Data Mining Infrastructure for AIMS Based Registry

Presenting Author: Hubert Kordylewski, PhD, Anesthesia Quality Institute (AQI), IL

Co-Authors: Benjamin Westlake, Anesthesia Quality Institute, IL, Richard Dutton, MD, MBA, Anesthesia Quality Institute, IL, Lance Mueller, MS, Anesthesia Quality Institute, IL

Introduction: AQI'S National Anesthesia Clinical Outcomes Registry includes more than 22 million anesthesia cases. AQI collects clinical, quality, and administrative with detailed 'sensory' perioperative data, medications administration, and unique events recorded in real time in electronic record systems. Thus, one case may consist of thousands of time dependent measurements. With the above constraints, the computational time required to develop and test data mining algorithms becomes a major issue. The abstract outlines AQI's data mining infrastructure (Figure 1).

Methods: Currently, when a problem arises, we re-use common functionality, which involves: (1) an extract of the data and (2) data analysis and data mining formulas (Figure 2). The formulas repository module in Figure 2 serves as a repository of implementations of two types of data mining formulas: (1) custom made algorithms designed and developed for specific problems and (2) batch procedures with substantial utilization of modules in commonly used data analysis packages (SQL modules and R packages). To solve requirements of performing data analysis at a reasonable time computational time, AQI's solution is a home grown implementation of the 'map-reduce inspired' approach in dividing computations among several linked SQL servers (Figure 2).

Results: Typical data mining task involves examination (and their correlations involve) of thousands of observations across fifty or more variables for each anesthesia case) on a large subset of NACOR's AIMS data (500,000 to 1,000,000 cases with vital signs, medication events). Within single server setup (16 cores at 3.8 Ghz, and 16GB EEC memory) we were able to analyze the subset of NACOR data (500,000 cases) in about thirty minutes in one pass (300 cases/second). Exporting the subset of NACOR data to a distributed computing environment on 6 SQL nodes (4 to 8 cores, 8GB memory) reduced the wait time between passes by a factor of four.

Conclusion: The majority of clinical registries or local hospital data centers do not have the need for a full blown 'Big Data' implementation. From the response we have received from sharing AQI's XML schema, and AQI's effort to standardize definitions for clinical outcomes, we see increasing demand for guidance and leadership in providing data related infrastructure. There are several areas in which we are planning to share the experience we acquired in our design and implementation of AQI's systems, including: (1)publishing our knowledge library for common categorization and filtering of cases (The library can also be

used by other clinical registries or data analysis teams), (2) Publishing implementation and of our data analysis infrastructure.

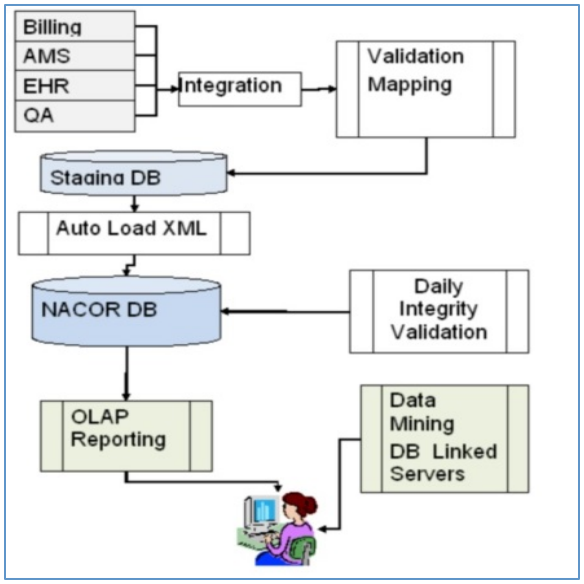

Figure1 Organization of data subsystems.

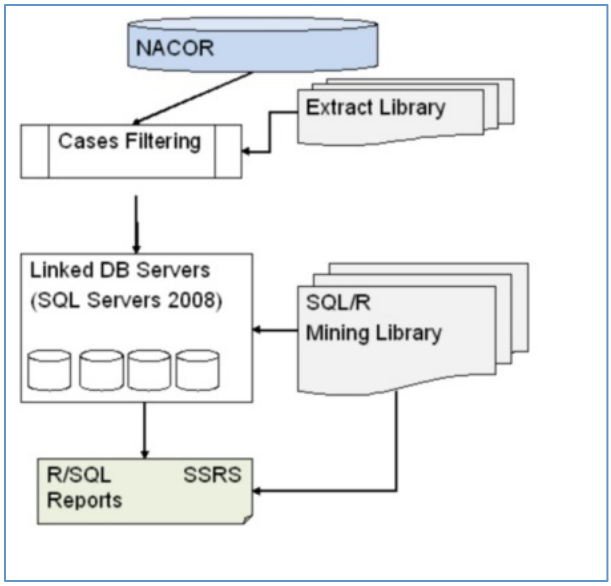

Figure2. Data Mining Infrastructure