COMBINING NOVEL DATASETS TO UNDERSTAND ILLNESS TRAJECTORY: THE CONDUIT DATABASE PROJECT
Maslove, David1; McGregor, Carolyn2; Dumontier, Michel3; Seely, Andrew4; Marshall, John1
1St. Michael's Hospital, Li Ka Shing Knowledge Institute, Toronto, Canada; 2University of Ontario Institute of Technology, Faculty of Health Sciences, Oshawa, Canada; 3Stanford University, Biomedical Informatics, Stanford, USA; 4The Ottawa Hospital, Thoracic Surgery & Critical Care Medicine, Ottawa, Canada

Introduction: Over the last decade, the field of critical care has witnessed an explosion of data, largely attributable to the increasing implementation of electronic medical records (EMR), emerging genomic technologies, and advances in physiologic waveform capture and analysis. These innovations are helping to bring critical care into the era of “big data”, in which enormous amounts of data will be generated and collected in the course of basic research and clinical care. Paradoxically, the shift in paradigm from evidence-based practice to personalized care is for the most part taking root in chronic conditions such as cancer, where the data describing disease processes are comparatively sparse. Transposing the principles of precision medicine to the realm of acute care will require new computational approaches that account for the temporal and spatial complexity of rapidly evolving, multi-focal diseases.

Objectives: The CONDUIT project aims to develop computational methods for the extraction, transformation, cleaning, and merging of large-scale data streams collected from critically ill patients. Based on an ontology of critical care concepts, the CONDUIT database schema will provide a universal framework to facilitate collaborative research across different health care and EMR systems. Further enabling this goal, CONDUIT is focused on an open data model, so that de-identified data will be available to a broad community of bioinformatics and clinical researchers.

Methods: Preliminary work has focused on methods of data transformation to support the use of machine learning algorithms with EMR data types. These include automated artifact and outlier detection, and discretization of continuous variables. Using partitional clustering and principal components analysis applied to publicly available data, we have explored differences in gene expression patterns at different times in the course of an acute illness, and from different tissue types.

Results: When analyzing EMR data, discretization of continuous variables, such as arterial blood gas and cardiac output measurements, enables their use in Bayesian learning algorithms. In the area of transcriptome analysis, the time of sampling, as well as cellular fraction used, affect gene expression profiles in acute illness, and do so differently in different patients.


![Table showing heart rate (HR), mean arterial pressure (MAP), respiratory rate (RR), oxygen saturation (O2 Sat), and temperature (Temp) at different times.

**Figure 2.** Clinical, genomic, and physiologic data represented as a feature vector.

\[ x = [x_1, x_2, x_3 \ldots x_n] \]

**FIGURE** – A model for the merging of EMR, genomic, and waveform data from critically ill patients. Data cleaning and transformation algorithms are used to transform values for each of the features contained within the individual data sets, which are then combined to form a feature vector. A mathematical representation of the feature vector, along with key metadata, can then be used in a variety of machine learning and data mining algorithms.