Tales from the EMR: Does a 21st-Century Data Warehouse Facilitate Clinical Research for Pancreatic Cancer?

Edward J. Arous  
*University of Massachusetts Medical School*

Jillian K. Smith  
*University of Massachusetts Medical School*

Sing Chau Ng  
*University of Massachusetts Medical School*

See next page for additional authors

Follow this and additional works at: [http://escholarship.umassmed.edu/ssp](http://escholarship.umassmed.edu/ssp)

Part of the Health Information Technology Commons, Health Services Research Commons, and the Surgery Commons

Repository Citation

[http://escholarship.umassmed.edu/ssp/114](http://escholarship.umassmed.edu/ssp/114)

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in Senior Scholars Program by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
Tales from the EMR: Does a 21st-Century Data Warehouse Facilitate Clinical Research for Pancreatic Cancer?

Authors
Edward J. Arous, Jillian K. Smith, Sing Chau Ng, Jennifer F. Tseng, and Theodore P. McDade

Comments
Medical student Edward Arous participated in this study as part of the Senior Scholars research program at the University of Massachusetts Medical School.

This poster is available at eScholarship@UMMS: http://escholarship.umassmed.edu/ssp/114
Introduction

The importance of an electronic medical record has been highlighted for both clinical care and research. In the current era, data warehouses and repositories have been established to serve the dual function of patient care and investigation. The aim of this study is to compare a newly developed institutional clinical data warehouse, linked with the hospital information system (HIS), to a prospectively maintained departmental database.

Methods

Databases

HIS-Linked Database

This novel HIS-linked institutional clinical data warehouse captures inpatient and outpatient clinical and billing information from a pool of over 2 million patients evaluated at an academic medical institution and its affiliates, since 1995. A cohort was identified; following Institutional Review Board approval, demographic and clinical data was obtained.

Surgical Oncology Database

A manually entered and prospectively maintained surgical oncology database of the same institution, tracking 394 patients since 1999 was also used for analysis.

Data Collection

Both databases were queried for 9 primary and secondary ICD-9-CM discharge diagnosis codes for pancreatic cancer. Duplicated patients, and those unique to either dataset, were flagged. Patients with diagnosis dates prior to 1999 were excluded to allow comparison over the same time period.

Statistical Analysis

For validation purposes, a 10% random sample of remaining patients unique to each dataset underwent manual review of medical records including clinic notes, admission/discharge notes, diagnostic imaging, and pathology reports.

Results

1107 patients were identified from the HIS-linked dataset with pancreatic neoplasm-associated diagnosis codes dating from 1999 to 2009. Of these, 254 (22.9%) were captured in both datasets, while 853 (77.1%) were only in the HIS-linked dataset (see Figure 1). Patients identified in each database had similar age, sex, and racial characteristics (see Table 2). Manual review of the 10% subset of the HIS-only group demonstrated that 55.6% of patients were without identifiable pancreatic pathology, suggesting miscoding, while 31.7% had diagnoses consistent with pancreatic neoplasm, and 12.7% with pseudocyst or pancreatitis (see Figure 2). Of the 394 patients tracked by surgical oncology, 254 (64.5%) were captured in both datasets, while 140 (35.5%) had not been captured in the HIS-linked dataset. Manual review of the 10% subset of the non-captured patients demonstrated 93.3% with pancreatic neoplasm and 6.7% with pseudocyst or pancreatitis. Lastly, a review of the 10% subset of the 254 patient overlap demonstrated that 87.5% of patients were with pancreatic neoplasm, 8.3% with pseudocyst or pancreatitis, and 4.2% without pancreatic pathology.

Conclusions

- A high degree of misclassification may be present if queries are based solely on ICD-9-CM diagnosis codes. For that reason, careful validation and data cleaning are critical steps prior to research use.
- The new HIS-linked system was unable to capture over one-third of the patients in the surgical oncology database.
- These results suggest cautious interpretation of national-level administrative data utilizing ICD-9-CM diagnosis codes.
- Current state-of-the-art data warehouses continue to require clinical correlation and validation through traditional retrospective mechanisms.

References