Towards better event log preparation with quality optimisation for hospital process mining

Ruihua Guo[1], Angus Richie[2], Yang Lu[3], Boris Choy [4], Ross Smith[1], Haeri Min[1], Qifan Chen[1] and Simon K. Poon[1]

[1]School of Computer Science, The University of Sydney, NSW, Australia

[2]Sydney Medical School, The University of Sydney, NSW, Australia

[3]School of Public Health, The University of Sydney, NSW, Australia

[4]Business School, The University of Sydney, NSW, Australia

Background

Hospital information systems (HIS) have facilitated our understanding of complex time-dependent patient behaviors using process mining techniques. The primary data input for process mining is an event log—a structured log file containing temporal information about a series of activities. However, real-world hospital data often contain missing, imprecise, invalid, or irrelevant records, which complicates the generation of high-quality event logs.

Quality issues in hospital event logs can arise from various sources and impact different data elements depending on the specific research question. The secondary use of electronic medical records (EMRs) for process-oriented research requires a systematic approach to quality assessment. Unlike general data-driven analyses, which use unit-level data as direct inputs, event log creation requires tailored data preparation strategies. A limited understanding of data elements and their semantic components within the complex HIS environment can misrepresent patient behaviors during event log formulation. Low-quality event logs can also lead to erroneous clinical conclusions, underperformance or ineffectiveness of process model mining (such as 'spaghetti effect').

Current strategies for assessing and enhancement of hospital event log quality for a specific health research topic remain limited and often inadequate.

Objective

In this study, we integrated a multi-layer evaluation approach to optimise the quality of hospital event logs from real-world EMRs, in the context of studying patients with heart failure (HF) at risk of readmission.

Contribution

The contributions of this study include demonstrating quality enhancement through a multi-layer evaluation framework in preparing real-world hospital event logs by optimizing source validity, data reliability, and log completeness while maintaining balanced complexity. Additionally, this study addresses the critical gap in hospital event log enhancement for HF patients at risk of readmission.

Methodology

Eligible patients were identified from DREAM, a multi-site hospital dataset encompasses routinely collected EMRs within a large metropolitan health system in Australia. Based on the readmission vulnerable window (i.e., the time between the index stay and the second admission), they have been classified into three subpopulation groups including 30-day, 90-

day and 180-day groups. The Weiskopf and Weng framework was used to evaluate the data quality across five dimensions—currency, correctness, completeness, concordance, plausibility, alignment with the study objective, and at multiple analytical levels. Results were benchmarked against the publicly available Medical Information Mart for Intensive Care IV (MIMIC-IV) hospital database. The quality of generated event logs was optimised using biodiversity frameworks and further compared to MIMIC-IV logs.

Results

Our findings showed that DREAM provided a timely, area-specific source of information with superior currency and source completeness compared to the benchmark database. The correctness and plausibility were comparable for both sources. For the completeness profile, both the DREAM and MIMIC logs showed higher *coverage* than *global completeness*. For the diversity profile, the MIMIC log exhibited higher diversity than the DREAM log, as measured by Species Richness, Expected Shannon Entropy, and Inverse Simpson Diversity. Fluctuations were observed across the three vulnerable periods in both the DREAM and MIMIC logs. In the DREAM logs, the 180-day readmission group showed the highest Expected Shannon Entropy compared to the other two periods (20.2 vs. 14.9 and 13.7). In contrast, in the MIMIC log, Expected Shannon Entropy decreased from the early to late vulnerable period (110.8, 82.4, and 79.0, respectively).

Conclusion

This study employed a multi-layer approach to assess and optimize the quality of event logs for mining and analyzing patients with HF at risk of readmission. The five-dimensional evaluation closely aligned DREAM data quality with the specific research question. Data element validity was assessed at multiple levels, including encounter, trace, event, label, and attributes. The multi-level evaluation approaches also investigated one of the most important data elements, timestamps, from its plausibility, completeness and timelessness. Adding an extra layer to event log quality evaluation enhanced both external validity and internal fairness, improving log comparability across different sources and within subpopulation groups. It significantly strengthened the log's validity in relation to the specific health research question. However, addressing treatment and process dynamics remains an area for future research.