

Matching Data Fragments with Imperfect Identifiers from Disparate Sources

Michael Craig*, Benjamin Moody, Sherman Jia, Mauricio Villarroel and Roger Mark

Harvard-MIT Division of Health Sciences and Technology,
Massachusetts Institute of Technology, Cambridge, MA, United States

Patient care in modern intensive care units (ICU) relies on a rich, yet disparate, set of clinical and physiologic data. To support research aimed at improving diagnosis and treatment of ICU patients, we have captured these data in the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-II) Database, using a customized data acquisition process that does not interfere with clinical practice. MIMIC-II includes (a) waveforms and derived parameters from bedside monitors, (b) clinical data from the ICU information system, and (c) data from other hospital laboratories and archives. These data come from devices that often do not retain detailed information regarding temporal relationships between parameters, as well as from highly aggregated sources. Assembling comprehensive records from fragments that may lack common identifiers is a problem that is likely to occur in any similar project. For projects such as MIMIC-II, in which many thousands of records must be constructed, automated solutions are essential tools.

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. Comparing recorded waveform data with nurse-reported vital signs, however, is less likely to lead to successful matching. Since the data we have collected are representative of ICU data, and can be analyzed only retrospectively, imprecision and inaccuracy become significant challenges, which we addressed using rule-based normalization and text edit-distance metrics. For patients whose records cannot be assembled automatically, we developed a visual verification tool. The entire process, both automatic and manual, currently matches almost 90% of the available waveform recordings to patients in the clinical database. We conclude by recommending how standards for recording both waveforms and clinical observations can be improved to increase the fraction of usable data.