Tales from the EMR: Does a 21st-century data warehouse facilitate clinical research for pancreatic cancer?

Edward J. Arous BS, Jillian K. Smith MD, MPH, Sing Chau Ng MS, Jennifer F. Tseng MD, MPH, Theodore P. McDade MD
University of Massachusetts Medical School, Surgical Outcomes Analysis & Research, Worcester, MA

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.

ICD-9-CM Code	Diagnosis
157.0	Malignant Neoplasm of Head of Pancreas
157.1	Malignant Neoplasm of Body of Pancreas
157.2	Malignant Neoplasm of Tail of Pancreas
157.3	Malignant Neoplasm of Pancreatic Duct
157.4	Malignant Neoplasm of Islets of Langerhans
157.8	Malignant Neoplasm of other unspecified sites of Pancreas
157.9	Malignant Neoplasm of Pancreas, part unspecified
211.6	Benign Neoplasm of Pancreas, except Islets of Langerhans
211.7	Benign Neoplasm of Islets of Langerhans

Table 1. *ICD-9-CM* diagnosis codes used in this study.

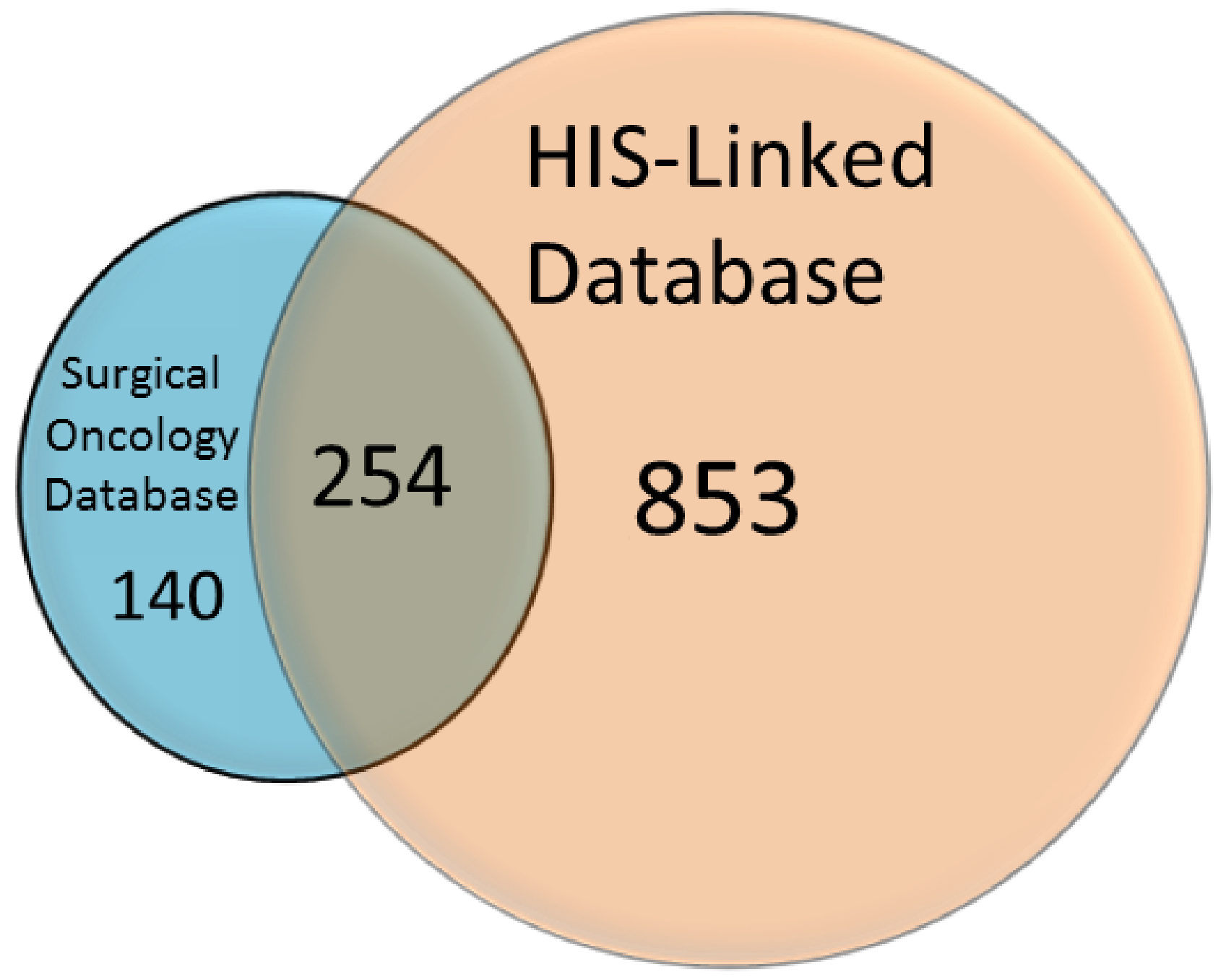

Figure 1. Patients obtained by *ICD-9-CM* query in the surgical oncology database and HIS-linked database.

