Delivering Data Mining Without the Warehouse: MUSC=s Experience
Myla D. Ebeling,1 Caroline K. Boyd,2 Deborah A. Agbor-Tabi3, Lawrence B. Afrin3
1Pediatric Epidemiology, 2Ctr. for Computing & Info. Tech. (CCIT), 3Hollings Cancer Ctr.
Medical University of South Carolina (MUSC), Charleston, South Carolina

Abstract. Many health care organizations use electronic clinical data repositories for direct patient care; fewer organizations have graduated to using their repositories for population-based analysis. We sought to develop such a capability for our institution in a time of severe resource constraints and discovered the need to address many technical questions and even more human systems questions. We describe the approaches used to ensure institution-wide comfort and satisfaction with this new capability prior to initial release.

Background. MUSC, like many large health care facilities, has a large clinical data repository (CDR) as the core of its electronic medical record, yet the only extraction method available to the average user is the clinical interface, providing data one patient at a time. Population-based extraction/analysis is desirable for many purposes, but many issues (e.g., extraction and analysis technologies, the very complexity of the CDR schema, data integrity, security, ethics) needed to be addressed before a useful data mining capability could be responsibly released to the general user community.

Some organizations1,2 have been fortunate to be able to develop warehouse-based data mining capabilities, but others, like ours, face resource constraints which preclude near-term establishment of warehouses. Still, the potential benefits of even a limited mining capability impelled us to develop a non-warehouse-based, generally available facility for our CDR.

Materials and Methods. In March 1999 MUSC=s Clinical Information Systems Steering Committee (CIS) decided to pursue mining and formed a subcommittee, the Data Mining Development Steering Committee (DMDSC, open to all interested users), to oversee the project. The DMDSC interacts with several other MUSC bodies. As formal data warehousing was fiscally impractical, the DMDSC focused its efforts on leveraging the existing CDR.

CCIT allocated start-up resources and made a (regularly updated) copy of the 105-GB, 540-table, 30,000-field Sybase CDR (Oacis Healthcare Systems; 75 million services on 573,000 patients in 9 years). Analytical packages were reviewed, the Cognos suite was selected, a small license set was purchased, and an initial set of cubes was developed. Cubes and other technology infrastructure were appropriately secured. The limitations of the data were investigated3 and documented. Meanwhile, university policy/procedure, IRB-approved protocol, and training/support resources were developed to allow users to perform web-based ad hoc queries of de-identified CDR data (identifier release allowed via separate IRB approval).

Results. Work is progressing toward an initial release scheduled for October 2002. Initially cubes were built slowly as the DMDSC developed an understanding of the process, but cube-building is now accelerating to provide a robust set of demographic/encounter-focused cubes for the initial release. DMDSC members are exploring the data to better understand its potential (unsuspected results already have been identified), verifying ability to enforce policy and procedure, and developing training material. Issues now being addressed include the need for more robust servers, options for financing user licenses, and methods for making the user community aware of the new capabilities.

A detailed review of the material developed for this project, and an update on the go-live experience, will be presented at the AMIA 2002 Annual Symposium.

Discussion. Delivering data mining capabilities to a diverse array of users is a complex endeavor requiring careful attention to both technological and human systems. Our pre-release experience suggests suboptimal but still useful analytical capabilities can be provided in a secure and confidential manner without formal data warehousing or extensive resource allocations. However, a cataloging of parties likely to be affected by the availability of mining capabilities, and efforts to align the interests of all such parties, must be pursued well in advance of go-live to improve the probability of a successful release.

Acknowledgements. Portions of this work were supported by Fast Track Systems, Inc.

References