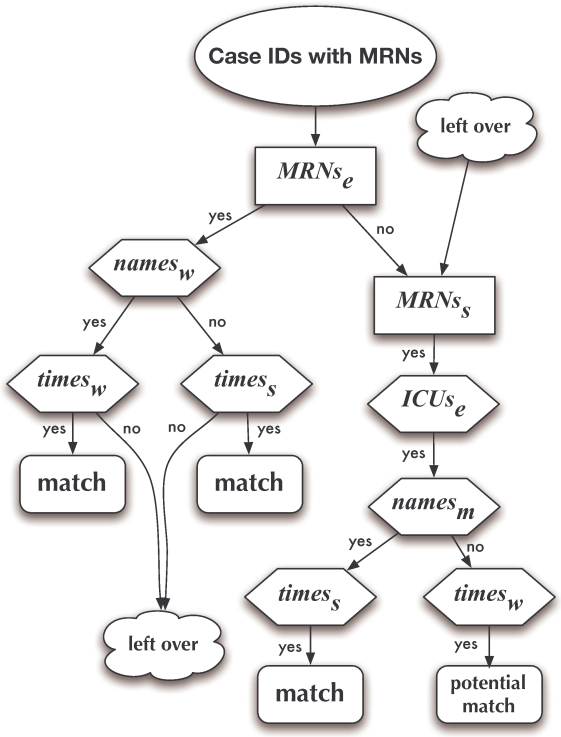

Figure 1. A flow-chart of the matching procedure using names, MRNs, and time intervals, in the clinical database.

In this figure, the leftmost branch contains the most confident matches: an exact MRN, a name that roughly matches, and an overlap between time intervals. The "left over" set, along with those Case IDs left out of the first join, have no exact MRN match; for these, similar, but not exact, MRNs are sought.

The data-flow for Case IDs without MRNs is simpler, and so not shown diagrammatically: here, we require very strongly matching names, along with some (medium-confidence) time-interval overlap, or else slightly more loosely matching names, but only with strong confidence in the times' overlap.

4. Results

The first generation of waveforms contains 2957 unique case IDs; 2350 were matched directly by MRN, and of the remaining 607 without MRNs, 353 were matched physiologically, with the manual, visual verification tool described. Of the 2350 direct-MRN matches, the same tool was used to verify those with a confidence rating less than 30%; 59 mismatches were found, and were corrected manually. In total, 2703, or 91.4% of the first generation, were matched. In the second generation, 1813 unique case IDs are under consideration,[5] of which 926 have MRNs. The automatic procedure described has matched 1202, or 66.2% of these. Though not yet used rigorously, another visual-verification tool was built and used to compare waveform trends and clinical hourly trends for a random set of 50 of these matches: no errors were detected, but it became clear that there were significant, unpredictable time shifts between otherwise well matching waveform and clinical data.

[4]This is true except for the fact that, unlike a true join in which two keys must match exactly, there may be some small distance between the keys, depending on the match function.

References

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

[2] Lee J, Mark RG. A hypotensive episode predictor for intensive care based on heart rate and blood pressure time series. In Computers in Cardiology 2010. Los Alamitos: IEEE Computer Society Press, 2010; to appear.

[3] Chen T, Mark RG, Clifford GD. The effect of signal quality on six cardiac output estimators. Computers in Cardiology September 2009;36:197–200.

[4] Shavdia D. Septic Shock: Providing Early Warnings Through Multivariate Logistic Regression Models. Master's thesis, Massachusetts Institute of Technology, 2007.

[5] Moody GB. WFDB Programmer's Guide, July 2010. http://www.physionet.org/physiotools/wpg/wpg.htm.

[6] Damerau FJ. A technique for computer detection and correction of spelling errors. Commun ACM 1964;7(3):171–176. ISSN 0001-0782.

[7] Levenshtein VI. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady February 1966;10:707–+.

[8] Friedl J. Mastering Regular Expressions. O'Reilly Media, Inc., 2006. ISBN 0596528124.

Address for correspondence:

Michael B. Craig
E25-505
Laboratory for Computational Physiology
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139 USA

[5]Waveforms continue to be collected under this generation, but their respective clinical data have not yet been made available, so matches for these have not yet been considered.