		Surgical Oncology Database Only	Both Databases	HIS-Linked Database Only
Total Patients		140	254	853
Age (in years)		64.0 (SD = 13.9)	63.0 (SD = 12.0)	62.8 (SD = 15.7)
Sex	M	51 (41.1%)	123 (50,0%)	412 (48.7%)
	F	73 (58.9%)	123 (50.0%)	434 (51.3%)
Race	White	109 (91.6%)	231 (90.9%)	748 (92.2%)
	Other	10 (8.4%)	23 (9.1%)	63 (7.8%)

Table 2. Demographic data of the patients identified solely in the surgical oncology or HIS-linked databases, as well as those captured in both.

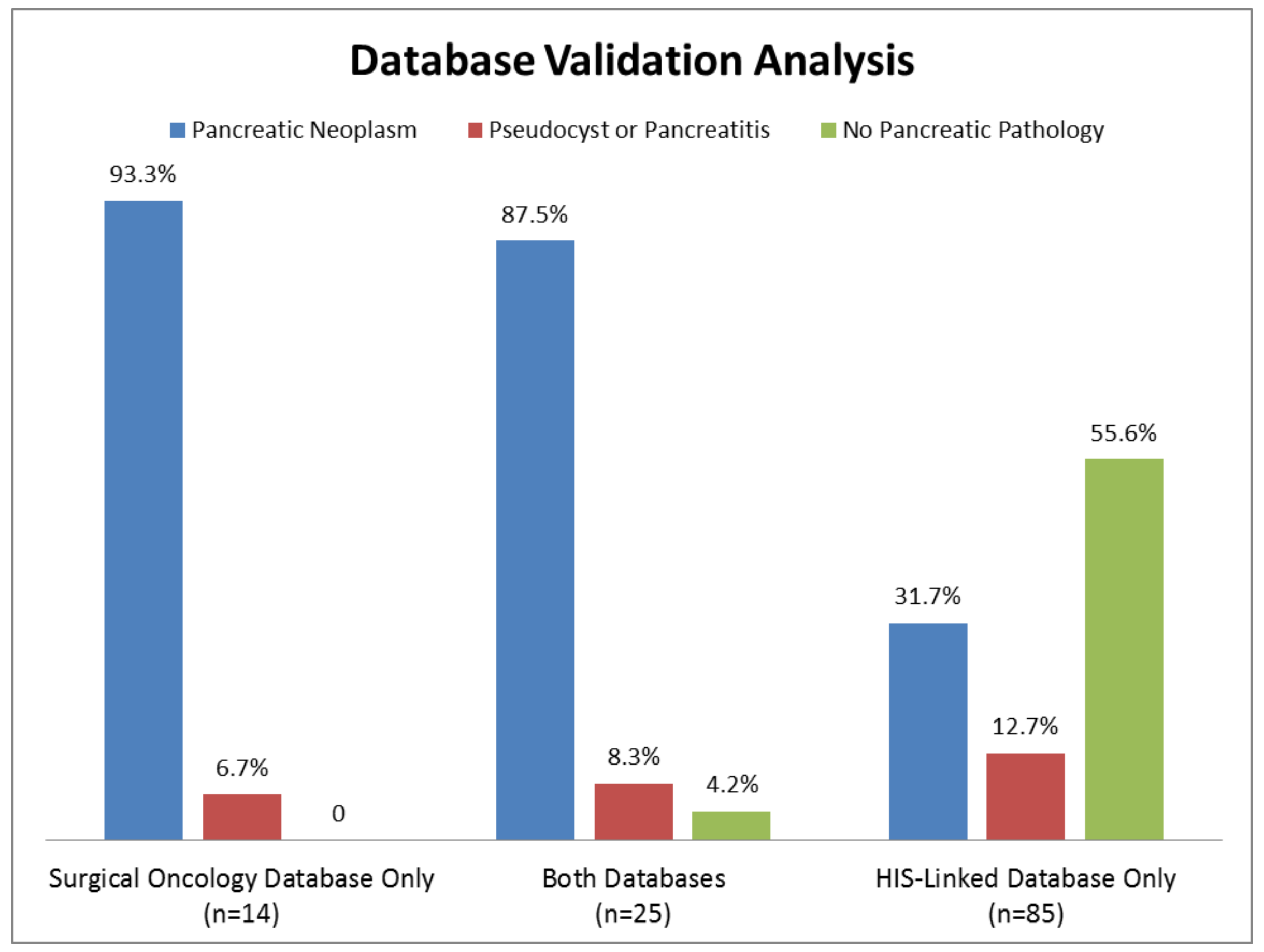

Figure 2. Validation analysis by manual review of a randomized 10% sample of each subset.

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

1. Murphy SN, Mendis ME, Berkowitz DA et al. Integration of clinical and genetic data in the i2b2 architecture. AMIA Annu Symp Proc. 2006; 1040.
2. Manasanch EE, Smith JK, Bodnari A et al. Tumor registry versus physician medical record review: a direct comparison of patients with pancreatic neuroendocrine tumors. *Journal of Oncology Practice* (in press).
3. *Massachusetts Integrated Clinical Academic Research Database*. 2009. Web. 20 Apr. 2011. <http://micard.umassmed.edu/>.

Supported in part by the Alpha Omega Alpha Carolyn L. Kuckein Student Research Fellowship.